Screening Overconfident Agents with Belief Heterogeneity in Moral Hazard

Master Thesis Presented to the

Department of Economics at the

Rheinische Friedrich-Wilhelms-Universität Bonn

In Partial Fulfillment of the Requirements for the Degree of

Master of Science (M.Sc.)

Supervisor: Prof. Botond Kőszegi

Submitted in October, 2025 by

Mine Nebahat Özer

Matriculation Number: 50064788

Abstract

We study a moral hazard problem with adverse selection by extending the framework of De la Rosa (2011): a risk-averse agent can possess private information about her subjective beliefs regarding the likelihood of success. The principal designs a menu of screening contracts, which we characterize by decomposing each contract into effort and risk-sharing elements. The analysis shows that overconfidence affects the incentive-insurance tradeoff through two different forces: the incentive effect, which reduces the cost of inducing desired effort, and the wager effect, which causes more risk exposure. Slightly overconfident agent, due to the incentive effect, opts for the incentive-compatible contract that provides insurance without destroying incentives; however, as a result of the wager effect, significantly overconfident agent self-selects the exploitative contract that exposes them to excessive risk. The results establish a monotone menu of screening contracts supported by the single-crossing and highlight how belief heterogeneity can simultaneously lower the cost of agency for slightly overconfident agent while causing exploitation for significantly overconfident agent.

Keywords: heterogeneous beliefs, overconfidence, screening, behavioural contract theory

Contents

1 Introduction

The study of contract problems began with the foundational work of Holmström (1979), who formulated how hidden effort creates incentive constraints that alter efficient allocation. Moreover, Grossman and Hart (1983) expanded this to demonstrate how incentive compatibility affects contracting under the moral hazard settings by a canonical model that describes the fundamental tradeoff between risk sharing and incentives under homogeneous beliefs by following the Pareto-optimal risk-sharing framework of Borch (1962).

While standard principal-agent models typically assume homogeneous beliefs, recent works suggest this assumption could be systematically violated in real. There is extensive psychological evidence that people are overconfident about their ability and future gains, Taylor and Brown (1988). In the real-world settings, like the labor market, agents often possess private information about their own abilities or effort effectiveness. Thus, even before exerting any effort, this systematic overestimation of success generates *heterogeneous beliefs* between the agent and the principal about the likelihood of success. This behavioral trait of the labor is especially relevant in the principal-agent problem, where the agent's effort influences a stochastic outcome, since both the effort and type are not fully observable. The principal, therefore, faces a dual information problem: **hidden type (adverse selection)** and **hidden action (moral hazard)** coexist.

This work examines a contract design problem that involves both moral hazard and adverse selection. It addresses a scenario where a principal (she) uses contracts to motivate an agent (he) to undertake costly hidden actions (moral hazard), while the agent also possesses private information about his beliefs regarding effort effectiveness (adverse selection). Therefore, our goal is to investigate the optimal menu of contracts for the principal in such situations. In designing the optimal menu, the principal must simultaneously (i) provide incentives for effort (moral hazard), (ii) induce truthful self-selection of types (adverse selection), and (iii) ensure participation. This combination reflects the screening approach with hidden action and hidden information. To analyse this screening setting, we employ the canonical moral hazard framework of De la Rosa (2011), the agent's effort directly influences the output level via two channels: the *wager*[1] and *incentive*[2] effects

[1] This wager effect will tend to push the equilibrium contract towards higher-powered incentives, representing a shift toward stronger motivation with greater performance-based pay but weaker insurance.

[2] This incentive effect will tend to push the equilibrium contract towards lower-powered incentives, representing a shift toward smaller incentive to induce the desired effort and enough degree of insurance for risk aversion.

of overconfidence (Propositions 3 and 4), and extend it to a screening environment with hidden types.

In our model, the risk-averse agent is capable of exerting the level of effort based on her distorted belief, and thereby, this belief heterogeneity introduces a new source of private information: agents differ in their subjective beliefs on effort effectiveness rather than in productivity or cost, unlike standard adverse-selection models. In this way, we focus on the scenario in which such belief heterogeneity alters both the cost of agency that arises from the incentive-insurance tradeoff and the the design of screening contracts. This given formulation is relevant to real-world scenarios where the principal must optimally contract with agents who hold biased beliefs about their own performance or effectiveness, such as employees who overestimate the impact of their effort on success. In these specific environment, firms, employers, or policymakers face the dual challenge of motivating effort while also screening individuals according to their biased beliefs regarding to success.

A prominent feature of our model is that the principal designs a menu of contracts to screen agents according to their reported type and thus maximizes expected profit in the presence of heterogeneity. This designed menu not only changes the structure of incentives for exerting effort but also opens the door for *exploitation*: overconfident agents accept contracts that are more riskier, allowing the principal to extract additional surplus under true probabilities at their expense. That is, when the agent is significantly overconfident, the incentive compatibility constraint becomes slack (Proposition 2 ii.). Hence, the principal can therefore reduce the the cost of agency by offering a contract that is riskier than necessary (Proposition 5). In this sense, overconfidence substitutes for risk aversion (Corollary 1), as biased beliefs make the agent willing to accept risk that a rational agent would otherwise avoid.

From a technical perspective, to analyze how belief heterogeneity shapes optimal screening contracts, we specify that the contract is decomposed into its fundamental **design elements**. Instead of providing the full optimization problem (that would be functional optimization, which is infinite-dimensional and analytically intractable), we aim for this decomposition to clarify how these distinct elements jointly determine the optimal screening contracts following the results of De la Rosa (2011) in a moral hazard problem. Every contract can thus be represented by two design elements: an *effort level element*, which determines the level of effort the principal wish to implement, and a *risk-sharing element*, which governs how the risk is allocated, which occurs due to uncertainty about success. With this riched and tractable formulation, it becomes possible to investigate that each agent type (rational, slightly overconfident, or highly overconfident), based on different levels of beliefs, voluntarily chooses the designed contract that fits well according to their own beliefs. Thereby, it enables the design of a clear, monotone screening menu is possible. The resulting *self-selection* satisfies, thus, the *single-crossing property* (Eliaz and Spiegler (2006)), which ensures that types' indifference curves cross at once. Following this reasoning, the principal constructs a monotone menu of screening contracts that separates agents according to their types.

Another significant implication with screening is that the different biased beliefs about the probability of success across agents lead to systematic variation. Compared to slightly overconfident agent, the agent who is significantly overconfident about the likelihood of success places disproportionately higher values on an exploitative contract that offers higher success-contingent pay, which ensures that their preferences over contracts are ordered consistently. This monotonicity encourages the *single-crossing* property is hold (Lemma 1), allowing the principal to design a simple, self-selecting menu of contracts in which significantly overconfident agents voluntarily choose the exploitative contract without demanding any compensation for risk (would be informational rent).

While significantly overconfident types choose *exploitative* (riskier) contracts without requiring informational rents, the situation differs for rational or slightly overconfident agents. For a rational type, inducing effort is too *costly* due to the risk aversion, so the optimal contract is *flat* and provides full insurance with no incentives, as the classical framework suggests. Suppose the low-type is only *slightly* overconfident. In that case, however, the principal may find it optimal to offer an *incentive-provision* contract that induces effort at a lower cost of agency. In this case, the feasible menu of the contracts reflects the De la Rosa (2011)'s result, as the contracts are sensitive to the extent of overconfidence. Hence, in the screening environment, the distinction between rational, slightly overconfident, and sufficiently overconfident types highlights the range of feasible contracts: from flat contracts with no incentives (full insurance), to minimally incentive-compatible contracts with desired effort, to fully exploitative contracts with effort that goes beyond the desired effort.

Related Literature: This work is closest to the literature that studies the principal-agent problem with moral hazard and heterogeneous beliefs, such as Santos-Pinto (2008), Gervais and Goldstein (2007), De la Rosa (2011), and the paper by Dumav et al. (2025). While Gervais and Goldstein (2007) considers the impact of overconfidence in a teamwork setting with complementarities between agents, Santos-Pinto (2008), Dumav et al. (2025), and De la Rosa (2011) analyze the optimal contract for a single agent. This work contributes to this literature by extending the analysis of De la Rosa (2011) to a screening environment, in which belief heterogeneity across types interacts with moral hazard, thereby altering both the cost of agency that results in exploitation-biased beliefs, and the structure of optimal screening menus.

This paper is also related to the literature on moral hazard problems with adverse selection(Laffont and Tirole (1986); Baron and Besanko (1987); Picard (1987); Melumad and Reichelstein (1989); Faynzilberg and Kumar (1997)). More recently,Gottlieb and Moreira (2017) study a moral hazard problem with binary outputs, binary effort levels, and two-dimensional private information, and discuss exclusion and distortion in the optimal menu. Castro-Pires et al. (2024) studies conditions under which moral hazard and adverse selection can be decoupled, thereby reducing the moral hazard problem with adverse selection to a pure screening problem.

Our work is organized as follows. Section 2 introduces the model environment, agent preferences and contract evaluation, and specifies the relevant constraints. Section 3 looks at the optimal contracts in the moral hazard case. Section 4 studies the screening problem with belief heterogeneity in moral hazard. Finally, Section 5 summarizes the main results of the analysis.

2 The Model

This framework is closely related to the analysis in De la Rosa (2011), which offers a particular formulation of heterogeneous beliefs and focuses on characterizing the shape of the optimal contract in moral hazard settings. Let the principal be *risk-neutral* and the agent be *risk-averse.* The principal cannot observe the agent's effort level $e \in \{0, 1\}$, they can observe the realized outcome $x \in \{x_0, x_1\}$. The agent receives an outcome-contingent payments (wage) $s \in \{s_0, s_1\}$ that is endogenously determined as a function of the realized outcome; $s : \{x_0, x_1\} \to \mathbb{R},$ i.e. $s_0 = s(x_0),\ s_1 = s(x_1)$ and let the cost of effort be given by $c(e) = c.e$, where $c > 0$ is a constant. Given the binary nature of effort, this cost function yields: $c(0) = 0 \quad \text{and} \quad c(1) = c$ for $e \in \{0, 1\}$

2.1 Environment and Beliefs

We consider a principal-agent environment where the principal interacts with agents characterized by the subjective belief profile over outcome. Let the true probability of high outcome (x_1) depend on the binary effort level $e \in \{0, 1\}$, such that

$$\mathbb{P}(x_1 \mid e) := q + ve.$$

When under effort e the principal's prior on x_1 is $(q + v.e)$. The agent assigns probability $\tilde{p} \in (0, 1)$ to state x_1 is:

$$\tilde{\mathbb{P}}(x_1 \mid e) := \tilde{q} + \tilde{v}e.$$

In this way, the agent does not necessarily agree with the principal on how likely success is, either at baseline level (without effort) or in how effective effort is perceived.

Type Definition. We assume that subjective evaluations of the outcome characterize the agent types. Specifically, we define θ-type which determines the pair $(\tilde{q}_\theta, \tilde{v}_\theta)$, where $\tilde{q}_\theta$ is the perceived (by agent) baseline probability of success and $\tilde{v}_\theta$ is the perceived marginal effectiveness of effort. Thus, each agent type (defined by index θ) is associated with a belief pair $(\tilde{q}_\theta, \tilde{v}_\theta)$.

We define a reference θ^*-type for all efforts $e \in \{0, 1\}$, corresponding to a **rational agent (or unbiased)**, whose beliefs match the true probabilities (q, v) known to the principal, i.e., $\tilde{q}_{\theta^*} = q$ and $\tilde{v}_{\theta^*} = v$, so that $\tilde{q}_{\theta^*} + \tilde{v}_{\theta^*}e = q + v.e$. An agent is said to be **overconfident** if she overestimates the likelihood of success:

$$\tilde{q}_\theta + \tilde{v}_\theta e \geq q + ve.$$

To establish the type structure, we consider a finite type space $\Theta = \{L, H\}$, where $\theta = H$ denotes a more overconfident agent than $\theta = L$. Formally,

$$\tilde{q}_H + \tilde{v}_H e > \tilde{q}_L + \tilde{v}_L e \geq q + ve \quad \text{for all } e \in \{0, 1\}.$$

Any positive distortions (e.g. biased upward) in the belief pair $(\tilde{q}_\theta, \tilde{v}_\theta)$ relative to the true values (q, v) correspond to overconfidence. We refer to an agent as **slightly overconfident** if her belief distortion is *small* and as **significantly overconfident** if the distortion is *large*. In our setting, the *low type* reflects slight overconfidence, whereas the *high type* reflects significant overconfidence.

