John Joaquim Sigaud Pease

# (Your) Ignorance is Bliss: Robust Moral Hazard

DISSERTAÇÃO DE MESTRADO

DEPARTAMENTO DE ECONOMIA
Programa de Pós–graduação em Economia

Rio de Janeiro
March 2016

Pontifícia Universidade Católica
do Rio de Janeiro

John Joaquim Sigaud Pease

# (Your) Ignorance is Bliss: Robust Moral Hazard



Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Economia of the Departamento de Economia , PUC–Rio as a partial fulfillment of the requirements for the degree of Mestre em Economia.

Advisor: Prof. Vinicius Nascimento Carrasco

Rio de Janeiro
March 2016

Pontifícia Universidade Católica
do Rio de Janeiro

John Joaquim Sigaud Pease

# (Your) Ignorance is Bliss: Robust Moral Hazard

Dissertation presented to the Programa de Pós–graduação em Economia of the Departamento de Economia , PUC–Rio as a partial fulfillment of the requirements for the degree of Mestre em Economia.

**Prof. Vinicius Nascimento Carrasco**
Advisor
Departamento de Economia — PUC–Rio

**Prof. Leonardo Bandeira Rezende**
Departamento de Economia - PUC-Rio

**Prof. Humberto Luiz Ataíde Moreira**
FGV-RJ

**Prof. Monica Herz**
Coordinator of the Centro de Ciências Sociais – PUC–Rio

Rio de Janeiro, March $23^{rd}$, 2016

**John Joaquim Sigaud Pease**

The author graduated in Economics from PUC-Rio in 2014.



Bibliographic data

Pease, John Joaquim Sigaud

(Your) Ignorance is Bliss: Robust Moral Hazard / John Joaquim Sigaud Pease; advisor: Vinicius Nascimento Carrasco. — Rio de Janeiro : PUC–Rio, Departamento de Economia, 2016.
(em Inglês)
v., 27 f: il. ; 29,7 cm

1. Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Economia.

Inclui referências bibliográficas.

1. Economia – Tese. 2. Risco Moral. 3. Incerteza. 4. Robustez. 5. Teoria dos Contratos. 6. Informação Assimétrica.
I. Carrasco, Vinicius Nascimento. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Economia. III. Título.
(em Português)

CDD: 330

## Acknowledgments



A longer list of thanks than I could possibly put together should ensue in order to give proper credit to all of those that made this master's degree possible. Firstly, I would like to thank my advisor, Vinicius Carrasco, for his continued guidance and advice. Secondly, I would like to thank the members of my committee, Leonardo Rezende and Humberto Moreira, for their thoughtful suggestions and corrections. I am also greatly indebted to my friends (and upon more than one occasion, mentors) Pedro Solti, Marcel Chamarelli, Leandro Gomes and Gustavo Albuquerque for their valuable insights and comments. I am thankful as well to the extraordinary Beatriz Figueiredo; her generosity and tenderness buoyed my drive and happiness even when I was certain that I would fail. Finally, I must express my infinite - yet still insufficient - gratitude to my parents, without whom this and most other endeavors of mine would have never been possible. Their unending kindness and support have been the brick and mortar of all that I have managed to build.

## Abstract

Pease, John Joaquim Sigaud; Carrasco, Vinicius Nascimento(advisor). **(Your) Ignorance is Bliss: Robust Moral Hazard**. Rio de Janeiro, 2016. 27p. MSc. Dissertation — Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

We consider an environment with moral hazard where a principal and agent have heterogeneous beliefs as to how actions map to output. We focus first on optimal contracts when the principal is at some level aware of the agent's biases, demonstrating that standard firm sale is generally suboptimal in such contexts. We then look at optimal contract design when a principal who is faced with total uncertainty regarding an agent's beliefs demands robustness to his own ignorance.



## Keywords

Moral Hazard; Uncertainty; Robustness; Contract Theory; Informational Asymmetry;

## Resumo

Pease, John Joaquim Sigaud; Carrasco, Vinicius Nascimento(orientador). **Robustez e Risco Moral**. Rio de Janeiro, 2016. 27p. Dissertação de Mestrado — Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

Consideramos um ambiente com risco moral onde um principal e um agente têm crenças heterogêneas sobre como ações levam a resultados. Focamos primeiramente em contratos ótimos quando o principal tem alguma noção sobre o viés do agente, demonstrando que a venda da firma é - em geral - subótima em tal contexto. Analisamos então o desenho do contrato ótimo quando o principal tem total incerteza sobre as crenças do agente e demanda robustez frente à sua ignorância.



## Palavras–chave

Risco Moral; Incerteza; Robustez; Teoria dos Contratos; Informação Assimétrica;

# Contents

# 1
# Introduction

In Sophocles' *Oedipus Rex*, King Laius of Thebes - acting upon the Delphic Oracles' prophecy that he would die by the hands of his child - delivered young Oedipus to his wife Jocasta so that she would kill him. Jocasta passed this duty on to a servant who left Oedipus to die from sun exposure at a mountaintop, where a shepherd found and saved the infant. Many years later, Oedipus unknowingly killed his biological father and committed incest with his mother.

In one of the first formal inquiries of moral Hazard, Bengt Holmström famously noted that *a problem of moral hazard may arise when individuals engage in risk sharing under conditions such that their privately taken actions affect the probability distribution of the outcome.* In Sophocles' tragedy, King Laius - who bore the entire risk of his son living - delegated Oedipus's death. Instead of killing the child outright - which perhaps would have a high empathic cost - the servant takes with ending Oedipus's life left him tied and bound at a mountaintop, somewhat reducing the probability of his death. Unlike Greek folklore would suggest, moral hazard - and not the fates - killed King Laius.

When treating optimal contracting under moral hazard, economists typically model both principal and agent as being fully aware of the agent's possible actions and of how each of these actions affects the probability distribution of the outcome. The former assumption has recently been relaxed by Carroll (2015); a setting is considered wherein a principal knows only a subset of an agent's possible actions and demands robustness to this uncertainty. Carroll finds that the optimal compensation scheme is linear under double-sided risk neutrality and limited liability, something Milgrom and Holmström (1981) previously purported should happen if robustness were to be satisfied.

The latter assumption - that both parties are fully informed as to how actions map onto the probability distribution of outcomes - has been slackened recently by Lopomo, Rigotti and Shannon (2011), de La Rosa (2005) and more recently by Carroll and Meng (2015). The first of these three papers examines a setting in which the agent has imprecise beliefs and demands robustness to

it. LRS find a simple set of conditions under which a principal facing an agent with demands for robustness can still design an optimal contract that has a simple (two-wage) structure. Although in the final part of our paper we are concerned with a principal who demands robustness to uncertainty (instead of an agent), the conceptual max-min problem is equivalent to theirs. However, while LRS try to find under which conditions optimal contracts will have a certain design, we are more concerned with what the optimal contract looks like under rather loose assumptions.

The second paper looks at the effect of agent overconfidence - and the Principal's awareness of belief heterogeneity - on the shape of optimal contracts and the welfare changes occurring from overconfidence in a setting of risk-aversion. What de La Rosa finds is that an agent's overconfidence can actually generate Pareto gains ex-post, given that it helps offset the agent's demand for higher rewards due to his risk aversion. Although we only take a small glimpse at what happens when an agent is risk averse, our paper also suggests that an agent's optimism may generate two-sided welfare gains.

The final paper - Carroll and Meng's - considers a situation in which the principal has a slight uncertainty of size $\epsilon$ about how the agent's actions translate into output (but the agent does not). In this context, the principal can demand robustness by refunding the agent with $\sqrt{\epsilon}$ of his profits. We believe this paper complements ours quite well, looking at what happens when local - instead of general - uncertainty is prevalent.

Our paper is fundamentally focused on what happens when a principal and an agent have heterogeneous beliefs. However reasonable it may be to assume in certain circumstances that all parties involved can perfectly assess the probabilistic consequences of every possible action in the universe of their concern, this hardly seems to be the most common occurrence in our world. Furthermore, it is unlikely that an agent is always (or even frequently) aware of his own ignorance. It is possible, for instance, that the servant tasked with killing Oedipus - oblivious to his own limitations - considered that the infant's death was absolutely certain be it through skull-bashing or abandonment with sun exposure and malnutrition. Quite clearly, it was *not*.

Experimental economics, pioneered by Daniel Kahneman and Amos Tversky, has consistently found that agents frequently deviate from what standard economic assumptions suggest they should do. In particular, there

exists mounting evidence that agents do not behave as Bayesian theory of judgment under uncertainty would prescribe they should, and instead often evaluate hypotheses in a biased manner. By using heuristics to assess information, agents sometimes ascribe wildly high probabilities to certain states of nature while completely ignoring the possibility of others. When betting on the success of their own effort, for example, individuals frequently overrate themselves. Again, a certain servant comes to mind.

However attractive behavioral assumptions may be, it is also easy to find reasons for which two Bayesian parties signing a contract would have distinct beliefs. On the one hand, a principal tends to engage in more contracts with agents than the other way around. If through her screening process a principal chooses only to engage with agents that have a similar technology, she may become very familiar with how the latter's actions affect the probability distribution of outcomes. On the other hand, since agents tend to engage in a smaller number of similar contracts, they might extrapolate imperfectly from other experiences when forming their judgments as to their own abilities, unable to understand ex-ante how small environmental differences might have large impacts on what the results of their actions are.

In this paper we aim to look at three different levels of informational asymmetry. We first consider what the optimal contract design is for the principal when she faces a biased agent whose beliefs she **knows** and whose actions she can perfectly observe or infer. Notice that this may happen even if a principal can solely observe results, as long as she knows that each action maps injectively to an outcome. When this occurs, the principal and agent engage in a wager where the former's ex-ante expected return is positive and unbounded.