Interpretation. Higher values of distortion are associated with *overconfident overall* beliefs (defined as $\tilde{p}_\theta(e) = \tilde{q}_\theta + \tilde{v}_\theta e$) in either dimension, for every possible effort level $e \in \{0, 1\}$:

- $q_{\theta'} > q_{\theta^*}$ means that type θ' is more confident about likelihood of success,
- $v_{\theta'} > v_{\theta^*}$ means that type θ' is more confident in the effectiveness of effort.

Belief pairs $(\tilde{q}_\theta, \tilde{v}_\theta)$ is private information known only to the agent, with $\tilde{q} + \tilde{v} < 1$[3]. They remain fixed and do not update during or after contracting. If beliefs were allowed to update in the process of contracting, the agent's perceived success probability would become endogenous and potentially affect her effort choice in response to contract outcomes. As shown by Heidhues et al. (2018), such a setting would require a dynamic learning model and introduce strategic considerations for the principal, who could manipulate belief updating over time. We restrict attention to fixed beliefs that remain constant throughout contracting, thereby isolating the effects of ex-post belief heterogeneity from dynamic learning. This assumption allows us to isolate the effect of ex-post belief heterogeneity on contract design without introducing the complexities of dynamic learning.

[3] By ensuring both beliefs and truth satisfy $\tilde{q} + \tilde{v} < 1$, this guarantees that both success and failure remain possible. That is what preserves the integrity of incentive design, which is that agents must evaluate success and failure, and the principal must actually design contracts to manage this uncertainty.

2.2 Timing

We define a simple principal-agent contracting model with discrete time $t \in \{0, 1, 2, 3\}$ to investigate a form of non-standard wage (payment) discrimination resulting from belief heterogeneity. First, at $t = 0$, agents privately shape their subjective beliefs for the chance of success (whether rational, slightly overconfident, or significantly overconfident), which specifies their types. At $t = 1$, the principal offers a menu of contracts for screening, from which the agent selects according to her private belief (type). At $t = 2$, the agent accepts or rejects the contract, with rejection leading to the outside option, which is normalized to zero: $U_0 = 0$. At $t = 3$, the agent exerts effort conditional on the chosen contract and the realized outcome determines the payment, s_1 in the case of success and s_0 otherwise (moral hazard).

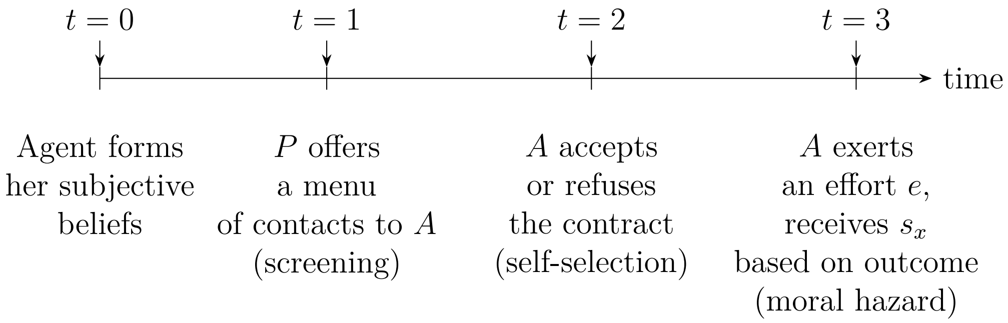
$t = 0$
$t = 1$
$t = 2$
$t = 3$
time
Agent forms her subjective beliefs
P offers a menu of contacts to A (screening)
A accepts or refuses the contract (self-selection)
A exerts an effort e, receives s_x based on outcome (moral hazard)

Figure 1: Timing of the Contractual Game Under Adverse Selection and Moral Hazard

2.3 Agent Preferences and Contract Evaluation

The project yields a binary outcome $x \in \{x_0, x_1\}$, where $x_1 \in \mathbb{R}_{++}$ denotes success and $x_0 \in \mathbb{R}_+$ denotes failure, with $x_1 > x_0$. These outcomes are publicly observable and form the basis for the agent's compensation for exerting effort: the agent receives s_1 if the project succeeds ($x = x_1$) and s_0 if it fails ($x = x_0$).

We define a contract as a pair[4] $s_x = (s_0, s_1) \in \mathbb{R}^2_+$, where the contract space, denoted S, is the set of all such outcome-contingent, non-negative transfers:

$$\mathcal{S} := \left\{ (s_0, s_1) \in \mathbb{R}^2_+ \,\middle|\, s_0 \geq 0, \ s_1 \geq 0 \right\}.$$

This set represents the principal's action space in the contract design problem and forms

[4] For simplicity, the notation omits type indices $\theta \in \Theta$ at this stage. They will be introduced later as $s_x^\theta = (s_0^\theta, s_1^\theta)$ when adverse selection is added to the model in Section 4.

the domain over which agent preferences, incentive constraints and participation constraints are defined.

We focus on contracts of the form $(s_0, s_1) \in \mathbb{R}^2_+$, where payments depend only on the observable outcome. Since the project outcome is the only mutually observable signal, restricting attention to outcome-contingent payments entails no loss of generality, as the principal cannot do better by conditioning on unobservable effort.

The risk-averse agent receives utility from monetary transfers (wages) $s_x \in \mathcal{S}$ and disutility from effort e, denoted by $c(e)$. We assume that the utility function is $U(s, e) = u(s) - c(e)$, where u characterises such as where

$$u(s_x) = \ln(s_x)$$

with $u(\cdot)$ being increasing and concave [5] $(u'(\cdot) > 0,\ u''(\cdot) < 0)$ and $x \in \{0, 1\}$. They exert effort $e \in \{0, 1\}$ at cost $c(e)$.

Contract Evaluation Under Biased Beliefs; Given effort level $e \in \{0, 1\}$, the agent evaluates contracts according to her subjective expected utility. For any contract $(s_0, s_1) \in \mathcal{S}$, the agent's perceived expected utility when planning to exert effort level e is:

$$U(s_0, s_1, e) = (\tilde{q} + \tilde{v}e)u(s_1) + (1 - \tilde{q} - \tilde{v}e)u(s_0) - c(e), \tag{1}$$

where utility is evaluated according to her biased belief $(\tilde{q}, \tilde{v})$. This expression makes explicit the key asymmetry in the model: the agent uses subjective probabilities $\tilde{p}(e) = \tilde{q} + \tilde{v}e$ to weight outcomes, rather than the objective probabilities $p(e) = q + ve$ known to the principal.

2.4 Constraints

This section specifies the conditions that must be satisfied for the principal's contract to be feasible and incentive compatible by covering individual rationality (IR), incentive compatibility for adverse selection (IC-AS) and incentive compatibility for moral hazard (IC-MH).

Notation. Hence, the principal designs the contract menu for screening, we use $s^\theta = (s^\theta_0, s^\theta_1)$ to denote the contract intended for type $\theta \in \{L, H\}$. Specifically:

[5] The logarithmic utility function $u(s) = ln(s)$ exhibits constant relative risk aversion (CRRA) with coefficient equal to 1.

(i) s_1^θ: payment to the θ-type agent upon success ($x = x_1$);

(ii) s_0^θ: payment to the θ-type agent upon failure ($x = x_0$);

(iii) $e^\theta \in \{0, 1\}$: effort level that contract s^θ is designed to implement.

Note that each contract $s^\theta = (s_0^\theta, s_1^\theta)$ can be decomposed into an effort element e^θ and a risk-sharing element $\Delta s^\theta := s_1^\theta - s_0^\theta$. The constraints below determine feasible $(e^\theta, \Delta u^\theta)$ pairs. This decomposition is formalized in Section 4.1 to characterize optimal screening.

Individual Rationality (IR) The contract must yield utility at least as high as the outside option, which is defined as $U_0 = 0$, for $\theta \in \{L, H\}$ and $e \in \{0, 1\}$[6] :

$$U_\theta(s_0^\theta, s_1^\theta, e) = (\tilde{q}_\theta + \tilde{v}_\theta e)u(s_1^\theta) + (1 - (\tilde{q}_\theta + \tilde{v}_\theta e))\, u(s_0^\theta) - c(e) \geq 0 \qquad (IR_\theta)$$

Incentive Compatibility for Adverse Selection ($(IC - AS)$) Each agent must prefer the contract intended for their own type over others, for all $\phi \in \{L, H\}$ and $e' \in \{0, 1\}$:

$$\begin{aligned} U_\theta(s_0^\theta, s_1^\theta, e) &= (\tilde{q}_\theta + \tilde{v}_\theta e)u(s_1^\theta) + (1 - (\tilde{q}_\theta + \tilde{v}_\theta e))\, u(s_0^\theta) - c(e) \\ &\geq (\tilde{q}_\theta + \tilde{v}_\theta e')u(s_1^\phi) + (1 - (\tilde{q}_\theta + \tilde{v}_\theta e'))\, u(s_0^\phi) - c(e') \\ &= U_\theta(s_0^\phi, s_1^\phi, e') \qquad (IC - AS_\theta) \end{aligned}$$

i.e., the high type prefers to report truthfully and pick (s_0^H, s_1^H) over mimicking the low-type contract where $\phi = L$. *For the significantly overconfident (high) type H:*

$$U_H(s_0^H, s_1^H, e) \geq U_H(s_0^L, s_1^L, e'), \qquad (IC - AS_H)$$

i.e., the low type prefers to report truthfully and pick (s_0^L, s_1^L) over mimicking the high-type contract where $\phi = H$. *For the slightly overconfident (low) type L:*

$$U_L(s_0^L, s_1^L, e) \geq U_L(s_0^H, s_1^H, e'), \qquad (IC - AS_L)$$

Incentive Compatibility for Moral Hazard ($(IC$-$MH)$) To induce the agent to exert the intended effort level $e = 1$ that mitigates moral hazard, the contract must

[6] For simplicity, the notation omits type indices $\theta \in \Theta$ at this stage. They will be introduced later as e^θ in Section 4.

satisfy the incentive-compatibility condition:

$$\begin{aligned} U_\theta(s_0^\theta, s_1^\theta) &= (\tilde{q}_\theta + \tilde{v}_\theta)u(s_1^\theta) + \left[1 - (\tilde{q}_\theta + \tilde{v}_\theta)\right] u(s_0^\theta) - c \\ &\geq \tilde{q}_\theta u(s_1^\theta) + \left[1 - \tilde{q}_\theta\right] u(s_0^\theta) = U_\theta(s_0^\theta, s_1^\theta) \end{aligned}$$

which simplifies to the condition

$$\tilde{v}_\theta \left(u(s_1^\theta) - u(s_0^\theta) \right) \geq c, \qquad (IC - MH_\theta)$$

(IC-MH) where s_1^θ and s_0^θ denote the payments upon success and failure, respectively, for θ-type and c represents the agent's cost of exerting effort. Note that IC-MH is a special case of IC-AS when $\phi = \theta$, where the agent compares different effort levels under her own contract rather than across types, but we state it separately for clarity.

The constraints jointly ensure (i) under **IC-AS**, a θ-type agent prefers the contract (s_0^θ, s_1^θ) designed for her own type over contracts intended for any other type ϕ and (ii) under **IC-MH**, the agent prefers to exert the intended effort level e rather than deviating to any alternative effort $e' \in \{0, 1\}$. Together, these conditions prevent both type misreporting of the private information and effort deviations by ensuring truthful self-selection and optimal effort choice.

3 Equilibrium Contracts under Moral Hazard

This section introduces the baseline model without adverse selection. In this baseline model, the agent chooses effort optimally given the contract and her biased belief profile $(\tilde{q}_\theta, \tilde{v}_\theta)$. This benchmark specifies how belief heterogeneity affects the structure of optimal incentive contracts by changing optimal risk-sharing under moral hazard (as analyzed by De la Rosa (2011)), before introducing the additional complexity of hidden types in Section 4.

To understand how belief heterogeneity affects optimal risk-sharing in the context of moral hazard, it is necessary to characterize the first-best implementation of any given effort level and the second-best implementation within a binary-action framework, where the principal makes a *take-it-or-leave-it* contract offer to the agent, who chooses the effort level $e \in \{0, 1\}$. In this setup, the principal's expected profit, [7]

$$\mathbb{E}[\Pi \mid e] = (q + ve)(x_1 - s_1) + [1 - (q + ve)](x_0 - s_0),$$

The agent accepts the contract only if her perceived expected utility is at least as high as her outside option $U_0 = 0$. This yields the individual rationality (IR) constraint:

$$(\tilde{q} + \tilde{v}e)u(s_1) + (1 - (\tilde{q} + \tilde{v}e))\, u(s_0) - c(e) \geq 0 \tag{IR}$$

When effort is not contractible in the second-best setting, the contract must also satisfy the incentive compatibility constraint under moral hazard (IC-MH):

$$\tilde{v}\,(u(s_1) - u(s_0)) \geq c, \tag{IC-MH}$$

This formulation shows that the principal must offer the contract so that both participation and incentive compatibility (effort provision) are satisfied under heterogeneous beliefs, De la Rosa (2011). To achieve this, the principal offers outcome-contingent payments (s_1, s_0) across success and failure states to maximize expected profit subject to these constraints. The complete derivation of *Principal's Optimization under Belief Heterogeneity* is provided in Appendix on page ix.

Even when effort is **observable** (*first-best*), disagreement in beliefs between agent and

[7] In the screening setup, μ_θ denotes the principal's ex-ante belief about the agent's type $\theta \in \{L, H\}$. Under pure moral hazard with a single agent, no type distribution is required and μ_θ is omitted.

principal leads to imperfect risk-sharing; relative to homogeneous beliefs, overconfident agents are more willing to accept risk because they overestimate the probability of the state of success. The following result, which extends Borch (1962) classic rule for *Pareto-optimal risk-sharing* to environments with heterogeneous beliefs, characterizes the first-best contract under heterogeneous beliefs distortion.

Proposition 1 (**First-Best Contract under Biased Beliefs, De la Rosa (2011)**). *A contract* $\langle s_1^{FB,e}, s_0^{FB,e}\rangle$ *is said to be an optimal first-best contract, which implement* $e \in \{0,1\}$, *if it satisfies the individual rationality (IR) condition*

$$(\tilde{q} + \tilde{v}e)\, u(s_1^{FB,e}) + \Big[1 - (\tilde{q} + \tilde{v}e)\Big]\, u(s_0^{FB,e}) - c(e) = 0,$$

and the following risk-sharing condition

$$\frac{\tilde{q} + \tilde{v}e}{1 - (\tilde{q} + \tilde{v}e)} \cdot \frac{u'(s_1^{FB,e})}{u'(s_0^{FB,e})} = \frac{q + ve}{1 - (q + ve)}.$$

All formal proofs are deferred to the appendix, where the complete derivations are provided. In the first-best implementation of any effort level, the principal chooses effort $e \in \{0,1\}$ and payment scheme directly, no need for incentive compatibility (IC), since there's no moral hazard. However, belief heterogeneity still affects the form of the contract, because the agent's perception of success affects how much insurance she wants. While effort is chosen directly by the principal, the contract must still account for the agent's beliefs to satisfy individual rationality. Overconfidence thus *distorts* the classical optimal insurance structure (under homogeneous beliefs) even in the absence of moral hazard.

When effort is **unobservable** (second-best) the principal must rely on incentives rather than direct effort enforcement. Thus, the impact of biased beliefs then becomes crucial in the second-best implementation, which is highlighted in the following.

Proposition 2 (**Second-Best Contract under Biased Beliefs, De la Rosa (2011)**). *A contract* $\langle s_1^{SB}, s_0^{SB}\rangle$ *is said to be an optimal second-best contract that implements effort* $e = 1$ *if the following conditions hold:*

(i) For a ***slightly overconfident agent****, the individual rationality (IR) constraint binds with equality and the incentive compatibility (IC) constraint also binds, i.e.*

$$(\tilde{q} + \tilde{v})u(s_1^{SB}) + \Big[1 - (\tilde{q} + \tilde{v})\Big]u(s_0^{SB}) - c = 0, \quad \textit{(IR)}$$

$$\tilde{v}\Big(u(s_1^{SB}) - u(s_0^{SB})\Big) = c. \quad \textit{(IC)}$$

(ii) For a ***significantly overconfident agent****, the individual rationality (IR) constraint binds with equality, but the incentive compatibility (IC) constraint is slack. In this case, the second-best contract coincides with the first-best contract, i.e.*

$$(\tilde{q} + \tilde{v})u(s_1^{SB}) + \Big[1 - (\tilde{q} + \tilde{v})\Big]u(s_0^{SB}) - c = 0, \quad \textit{(IR)}$$

$$\tilde{v}\Big(u(s_1^{SB}) - u(s_0^{SB})\Big) > c. \quad \textit{(IC)}$$

In the second-best effort implementation, the contract must satisfy incentive compatibility to ensure the desired effort is exerted through outcome-contingent payments. With *slight* overconfidence, both the individual rationality (IR) and incentive compatibility (IC) constraints bind. One observes that this coincides with a standard moral hazard tradeoff between insurance and incentives (see Grossman and Hart (1983)): *the risk averse agent cannot be fully insured and must be provided payments to maintain incentive compatibility*[8]. On the other hand, being a *significantly* overconfident agent results in that incentive compatibility (IC) constraint being slack. As a result, this leads to effort being exerted voluntarily, and, information rents are eliminated entirely. More precisely, the second-best contract that intends to implement effort provides excessively powerful incentives in the sense that they are more powerful than necessary to implement given effort; thus, the (IC) constraint no longer binds, and no need for information rent to induce desired effort is that case.

When overconfidence is sufficiently high, her biased beliefs about effort effectiveness replace the usual *incentive-insurance mechanism* in determining the optimal contract under moral hazard. As the agent already believes that her exerted effort almost guarantees the state of success, she is motivated without requiring any extra payment (information rents), which is highly dependent on performance. This feature is one of the main insights of De la Rosa (2011), as it reveals how belief distortions can reduce the cost of implementing effort, even in the presence of moral hazard. In this case, the agent exerts desired effort even in the absence of a strong incentive.

The discussion above shows how the extent of overconfidence alters the structure of risk-sharing. To solidify the foundations of this result, it is useful to justify how

[8] In the second-best with identical beliefs, incentives require outcome-contingent payment to maintain: $v\,[u(s_1) - u(s_0)] = c$. This exposes the risk-averse agent to income risk, so to satisfy participation, it causes an extra payment called the *cost of agency*, the inefficiency from hidden action and risk aversion.

overconfidence reshapes the trade-off between incentives and insurance in moral hazard: stronger incentive provision reduces insurance, and more insurance weakens incentives to exert desired effort. Overconfidence distorts the structure of this incentive-insurance tradeoff through two different tools: the *wager effect* and the *incentive effect.*

Proposition 3 (**Wager Effect of Overconfidence, De la Rosa (2011)**). *In a first-best setting (no moral hazard), the optimal contract* $\langle s_1^{FB,e}, s_0^{FB,e}\rangle$ *becomes riskier as overconfidence and optimism increases. Formally,*

$$\frac{d}{d\tilde{q}}\Big(u(s_1^{FB}) - u(s_0^{FB})\Big) > 0 \quad \textit{and} \quad \frac{d}{d\tilde{v}}\Big(u(s_1^{FB}) - u(s_0^{FB})\Big) > 0,$$

when

$$\frac{1}{u'(s)}\left(-\frac{u''(s)}{u'(s)}\right)$$

is a non-decreasing function.

Under belief heterogeneity, the principal takes advantage of the agent's willingness to bear risk by the wager effect, even if the agent is *risk-averse.* Since an agent who is more confident in success values outcome-contingent payments more likely. The principal reduces insurance (i.e., increasing performance-based payment), while still satisfying individual participation. The following corollary formally demonstrates how optimism, even without the effect of chosen level effort (i.e., the incentives are not mainly driven to induce effort), can result in *riskier* payment structures.

Corollary 1 (**Optimism implies risk-bearing**). *If the agent is relatively overall overconfident,* $\tilde{q} > q$ *and* $\tilde{v} > v$*, then, in equilibrium, it is first-best optimal to shift payment toward success, which implies* $s_1^{FB,e} > s_0^{FB,e}$*, thereby exposing the agent to risk.*

Note that in the case of homogeneous beliefs, the first-best result implies $s_1^{FB,e} = s_0^{FB,e}$. Under heterogeneous beliefs, the optimal risk-sharing condition is given by $\frac{\tilde{q}+\tilde{v}e}{1-(\tilde{q}+\tilde{v}e)} = \frac{q+ve}{1-(q+ve)}$, which yields "*full insurance*" such that $s_1^{FB,e} = s_0^{FB,e} \equiv s^{FB,e}$ such that $u(s^{FB,e}) = \bar{u} + c(e)$, (see De la Rosa (2011)). However, when the agent is overall overconfident, the principal then offers a contract that rewards success with a greater payment level, which increases the agent's expected payment according to his own biased beliefs, thus the agent willingly accepts this additional risk due to her biased belief about success. As a result, the wager effect pushes the equilibrium contract towards higher-powered incentives in the case of agent overconfidence. This represents a fundamental shift to classical risk-sharing[9]: belief heterogeneity alone results in inefficient insurance. In this sense, overconfidence substitutes for risk aversion.

When effort is not observable, and the agent is biased (e.g., overconfident), the principal must design a contract that motivates effort. This effect is formally captured in the following result of De la Rosa (2011):

Proposition 4 (Incentive Effect of Overconfidence, De la Rosa (2011)). *In a second-best setting, the contract $\langle s_1^{SB,e}, s_0^{SB,e} \rangle$ that implements effort becomes less risky in the degree of agent overconfidence, such that*

$$\frac{d}{d\tilde{v}} \left(u(s_1^{SB,e}) - u(s_0^{SB,e}) \right) < 0$$

whereas agent optimism does not affect the power of incentives;

$$\frac{d}{d\tilde{q}} \left(u(s_1^{SB,e}) - u(s_0^{SB,e}) \right) = 0$$

when

$$\frac{1}{u'(s)} \left(\frac{-u''(s)}{u'(s)} \right) \quad \text{is a non-decreasing function.}$$

When the agent overestimates the effectiveness of effort[10], based on *incentive effect*, the agent requires lower-powered incentives to be motivated to exert desired effort. Consequently, any contract that implements desired effort necessarily exposes the agent to risk, but under slight overconfidence, the incentive-insurance tradeoff still remains a binding concern (i.e. incentive compatibility (IC) binds), meaning the principal must still assign some incentives to satisfy incentive compatibility. However, because the agent overestimates the likelihood of exerting effort, the required payments spread is smaller, so the contract entails more insurance than under unbiased beliefs. The incentive effect,thus, dominates the wager effect in this scenario by ensuring that the principal can reduce the cost of inducing effort while still inducing the target effort level. The following remark clarifies this result.

Remark 1. *This follows directly from Proposition 4 by noting that $\tilde{v} > v$, the incentive-compatibility condition is*

$$\tilde{v} \left(u(s_1) - u(s_0) \right) = c,$$

[9] In a first-best (efficient) contract under homogeneous beliefs, a risk-neutral principal would bear all the risk, and a risk-averse agent would be fully insured $s_1^{FB,e} = s_0^{FB,e}$.

[10] Optimism, in contrast, does not affect the structure of incentives in this setting.

which implies

$$u(s_1) - u(s_0) = \frac{c}{\tilde{v}}.$$

Since $\tilde{v} > v$, it follows that

$$\frac{c}{\tilde{v}} < \frac{c}{v}.$$

Thus, the minimum payment spread needed to induce effort is smaller than under the unbiased benchmark. This allows the principal to provide more insurance while still satisfying incentives, thereby lowering the expected cost of implementation. Slight overconfidence therefore reduces the cost of agency by making effort cheaper to sustain.

Combining the wager and incentive effects characterizes a key insight from De la Rosa (2011): the power of incentives in second-best contracts is non-monotonic in the degree of overconfidence. While slight overconfidence *relaxes* the incentive constraint (becomes easier to satisfy) and reduces the cost of implementing desired effort, significant overconfidence makes the constraint *slack*, leading to riskier contracts that expose the agent to a greater gap in outcome-contingent payments.