Our second concern is with the less simplistic situation in which a principal knows what an agent's beliefs are, but cannot observe actions. This is the case for a large part of insurance providers, who are unable to track an agent's actions after insurance has been provided. When actions are unobservable, the principal's optimal contract will be dependent on the agent's direction of bias. When the agent is more optimistic than the principal, the latter can once again infinitely exploit the former. When the agent is more pessimistic, however, the principal's gains will be bounded.

Our final concern is also our most general: what does a principal who demands robustness do when she is faced with an agent whose beliefs are

unknown to her? If said principal evaluates outcomes ex-ante by their worst case - that is, if she has max-min preferences - she will come to one of two conclusions. Under a very general set of conditions, no action except that which is least costly for the agent will be implementable. This occurs because the agent can always believe that two actions have the same map to outcomes. Therefore, if a principal cannot observe an agent's preferences, she will not be able to satisfy incentive compatibility constraints.

Under a less general framework - if the principal at least knows that she and the agent share the same beliefs over increases in the aggregate surplus of the economy - a new set of actions becomes implementable. Under this scenario, the principal knows that any action which increases output net of costs will also do so under any of the agent's possible belief sets. Optimality for the principal under this set of conditions can be reached through the firm's sale at a low price - one which depends on the difference between the lowest possible outcome and the least costly action. We will proceed to show that, in spite of implementability, no action more expensive than cheapest one is desirable to the principal.

# 2
# The General Model

In our general setup, a principal wishes to enter into a contract with an agent whose actions stochastically affect output. In our model, we consider that the set of possible outcomes $Y \subset \mathbb{R}^+$ is common knowledge, as is the agent's set of possible actions (his technology) $A$ and their underlying costs $c \in \mathbb{R}^+$. We assume all sets to be compact. We analyze a framework under which both parties have linear utilities over wealth, but where the principal has max-min preferences when faced with Knightian uncertainty (henceforth, ambiguity). We normalize reserve utility to zero without loss of generality.

Let $\pi$ and $\phi \in \Phi$ represent the principal and agent's beliefs over some discrete set of possibilities, respectively. We will assume, through the course of this paper, that a simple non-triviality condition holds, such that both agent and principal always find at least one action to be desirable. This condition can be stated as follows:

$$\exists\ \underline{a} \in A\ s.t.\ E_\phi[y|\underline{a}] - c(\underline{a}) > 0\ \forall\ \phi \in \Phi$$

If we assume no structure over the set of possible beliefs $\Phi$, this directly implies that the least costly implementable action to the agent is financially smaller than the lowest level of output possible, that is:

$$\underline{y} > c(\underline{a})$$

The assumption that there exists some action that is profitable in aggregate under any state of nature can be somewhat strong. However, since there is a trade-off between imposing structure over costs or structures over beliefs - and the concern of this paper is mainly the effects of belief

heterogeneity - we will maintain it throughout our analysis. With this caveat in mind, let us now proceed to the timing of our game:

1. The principal offers a contract $w$.

2. The agent, upon learning $w$, chooses an action $a \in A$

3. Output $y \sim \pi$ is realized.

4. Payoffs are received: $y - w(y)$ to the principal, $w(y) - c(a)$ to the agent.

We will now focus on the principal's problem of maximizing her expected payoff subject to the possibility of heterogeneous beliefs in a plethora of settings. The next section focuses on the first of these.



## 2.1 An Omniscient Principal

Our first concern is with a principal who not only knows the agent's beliefs $\phi$, but who also observes which action the agent takes. Although it is rare that a principal has such encompassing knowledge - only God and Google come to mind - it is still of theoretical interest to understand the base case of contracting under heterogeneous beliefs before adding Moral Hazard to the equation. Hence, under this set of conditions, the principal's problem can be written as:

$$\max_{w(y,a)} E_\pi[y - w(y,a)|a] \quad \textbf{s.t.}$$
$$E_\phi[w(y)|a] - c(a) \geq 0 \qquad \text{(IR)}$$

It is not difficult to see how the principal will proceed under these conditions. Since she can observe actions, the optimal contract can be made contingent on them, eliminating any incentive compatibility constraints that could otherwise show up. Therefore, if the principal wishes to implement

action $a'$, she can set $w(y, a)$ as such:

$$w(y,a) = \begin{cases} w(y) & \text{if } a = a' \\ -\infty & \text{if } a \neq a' \end{cases}$$

Specifying $w(y)$ is also not difficult. If the agent has beliefs that do not coincide perfectly with the principal's, this implies that for some action $a' \in A$ and for some $y$:

$$\phi(y|a') \neq \pi(y|a')$$

Which means that $\exists\ y_1, y_2$ such that:

$$\phi(y_1|a') < \pi(y_1|a')$$
$$\phi(y_2|a') > \pi(y_2|a')$$

Which gives us our first theorem:

**Teorema 2.1** *The optimal contract $w^*(y|a')$ for an omniscient principal is:*

$$w^*(y) = \begin{cases} \alpha + c(a') & \text{if } y = y_1 \\ -\alpha\dfrac{\phi(y_1|a')}{\phi(y_2|a')} + c(a') & \text{if } y = y_2 \\ c(a') & \text{if } y \neq y_1, y_2 \end{cases}$$

In order to prove that this contract is optimal, let us first show that it satisfies the agent's incentive rationality constraint. The contract's expected payoff under $w(y, a')$ is:

$$E_\phi[w(y,a')|a'] = E_\phi[w(y)|a'] = c(a')$$

Where the equality between the second and third terms is guaranteed by showing that $E_\phi[w(y)|a'] = c(a')$. Therefore, regardless of the value of $\alpha$, the agent's participation is guaranteed. The principal's expected payoff $V_p(w,a')$, on the other hand, is *not* independent of $\alpha$:

$$V_P(w,a') = E_\pi[y|a'] + \alpha\left(\frac{\pi(y_2|a')\phi(y_1|a')}{\phi(y_2|a')} - \pi(y_1|a')\right) - c(a')$$

The heterogeneity conditions set above guarantee that the term in parentheses is strictly positive, since the agent is underestimating the probability of state $y_2$ where he must pay the principal and overestimating the probability of state $y_1$ where he is paid. Therefore, by increasing $\alpha$, the principal can increase his own expected payoff infinitely:

$$\alpha \to +\infty \implies V_P(w,a') \to +\infty$$

This guarantees that this contract is optimal for the principal since he can do no better from an ex-ante perspective than to provide himself with infinite utility.

By providing the agent with a contract where a probabilistically-weighted fee $\alpha$ is charged in one state of nature and paid in another, the principal can use the agent's biases to contract and exploit him. Therefore, under any situation in which actions are observable and both parties disagree about the conditional probability of some event, the "wage effect" takes over. The optimal contract, in this scenario, effectively becomes a bet.

Even though this contract is theoretically possible, in reality it is very unlikely that the agent's beliefs would not be affected by the contract offered

to him. Moreover, it seems rather unrealistic that any two parties signing a contract would be willing to sign "rich-or-ruin" contracts where arbitrarily large sums are paid with positive probability.

One way to make this problem somewhat more realistic is to assume that the principal cannot charge the agent under any contingency. Under the condition of limited liability, the optimal contract would not longer take the shape that it did before, since the principal's problem changes slightly to:

$$\begin{aligned} &\max_{w(y)} E_\pi[y - w(y)|a'] \textbf{ s.t.} \\ &E_\phi[w(y)|a'] \geq c(a') \qquad \text{(IR)} \\ &w(y) \geq 0 \forall y \in Y \qquad \text{(LL)} \end{aligned}$$

Since payoffs are linear and the principal need only convince the agent to participate, this problem can be simplified to choosing a single state of nature under which the agent should be paid. Mathematically, the problem becomes:

$$\min_y \quad c(a')\frac{\pi(y|a')}{\phi(y|a')}$$

And the agent gets paid only in the state of nature whose probability he most overestimates relative to the principal's beliefs.

**Teorema 2.2** *Let* $y' \in \underset{y}{argmin}\frac{\pi(y|a')}{\phi(y|a')}$. *The optimal contract* $w^*(y|a')$ *for an omniscient principal under limited liability is:*

$$w^*(y) = \begin{cases} \dfrac{c(a')}{\phi(y'|a')} & \text{if } y = y' \\ 0 & \text{if } y \neq y' \end{cases}$$

The proof is very simple. It is easy to see that both constraints that the principal has to respect are met. Therefore, we need only show that there is no contract that simultaneously satisfies both constraints and increases the principal's expected payoff. To do this, let us consider that one such contract exists, that is:

$$\exists\ w'(y) \neq w^*(y) \text{ such that } E_\pi[w'(y)|a'] < E_\pi[w^*(y)|a'] = c(a')\frac{\pi(y|a')}{\phi(y'|a')}$$

If this is true, then a contract must be specified where less is paid under state $y'$ and compensated under some other state or junction of states in order to meet the agent's incentive rationality constraint. Let's assume the reduction of $\Delta$ in payment under $y'$ is equalized as a payment $\gamma$ under some other state $y''$. In order to guarantee the agent's participation, $\gamma$ would have to be such that:

$$\gamma\phi(y''|a') \geq c(a') - \phi(y'|a')(\frac{c(a')}{\phi(y'a|a')} - \Delta)$$

And the principal's payment changes by:

$$E_\pi[w'(y)|a'] - E_\pi[w^*(y)|a'] = \pi(y''|a')\Delta\frac{\phi(y'|a')}{\phi(y''|a')} - \pi(y'|a')\Delta$$