Observation 1 (Non-Monotonicity of Incentive Power). *Let **the power of incentive** (simply the gap in the outcome-contingent payment) be defined as*

$$\Delta u(\tilde{v}) := u(s_1^{SB}, e) - u(s_0^{SB}, e),$$

*where $\tilde{v}$ is the agent's perceived marginal return to effort. Then $\Delta u(\tilde{v})$ is **non-monotonic** in $\tilde{v}$:*

*(i) It **decreases** when $\tilde{v}$ is slightly above v:*

$$\frac{d}{d\tilde{v}} \Delta u(\tilde{v}) < 0,$$

*(ii) and **increases** when $\tilde{v}$ becomes sufficiently large:*

$$\frac{d}{d\tilde{v}} \Delta u(\tilde{v}) > 0.$$

Justification. Observation 1 follows directly from Proposition 3 and 4. When the agent is only slightly overconfident, both (IR) and (IC) bind, and the binding (IC) constraint as follows

$$\tilde{v}\Big(u(s_1^{SB}) - u(s_0^{SB})\Big) = c$$

that yields

$$\frac{d}{d\tilde{v}}\Delta u(\tilde{v}) < 0,$$

which reflecting the standard *incentive effect.* However, when the agent becomes significantly overconfident (such an increase in the $\tilde{v}$), the (IC) constraint becomes slack, so given effort is exerted voluntarily and $\Delta u(\tilde{v})$ rises with $\tilde{v}$, the *wager effect.* Hence, the power of incentives that is defined by $\Delta u(\tilde{v})$ first decreases, and then increases in the parameter $\tilde{v}$ that satisfies a non-monotonicity relationship.

This relationship between power of incentive and belief profile (more specifically parameter $\tilde{v}$) carries a significant implication. When belief heterogeneity is present, it can even enable the principal to induce effort level that exceed those possible in the first-best implementation. Despite an outcome-contingent payment often serving to overcome moral hazard, sufficiently higher (or significantly) overconfidence allows the principal to achieve effort without providing any marginal incentives (extra payment for success as an information rent) by eliminating agency costs altogether.

Observation 2. ***(Effort Exceeds First-Best under Significant Overconfidence)*** *When the agent is sufficiently overconfident (i.e., $\tilde{v} \gg v$), the incentive compatibility constraint becomes slack. As a result, the principal can achieve or even exceed first-best effort levels under second-best implementation without incurring the usual incentive costs.*

Justification. This observation follows directly from Proposition 3. When the agent is exhibiting significantly overconfidence, she perceives that their effort will produce much better results than it actually will. In this case, the second-best contract that implements effort applies excessively powerful incentives in the sense that they are more powerful than necessary to implement desired effort (i.e. the (IC) constraint no longer binds). As a result, effort is exerted voluntarily, and the principal no longer needs to distort payments to induce it by providing information rents.

In such settings, belief distortions become a strategic tool that the principal can *exploit* by adapting the incentive structure to the agent's belief-driven effort. As a result, even though incentives are no longer needed to motivate effort, the *riskier* contracts turn *exploitative*[11]: The principal extracts first-order gains from a lower expected payment, while the agent incurs only second-order losses. This asymmetry explains why exploitation is profitable to the principal. The following proposition demonstrates how the principal's

[11] While outcome-contingent payments are typically required to overcome moral hazard, sufficiently strong overconfidence allows the principal to induce effort without offering any marginal incentives, thereby creating an exploitation motive.

minimization of expected payment captures the exploitation motive under the higher overconfidence level. To derive explicit comparative statics, we now adopt mean-variance preferences rather than logarithmic utility.

Proposition 5 (**Exploitation Motive under Significantly Overconfidence**). *Let $q+v$ be the true success probability under $e=1$ and $\tilde{q}+\tilde{v}$ the agent's belief, and $\tilde{v} \gg v$. The principal is risk-neutral and pays the expected transfer (payment) to the agent* [12]. *In this case, the principal's problem is*

$$min \ (q+v)s_1 + (1-q-v)s_0,$$

s.t.

$$(\tilde{q}+\tilde{v})\, s_1 + \Big(1-(\tilde{q}+\tilde{v})\Big)\, s_0 - (\tilde{q}+\tilde{v})\Big(1-(\tilde{q}+\tilde{v})\Big)\,(s_1-s_0)^2 \geq 0.$$

Solving this optimization gives the optimal payment spread:

$$s_1 - s_0 \ = \ \frac{(\tilde{q}+\tilde{v})-(q+v)}{2(\tilde{q}+\tilde{v})(1-\tilde{q}-\tilde{v})} \ \equiv \ \Delta^*. \tag{1}$$

Given $s_H - s_L$, the principal sets the levels of the two payments so that PC binds.

In the given formulation, the payment spread depending on the range between the biased belief and the true probability, equation (1) reproduces a basic *exploitation motive* identified by Köszegi (2014) for overconfident agents: if the agent is too overconfident about that success will occur, so the principal can decrease expected payments by paying her a little more for high outcome and much less for low outcome. Under classical moral hazard, the principal offers a contract as a remedy to induce desired effort, but here it serves instead to exploit the agent's distorted beliefs. In particular, the principal adjusts the payment spread to minimize the *expected payment* from her perspective, exploiting the agent's bias, while ensuring that the (IR) remains binding.

When overconfidence is only slight, the principal's ability to exploit biased belief is not plausible implying that in such environments, slightly overconfident agents are unlikely to be offered exploitative contracts. To verify that, recall the incentive-compatibility constraint (IC-MH), and let Δ^{**} denote the minimal payment spread necessary to satisfy

[12] In this proposition, we adopt mean-variance preferences of the form $\tilde{p}s_1 + (1-\tilde{p})s_0 - \tilde{p}(1-\tilde{p})(s_1-s_0)^2$, following Köszegi (2014). This differs from the logarithmic utility $u(s) = \ln(s)$ used elsewhere but provides closed-form solutions. The mean-variance formulation can be viewed as a local second-order Taylor approximation to logarithmic utility around mean pay $\bar{s}$: $\mathbb{E}[\ln(s)] \approx \ln(\bar{s}) - \frac{1}{2\bar{s}^2}\mathrm{Var}(s)$. The qualitative insight, that overconfidence enables exploitation, holds under both specifications.

this constraint (IC) in order to incentivise the agent for desired effort. Indeed, the IC condition requires that, where $\tilde{v}$ represents the agent's perceived belief about the marginal effect of effort on the likelihood of success;

$$\tilde{v}\Big(u(s_1) - u(s_0)\Big) = c,$$

The comparison of Δ^* and Δ^{**} determines the structure of the contracts in terms of the power of incentives. If $\Delta^* > \Delta^{**}$, then the optimal contract sets $s_1 - s_0 = \Delta^*$, in which case the IC constraint is *slack*, since the payment spreads more power to implement effort (i.e. higher-powered incentives). The contract has therefore primarily exploitative motive, as its main purpose is to take advantage of the agent's mistaken beliefs. By contrast, if $\Delta^* < \Delta^{**}$, then the optimal contract sets $s_1 - s_0 = \Delta^{**}$, ensuring that the IC constraint is satisfied at the optimum. In this case, the contract is not primarily exploitative but instead mandated to induce desired effort level by the principal.

To formalize the differences between the contracts' forms under biased beliefs, we first define the notion of an *incentive-compatible contract*, which provides the minimal payment spread Δ^{**} (lower-powered incentives) required to induce desired effort, and contrast it with an *exploitative contract*, which exposes the agent to excessive risk (higher-powered incentives) by exploiting perceived belief about the success.

Definition 1 (**Incentive-Compatible Contract**). *A contract is said to be an* ***incentive-compatible contract*** *if the payment spread $s_1 - s_0 = \Delta^{**}$ is set at the minimal level required to satisfy the agent's incentive-compatibility (IC) constraint. Formally,*

$$u(s_1) - u(s_0) = \frac{c}{\tilde{v}},$$

so that the contract provides just enough performance-based pay to induce desired effort without any surplus extraction from biased beliefs. Its purpose is solely to motivate desired effort as a remedy for moral hazard.

Observe that the incentive-compatible contract is consistent with the second-best implementation contract used to induce desired effort under hidden action when the agent is slightly overconfident (see Proposition 2). It provides exactly the minimum outcome-contingent payment (sufficient to induce effort) by providing more insurance that is required to satisfy the agent's IC constraint by balancing the tradeoff between incentives and insurance.

Definition 2 (**Exploitative Contract**). *A contract is said to be an* ***exploitative*** *if the payment spread $s_1 - s_0 = \Delta^*$ exceeds the level required to satisfy the agent's incentive-compatibility (IC) constraint. Formally,*

$$u(s_1) - u(s_0) > \frac{c}{\tilde{v}}, \tag{2}$$

and the agent's indirect utility from exerting high effort $e = 1$ is;

$$U(s_0, s_1, 1) = (\tilde{q} + \tilde{v})u(s_1) + \Big(1 - (\tilde{q} + \tilde{v})\Big)u(s_0) - c \geq 0. \tag{3}$$

while the corresponding direct (objective) utility satisfies,

$$(q + v)u(s_1) + (1 - q - v)u(s_0) - c < 0. \tag{4}$$

Observe that the *exploitative contract* is identical to the first-best implementation under disagreement (see Proposition 2 and Corollary 1), when the contract implements $e = 1$, because the agent's bias removes the incentive constraint entirely. When $s_1^{FB} > s_0^{FB}$, we have

$$u(s_1^{FB}) - u(s_0^{FB}) > \frac{c}{\tilde{v}}.$$

which satisfies the condition of an exploitative contract. In this case, the incentive-compatibility constraint is not binding, only the participation constraint binds, and effort is implemented without additional *payment* (rent) because the agent's overconfidence makes him bear more risk than is strictly necessary.

Overall, the aim of offering the incentive-compatible contract is not for exploitation because it arises from mitigating the moral hazard problem. In contrast, when overconfidence itself provides sufficient motivation, offering the exploitative contract arises from eliminating the cost of agency. These differences demonstrate how belief heterogeneity systematically affects the structure of optimal contracts in moral hazard. Since overconfidence raises the agent's perceived belief level about the likelihood of success, it lowers the principal's cost of agency and thereby affects both the structure and efficiency of the second-best implementation. In the following section, we will use these results in the presence of hidden types to investigate how overconfidence shapes incentives and contract design when screening becomes relevant.

4 Screening Overconfidence with Moral Hazard

In the De la Rosa (2011)'s analysis, the principal holds perfect information about the agent's biases on the likelihood of success. However, once adverse selection is introduced, perfect information no longer holds, since contract choice depends on private information about the agent's type, which specifies his biased belief profile. Thus, the principal must design contracts that both incentivize effort and screen cognitive heterogeneity in beliefs in a setting with both hidden types and hidden action are relevant.

The present analysis builds on the standard contract-theoretic framework in which adverse selection is followed by moral hazard: we consider a situation where moral hazard affects the random profit of a risk-neutral principal. Thus, the agent first selects a contract based on private information about her type, and subsequently exerts the implemented level of effort that is specified by the contract. We adopt this model by allowing the agent first selects a contract (s_0^θ, s_1^θ) based on private information of type $\theta \in \{L, H\}$, which gives rise to **screening** concerns, and then exerted the desired effort $e^\theta \in \{0, 1\}$ which is associated to moral hazard problem. Since agents' private belief profiles $(\tilde{q}_\theta, \tilde{v}_\theta)$ differ across types $\theta \in \{L, H\}$, the principal cannot directly observe these beliefs and to maximize the expected profit, she must design a screening menu of contracts $\{(s_0^\theta, s_1^\theta, e^\theta)\}_{\theta \in \{L,H\}}$ that induces *self-selection* by type.

4.1 Principal's Problem and Feasible Menu

In the problem with adverse selection, the principal aims to maximize her profit by optimally designing a menu $\{(s_0^\theta, s_1^\theta, e^\theta)\}_{\theta \in \{L,H\}}$ that satisfies the moral hazard (MH), the individual rationality (IR) and incentive compatibility (IC) constraints. Moral hazard requires that the effort level e^θ can be implemented through the contract (s_0^θ, s_1^θ) by ensuring that the agent finds exerting the desired effort e^θ given her biased belief profile. Individual rationality mandates that both contracts are non-negative, whereas incentive compatibility ensures that it is optimal for the agent to truthfully report his type and to exert the desired effort level.

Rather than optimizing over the full contract space by simultaneously choosing both payment levels (s_0, s_1) for each θ-type agent, which would require solving an infinite-dimensional functional optimization problem where the function of the realized outcome determined by $s : \{x_0, x_1\} \to \mathbb{R}$, i.e. $s_0 = s(x_0)$, $s_1 = s(x_1)$. We will formulate the set of feasible menus by decomposing a contract into design elements. We will focus on designing these elements instead of designing the contracts directly, which makes the analysis feasible and provides economic intuitions.

Design Elements: are the fundamental components that define the effort induced by the contract and how risk is allocated across two parties; agent and principal. In particular, for each type $\theta \in \{L, H\}$, a contract is fully characterized by the following design elements:

(i) an *effort level element* $e^\theta \in \{0, 1\}$, specifying the level of effort the contract intends (wishes) to implement;

(ii) a *risk-sharing element*; $\Delta s^\theta := s_1^\theta - s_0^\theta$ (equivalent to the payment spread), describing how risk is distributed and thereby determines the power of incentives[13] that the risk-averse agent faces.

This decomposition serves two critical purposes in our screening framework.

First, it makes the analytical problem tractable. By focusing on $(e^\theta, \Delta s^\theta)$ pairs instead of (s_0^θ, s_1^θ) directly, we reduce the dimensionality of the optimization and can characterize the optimal menu through closed-form conditions on the incentive compatibility and participation constraints.

Second, this decomposition enables us to establish the monotone menu structure that distinguishes our screening problem. By characterizing feasible $(e^\theta, \Delta s^\theta)$ pairs for each type, we can directly compare risk exposure across types and show that more overconfident agents must receive riskier contracts, a result that follows from the single-crossing property and would be obscured in the full contract space.

To see the relevant single-crossing property, observe that for an outcome-contingent payment as a contract $s = (s_0, s_1)$[14] and effort $e \in \{0, 1\}$, under $\tilde{p}_\theta(e) = \tilde{q}_\theta + \tilde{v}_\theta e$ the θ-type's perceived expected utility is

$$U_\theta(s, e) = \tilde{p}_\theta(e)\, u(s_1) + \Big(1 - \tilde{p}_\theta(e)\Big)\, u(s_0) - c(e),$$

[13] The payment spread determines the power of incentives $\Delta u^\theta = u(s_1^\theta) - u(s_0^\theta)$, capturing the trade-off between insurance and incentive.

[14] For notational simplicity, we write $s = (s_0, s_1)$ instead of $s^\theta = (s_0^\theta, s_1^\theta)$ in what follows, as the analysis applies to any point in the contract space.

so along an indifference curve $dU_\theta = 0$ its slope in (s_0, s_1) is

$$\frac{ds_1}{ds_0}\bigg|_{dU_\theta=0} = -\,\frac{(1-\tilde{p}_\theta(e))\,u'(s_0)}{\tilde{p}_\theta(e)\,u'(s_1)}.$$

Because $u' > 0$ and $u'' < 0$, this slope is strictly increasing in $(1-\tilde{p}_\theta)/\tilde{p}_\theta$. Hence, when the high type is overall more overconfident, $\tilde{p}_H(e) > \tilde{p}_L(e)$, her indifference curves are *flatter* than those of the low type at every point (s_0, s_1). Indifference curves therefore cross at most once, which delivers single-crossing and the usual monotone (downward) mimicking incentives: the high type values an additional unit of success-contingent pay more than the low type. This is the precise sense in which belief heterogeneity substitutes for differences in risk aversion in standard screening models (see Mirrlees (1971); Myerson (1982)). We now formalize this as Lemma 1.

Lemma 1 (**Single-Crossing with Belief Heterogeneity**). *Consider two types $\theta \in \{L, H\}$ who differ only in subjective beliefs about success, with perceived success probability under effort $e \in \{0,1\}$ given by $\tilde{p}_\theta(e) = \tilde{q}_\theta + \tilde{v}_\theta e$. Assume (i) $u(\cdot)$ is increasing and concave ($u'(\cdot) > 0$, $u''(\cdot) < 0$) and (ii) high-type is overall overconfident in both dimensions than low-type, s.t.*

$$\tilde{q}_H > \tilde{q}_L \quad \textit{and} \quad \tilde{v}_H \geq \tilde{v}_L.$$

For any outcome contingent contract $s = (s_0, s_1)$ and effort $e \in \{0,1\}$, the θ-type perceived expected utility is

$$U_\theta(s,e) = \tilde{p}_\theta(e)\,u(s_1) + \big(1-\tilde{p}_\theta(e)\big)u(s_0) - c(e).$$

Then indifference curves in (s_0, s_1) satisfy the single crossing property: for every (s_0, s_1) and e,

$$\frac{\partial U_H/\partial s_1}{\partial U_H/\partial s_0} > \frac{\partial U_L/\partial s_1}{\partial U_L/\partial s_0}.$$

Proof. By concavity, $\partial U_\theta/\partial s_1 = \tilde{p}_\theta(e)u'(s_1)$ and $\partial U_\theta/\partial s_0 = (1-\tilde{p}_\theta(e))u'(s_0)$. Hence, the indifference-curve slope (the MRS between s_1 and s_0) is

$$\mathrm{MRS}_\theta(s_0, s_1; e) = \frac{\partial U_\theta/\partial s_1}{\partial U_\theta/\partial s_0} = \frac{\tilde{p}_\theta(e)}{1-\tilde{p}_\theta(e)} \cdot \frac{u'(s_1)}{u'(s_0)}.$$

The term $\frac{u'(s_1)}{u'(s_0)}$ is common across types at a given point (s_0, s_1). Since $\tilde{p}_H(e) > \tilde{p}_L(e)$ by

$\tilde{q}_H > \tilde{q}_L$ and $\tilde{v}_H \geq \tilde{v}_L$, we have

$$\frac{\tilde{p}_H(e)}{1-\tilde{p}_H(e)} > \frac{\tilde{p}_L(e)}{1-\tilde{p}_L(e)},$$

with strict inequality whenever at least one of the two is strict. Therefore, $\text{MRS}_H > \text{MRS}_L$, so the high-type indifference curve is everywhere steeper than the low-type curve, and the single crossing property holds. □

Observation 3 (**Monotone Menu**). *Under Lemma 1 and Proposition 5, any incentive compatible separating menu assigns more insurance to the less overconfident type:*

$$0 \leq (s_1^L - s_0^L) \ \leq \ (s_1^H - s_0^H),$$

that is, the contract intended for high-type is more riskier (larger success-failure spread) than the contract intended for low-type.

The feasible menu separates types by offering the high-type agent the exploitative contract with a lack of insurance, which also prevents her from mimicking the low-type agent. For the low-type, the contract includes more insurance to achieve the desired effort, consistent with De la Rosa (2011). This structure follows from single-crossing and thus monotonicity, ensuring that the low-type's incentive-compatibility constraint binds.

This separation mechanism relies on choosing contract design elements, for each type $\theta \in \{L, H\}$, the principal chooses an effort level e^θ and a payment spread Δs^θ. By appropriately selecting design elements $(e^\theta, \Delta s^\theta)$ for each type, the principal constructs a menu to separate types based on biased beliefs. The following definition formalizes which menus can successfully imply this screening.

Definition 3 (**Feasible Menu**). *A menu of contracts $\{(s_0^L, s_1^L, e^L), (s_0^H, s_1^H, e^H)\}$ is said to be* feasible *if it simultaneously satisfies the following conditions, where the agent's subjective success probability is given by $\tilde{p}_\theta(e) = \tilde{q}_\theta + \tilde{v}_\theta e$:*

(i) ***Participation (IR).*** *For each θ-type, expected utility under its own contract must be non-negative:*

$$\tilde{p}_\theta(e^\theta)u(s_1^\theta) + (1 - \tilde{p}_\theta(e^\theta))u(s_0^\theta) - c(e^\theta) \ \geq \ 0.$$

(ii) ***Moral Hazard (IC-MH).*** *For each type $\theta \in \{L, H\}$, the induced effort choice*

$e^\theta \in \{0,1\}$ must solve the best-response problem

$$e^\theta \in \arg\max_{e\in\{0,1\}} \Big[\tilde{p}_\theta(e)u(s_1^\theta) + (1-\tilde{p}_\theta(e))u(s_0^\theta) - c(e)\Big].$$

Equivalently, if $e^\theta = 1$, then the wage spread must satisfy

$$u(s_1^\theta) - u(s_0^\theta) \;\geq\; \frac{c}{\tilde{v}_\theta},$$

while if $e^\theta = 0$, the spread collapses to

$$u(s_1^\theta) - u(s_0^\theta) = 0.$$

(iii) ***Adverse Selection (AS-IC).*** *For each pair $\theta \neq \phi$, θ-type must weakly prefer its own contract to mimicking the contract of type ϕ:*

$$\tilde{p}_\theta(e^\theta)u(s_1^\theta) + (1-\tilde{p}_\theta(e^\theta))u(s_0^\theta) - c(e^\theta) \;\geq\; \tilde{p}_\theta(e_\phi)u(s_1^\phi) + (1-\tilde{p}_\theta(e_\phi))u(s_0^\phi) - c(e_\phi).$$

(iv) ***Monotonicity (follows from Single-Crossing).*** *Incentive compatibility further implies the monotonicity condition*

$$\Big(u(s_1^H) - u(s_0^H)\Big) \geq \Big(u(s_1^L) - u(s_0^L)\Big) \geq 0,$$

ensuring that the high type is always offered at least as strong incentives as the low type.

Define the set of feasible menus[15] by

$$\mathcal{F} := \Big\{\{(s_0^\theta, s_1^\theta, e^\theta)\}_{\theta\in\{L,H\}} : \text{(IC-MH), (IR), (AS-IC), (MON) are all satisfied}\Big\}.$$

Principal's Problem (P); The principal's problem is then to choose a feasible menu $S \in \mathcal{F}$ that maximizes her expected profit:

$$\max_{S\in\mathcal{F}} \mathbb{E}[\Pi_\theta(x)] = \sum_{\theta\in\{L,H\}} \mu_\theta\Big[(q+ve^\theta)(x-s_1^\theta) + \Big(1-(q+ve^\theta)\Big)(x-s_0^\theta)\Big]. \tag{4.1}$$

[15] The feasible set $\mathcal{F}$ is non-empty. A feasible menu can be constructed by offering the High-type the first-best contract under disagreement with $e^H = 1$, and offering the Low-type a flat contract with $s_0^L = s_1^L = 1$ and $e^L = 0$. This menu satisfies all four constraints when overconfidence is ordered ($\tilde{q}_H + \tilde{v}_H > \tilde{q}_L + \tilde{v}_L$) and effort cost is sufficiently small.

Having characterized the feasible set of contracts, the principal's problem now reduces into selecting a menu that maximizes her expected profit while respecting the constraints that ensure truthful self-selection and correct effort provision. The existence of an optimal menu to problem (P) is also guaranteed under standard conditions.[16]. The following result provides a complete characterization of optimal menu structure.

Result 1 (**Optimal Screening Menu Structure**). *In the principal's screening problem with belief heterogeneity and a finite type space $\Theta = \{L, H\}$, where H denotes the significantly overconfident type, and L for slightly overconfident, the structure of the optimal menu $\{(s_0^L, s_1^L, e^L), (s_0^H, s_1^H, e^H)\}$ is characterized by corresponding design elements as follows.*

Case 1: for High-Type Agent (Significantly Overconfident),

The high-type agent is always induced to exert effort, that is, $e^H = 1$. The optimal contract for the high-type takes the form of an exploitative contract as defined in Definition 2, where the payment spread $\Delta s^H := s_1^H - s_0^H$ satisfies:

$$u(s_1^H) - u(s_0^H) > \frac{c}{\tilde{v}_H}.$$

Consequently, $s_1^H > s_0^H$, making this contract riskier than the low-type's contract, consistent with the wager effect established in Proposition 3.

Case 2: Low-Type Agent with Rational Beliefs[17]

If the low-type L exhibits rational beliefs (i.e., $\tilde{q}_L = q$ and $\tilde{v}_L = v$), inducing effort is prohibitively costly due to the risk-aversion motive. The optimal contract is therefore flat:

$$s_0^L = s_1^L =: \bar{s}^L, \quad e^L = 0.$$

This contract provides full insurance with zero incentive provision and satisfies the individual rationality constraint with equality:

$$u(\bar{s}^L) = 0.$$

Case 3: Low-Type Agent with Slight Overconfidence

[16] Formally, existence follows from the being nonempty and compact of the feasible set $\mathcal{F}$ and the continuity of the principal's expected profit function in $(s_0^\theta, s_1^\theta, e^\theta)$. Since the agent's utility function $u(\cdot)$ is continuous and strictly concave, and the type space $\Theta = \{L, H\}$ is finite, the maximization problem admits at least one solution (see Myerson (1982); Laffont and Martimort (2002))

[17] Following De la Rosa (2011): when overconfidence is insufficient, inducing effort is too costly and thus $e = 0$ optimal.