For the principal to benefit, the expected value of the first contract must be smaller than the second to the principal. However, this would imply:

$$\pi(y''|a')\Delta\frac{\phi(y'|a')}{\phi(y''|a')} - \pi(y'|a')\Delta < 0 \implies \frac{\pi(y''|a')}{\phi(y''|a')} < \frac{\pi(y'|a')}{\phi(y'|a')}$$

Which is absurd, since $y' \in \underset{y}{\text{argmin}}\frac{\pi(y|a')}{\phi(y|a')}$. Interestingly, this proves that the optimal contract has a bonus structure *even in the absence of moral hazard*, so long as principal and agent have different beliefs over the conditional probability distribution of outcomes. This might suggest that principals should shift part of their payment structure to bonuses even when there are no conflicts of incentive if agents exhibit overconfidence in their own

abilities.

In a setting with limited liability, the principal's gain depends on how much her opinions diverge from the agent's. Her expected payoff, after all, is:

$$E_\pi[y - w^*(y)|a'] = E_\pi[y|a'] - c(a')\frac{\pi(y'|a')}{\phi(y'|a')}$$

If the two parties have marginally different beliefs, not much can be gained over the optimal contract in a setting with homogeneous beliefs (where $\phi(y|a') = \pi(y|a')$ and the optimal contract has expected payment $c(a')$). On the other extreme, if $\phi(y'|a') > 0,\ \pi(y'|a') > 0$, then the principal gets to keep an extra $c(a')$ for herself. Limited liability, therefore, also limits the principal's gains from exploitation to simply not having to pay the agent.



## 2.2 Moral Hazard Under Belief Heterogeneity

Consider now a situation where an agent's beliefs are common knowledge but his actions cannot be observed. If beliefs are homogeneous, our problems becomes one of moral hazard with risk-neutral players. Under these circumstance, selling the firm at cost $E_p i[y|a'] - c(a')$ is optimal, since it aligns the agent's incentives with the principal's by making him bear all outcome-related risk.

A more interesting situation arises if the principal and agent disagree as to how actions map to outcome - a situation that seems to be more common in practice. Consider, for instance, a thesis advisor and a prospective advisee. The advisor, having counseled numerous students before, and having been a student herself, knows how much effort an advisee of particular skill needs to put into his work in order to produce results that are, on average, decent. The student, on the other hand, might be overconfident or excessively insecure. In order to convince the student to take her as an advisor, yet to produce the best work possible, the advisor needs to provide incentives for the student to exert effort which she cannot observe with scaring him off the program. It is this type of situation which we wish to characterize.

The principal's problem when faced with moral hazard and an agent with beliefs different from her own can be expressed as:

$$\begin{aligned}
&\max_{w(y)} E_\pi[y - w(y)|a'] \textbf{ s.t.} \\
&E_\phi[w(y)|a'] - c(a') \geq 0 \qquad \text{(IR)} \\
&E_\phi[w(y)|a'] - c(a') \geq E_\phi[w(y)|a] - c(a) \qquad \text{(IC)}
\end{aligned}$$

Let us again consider that the agent's beliefs are biased for at least two states of nature $y_1, y_2$ for some action $a'$ that the principal wishes to implement. Recall the contract $w^*(y)$ from **Theorem 1**:

$$w^*(y) = \begin{cases} \alpha + c(a') & \text{if } y = y_1 \\ -\alpha\dfrac{\phi(y_1|a')}{\phi(y_2|a')} + c(a') & \text{if } y = y_2 \\ c(a') & \text{if } y \neq y_1, y_2 \end{cases}$$

**Teorema 2.3** *$w^*(y|a')$ is optimal under belief heterogeneity with unobservable actions if and only if:*

$$\frac{\phi(y_1|a')}{\phi(y_2|a')} \geq \frac{\phi(y_1|a)}{\phi(y_2|a)} \tag{2-1}$$

The condition above says that for the principal to exploit the agent's biases infinitely - as she did in the absence of limited liability when actions were observable - the ratio of payment-to-charge under action $a'$ must exceed or equal that same ratio under any alternative action $a \in A$ available to the agent.

The proof of this theorem is also quite straightforward. We have shown previously that this contract can give the principal infinite utility and that it satisfies the agent's IR constraint. Therefore, all we need show is that it satisfies the agent's incentive compatibility constraint, that is:

$$E_\phi[w(y)|a'] - c(a') \geq E_\phi[w(y)|a] - c(a) \implies$$

$$0 \geq c(a') - c(a) + \alpha\phi(y_1|a) - \alpha\frac{\phi(y_1|a')}{\phi(y_2|a')}\phi(y_2|a) \; \forall \; a \in A$$

With some simple algebraic manipulation, we can get to:

$$\frac{\phi(y_1|a')}{\phi(y_2|a')} \geq \frac{\phi(y_1|a)}{\phi(y_2|a)} + \frac{c(a') - c(a)}{\alpha\phi(y_2|a)}$$

Notice that the third term in the equation converges to zero as $\alpha \to \infty$, allowing us to get to condition (1).

The reasoning behind this result is quite intuitive: if the agent believes that the relative likelihood of getting paid is greater under $a'$ than under $a$, then it is more profitable ex-ante to undertake action $a'$. Moreover, since the IR constraint is binding under $a'$ with contract $w^*(y)$, $a$ is not profitable for the agent as $\alpha \to \infty$. Finally, notice that the likelihood ratio conditions is exactly what allows the principal to exploit the agent, since the latter does not realize that:

$$\frac{\phi(y_1|a')}{\phi(y_2|a')} \geq \frac{\pi(y_1|a')}{\pi(y_2|a')} + \frac{c(a') - c(a)}{\alpha\phi(y_2|a)}$$

This theorem, thence, has an intuitive appeal. Whenever the principal can clearly see that the agent is so biased about the relative likelihood of some state of nature for action $a'$ that he believes it is both greater than it would be under any other action $a \in A$ and under the principal's own beliefs, the agent will be infinitely exploited. This provides us with the following corollary:

**Corolário 2.4** *If there exists $y_2$ such that for all $\phi(y_2|a') = 0$ we have $\pi_2(y_2|a') > 0$, the agent will accept contract $w^E(y)$ and be infinitely exploited, where:*

$$w^E(y) = \begin{cases} c(a') & \text{if } y \neq y_2 \\ -\infty & \text{if } y = y_2 \end{cases}$$

The proof is again trivial, given that $\phi(y_2|a') = 0$ implies:

$$\frac{\phi(y_1|a')}{\phi(y_2|a')} \to +\infty \implies \frac{\phi(y_1|a')}{\phi(y_2|a')} \geq \frac{\phi(y_1|a)}{\phi(y_2|a)} \text{ and } \frac{\phi(y_1|a')}{\phi(y_2|a')} > \frac{\pi(y_1|a)}{\pi(y_2|a)}$$

And our corollary, hence, demonstrates that if for some group of actions $A' \subseteq A$ the support of $\phi(y|a)$, $a \in A'$ does not contain the support of $\pi(y|a), a \in A'$, the principal can once again offer the agent a contract that will exploit him infinitely. As such, it is clear that as the number of states of nature whose probability the principal and agent disagree on increases, it becomes ever more likely that the principal will have a greater chance to implement the exploiting contract.

## 2.3 Robust Moral Hazard

In the real world, there are many situations under which a principal does not know what an agent's beliefs are. An insurance company does not know if an agent it contracts with believes he is extraordinarily unlucky or wildly fortunate. A company hiring a working can sometimes only partially infer the employee's thoughts about how his effort interacts with his skill. A king giving a servant the task of killing a baby does not know how the probability distribution of outcomes that the servant ascribes to his actions.

The natural next step in our analysis is therefore to consider what happens when a principal is unaware of an agent's beliefs. To do this, we will consider that when the principal is faced with unquantifiable (or Knightian) uncertainty, she will require contracts to be robust in a max-min sense. In other words, our principal has infinite ambiguity aversion, thus evaluating payoffs by their worst-case scenario. Moreover, in this section we will restrict our attention to the case in which, if action $a'$ has cost greater than action $a$, then $\pi(y|a')$ second-order stochastically dominates $\pi(y|a)$, that is:

$$c(a') > c(a) \implies E_\pi[y|a'] > E_\pi[y|a]$$

Let us first consider what happens under the most general of settings, where the principal has no information about the set $\Phi$ of beliefs that an

agent can have. In this scenario, our principal must solve the following problem:

$$\max_{w(y)} \min_{\phi \in \Phi} E_\pi[y - w(y)|a'] \textbf{ s.t.}$$
$$E_\phi[w(y)|a'] - c(a') \geq 0 \qquad (IR)$$
$$E_\phi[w(y)|a'] - c(a') \geq E_\phi[w(y)|a] - c(a) \ \forall \ a \in A \quad (IC)$$

A rather savorless result follows:

**Teorema 2.5** *Only the least costly action $\underline{a} \in A$ is implementable when the principal has max-min preferences and no knowledge of the agent's beliefs.*

The proof is - once more - trivial.

Let two actions $a$ and $a'$ with costs $c(a) > c(a')$ have the same probability distribution $\phi(y|a) = phi(y|a')$. If actions are not observable and contracts are only contingent upon results, the incentive compatibility constraints will never be met for the more expensive action, since:

$$E_\phi[w(y)|a] - c(a) < E_\phi[w(y)|a'] - c(a')$$

Since the agent's beliefs are absolutely unrestricted, the principal will thence be unable to implement any action except the least expensive one.

The result above is somewhat uninteresting. If the principal is limited to hiring the agent to perform the least costly action, he will pay the agent a flat fee equal to the agent's cost $c(\underline{a})$ and receive payoff $E_\pi[y|\underline{a}] - c(\underline{a})$. Notice that this result holds as well under limited liability.

If we assume some, albeit not much, structure over the agent's beliefs, we can "impose" the implementability of other actions without much loss of generality.

**Definição 2.6** *We say that the agent and principal have* ***aligned beliefs over aggregate surplus*** *if, for all $\phi \in \Phi$:*

$$E_\pi[y|a'] - c(a') \geq E_\pi[y|a] - c(a) \implies$$
$$E_\phi[y|a'] - c(a') \geq E_\phi[y|a] - c(a)$$

Aligned beliefs over aggregate surplus (henceforth ABAS) are not the most restrictive of conditions. All that this condition states is that, if the principal finds that some action $a \in A$ increase the total surplus of the economy, so does the agent. Effectively, this conditions reduces the region in which $\phi(y|a')$ can be for any action other than $\underline{a}$, since $E_\phi[y|a'] > \underline{y} + c(a') - c(\underline{a})$.

This brings us to our most important theorem. When faced with an agent whose beliefs she is almost entirely unaware of, the principal can still implement any action $a' \neq \underline{a}$ that increases the economy's aggregate surplus under her own beliefs.

**Teorema 2.7** *When ABAS is valid, the optimal robust contract $w^r(y)$ when implementing $a' \neq \underline{a}$ is:*

$$w^r(y) = y + c(\underline{a}) - \underline{y} \tag{2-2}$$

Theorem 5 states that if the principal wishes to implement some action other than the cheapest one, the optimal contract is to sell the firm at the very low price $y - c(\underline{a})$. The proof is as follows:

First, let us verify that $w^r(y)$ satisfies the agent's incentive constraints. Let us start with the agent's participation:

$$E_\phi[w^r(y)|a'] - c(a') = E_\phi[y|a'] - \underline{y} - c(a') + c(\underline{a})$$

Which we know is positive due to ABAS, since:

$$E_\phi[w^r(y)|a'] - c(a') \geq E_\phi[y|\underline{a}] - c(\underline{a}) \geq \underline{y} - c(\underline{a}) \geq 0$$

Where the last inequality holds due to non-triviality. Now, let us look at the incentive compatibility constraint:

$$E_\phi[w^r(y)|a']-c(a')-E_\phi[w^r(y)|a]-c(a) = E_\phi[y|a']-\underline{y}-c(a')+c(\underline{a})-E_\phi[y|a]-\underline{y}-c(a)+c(\underline{a}) \implies$$

$$E_\phi[y|a'] - c(a') - E_\phi[y|a] - c(a) \geq 0$$

Where the final inequality holds due to ABAS.

Since the agent's constraints are satisfied, all that we must check now is the optimality of $w^r(y)$. To do so, let us assume by contradiction that there exists $w'(y)$ that is cheaper for the principal than $w^r(y)$. That is:

$$E_\pi[w'(y)|a'] < E_\pi[w^r(y)|a'] = E_\pi[y|a'] + c(a') - \underline{y} \tag{2-3}$$

For contract $w'(y)$ to be optimal, we know that it must satisfy the agent's constraints for any beliefs $\phi \in \Phi$ as well. We will show that it does not by looking at only two pairs of probability distributions. Let $\phi(y|a') = \pi(y|a') = u^P$, where $u^P$ is a vector of probabilities.

By the convexity of probability simplexes, we know that there exists $\alpha$ such that $\phi(y|\underline{a}) = v = \alpha \cdot u^P + (1-\alpha) \cdot \delta_{\underline{y}}$ satisfies ABAS with equality for action $\underline{a}$, that is:

$$y \cdot (u^P - v) = c(a') - c(\underline{a})$$

Moreover, for any action $a \neq a', \underline{a}$ whose aggregate surplus $E_\pi[y|a] - c(a)$ lies between that of any two actions (in particular, between the surplus of action $a$ and action $\underline{a}$), we can also guarantee that there exists a conditional probability distribution that satisfies ABAS. All we need notice is that there exists $\beta$ that satisfies the equation below.

$$y \cdot (1-\beta) \cdot (u^P - \delta_y) = c(a') - c(a)$$

Given that there exists a distribution $\phi(y|a) = \beta u^P + (1-\beta)\delta_{\underline{y}}$ under which the condition above is met for $\beta \in (0,1)$ and given that if $c(a) > c(\underline{a})$ then $\pi(y|a)$ SOSD $\pi(y|\underline{a})$, it is easy to see that for any action ABAS can be satisfied by picking distributions such as the one explored above.

Now we must guarantee also that If $w'(y)$ is optimal, then it must satisfy an IC constraint for $(u^P, v)$, that is:

$$w' \cdot (u^P - v) = (1-\alpha) \cdot w' \cdot (u^P - \delta_{\underline{y}}) = w' \cdot u^P - w_{\underline{y}} \geq c(a') - c(a) \implies$$

$$w_{\underline{y}} \leq \frac{c(a) - c(a')}{1-\alpha} + w' \cdot u^P < \frac{c(a) - c(a')}{1-\alpha} + w^r \cdot u^P$$

Where the last condition comes from condition (4) rewritten as a dot product.

Let us now consider that the agent has a different set of beliefs. Consider $\phi(y|\underline{a}) = \delta_{\underline{y}}$. Again, by the convexity of probability simplexes, we know that there exists $u'$ such that $u' = \gamma \cdot u^P + (1-\gamma) \cdot \delta_{\underline{y}}$ satisfies ABAS with equality (and that the same can be extended to all $a \in A$ by choosing a different value for $\gamma$):

$$y \cdot (u' - \delta_{\underline{y}}) = c(a') - c(\underline{a})$$

Which implies $\gamma = 1 - \alpha$. Therefore, $u' = (1-\alpha) \cdot u^P - \alpha \cdot \delta_{\underline{y}}$.

For $w'(y)$ to be optimal, it must also satisfy an IR constraint for $u'$. That is:

$$w' \cdot u' = w' \cdot ((1-\alpha) \cdot u^P + \alpha \cdot \delta_{\underline{y}}) - c(a') \geq 0 \implies$$

$$w_{\underline{y}} \geq \frac{(\alpha - 1) \cdot w' \cdot u^P + c(a')}{\alpha} > \frac{(\alpha - 1) \cdot w^r \cdot u^P + c(a')}{\alpha}$$

By merging together the IC and IR constraints that we have put together, we find that $w_{\underline{y}}$ must be such that:

$$\frac{(\alpha - 1) \cdot w^r \cdot u^P + c(a')}{\alpha} < w_{\underline{y}} < \frac{c(a) - c(a')}{1-\alpha} + w^r \cdot u^P \qquad (2\text{-}4)$$

Which, after a bit more algebraic manipulation, will be found to be absurd.

What Theorem 5 shows us is simply that no contract can simultaneously satisfy ABAS and yield superior ex-ante utility to the Principal than the sale of the firm at price $\underline{y} - c(\underline{a})$ without breaching one of the agent's constraints

for some belief of his.

Two corollaries immediately follow.

**Corolário 2.8** *The firm's sale at price $\underline{y} - c(\underline{a})$ is also optimal under limited liability.*

The proof follows directly from the proof of theorem 5. All that is left to show is that the agent is never charged anything by the principal under any state of nature. To do this, we need only look at what happens to him under the worst possible state of nature. To do this, we need only look at what happens to him under the worst possible state of nature $\underline{y}$:

$$w^r(\underline{y}) = c(\underline{a}) > 0$$

Our next corollary demonstrates that implementing any action $a \neq \underline{a}$ is almost surely suboptimal for the principal:

**Corolário 2.9** *It is almost surely more beneficial for the principal to implement action $\underline{a}$ than any other implementable action.*

Again, the proof is direct. The principal's payoff under action $\underline{a}$ by paying the agent $c(\underline{a})$ is:

$$E_\pi[y - w(\underline{y})|a] = E_\pi[y|a] - c(\underline{a}) \geq \underline{y} - c(\underline{a})$$

With the inequality holding strictly whenever $E_\pi[y|a] > \underline{y}$. Therefore, the principal is never strictly better off by implementing any action $a \neq \underline{a}$, and is worse off whenever he does so and $E_\pi[y|a] > \underline{y}$. Belief uncertainty, even when slightly mitigated, leads to a Pareto deterioration of payoffs when compared to payoffs generated under homogeneous beliefs and common knowledge.

# 3
# Conclusion

Our paper has focused on the impact of belief heterogeneity in a setting with moral hazard and two-sided risk neutrality. In all three scenarios analyzed, either the optimal shape of contracts or their profitability was changed relative to similar frameworks where the principal and agent had homogeneous beliefs. contract shapes were also changed relative to their benchmark when limited liability was introduced.

Our foray into moral hazard with heterogeneous beliefs began by looking at a situation where informational asymmetries were absent, but belief heterogeneity was present. We observed that, in this situation of omniscience, a principal could always exploit the agent he was contracting with if they had dissimilar beliefs. Under the same setup with limited liability, we found that the agent could no longer be explored beyond a certain extent, and that his biases did not always lead to the same amount of loss relative to the contract they would receive under belief homogeneity.

We proceeded with our analysis by delving into a world where moral hazard exists - actions are not observable - but belief heterogeneity persists. We found that when an agent's beliefs were "optimistic" relative to those of the principal, he would be exploited infinitely. We continued by looking at what happened when the agent's beliefs did not have the same support as the principal's and looked at when the agent was "safe" from absolute exploitation.

Our final - and perhaps most interesting - concern was related to a situation of complete uncertainty where the principal did not know what the agent's beliefs were. We finished by describing the optimal contract under three separate conditions, showing how varying these made different sets of actions possible and attractive, and how introducing uncertainty and the necessity for robustness leads to a Pareto deterioration in welfare.