If the low-type L is slightly overconfident, the principal may optimally implement an incentive-compatible contract that induces effort, that is, $e^L = 1$. In this case, the resulting contract coincides with the characterization in Definition 1, where the payment spread is set at the minimal level required by incentive compatibility:

$$u(s_1^L) - u(s_0^L) = \frac{c}{\tilde{v}_L}.$$

This characterization makes clear that the nature of the biased beliefs critically shapes the contract design in screening: rational agents receive full insurance with no incentives, while slightly overconfident agents are induced to exert effort through an incentive-provision contract. The high-type, by contrast, is always induced to provide effort under an exploitative, riskier contract. These insights provide the foundation for the concluding discussion, where the broader implications of belief heterogeneity for contract theory are examined. A detailed summary of the optimal contracts in terms of their design elements is provided in Appendix A.3.

5 Conclusion

We attempt to examine an optimal screening contract design that involves both moral hazard and adverse selection when principals face agents with heterogeneous beliefs about success probabilities and effort effectiveness. By revisiting the canonical moral hazard framework of De la Rosa (2011), we shows that overconfidence has two different effects: the *wager effect*, which supports bearing risk voluntarily, and the *incentive effect*, which reduces the payment spread necessary to induce desired effort. When overconfidence is sufficiently high, the incentive constraint becomes slack (wager effect dominates), enabling exploitation: principals offer the contract riskier (i.e. exploitative contract) than necessary. When overconfidence is slight, at lower agency costs than under rational beliefs, the contract (i.e. incentive-compatible contract) balances the trade-off between insurance and incentive since incentive compatibility is still a binding concern (incentive effect dominates).

Carrying over the moral hazard framework to adverse selection environments, it is proven that the single-crossing property is satisfied when the belief heterogeneity is present. In terms of characterizing a monotone separating menu; the significantly overconfident agents self-select into exploitative contracts, inducing $e = 1$, whereas slightly overconfident agents opt for incentive-compatible contract that implements the second-best $e = 1$, and rational agents tend to choose full insurance with no effort provision $e = 0$.

Taking a broader view of the results, our work contributes in three ways: Theoretically, we extend moral hazard with belief heterogeneity to adverse selection settings where the degree of belief heterogeneity (overconfidence) is now private information; Analytically, decomposing contracts into effort and risk-sharing elements makes the screening problem tractable; Normatively, it is revealed that high level of overconfidence profile about the likelihood of success could be exploited by risk-neutral principal that raising welfare concerns in markets where belief heterogeneity are common.

A Appendices

A.1 Belief Heterogeneity and the Risk Structure of Optimal Contracts under Moral Hazard

Proposition 1 (First-Best Contract under Biased Beliefs, De la Rosa (2011))

A contract $\langle s_1^{FB,e}, s_0^{FB,e}\rangle$ is said to be an optimal first-best contract, which implement $e \in \{0,1\}$, if it satisfies the individual rationality (IR) condition

$$(\tilde{q}+\tilde{v}e)\,u(s_1^{FB,e}) + \Big[1-(\tilde{q}+\tilde{v}e)\Big]\,u(s_0^{FB,e}) - c(e) = 0,$$

and the following risk-sharing condition

$$\frac{\tilde{q}+\tilde{v}e}{1-(\tilde{q}+\tilde{v}e)} \cdot \frac{u'(s_1^{FB,e})}{u'(s_0^{FB,e})} = \frac{q+ve}{1-(q+ve)}.$$

Proof. The first-best contract that implements effort level $e \in \{0,1\}$ is the pair of payments which maximizes the principal's expected profit conditional on effort level e being implemented,

$$\mathbb{E}[\pi \mid e] = (q+ve)(x_1-s_1) + [1-(q+ve)](x_0-s_0),$$

subject only to the agent's individual-rationality (IR) constraint.

First, note that (IR) must bind; otherwise, the principal could reduce s_0 or s_1 and thus increase her expected profit. So

$$(\tilde{q}+\tilde{v}e)u(s_1^{FB,e}) + [1-(\tilde{q}+\tilde{v}e)]u(s_0^{FB,e}) - ce = \bar{u}.$$

We can write the Lagrangian as

$$\mathcal{L} = (q+ve)(x_1-s_1)+[1-(q+ve)](x_0-s_0)+\lambda\Big[(\tilde{q}+\tilde{v}e)u(s_1)+[1-(\tilde{q}+\tilde{v}e)]u(s_0)-ce-\bar{u}\Big].$$

The necessary first-order conditions with respect to s_1 and s_0 yield

$$\frac{\tilde{q}+\tilde{v}e}{1-(\tilde{q}+\tilde{v}e)} \cdot \frac{u'(s_1^{FB,e})}{u'(s_0^{FB,e})} = \frac{q+ve}{1-(q+ve)}. \tag{A1}$$

□

Proposition 2 (Second-Best Contract under Biased Beliefs, De la Rosa (2011)).

(i) For a **slightly overconfident agent**, the individual rationality (IR) constraint binds with equality and the incentive compatibility (IC) constraint also binds, i.e.

$$(\tilde{q}+\tilde{v})u(s_1^{SB}) + \Big[1-(\tilde{q}+\tilde{v})\Big]u(s_0^{SB}) - c = 0, \quad \text{(IR)}$$

$$\tilde{v}\Big(u(s_1^{SB}) - u(s_0^{SB})\Big) = c. \quad \text{(IC)}$$

(ii) For a **significantly overconfident agent**, the individual rationality (IR) constraint binds with equality, but the incentive compatibility (IC) constraint is slack, i.e.

$$(\tilde{q}+\tilde{v})u(s_1^{SB}) + \Big[1-(\tilde{q}+\tilde{v})\Big]u(s_0^{SB}) - c = 0, \quad \text{(IR)}$$

$$\tilde{v}\Big(u(s_1^{SB}) - u(s_0^{SB})\Big) > c. \quad \text{(IC)}$$

A contract $\langle s_1^{SB}, s_0^{SB}\rangle$ *is said to be an optimal second-best contract that implements effort* $e = 1$ *if the following conditions hold:*

(i) For a ***slightly overconfident agent****, the individual rationality (IR) constraint binds with equality and the incentive compatibility (IC) constraint also binds, i.e.*

$$(\tilde{q}+\tilde{v})u(s_1^{SB}) + \Big[1-(\tilde{q}+\tilde{v})\Big]u(s_0^{SB}) - c = 0, \quad \textit{(IR)}$$

$$\tilde{v}\Big(u(s_1^{SB}) - u(s_0^{SB})\Big) = c. \quad \textit{(IC)}$$

(ii) For a ***significantly overconfident agent****, the individual rationality (IR) constraint binds with equality, but the incentive compatibility (IC) constraint is slack, i.e.*

$$(\tilde{q}+\tilde{v})u(s_1^{SB}) + \Big[1-(\tilde{q}+\tilde{v})\Big]u(s_0^{SB}) - c = 0, \quad \textit{(IR)}$$

$$\tilde{v}\Big(u(s_1^{SB}) - u(s_0^{SB})\Big) > c. \quad \textit{(IC)}$$

Proof. **Claim (i);** In the second-best problem that implements $e = 1$, when the agent is only *slightly* overconfident, the optimal contract (s_1^{SB}, s_0^{SB}) satisfies

$$(\tilde{q}+\tilde{v})\,u(s_1^{SB}) + \Big[1-(\tilde{q}+\tilde{v})\Big]u(s_0^{SB}) - c = 0 \quad \text{and} \quad \tilde{v}\Big(u(s_1^{SB}) - u(s_0^{SB})\Big) = c.$$

In words: **IR binds and IC binds.** The second-best contract is the pair of payments which maximizes the principal's expected profit conditional on effort being implemented,

$$\mathbb{E}[\pi \mid e = 1] = (q + v)(x_1 - s_1) + [1 - (q + v)](x_0 - s_0),$$

subject to the agent's individual-rationality (IR) and the agent's incentive-compatibility (IC) constraint,

$$(\tilde{q} + \tilde{v})u(s_1) + [1 - (\tilde{q} + \tilde{v})]u(s_0) - c \geq 0,$$

$$\tilde{v}\Big(u(s_1) - u(s_0)\Big) \geq c.$$

First, note that (IR) must bind; otherwise, the principal could reduce s_0 and thus increase her expected profit without destroying incentives. We will show that the (IC) constraint binds if the agent is *slightly overconfident overall*, which implies that the binding (IR) and (IC) constraints define the second-best contract $\langle s_1^{SB}, s_0^{SB} \rangle$.

In order to show this, we prove that if the agent is slightly overconfident overall, that is, if

$$\frac{\tilde{q} + \tilde{v}}{1 - (\tilde{q} + \tilde{v})} \frac{u'(s_1^{SB})}{u'(s_0^{SB})} \leq \frac{q + v}{1 - (q + v)},$$

then any change in the offered contract away from $\langle s_1^{SB}, s_0^{SB} \rangle$ that increases the principal's expected profit and maintains (IR) and (IC) as binding.

By totally differentiating the binding (IR) constraint,

$$(\tilde{q} + \tilde{v})u(s_1^{SB}) + [1 - (\tilde{q} + \tilde{v})]u(s_0^{SB}) - c = 0,$$

we find the set of marginal changes (ds_1, ds_0) that maintain a binding (IR):

$$(\tilde{q} + \tilde{v})u'(s_1^{SB})ds_1 + [1 - (\tilde{q} + \tilde{v})]u'(s_0^{SB})ds_0 = 0,$$

which can be written as

$$ds_0 = -\frac{\tilde{q} + \tilde{v}}{1 - (\tilde{q} + \tilde{v})} \frac{u'(s_1^{SB})}{u'(s_0^{SB})} ds_1. \tag{A1}$$

Note that the principal's expected payment to the agent is $(q+v)s_1 + \Big[1 - (q+v)\Big]s_0$. Thus, to increase expected profit [18], the total differential of the principal's payoff must

[18] Since the principal's profit equals output minus payment, the total differential of the principal's payoff is $d\Pi = \mathit{Output} - dE[\text{Payment}] = \mathit{Output} - (q + v)ds_1 - [1 - (q + v)]ds_0$. For profit to increase, it must hold that $d\Pi > 0$.

be positive:

$$-(q+v)ds_1 - [1-(q+v)]ds_0 > 0.$$

Substituting for ds_0 gives

$$\left[-(q+v) + \frac{\tilde{q}+\tilde{v}}{1-(\tilde{q}+\tilde{v})}\frac{u'(s_1^{SB})}{u'(s_0^{SB})}[1-(q+v)]\right]ds_1 > 0,$$

or equivalently,

$$\left(\frac{\tilde{q}+\tilde{v}}{1-(\tilde{q}+\tilde{v})}\frac{u'(s_1^{SB})}{u'(s_0^{SB})} - \frac{q+v}{[1-(q+v)]}\right)ds_1^{SB} > 0. \tag{A2}$$

Where the agent is *slightly overconfident overall*

$$\frac{\tilde{q}+\tilde{v}}{1-(\tilde{q}+\tilde{v})}\frac{u'(s_1^{SB})}{u'(s_0^{SB})} \leq \frac{q+v}{1-(q+v)},$$

which implies that $ds_1^{SB} < 0$ and $ds_0^{SB} > 0$. Thus, for a slightly overconfident agent, any reduction in s_1 (and corresponding increase in s_0) that maintains a binding (IR) would decrease the power of incentives, *violating (IC)*. Hence, (IC) must *bind* at the optimum. Together with the binding (IR), this characterizes the second-best contract $\langle s_1^{SB}, s_0^{SB}\rangle$.

Claim (ii); In the second-best problem that implements $e = 1$, when the agent is *significantly* overconfident, the optimal contract (s_1^{SB}, s_0^{SB}) coincides with the first-best contract (s_1^{FB}, s_0^{FB}). In this case, the individual rationality (IR) constraint binds with equality, while the incentive compatibility (IC) constraint is slack, i.e.

$$(\tilde{q}+\tilde{v})\,u(s_1^{SB}) + \Big[1-(\tilde{q}+\tilde{v})\Big]u(s_0^{SB}) - c = 0, \quad \text{(IR)}$$

$$\tilde{v}\Big(u(s_1^{SB}) - u(s_0^{SB})\Big) > c. \quad \text{(IC)}$$

When the agent is *significantly overconfident overall*,

$$\frac{\tilde{q}+\tilde{v}}{1-(\tilde{q}+\tilde{v})}\frac{u'(s_1^{SB})}{u'(s_0^{SB})} > \frac{q+v}{1-(q+v)}.$$

Refer to condition (A2) above, which defines a change in the contract $\langle s_1^{SB}, s_0^{SB}\rangle$ that both maintains a binding (IR) constraint and increases expected profit for the principal. If the agent is significantly overconfident overall, then such a change has $ds_1^{SB} > 0$ and $ds_0^{SB} < 0$. Such a change increases the power of incentives with respect to $\langle s_1^{SB}, s_0^{SB}\rangle$, so it (strictly) satisfies incentive compatibility. Given that there exists a change in the $\langle s_1^{SB}, s_0^{SB}\rangle$ contract that strictly satisfies the (IC) constraint, maintains the (IR) constraint, and increases expected profit for the principal, it must be the case that (IC) is slack at the optimum.

We can thus solve the problem ignoring that constraint, so the first-best contract that implements effort, $\langle s_1^{FB}, s_0^{FB} \rangle$, is incentive compatible and thus also the second-best contract that implements effort.

Alternatively, one could directly prove that $\langle s_1^{FB}, s_0^{FB} \rangle$ is strictly incentive compatible if the agent is significantly overconfident overall. Note that both $\langle s_1^{SB}, s_0^{SB} \rangle$ and $\langle s_1^{FB}, s_0^{FB} \rangle$ satisfy the binding (IR) constraint, that $\langle s_1^{SB}, s_0^{SB} \rangle$ satisfies the binding (IC) constraint, and that $u'' < 0$. If the agent is significantly overconfident overall, $\frac{u'(s_1^{SB})}{u'(s_0^{SB})} > \frac{u'(s_1^{FB})}{u'(s_0^{FB})}$, and therefore $\tilde{v}(u(s_1^{FB}) - u(s_0^{FB})) > c$.

□

Proposition 3 (Wager Effect of Overconfidence, De la Rosa (2011))
In a first-best (no moral hazard) setting, the contract $\langle s_1^{FB,e}, s_0^{FB,e} \rangle$ becomes riskier as overconfidence and optimism increase, such that

$$\frac{d}{d\tilde{q}}\left(u(s_1^{FB}) - u(s_0^{FB})\right) > 0 \quad \textit{and} \quad \frac{d}{d\tilde{v}}\left(u(s_1^{FB}) - u(s_0^{FB})\right) > 0$$

when

$$\frac{1}{u'(s)}\left(-\frac{u''(s)}{u'(s)}\right) \textit{ is a non-decreasing function}$$

Proof. In the first-best, the principal chooses s_0^{FB}, s_1^{FB} and $e \in \{0,1\}$ to maximize profit. The agent's participation constraint binds:

$$(\tilde{q} + \tilde{v}e)u(s_1^{FB}) + (1 - \tilde{q} - \tilde{v}e)u(s_0^{FB}) - c(e) = 0 \tag{IR}$$

The risk-sharing condition is derived from the principal's FOC (using Lagrangian):

$$\frac{u'(s_1^{FB})}{u'(s_0^{FB})} = \frac{1 - (\tilde{q} + \tilde{v}e)}{\tilde{q} + \tilde{v}e} \cdot \frac{q + ve}{1 - (q + ve)} \tag{RS}$$

Let $\gamma := \frac{u'(s_1^{FB})}{u'(s_0^{FB})}$. Then, as $\tilde{q}$ and $\tilde{v}$ increases, the factor

$$\frac{1 - (\tilde{q} + \tilde{v}e)}{\tilde{q} + \tilde{v}e}$$

strictly decreases. To maintain (RS), γ must decrease. Since u' is decreasing and strictly positive, a decrease in γ implies:

$$s_1^{FB} - s_0^{FB} \quad \text{increases.}$$

From the concavity of u, $u(s_1^{FB}) - u(s_0^{FB})$ increases with $s_1^{FB} - s_0^{FB}$, and hence

$$\frac{d}{d\tilde{q}}\left(u(s_1^{FB}) - u(s_0^{FB})\right) > 0, \quad \frac{d}{d\tilde{v}}\left(u(s_1^{FB}) - u(s_0^{FB})\right) > 0.$$

These monotonicities hold under the sufficient condition that $A(s)$ is non-decreasing (i.e., relative risk aversion is increasing or constant), which guarantees that the mapping from γ to $(s_1 - s_0)$ is monotonic. □

Corollary 1 (Optimism implies risk-bearing). *If the agent is relatively overall overconfident, $\tilde{q} > q$ and $\tilde{v} > v$, then, in equilibrium, it is first-best optimal to shift payment toward success, which implies $s_1^{FB,e} > s_0^{FB,e}$, thereby exposing the agent to risk.*

Proof. Let the agent be relatively optimistic, such that, $\tilde{q} > q$. The first-best risk-sharing condition suggests:

$$\frac{\tilde{q} + \tilde{v}e}{1 - (\tilde{q} + \tilde{v}e)} \cdot \frac{u'(s_1^{FB,e})}{u'(s_0^{FB,e})} = \frac{q + ve}{1 - (q + ve)}.$$

Since $\tilde{q} > q$, the left-hand side increases. To maintain risk-sharing condition equal, $\frac{u'(s_1^{FB,e})}{u'(s_0^{FB,e})}$ must decrease. Because $u'(\cdot)$ is strictly decreasing by concavity, a decrease in the ratio implies $s_1^{FB,e} > s_0^{FB,e}$. Hence, the optimal contract shifts payment toward success, exposing the agent to more risk than usual under homogeneity. □

Proposition 4 (Incentive Effect of Overconfidence, De la Rosa (2011)). *In a second-best setting, the contract $\langle s_1^{SB,e}, s_0^{SB,e}\rangle$ that implements effort becomes less risky in the degree of agent overconfidence, such that*

$$\frac{d}{d\tilde{v}}\left(u(s_1^{SB,e}) - u(s_0^{SB,e})\right) < 0$$

whereas agent optimism does not affect the power of incentives;

$$\frac{d}{d\tilde{q}}\left(u(s_1^{SB,e}) - u(s_0^{SB,e})\right) = 0$$

when

$$\frac{1}{u'(s)}\left(-\frac{u''(s)}{u'(s)}\right) \text{ is a non-decreasing function}$$

Proof. The second-best contract $\langle s_1^{SB}, s_0^{SB}\rangle$ is characterized, in particular, by the binding (IC) constraint which we can rewrite as

$$\tilde{v}\left(u(s_1^{SB}) - u(s_0^{SB})\right) = c.$$

Given $u(s_1^{SB}) - u(s_0^{SB}) = \frac{c}{\tilde{v}} > 0$, it follows immediately that $\frac{d}{d\tilde{v}}\left(u(s_1^{SB}) - u(s_0^{SB})\right) < 0$ and $\frac{d}{d\tilde{q}}\left(u(s_1^{SB}) - u(s_0^{SB})\right) = 0$. □

Proposition 5 (Exploitation Motive under Significantly Overconfidence). *Let $q+v$ be the true success probability under $e = 1$ and $\tilde{q}+\tilde{v}$ the agent's belief, and $\tilde{v} \gg v$. The principal is risk-neutral and pays the expected transfer (payment) to the agent. In this case, the principal's problem is*

$$min\ (q+v)s_1 + (1-q-v)s_0,$$

s.t.

$$(\tilde{q}+\tilde{v})\,s_1 + \left(1-(\tilde{q}+\tilde{v})\right)s_0 - (\tilde{q}+\tilde{v})\left(1-(\tilde{q}+\tilde{v})\right)(s_1-s_0)^2 \geq 0.$$

Solving this optimization gives the optimal wage spread:

$$s_1 - s_0 \ = \ \frac{(\tilde{q}+\tilde{v})-(q+v)}{2\tilde{v}(1-\tilde{q}-\tilde{v})} \ \equiv \ \Delta^*. \tag{1}$$

Given $s_H - s_L$, the principal sets the levels of the two payments so that PC binds.

Proof. In the case where $p = q+v$ and $\tilde{p} = \tilde{q}+\tilde{v}$, the principal's problem is

$$\begin{aligned} &\min\ ps_1 + (1-p)s_0 \\ \text{s.t.}\quad &\tilde{p}s_1 + (1-\tilde{p})s_0 - \tilde{p}(1-\tilde{p})(s_1-s_0)^2 \geq 0. \end{aligned} \tag{IR}$$

Noting that

$$ps_1 + (1-p)s_0 = \tilde{p}s_1 + (1-\tilde{p})s_0 - (\tilde{p}-p)(s_1-s_0),$$

The participation constraint (IR) binds and follows that at the optimum; $\tilde{p}s_1 + (1-\tilde{p})s_0 = \tilde{p}(1-\tilde{p})(s_1-s_0)^2$, then we can plug (IR) into the maximand to get

$$\min\ (\tilde{q}+\tilde{v})(1-(\tilde{q}+\tilde{v})(s_1-s_0)^2 - ((\tilde{q}+\tilde{v})-(q+v))(s_1-s_0).$$

The objective is a strictly convex quadratic since $\tilde{p}(1-\tilde{p}) > 0$. The objective is a function of $d := s_1 - s_0$:

$$f(d) = (\tilde{q}+\tilde{v})\Big(1-(\tilde{q}+\tilde{v})\Big)d^2 - \Big((\tilde{q}+\tilde{v}) - (q+v)\Big)d.$$

The first-order condition is

$$f'(d) = 2(\tilde{q}+\tilde{v})\Big(1-(\tilde{q}+\tilde{v})\Big)d - \Big((\tilde{q}+\tilde{v}) - (q+v)\Big) = 0,$$

which yields

$$d^* = s_1 - s_0 = \frac{(\tilde{q}+\tilde{v}) - (q+v)}{2(\tilde{q}+\tilde{v})\Big(1-(\tilde{q}+\tilde{v})\Big)} = \frac{\tilde{p}-p}{2\tilde{p}(1-\tilde{p})}.$$

The second-order condition,

$$f''(d) = 2(\tilde{q}+\tilde{v})\Big(1-(\tilde{q}+\tilde{v})\Big) > 0,$$

confirms that this is a minimum, and which implies the optimal spread

$$s_1 - s_0 = \frac{\tilde{q}+\tilde{v}-(q+v)}{2(\tilde{q}+\tilde{v})(1-\tilde{q}-\tilde{v})} \equiv \Delta^*. \tag{5}$$

□

A.2 Principal's Optimization under Belief Heterogeneity

In a moral hazard context, and the agent is risk-averse, the principal's program is written as under the heterogeneous belief about the outcomes:

$$\mathbb{E}[\Pi \mid s_x] = (q + v.e)(x_1 - s_1) + [1 - (q + v.e)]\,(x_0 - s_0).$$

The agent is considered optimistic when he believes that success is more likely than the principal does $\tilde{q} > q$ and overconfident if he overestimates the additional effect of his own effort on the likelihood of success $\tilde{v} > v$

The risk averse agent receives utility from monetary transfers s_x and disutility from effort e. We assume that the utility function is $U(s, e) = u(s) - c(e)$, where u characterises such as

$$u(s_x) = \ln(s_x)$$

with $u(\cdot)$ being increasing and concave ($u'(\cdot) > 0$, $u''(\cdot) < 0$).

Let the agent chooses $e \in \{0, 1\}$ to maximize his expected utility:

$$\tilde{\mathbb{E}}[u(w_x) \mid e] - c(e) = (\tilde{q} + \tilde{v}e).u(s_1) + (1 - (\tilde{q} + \tilde{v}e)).u(s_0) - ce$$

Assume the principal aims to implement effort ($e^{SB} = 1$) in the second-best scenario. When the agent is risk-averse, the principal's problem can be expressed as:

$$\mathbb{E}[\Pi] = (q + v.e)(x_1 - s_1) + [1 - (q + v.e)]\,(x_0 - s_0).$$

$$(\tilde{q} + \tilde{v}e).u(s_1) + (1 - (\tilde{q} + \tilde{v}e)).u(s_0) - c \geq 0, \tag{PC}$$

$$(\tilde{q} + \tilde{v}e).u(s_1) + (1 - (\tilde{q} + \tilde{v}e)).u(s_0) - c \geq \tilde{q}u(s_1) + [1 - \tilde{q}]u(s_0) \tag{IC}$$

Since it is not immediately evident that the program is concave, such that the first-order Kuhn-Tucker conditions are both necessary and sufficient for optimality. Let define $\bar{u} = u(s_1) = ln(s_1)$ and $\underline{u} = u(s_0) = ln(s_0)$, or equivalently, let $s_1 = exp(\bar{u})$ and $s_0 = exp(\underline{u})$.These new variables are the levels of ex post utility obtained by the agent in both states of nature. The new maximization problem of principal becomes:

$$\mathbb{E}_{\underline{u},\bar{u}}[\Pi] = (q+v)(x_1 - e^{\bar{u}}) + [1-(q+v)]\,(x_0 - e^{u})$$

$$(\tilde{q}+\tilde{v})\bar{u} + [1-(\tilde{q}+\tilde{v})]\underline{u} - c \geq 0 \tag{PC}$$

$$(\tilde{q}+\tilde{v})\bar{u} + [1-(\tilde{q}+\tilde{v})]\underline{u} - c \geq \tilde{q}\bar{u} + [1-\tilde{q}]\underline{u} \tag{IC}$$

$$\Rightarrow \tilde{v}(\bar{u} - \underline{u}) \geq c$$

Observe that the principal's objective function is now strictly concave with respect to $(\bar{u}, u)$ *because the function exp*$(\cdot)$ *is strictly convex. Since the constraints are linear, the problem becomes a concave optimization problem meaning that the Kuhn-Tucker conditions are both necessary and sufficient for identifying an optimal solution* $(\bar{u}^{SB}, \underline{u}^{SB})$.

Let λ *and* μ *be the non-negative lagrange multipliers corresponding to constraints (IC) and (PC), respectively and the lagrangian is;*

$$\begin{aligned}\mathcal{L} = &(q+v)(x_1 - e^{\bar{u}}) + [1-(q+v)](x_0 - e^{\underline{u}}) \\ &+ \lambda[\tilde{v}(\bar{u} - \underline{u}) - c] + \mu[(\tilde{q}+\tilde{v})\bar{u} + (1-(\tilde{q}+\tilde{v}))\underline{u} - c]\end{aligned}$$

The first-order conditions for this program can then be stated as:

$$\frac{\partial \mathcal{L}}{\partial \bar{u}} = -(q+v)e^{\bar{u}} + \lambda\tilde{v} + \mu(\tilde{q}+\tilde{v}) = 0$$

$$\frac{\partial \mathcal{L}}{\partial \underline{u}} = -[1-(q+v)]e^{\underline{u}} - \lambda\tilde{v} + \mu(1-(\tilde{q}+\tilde{v})) = 0$$

$$-(q+v)e^{\bar{u}} + \lambda\tilde{v} + \mu(\tilde{q}+\tilde{v}) = \boxed{-\frac{q+v}{u'(s_1^{SB})} + \lambda\tilde{v} + \mu(\tilde{q}+\tilde{v}) = 0},$$

$$((q+v)-1)e^{\underline{u}} - \lambda\tilde{v} + \mu(1-(\tilde{q}+\tilde{v})) = \boxed{\frac{(q+v)-1}{u'(s_0^{SB})} - \lambda\tilde{v} + \mu(1-(\tilde{q}+\tilde{v})) = 0}.$$

[19] *where* s_1^{SB} *and* s_0^{SB} *are the second-best optimal transfers. Since* $u(s) = \ln(s)$, *we get by rearranging terms:*

$$\frac{1}{u'(s_1^{SB})} = \frac{\lambda\tilde{v} + \mu(\tilde{q}+\tilde{v})}{q+v} \tag{1}$$

[19] Define $h(u) = \exp(u)$, so $u(h(\bar{u})) = \bar{u}$, differentiating both sides with respect to $\bar{u}$ gives $\frac{d}{d\bar{u}}u(h(\bar{u})) = \frac{d}{d\bar{u}}\bar{u}$. By the chain rule, $u'(h(\bar{u}))\cdot h'(\bar{u}) = 1$. Solving for $h'(\bar{u})$ yields $h'(\bar{u}) = \frac{1}{u'(h(\bar{u}))} = \frac{1}{\frac{1}{exp(\bar{u})}} = e^{\bar{u}}$.

$$\frac{1}{u'(s_0^{SB})} = \frac{\lambda\tilde{v} - \mu(1-(\tilde{q}+\tilde{v}))}{(q+v)-1} \tag{2}$$

The four variables $(s_1^{SB}, s_0^{SB}, \lambda, \mu)$ are determined simultaneously as the solution to a system of four equations (IC), (PC), (1) and (2).

By multiplying (1) by $(q-v)$ and (2) by $((q+v)-1)$ and adding those two modified equations, we obtain:

$$\frac{(q+v)}{u'(s_1^{SB})} = \lambda\tilde{v} + \mu(\tilde{q}+\tilde{v})$$

$$\frac{(1-(q+v))}{u'(s_0^{SB})} = -\lambda\tilde{v} + \mu(1-(\tilde{q}+\tilde{v}))$$

$$\frac{(q+v)}{u'(s_1^{SB})} + \frac{(1-(q+v))}{u'(s_0^{SB})} = \mu \tag{3}$$

Since $u'(\cdot) > 0$ then $\mu > 0$, hence, the participation constraint (PC) is necessarily binding. Using (3) and (1), we obtain also:

$$\lambda = \frac{\frac{(q+v)(1-\tilde{q}-\tilde{v})}{u'(s_1^{SB})} - \frac{(1-q-v)(\tilde{q}+\tilde{v})}{u'(s_0^{SB})}}{\tilde{v}} \tag{4}$$

To determine the sign of λ, it is necessary to check the numerator

$$\frac{(q+v)(1-\tilde{q}-\tilde{v})}{u'(s_1^{SB})} - \frac{(1-q-v)(\tilde{q}+\tilde{v})}{u'(s_0^{SB})}$$

By the definition of slightly overconfident, we know that

$$\frac{\tilde{q}+\tilde{v}}{1-(\tilde{q}+\tilde{v})} \cdot \frac{u'(s_1^{SB})}{u'(s_0^{SB})} \leq \frac{q+v}{1-(q+v)}$$

this implies $\lambda \geq 0$.Indeed, from (IC), we have $\bar{u}^{SB} - u^{SB} \geq \frac{c}{\tilde{\nu}} > 0$ and thus $u(s_1^{SB}) > u(s_0^{SB})$. Since $u(\cdot)$ is increasing and concave ($u'(\cdot) > 0$, $u''(\cdot) < 0$). Thus, $s_1^{SB} > s_0^{SB}$, implying that the right-hand side of (4) is strictly positive since $u'' < 0$.

Conversely, by the definition of significantly overconfident, we know that

$$\frac{\tilde{q}+\tilde{v}}{1-(\tilde{q}+\tilde{v})} \cdot \frac{u'(s_1^{SB})}{u'(s_0^{SB})} > \frac{q+v}{1-(q+v)}$$

which implies $\lambda < 0$, indicating that the incentive compatibility constraint (IC) is slack when the agent exhibits significant overconfidence.

As a result,the sign of the Lagrange multiplier λ, which corresponds to the incentive compatibility (IC) constraint, depends on the degree of the agent's overconfidence. Specifically, when the agent is slightly overconfident overall*, the constraint binds and $\lambda \geq 0$. However, when the agent is* significantly overconfident overall*, which implies that the IC constraint is slack, and therefore $\lambda < 0$. This means that the agent's excessive belief in the effectiveness of his own effort leads him to exert effort voluntarily. Consequently, the optimal second-best contract coincides with the first-best allocation.*

A.3 Summary of Optimal Screening Contracts

This appendix provides a detailed summary of the optimal screening contracts presented in Result 1, restated in terms of the two design elements introduced in Section 4.1: The table below summarizes the corresponding contract structure for each agent type.

Summary of Result 1 in Terms of Design Elements. *The optimal menu*

$$\{(s_0^L, s_1^L, e^L), (s_0^H, s_1^H, e^H)\}$$

is characterized by distinct combinations of the two design elements—the effort element e^θ *and the risk-sharing element* $\Delta s^\theta = s_1^\theta - s_0^\theta$*, which reveal how belief heterogeneity shapes the contract structure.*

1. ***High-Type Agent (Significantly Overconfident)***

   Effort element: $e^H = 1$.

   Risk-sharing element: $\Delta s^H = s_1^H - s_0^H > \frac{c}{\tilde{v}_H}$.

   This contract is exploitative, as the incentive-compatibility constraint is slack. The high-type agent exerts effort voluntarily due to being significantly overconfident, and the principal offers a riskier contract with a larger outcome spread, this corresponds to the wager effect.

2. ***Low-Type Agent with Rational Beliefs***

   Effort element: $e^L = 0$.

   Risk-sharing element: $\Delta s^L = 0$.

   The contract is flat, providing full insurance and no incentives. Since effort is too costly to induce under rational expectations, the individual-rationality constraint binds, and the principal offers complete insurance satisfying $u(\bar{s}_L) = 0$.

3. ***Low-Type Agent with Slight Overconfidence***

   Effort element: $e^L = 1$.

   Risk-sharing element: $\Delta s^L = s_1^L - s_0^L = \frac{c}{\tilde{v}_L}$.

   This is an incentive-compatible contract, where both the participation and incentive constraints bind. Slightly overconfidence lowers the required incentive spread, allowing the principal to implement effort at a lower agency cost, this corresponds to the incentive effect.

References

Baron, D. P. and Besanko, D. (1987). Monitoring, moral hazard, asymmetric information, and risk sharing in procurement contracting. The RAND Journal of Economics*, 18(4):509–532.*

Borch, K. (1962). Equilibrium in a reinsurance market. Econometrica*, 30(3):424–444.*

Castro-Pires, H., Chade, H., and Swinkels, J. (2024). Disentangling moral hazard and adverse selection. American Economic Review*, 114(1):1–37.*

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

Dumav, M., Khan, U., and Rigotti, L. (2025). Optimal contracts when the players think differently. Economic Theory.

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

Faynzilberg, P. S. and Kumar, P. (1997). Optimal contracting of separable production technologies. Games and Economic Behavior*, 21(1):15–39.*

Gervais, S. and Goldstein, I. (2007). The positive effects of biased self-perceptions in firms. Review of Finance*, 11(3):453–496.*

Gottlieb, D. and Moreira, H. (2017). Simultaneous adverse selection and moral hazard. Wharton School and EPGE/FGV Working Paper.

Grossman, S. J. and Hart, O. D. (1983). An analysis of the principal-agent problem. Econometrica*, 51(1):7–45.*

Heidhues, P., Köszegi, B., and Strack, P. (2018). Unrealistic expectations and misguided learning. Econometrica*, 86(5):1159–1214.*

Holmström, B. (1979). Moral hazard and observability. The Bell Journal of Economics*, 10(1):74–91.*

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

Laffont, J.-J. and Martimort, D. (2002). The Theory of Incentives: The Principal-Agent Model. *Princeton University Press, Princeton, NJ.*

Laffont, J.-J. and Tirole, J. (1986). Using cost observation to regulate firms. Journal of Political Economy*, 94(3):614–641.*

Melumad, N. D. and Reichelstein, S. (1989). Value of communication in agencies. Journal of Economic Theory*, 47(2):334–368.*

Mirrlees, J. A. (1971). An exploration in the theory of optimum income taxation. The Review of Economic Studies*, 38(2):175–208.*

Myerson, R. B. (1982). Optimal coordination mechanisms in generalized principal-agent problems. Journal of Mathematical Economics*, 10(1):67–81.*

Picard, P. (1987). On the design of incentive schemes under moral hazard and adverse selection. Journal of Public Economics*, 33(3):305–331.*

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

Taylor, S. E. and Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. Psychological Bulletin*, 103(2):193–210.*

Statement of Authorship

"I hereby confirm that the work presented has been performed and interpreted solely by myself except for where I explicitly identified the contrary. I assure that this work has not been presented in any other form for the fulfillment of any other degree or qualification. Ideas taken from other works in letter and in spirit are identified in every single case."

Place and date: Göttingen, 15.10.2025

Signature: