Discussion Paper Series – CRC TR 224

Discussion Paper No. 521
Project B 02

Confidence and Organizations

Andres Espitia[1]

March 2024

[1]University of Bonn, Email: aespitia@uni-bonn.de

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Collaborative Research Center Transregio 224 - www.crctr224.de
Rheinische Friedrich-Wilhelms-Universität Bonn - Universität Mannheim

Confidence and Organizations*

Andrés Espitia†

March 6, 2024

Abstract

Miscalibrated beliefs generally compromise the quality of workers' decisions. Why might a firm prefer to hire an individual known to be overconfident? In this paper, I explore the role of such biases when members of the organization disagree about the right course of action. I present a model in which an agent uses his private information to make a choice on behalf of a principal. In this setting, I consider what I call *the belief design problem*: how would the principal like the agent to interpret his observations? I provide conditions under which the solution indicates a preference for a well-calibrated, an underconfident, or an overconfident agent. A well-calibrated agent is preferred if and only if his information does not affect the expected difference in the players' preferred actions. Overconfidence is optimal when the principal seeks to adjust actions beyond what a well-calibrated agent would do.

JEL Codes: D82, D83, D91

Keywords: principal-agent, overconfidence, belief design

*I am grateful to David Besanko, Luis Rayo, Daniel Barron, and Wojciech Olszewski for their guidance and encouragement. I would also like to thank Sarah Auster, Henrique Castro-Pires, Francesc Dilme, Théo Durandard, Alkis Georgiadis-Harris, Benjamin Golub, Botond Kőszegi, David Laibson, Ulrike Malmendier, Andrei Matveenko, Edwin Muñoz-Rodríguez, Michael Powell, Rosina Rodríguez-Olivera, Bruno Strulovici, Alvaro Sandroni, Jeroen Swinkels, Boli Xu, Gabriel Ziegler, as well as participants at various seminars and conferences for their insightful comments and suggestions. Support by the Deutsche Forschungsgemeinschaft (German Research Foundation) through grant CRC TR 224 (Project B02) is gratefully acknowledged.

†Institute for Microeconomics - University of Bonn. Email: aespitia@uni-bonn.de.

1 Introduction

There is extensive evidence about the pervasiveness of overconfidence in organizations (Malmendier and Taylor, 2015). CEOs, executives, and managers (Malmendier and Tate, 2005; Ben-David et al., 2013; Adam et al., 2015; Barrero, 2022); traders and investors (Daniel and Hirshleifer, 2015); lawyers (Goodman-Delahunty et al., 2010); medical doctors (Berner and Graber, 2008; Croskerry and Norman, 2008); and entrepreneurs (Koellinger et al., 2007) have all been found to exhibit some degree of overconfidence. This behavioral bias is often perceived to compromise the quality of employees' decisions, and being detrimental to the firms' performance.[1] Given the costs that organizations suffer due to overconfidence, why might employees *known* to be overconfident be systematically hired and retained, even when their unbiased counterparts are available? Additionally, how would organizations use their knowledge about applicants' confidence to select their employees optimally?

In this paper, I explore the role of biased beliefs as an instrument to alleviate *agency frictions* (those arising from different preferences among the members of an organization). I focus on a particular manifestation of overconfidence: *overprecision* – an exaggerated faith in one's information. Informally speaking, individuals suffer from overprecision whenever they think that they are more informed than they actually are.[2] Intuitively, an overconfident employee is desired by the firm when their well-calibrated counterpart is too *unresponsive* to new information. Overprecision leads to over-updating (more extreme beliefs) and to overreaction (more extreme actions), which helps the firm, for example, when well-calibrated employees would otherwise stay *too* close to some reference action.

In order to examine the implications of this mechanism, I develop a model in which an employee makes a decision that affects both himself and the firm. While there is disagreement about the right course of action, the employee has private access to relevant but unverifiable information. In this setting, I introduce a novel feature: the firm can choose how the employee interprets his observations. I refer to this decision by the firm as the *belief design problem*.

Formally, belief design corresponds to an optimization problem in which the firm chooses the employee's posterior beliefs after each signal realization, subject to the constraint that they both agree ex-ante on the distribution of the state and signals. In practice, it can be understood as a process in which the firm selects employees with the desired beliefs among a pool of candidates with sufficiently diverse views of the world. I focus on how the employee interprets information

[1] In the words of Nobel laureate Daniel Kahneman, "an unbiased appreciation of uncertainty is a cornerstone of rationality – but it is not what people and organizations want" (see Kahneman, 2011, chap. 24). Furthermore, "Kahneman recently told an interviewer that if he had a magic wand that could eliminate one human bias, he would do away with overconfidence" (Malmendier and Taylor, 2015, pg. 1).

[2] Other forms of overconfidence include overestimation (thinking that one's performance and abilities are above their actual level) and overplacement (erroneously thinking that one has outperformed others or that one's abilities are above those of other individuals). See Moore and Healy (2008) for a discussion on the connections and differences between these manifestations of overconfidence.

exogenously produced. This is in contrast with the information design approach (see Kamenica, 2019) which focuses on the actual provision of information to the employee.

The contributions of this paper are threefold. First, the main contribution is to provide specific conditions under which the firm prefers a well-calibrated, underconfident, or overconfident employee, which I discuss below. The second contribution corresponds to the methodology used to establish these conditions. Belief design is a flexible and tractable approach to study belief-based biases. A crucial step is to interpret the beliefs optimally chosen by the firm. I propose a definition of overprecision based on the *concordance stochastic order* that applies to a general class of information structures. This allows me to expand on the standard approach where we would assume a particular joint distribution and restrict biased-beliefs to belong to the same parametric family. Two commonly assumed structures are the bi-variate normal distribution and the *truth-or-noise* (in which the signal reveals the states with a certain probability and it is noise otherwise). This last case is discussed in Appendix B. Third, I study the interaction between belief-based selection of employees and other tools that the firm could use to reduce the costs of agency frictions. In particular, I allow the firm to commit to action-contingent transfers as well as to centralize decision making.

A central object to describe the optimal employee's characteristics is the difference in the actions preferred by each player in each state, which I refer to as the *conflict of interests*. Since the employee is ex-ante well-calibrated, his actions will be biased *on average*. Under the standard assumption of quadratic-loss preferences, the firm dislikes changes in the employee's actions that are not justified by his private information. I show that the optimal employee agrees with the firm on the *responses* to his information, i.e., the bias on his actions does not change with signal realizations. As a result, the firm prefers a well-calibrated employee if and only if the expected conflict of interest is invariant in the signal realization (first part of Proposition 1, generalized in Proposition 2).

In contrast, if the signal affects the expected conflict of interest, evenly distributing the employee's bias requires that the optimal employee takes lower actions than his well-calibrated counterpart after signal realizations that induce high expected conflict of interests. The critical condition for the optimality of overconfidence is that the signal moves the conditional expectation of the conflict of interest and the employee's preferred action in *opposite* directions. If this is the case, the optimal agent takes more extreme actions than the well-calibrated one, which is a manifestation of overconfidence. Analogously, an underconfident employee is optimal if the expected conflict of interests and the employee's preferred action move in the *same* direction.

Optimality of overconfidence arises in situations in which, for example, the employee's preferred action increases less than proportionally than the firm's preferred action. This may occur because it is costly for the employee to adjust to the current conditions or because he is subject to some degree of status quo bias.[3] Thus, the demand for overconfidence may arise as a strategy

[3]It is sometimes assumed that adjustment costs are paid by the firm (see Barrero, 2022). Absent any additional

to mitigate the effects of other pervasive behavioral biases, status quo bias in decision-making being a salient example.

I additionally explore the effect of alternative tools that organizations may use to alleviate agency frictions. First, I study the interaction between belief design and action-contingent transfers. I discuss conditions under which the use of transfers does not change the optimal beliefs (Proposition 4). Moreover, in the optimum, these tools are used for different purposes: beliefs are used to spread the employee's bias across signal realizations, while transfers are used to decrease his average bias. Interestingly, belief design leads to "flatter" compensation contracts: all equilibrium actions taken by the optimal employee yield the same transfer. In contrast, when the employee is restricted to be well-calibrated, the optimal transfers vary with the actions he takes in equilibrium.

Finally, while modeled symmetrically in this paper, overconfidence and underconfidence have a stark difference: there is a natural substitute for extreme underconfidence. Namely, the firm can make the choice without the information held by the employee. When the firm allocates decision rights to the employee, it is only with the hope of using their private information. When the optimal employee is not sufficiently confident, the firm is better off avoiding the conflict of interests altogether by centralizing decision-making (Proposition 5).

Related literature. The interest in the economic effects of overconfidence is long-standing.[4] The term, however, has been used to describe quite distinct phenomena (see Moore and Healy, 2008). In this paper I entirely focus on *overprecision*, which is considered the most robust and least understood form of overconfidence (Haran et al., 2010; Moore and Schatz, 2017).

The potential benefits of overconfidence (broadly defined) have been discussed in some specific settings. Such benefits can be classified roughly into three groups. First, overconfidence may serve as a strategic commitment device, allowing the firm or some of its members to implement strategies that would not otherwise be credible (Kyle and Wang, 1997; Rotemberg and Saloner, 2000; Van den Steen, 2005; Gervais and Goldstein, 2007; Englmaier, 2011; Bolton et al., 2013; Englmaier and Reisinger, 2014; Phua et al., 2018; Ba and Gindin, 2023). Second, it may permit more favorable acquisition, revelation, and aggregation of private information (Bernardo and Welch, 2001; Blanes-i Vidal and Möller, 2007; Che and Kartik, 2009; Van den Steen, 2010; Levy and Razin, 2015; Hestermann and Le Yaouanq, 2020; Ilinov et al., 2022; Ostrizek, 2022). Finally, it may allow the firm to have better control over risk-taking and sharing as well as facilitating diversification (Goel and Thakor, 2008; Santos-Pinto, 2008; De la Rosa, 2011; Gervais et al., 2011; Palomino and Sadrieh, 2011; Heller, 2014). I contribute to this literature by linking the characteristics of the optimal employee to the conflicting preferences of the members of the organization. This approach allows us to study the drivers of the demand for overconfidence (manifested as overprecision) without needing to specify the details of the idiosyncratic

frictions, my results support the conclusion that any bias in the beliefs would be detrimental for the firm.

[4] Adam Smith stated that "the over-weening conceit which the greater part of men have of their own abilities, is an ancient evil remarked by the philosophers and moralists of all ages" (Smith, 1776, chap. X, book I).

interactions among the members of the organization.

While a significant portion of the literature takes the biases in beliefs as given, there is a growing interest in making those misspecifications endogenous. The study of persuasion through the provision of *narratives* (the relationship between observed signals and outcomes) is particularly related (Eliaz and Spiegler, 2020; Schwartzstein and Sunderam, 2021; Jain, 2023; Ispano, 2023; Aina, 2024). In the context of the current paper, the choice variable for the firm (i.e., how the employee interprets his private observation) is itself a narrative. However, two differences emerge. First, the common approach is that narratives can be made contingent of the realized data. An exception is Ispano (2023), who compares that approach with narratives that are provided before the information is available. In that case, the persuader faces similar constraints as the ones used in the belief design problem. Second, I provide an interpretation for the optimal narrative in the light of one of the most ubiquitous type of behavioral bias, i.e., overconfidence.

Similar to the literature on narratives, Niu (2023) also makes the employee's misspecifications endogenous. In that case, the firm can distort the employee's perception about the difficulty or quality of the task or project he is working on. Those distortions affect the inferences the employee makes about his own ability from the history of outcomes. The resulting biases can be interpreted as a degree of overoptimism on the project.

Finally, I illustrate how belief-based selection can be used as an indirect source of incentives. Numerous tools that organizations can use to deal with conflicting preferences have been studied. One possibility consists of exploiting individual characteristics to improve outcomes (Prendergast and Stole, 1996; Prendergast, 2007, 2008). I contribute to this literature by considering beliefs as part of those characteristics. The literature on information design (Rayo and Segal, 2010; Kamenica and Gentzkow, 2011; Kamenica, 2019) shares the same object of choice, but focuses on the information that employees actually observe rather than on how they interpret exogenously available information. Alternatively, the delegation literature (Holmström, 1977, 1984; Alonso and Matouschek, 2008) studies the use of rules on the set of available actions from which the employee can choose. I build on a similar framework: an organization formed by two individuals with different preferences over possible decisions, where there is a mismatch between authority and information, and where contingent transfers are infeasible. In contrast to that literature, I focus on employees who are not fully informed and may have misspecified beliefs about the amount of information they posses.

2 Model

Preliminaries. For any finite set $X \subset \mathbb{R}$, I use $\Delta(X)$ to denote the set of probability mass functions over X. Subscripts on operators are used to explicitly specify the probability mass function being used, e.g., $\mathbb{E}_g$ indicates that the expectation is taken according to the distribution $g \in \Delta(X)$. For any joint distribution $g \in \Delta(X \times X')$, let $g_X \in \Delta(X)$ and $g_{X'} \in \Delta(X')$ denote its marginals. Finally, all variables with a tilde are random variables.

Players and actions. An agent (*he*) makes a decision $x \in \mathbb{R}$ that also affects a principal (*she*). Payoffs depend on a state of the world $\theta \in \Theta := \{\theta_1, \dots, \theta_n\} \subset \mathbb{R}$, with $n \geq 2$ finite. States are labeled such that $\theta_1 < \cdots < \theta_n$. Ex-post payoffs are given by $-(x-\theta)^2$ for the principal and by $-(x-y(\theta))^2$ for the agent. That is, the state represents the principal's preferred action. On the other hand, the agent's preferred action is given by the *bias function* $y : \Theta \to \mathbb{R}$. I assume that y is strictly increasing. This provides some minimum degree of alignment in the players' preferences. The difference between the players' preferred actions in a given state is denoted by $c(\theta) := y(\theta) - \theta$ and I refer to it as the *conflict of interest* in state θ.

Information. The agent has private and non-verifiable information about the state of the world. He observes a signal realization $s \in S := \{s_1, \dots, s_m\}$, with $m \geq 2$ finite. From the point of view of the principal, states and signals are distributed according to $f \in \Delta(\Theta \times S)$, which is an $n \times m$ matrix with ij-th entry equal to $f(\theta_i, s_j) := Pr_f[\tilde{\theta} = \theta_i, \tilde{s} = s_j]$. I refer to f as the *true distribution* and assume it has full-support, i.e. $f(\theta, s) > 0$ for all $(\theta, s) \in \Theta \times S$. Higher states are more likely after higher signal realizations, in particular, signals are labeled such that $s_j = \mathbb{E}_f[\tilde{\theta} | s_j]$.

Belief design. The key feature of the model is that the principal is allowed to choose how the agent interprets his private information. Specifically, she chooses a distribution $g \in \Delta(\Theta \times S)$ such that after observing a given signal realization the agent computes his posterior beliefs according to g (instead than according to f). This step can be interpreted as a selection or hiring process in which the principal chooses an agent that already posses the desired beliefs. Alternatively, it can also reflect the principal's ability to (costlessly) train the agent on how to interpret his information.[5]

I impose two restrictions on the set of joint distributions the principal can select. First, I assume that players *ex-ante* agree on the distribution of the state, i.e., $g_\Theta = f_\Theta$. Additionally, I require the agent to be well-calibrated about the frequency of signal realizations, i.e., $g_S = f_S$. These restrictions allows us to focus on the information the agent perceives after each signal realization. Let $\mathcal{G} := \{g \in \Delta(\Theta \times S) : g_\Theta = f_\Theta, g_S = f_S\}$ denote the set of feasible choices for the principal.

[5] See Gervais and Odean (2001), Haran et al. (2010) and Meikle et al. (2016) for practices that can mitigate or exacerbate miscalibration.

Timing. The timing of events is as follows:

1. Belief design: the principal chooses $g \in \mathcal{G}$.
2. Nature draws (θ, s) according to f.
3. The agent observes s, interpreting it according to g, and chooses $x \in \mathbb{R}$.

Note that the description of the timing assumes that the agent learns nothing from the principal's choice. This is, interpreting belief design as a selection process, candidates' beliefs remain the same whether they are selected by the principal or not.

Applications. I now discuss two specific applications. The purpose is to illustrate the type of situations in which the forces captured by the model are relevant. We will revisit these applications to illustrate the implications of the results.

First, consider a CEO (the principal) selecting a middle manager (the agent) in charge of compensating a subordinate. Specifically, the manager chooses a level of reward $x \in \mathbb{R}$ for his subordinate after having privately observed a signal of the subordinate's actual performance. The state represents the CEO's ideal level of reward given the subordinate's actual performance, which is never directly observed. The signal is normalized to represent the CEO's ideal level of reward given some noisy measure of the subordinate's performance.

A plausible concern for the CEO is that the manager would be overly reluctant to provide low rewards to his subordinate. In other words, while the two parties may be relatively aligned when the subordinate deserves a bonus or a promotion, the manager may be averse to fire the subordinate, even when the CEO wants exactly that.

Second, consider the director of a nuclear power plant (the principal) selecting a risk manager (the agent) in charge of monitoring and controlling the safety in the plant's operations. The manager chooses a level of risk abatement $x \in \mathbb{R}$ after having privately observed a signal about the safety conditions of the operations. The state represents the director's ideal abatement level given the actual performance of the plant. As in the previous application, the signal represents the director's ideal abatement level given some noisy measure of the performance of the plant. In this case, both individuals would agree about the right course of action when the risk is high: avoiding an accident is a mutual goal when problems are likely to occur. On the other hand, when the risk is sufficiently low, the manager does not see the need to keep the abatement at the levels desired by the director.

In both contexts we ask the following question: when is an over/under-confident manager optimal? This question, however, is not entirely meaningful without a definition of overconfidence. This is precisely the next step.

Defining overconfidence. Our goal is to provide conditions under which the optimal agent displays beliefs that can be interpreted as over/under-precision. If the optimal agent's beliefs coincide with the ones prescribed by the true distribution f, then we say that he is *well-calibrated*. Otherwise, we need to compare a given solution g^* with f. I propose using a well-known stochastic order to make such comparisons in a way that reflects the notion of over/under-precision. This order is based on the concept of concordance, which informally corresponds to large values of the state going together with large values of the signal. Thus, an increase in the concordance between these two random variables can be interpreted as the signal "revealing more" about the state.

Definition 1 (Concordance Order) *$g \in \mathcal{G}$ dominates f in the **concordance order**, denoted $g \succeq f$, if and only if for all $k \in \{1,\dots,n\}$ and $l \in \{1,\dots,m\}$ we have $\sum_{i=1}^{k}\sum_{j=1}^{l} g(\theta_i, s_j) \geq \sum_{i=1}^{k}\sum_{j=1}^{l} f(\theta_i, s_j)$.*[6]

The concordance order ranks two distributions by comparing their cumulative functions pointwise. When $g \succeq f$, the probability that the realizations of the state and the signal are both "small" is higher under g than under f. Since g and f have the same marginals, it is also true that the probability that the realizations of the state and the signal are both high increases when their distribution change from f to g (Epstein and Tanny, 1980, theorem 3). Therefore, an agent with beliefs $g \succeq f$ interprets higher signal realization as *stronger* evidence of higher states than his well-calibrated counterpart. To the extend that overprecision is informally understood as an excessive faith in one's information then this definition captures exactly that.[7]

Additionally, $g \succeq f$ implies that the (Pearson) correlation coefficient, the Kendall's τ, and the Spearman's ρ between the state and the signal are all higher under g than under f (Tchen, 1980). The converse is not true in general. Thus, the concordance order is more conservative than, for instance, comparing covariances as a criteria to define overprecision (which has the additional problem of lacking a strong justification beyond the multivariate-normal case).

Finally, note that we can have a distribution $g \preceq f$ that reverses the relationship among states and signal realizations, such that a high signal ends up being evidence of a low state. This extreme change in the interpretation of the signal is not compatible with the idea of overprecision. However, given the maintained assumptions, this problem does not arise in the current setting. In particular, $y(\cdot)$ being increasing guarantees that the solution to the belief design problem is always above (in the concordance order) the independent distribution, i.e., any solution g^* satisfies $g^*(\theta, s) \geq f_\Theta(\theta) f_S(s)$ for all (θ, s).

[6] See Tchen (1980), Epstein and Tanny (1980), Shaked and Shanthikumar (2007, Chapter 9), Meyer and Strulovici (2012), and Mekonnen and Leal-Vizcaíno (2022).

[7] In the present two-dimensional setting other orders – the supermodular stochastic order, greater weak association, the convex-modular order, and the dispersion order – coincide with the concordance order (see Meyer and Strulovici, 2012).

Thus, if we find a solution $g^* \neq f$ with $g^* \succeq f$, we can say that the optimal agent is *overconfident*. Analogously, when $f \succeq g^*$ we can say that he is *underconfident*.

3 Analysis

For any $g \in \Delta(\Theta \times S)$ and signal realization $s \in S$, the agent optimally chooses $\mathbb{E}_g[y(\tilde{\theta})|s]$. Note that all feasible beliefs yield the same average action:

$$\mathbb{E}_f[\mathbb{E}_g[y(\tilde{\theta})|s]] = \mathbb{E}_g[\mathbb{E}_g[y(\tilde{\theta})|s]] = \mathbb{E}_g[y(\tilde{\theta})] = \mathbb{E}_f[y(\tilde{\theta})]$$

where the first equality follows from $f_S = g_S$, the second from the law of iterated expectations, and the third from $g_\Theta = f_\Theta$.

Knowing the agent's actions, we can express the principal's expected payoff as a function of any belief g as $U(g) := -\mathbb{E}_f[(\mathbb{E}_g[y(\tilde{\theta})|\tilde{s}] - \tilde{\theta})^2]$. As a result, belief design corresponds to the following optimization problem:

$$\max_{g \in \mathcal{G}} U(g).$$

I will divide the analysis of this problem into two parts. As a first step, I focus on the simplest version of the model, in which the state and signal spaces are binary. Then, I discuss the key insights that extend to the general version of the model. Proofs are provided in Appendix A.

3.1 Binary states and signals

Assume that $n = m = 2$. For any true distribution f, the set of feasible choices for the principal can be characterized by a single scalar. Formally, any $g \in \mathcal{G}$ can decomposed as $f + \varepsilon B$ where $\varepsilon \in \mathbb{R}$ is a scalar and $B \in \mathbb{R}^{2\times 2}$ is a matrix of constants.[8] To see why this is true, suppose we want to modify the probability of some pair of states and signals, say (θ_1, s_1), by adding some amount ε. Since marginal probabilities must remain constant, we need to adjust the probability of (θ_1, s_2) and (θ_2, s_1) by adding $-\varepsilon$. As a consequence, we also need to add ε to the probability of (θ_2, s_2).[9]

Furthermore, ε is constrained by the fact that the resulting g must be a well-defined probability mass function, i.e., $g(\theta_i, s_j) \in [0, 1]$. Consequently, ε must belong to some compact interval

[8] Specifically, this decomposition can be seen as

$$g = \begin{bmatrix} f(\theta_1, s_1) + \varepsilon & f(\theta_1, s_2) - \varepsilon \\ f(\theta_2, s_1) - \varepsilon & f(\theta_2, s_2) + \varepsilon \end{bmatrix} = \begin{bmatrix} f(\theta_1, s_1) & f(\theta_1, s_2) \\ f(\theta_2, s_1) & f(\theta_2, s_2) \end{bmatrix} + \varepsilon \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = f + \varepsilon B.$$

[9] Some authors refer to ε as an *elementary transformation* (see Epstein and Tanny, 1980).

$[\underline{\varepsilon}, \bar{\varepsilon}]$.[10] The full support assumption guarantees that $\underline{\varepsilon} < 0 < \bar{\varepsilon}$. In other words, any marginal deviation from the true distribution is feasible.

As a result, the belief design problem can be seen as as choosing $\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon}]$ to maximize $U(f+\varepsilon B)$, which is a well-behaved concave maximization program. Before discussing its solution, note that an increase in ε corresponds to a shift in probability mass from the off-diagonal of g to its diagonal. That is, the agent's belief that the state is θ_i after observing s_i increases with ε. Therefore, we can interpret ε as the agent's *level of confidence*: an agent with $\varepsilon > 0$ is *overconfident*, with $\varepsilon < 0$ is *underconfident*, and *well-calibrated* otherwise. This is both intuitive and consistent with the definition of overconfidence proposed in section 2.

Increasing ε makes the agent's actions more extreme: as ε grows the agent's action after observing s_i gets closer to $y(\theta_i)$. In other words, the agent's action increases with ε after s_2 and decreases after s_1. Therefore, increasing ε benefits the principal when a well-calibrated agent would take actions that are *too* low after s_2 or *too* high after s_1. Thus, $c(\theta_1) > c(\theta_2)$ would lead to overprecision to be optimal. Equivalently, overprecision is beneficial when $\theta_2 - \theta_1 > y(\theta_2) - y(\theta_1)$, i.e., when the principal would adjust the action beyond what the agent would like to do. This formalizes the intuition that overconfidence helps when there is unresponsiveness by the agent.

The following result describes the optimal agent in the binary case.

Proposition 1 *The unique optimal agent is*

1. *well-calibrated if and only if* $\theta_2 - \theta_1 = y(\theta_2) - y(\theta_1)$.
2. *overconfident if and only if* $\theta_2 - \theta_1 > y(\theta_2) - y(\theta_1)$.
3. *underconfident if and only if* $\theta_2 - \theta_1 > y(\theta_2) - y(\theta_1)$.

Alternatively, the correlation between the conflict of interests and the agent's preferred action determine the confidence of the optimal agent. In particular, the optimal agent is overconfident when such correlation is negative.

This result implies that a well-calibrated agent is optimal if and only if his bias is additive, i.e., $y(\theta) = \theta + b$. This extends to the optimal agent too, he would ideally behave *as if* his bias was additive. However, the bounds on ε may prevent the principal from getting that far.

[10] The bounds on the interval are as follows

$$\underline{\varepsilon} := -\min\{f(\theta_1, s_1), 1 - f(\theta_1, s_2), 1 - f(\theta_2, s_1), f(\theta_2, s_2)\}$$
$$\bar{\varepsilon} := \min\{1 - f(\theta_1, s_1), f(\theta_1, s_2), f(\theta_2, s_1), 1 - f(\theta_2, s_2)\}.$$

In addition, the interior optimum confidence level is given by

$$\varepsilon^* = -\frac{Cov_f(c(\tilde{\theta}), y(\tilde{\theta}))}{Var_f(y(\tilde{\theta}))}|f| = [\frac{Cov_f(\tilde{\theta}, y(\tilde{\theta}))}{Var_f(y(\tilde{\theta}))} - 1]|f|$$

where $|f|$ denotes the determinant of the matrix f.

Since signal realizations are labeled by the conditional expectation over states they induce, it follows that $|f| > 0$. Assuming that $y(\cdot)$ is increasing directly implies that $\varepsilon^* > -|f| \in (\underline{\varepsilon}, 0)$. This is, since an agent whose confidence equals $-|f|$ acts as if the signal was uninformative, the optimal agent always extracts some information from the signal.

Moreover, the optimal confidence level increases with the slope of the best affine predictor of the principal's preferred action, $\tilde{\theta}$, as a function of the agent's preferred action, $y(\tilde{\theta})$. Lastly, if for given primitives we have $\varepsilon^* > \bar{\varepsilon}$ then the optimal agent is maximally overconfident.

Now I can illustrate the implications of the result for the applications described in section 2. First, the CEO concerns about the manager's reluctance to punish his subordinate is captured by assuming $y(\theta_1) > \theta_1$ and $y(\theta_2) \approx \theta_2$. It follows that $c(\theta_1) > c(\theta_2)$, and by Proposition 1 we conclude that the CEO would strictly prefer to hire an *overconfident* manager.

On the other hand, in the nuclear power plant example, the friction is due to the manager's low willingness to mitigate risks when an accident is unlikely to occur. This can be represented by assuming $y(\theta_1) < \theta_1$ and $y(\theta_2) \approx \theta_2$. It follows that $c(\theta_1) < c(\theta_2)$, and we conclude that the optimal risk manager would be *underconfident*.

3.2 General case

In this subsection, I lift the assumption on the number of states and signals. We start by observing that the objective function in the belief design problem can be decomposed as follows

$$U(g) = -\mathbb{E}_f[(\mathbb{E}_f[\tilde{\theta}|\tilde{s}] - \tilde{\theta})^2] - \mathbb{E}_f[c(\tilde{\theta})]^2 - Var_f(\mathbb{E}_g[y(\tilde{\theta})|\tilde{s}] - \mathbb{E}_f[\tilde{\theta}|\tilde{s}]).$$

The first term represents the payoff that the principal could obtain if she was informed and in charge of choosing the action. It corresponds to a loss due to the residual uncertainty in the environment. The second term reflects a loss due to the average bias that the agent introduces with his choice. Finally, agent's beliefs affect the objective only through the last term, which represents a loss due to variance in the agent's actions beyond the adjustments that the principal herself would make.

This decomposition illustrates that the principal ideally wants $\mathbb{E}_g[y(\tilde{\theta})|\tilde{s}] - \mathbb{E}_f[\tilde{\theta}|\tilde{s}]$ to be constant. In other words, the ideal agent is one who acts *as if* his bias was additive. Therefore, if the expected conflict of interests is invariant in the signal, i.e. $\mathbb{E}_f[y(\tilde{\theta})|\tilde{s}] = \mathbb{E}_f[\tilde{\theta}|\tilde{s}] + b$, the

well-calibrated agent would be optimal since he is feasible and already behaves as the principal ideally wants. It turns out that this condition is also necessary. Whenever the conditional expectation of the conflict of interest varies with the signal, there is a marginal deviation from the true distribution that strictly increases the principal's expected payoff. This is achieved by decreasing the agent's action after a signal realization leading to higher expected conflict of interests, while increasing the agent's action after a realization inducing lower expected conflict of interests. The construction of a distribution that improves upon the true one follows the same logic as in the binary case, the details are discussed in the proof of the following proposition (see appendix A).

Proposition 2 *The optimal agent is well-calibrated if and only if* $\mathbb{E}_f[y(\tilde{\theta}) - \tilde{\theta}|\tilde{s}]$ *is constant.*

Moreover, because the marginals of g equal those of f, $\mathbb{E}_g[y(\tilde{\theta})|\tilde{s}] - \mathbb{E}_f[\tilde{\theta}|\tilde{s}]$ can only equal one constant, which is $\mathbb{E}_f[c(\tilde{\theta})]$. Therefore, for this ideal agent we have that $\mathbb{E}_g[y(\tilde{\theta})|\tilde{s}] - \mathbb{E}_f[y(\tilde{\theta})|\tilde{s}] = \mathbb{E}_f[c(\tilde{\theta})] - \mathbb{E}_f[c(\tilde{\theta})|\tilde{s}]$. When $\mathbb{E}_f[c(\tilde{\theta})|s]$ is decreasing in s, the ideal agent's optimal action would be below the well-calibrated agent's action for *low* signals realizations, while the opposite is true for *high* ones. Since $\mathbb{E}_f[y(\tilde{\theta})|s]$ is increasing, this pattern corresponds to more extreme actions by the agent, a direct manifestation of overconfidence. As a result, the idea that overconfidence is optimal when the principal would adjust the action more than agent generalizes beyond the binary special case. The following result formalizes this intuition.

Proposition 3 *There exists* $\bar{\alpha} > 0$ *such that if* $|\mathbb{E}_f[c(\tilde{\theta})|s] - \mathbb{E}_f[c(\tilde{\theta})]| \leq \bar{\alpha}$ *for all* $s \in S$ *and* $\mathbb{E}_f[y(\tilde{\theta})|s'] - \mathbb{E}_f[y(\tilde{\theta})|s] < \mathbb{E}_f[\tilde{\theta}|s'] - \mathbb{E}_f[\tilde{\theta}|s]$ *for all* $s' > s$*, then any optimal agent acts as an overconfident agent. On the other hand, if* $\mathbb{E}_f[y(\tilde{\theta})|s'] - \mathbb{E}_f[y(\tilde{\theta})|s] > \mathbb{E}_f[\tilde{\theta}|s'] - \mathbb{E}_f[\tilde{\theta}|s]$ *for all* $s' > s$*, any optimal agent acts as an underconfident agent.*

The previous result provides sufficient conditions for at least one optimal agent to be strictly ranked above or below the true distribution according to the concordance order. The uniqueness from the binary case is necessarily lost since there exist several (typically a continuum of) distributions with the same marginals and conditional expectations. However, all optimal agents take the same actions. Moreover, all optimal agents must be unranked among themselves, i.e., if g and g' solve the belief design problem it cannot be that $g \succeq g'$.

The proof of this proposition parallels that of proposition 1. We start by changing the principal's choice from $g \in \mathcal{G}$ to a matrix of *elementary transformations* $t \in \mathbb{R}^{(n-1)\times(m-1)}$. Let t_{kl} denote a typical entry of the matrix t. Note that $t_{kl} > 0$ moves probability mass from θ_{k+1} to θ_k after s_l is realized, while the opposite happens after s_{l+1}. Informally, a positive elementary transformation moves mass from discordant pairs of states and signal realizations to the adjacent concordant pairs. The next step is to analyze the first order conditions that optimal elementary transformations need to satisfy. Then, I propose a family of transformations that satisfy these

conditions. When all entries of t are positive, we get a distribution that dominates the initial one in the concordance order. The assumptions in proposition 3 guarantee that the proposed solution is both feasible and positive (or negative), which proves the existence of an overconfident (or underconfident) optimal agent.

In what follows, I explore the implications of additional tools available to the principal. I go back to the assumption that $n = m = 2$. First, I will study how belief design is affected by availability of action-contingent transfers. I will argue that transfers do not affect optimal beliefs as long as the expected conflict of interests is not too far away from zero. Moreover, in the optimum each tool (belief and contract design) is used for different purposes. Additionally, I consider the possibility that the principal can make the choice herself. This imposes some restrictions on the characteristics of an agent who is actually allowed to make the choice: for the principal to delegate the choice, agent's confidence must be sufficiently high.

4 Transfers

Conflicting preferences among its members is a prominent challenge that organizations face. Provision of monetary incentives is a particularly relevant tool that can be used to mitigate the pernicious effects of agency frictions. In this section, I consider the interaction between belief design and action-contingent transfers. In particular, in addition to belief design, the principal is also allowed to commit to non-negative payments contingent on the agent's action. We focus again on the $n = m = 2$ case. The timing is as follows:

1. Belief and contract design: the principal chooses $g \in \mathcal{G}$ and $w : \mathbb{R} \to \mathbb{R}_+$.
2. Nature draws (θ, s) according to f.
3. The agent observes s and chooses $x \in \mathbb{R}$.

Payoffs are given by $-(x-\theta)^2 - w(x)$ for the principal and by $-(x-y(\theta))^2 + w(x)$ for the agent.

As a first step, we can think of the problem as the principal recommending action x_i after signal realization s_i, paying w_i after that action is observed, and paying zero when a non-recommended action is observed. Additionally, in the binary case, belief design can be thought as choosing a confidence level $\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon}]$. Therefore, it suffices for the the principal to consider tuples $(x_1, x_2, w_1, w_2, \varepsilon)$ consisting of recommended actions, payments for those actions, and a confidence level subject to obedience constraints.

The recommended actions must be incentive compatible given promised transfers and agent's confidence. After each signal realization two deviations are relevant: to the other recommended action and to the best action among those that were not recommended. The best deviation to

an action yielding no transfer corresponds to choosing $\mathbb{E}_\varepsilon[y(\tilde{\theta})|s_i]$ after signal s_i, which yields an expected payoff of $-Var_\varepsilon(y(\tilde{\theta})|s_i)$.

The main result in this section provides conditions under which the optimal beliefs coincide with those described in subsection 3.1 in which transfers were not available. In the optimum, each tool plays a different role: transfers are used to decrease the average bias in the agent's action, while beliefs are used to distribute such bias across signal realizations.

Proposition 4 *Assume that* $|\mathbb{E}_f[c(\tilde{\theta})]| \leq \mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]$. *The availability of transfers does not change the optimal beliefs (thus, proposition 1 applies) and* $w_1^* = w_2^* = \mathbb{E}_f[c(\tilde{\theta})]^2$.

This result also highlights the effects of belief design on optimal transfers. When belief are optimally chosen, wages do not change with the actions the agent takes in equilibrium. On the other hand, when only the well-calibrated agent is available, the optimum for the principal is given by

$$\begin{aligned} x_j &= \frac{1}{2}[\mathbb{E}_f[\tilde{\theta}|s_j] + \mathbb{E}_f[y(\tilde{\theta})|s_j]], \\ w_j &= \frac{1}{4}\mathbb{E}_f[c(\tilde{\theta})|s_j]^2. \end{aligned}$$

Therefore, unless $\mathbb{E}_f[c(\tilde{\theta})|s_j]$ is constant (in which case the well-calibrated agent is indeed optimal), wages are different for both recommended actions. This is, optimal belief-based selection leads to "flatter" compensation schemes.

It is worth noting that Ashworth and Sasso (2019) study the interaction between optimal delegation sets and transfers in a similar setting. When the principal and the agent only differ in their level of confidence, the optimal mechanism does not use transfers.

5 Delegation

Consider the case in which the principal can also decide whether to delegate the choice or to centralize it. In this section, I discuss when it is the case that the principal actually prefers delegation and its implications about the optimal agent's confidence.

The main takeaway is that centralization is a substitute for *extreme* underconfidence. The intuition is simple: if the optimal agent is not *sufficiently* using the information, it would be better for the principal to make the choice herself in order to avoid the conflict of interests. The principal would only rely on the agent to use his private information; otherwise, she would be better off avoiding the bias that the agent introduces to the decision.

The following result shows that the optimal agent must be sufficiently confident for the choice to be delegated to him.

Proposition 5 *The choice is delegated to the optimal agent if and only if*

$$\varepsilon^* \geq \frac{f_S(s_1)f_S(s_2)\mathbb{E}_f[c(\tilde{\theta})]^2}{|f|(y(\theta_2)-y(\theta_1))(\theta_2-\theta_1)} - |f|.$$

Naturally, an agent that is on average unbiased (i.e. $\mathbb{E}_f[c(\tilde{\theta})] = 0$) would act just as the principal in the absence of any information. Therefore, when $\mathbb{E}_f[c(\tilde{\theta})] = 0$ the choice would be delegated even to a maximally underconfident agent (one who thinks that signals are uninformative). On the other hand, whenever $\mathbb{E}_f[c(\tilde{\theta})] \neq 0$ the agent's confidence must be strictly above $-|f|$ (the confidence of a maximally underconfident agent) for delegation to be optimal.

While extreme underconfidence never leads to delegation to an agent that is on average biased, it can be optimal to delegate to an extremely overconfident agent.[11] Therefore, while the two phenomena are modeled symmetrically, the possibility to centralize decision making highlights a key difference between over- and underconfidence.

6 Conclusions

There is abundant evidence about the pervasiveness of overconfident employees. This paper discusses situations in which a firm is willing to select an employee not *despite* of his overconfidence but precisely *because* of it. Thus, a reason for overconfidence to be ubiquitous is that the conditions that lead to its optimality are relevant in many environments. Under-responsiveness to the firm's interests is "bread-and-butter" when employees prioritize a quiet life over adjusting their behavior to current conditions and it can be (at least partially) alleviated by the employee's misperception of being more informed.

Moreover, belief-based selection interacts meaningfully with some other measures that the firm could take to mitigate agency frictions. For example, an employee with optimal beliefs would face "flatter" compensations schemes than his well-calibrated counterpart. Additionally, centralizing decision-making is a natural substitute for extreme underconfidence, while it can be optimal to delegate to a "fully" overconfident employee.

[11] Consider the following example: $\theta_1 = 0$, $\theta_2 = 10$, $y(\theta) = 3 + \theta/3$, $f(\theta_i, s_i) = 0.4$, and $f(\theta_i, s_{-i}) = 0.1$. We have that $\varepsilon^* = 0.3 > \bar{\varepsilon} = 0.1$, which implies that the optimal agent is maximally overconfident, interpreting signals at face value (after seeing s_i this agent would be convinced that the state is θ_i). The expected payoff from delegation is -17.89, while the expected payoff from centralization is -25. Therefore, it is optimal to delegate to the maximally overconfident agent.

The main methodological innovation, the belief design problem, allows the firm to exploit the heterogeneity in applicants' confidence by selecting the candidate with most favorable beliefs (keeping fixed all other individual characteristics). In this context, I show that belief-based selection leads to employees with a common feature: they tend to act *as if* their disagreement with the firm was invariant to their private information. As a result, the firm prefers a well-calibrated agent if and only if this disagreement does not change with the employee's observations to begin with. On the other hand, overconfidence helps when this conflict of interests moves in the opposite direction than the employee's preferred action. As a tool, belief-design is flexible and can be adjusted to systematically think about other belief-based biases, such as overoptimism.

Two assumptions were maintained through this exercise: that the potential employees are sufficiently diverse in their beliefs and that their characteristics are perfectly observed by the firm. The first assumption is made to provide a benchmark. The second is motivated by evidence suggesting that firms can learn quite rapidly about its employees' characteristics (Lange, 2007; Hansen et al., 2021). Issues of a constrained set of available employees as well as asymmetric information about the applicants' characteristics are left for future research.

These findings illustrate interesting ways in which personality traits, such as confidence, can impact labor market outcomes, such as wages and career choices (see Schulz and Thöni, 2016). Similarly, they help explain several documented managerial and organizational practices targeted towards altering employees' perceptions (see Haran et al., 2010; Meikle et al., 2016). Additionally, if confidence has the potential to affect expected outcomes (in the labor market or otherwise), we may expect individuals to invest in "adjusting" this personal characteristic according to their goals.[12]

Finally, while this paper specifically focuses on the role for overconfidence inside organizations, the main mechanism is likely to operate in several other types of social interactions where conflicting preferences and private information play a significant role (such as friendships and romantic relationships).

[12]See Kreps (2019) for a discussion about the role of business schools in boosting students' confidence.

References

T. R. Adam, C. S. Fernando, and E. Golubeva. Managerial overconfidence and corporate risk management. *Journal of Banking & Finance*, 60:195–208, 2015.

C. Aina. Tailored stories. *Working Paper*, 2024.

R. Alonso and N. Matouschek. Optimal delegation. *The Review of Economic Studies*, 75(1): 259–293, 2008.

S. Ashworth and G. Sasso. Delegation to an overconfident expert. *The Journal of Politics*, 81 (2):692–696, 2019.

C. Ba and A. Gindin. A multi-agent model of misspecified learning with overconfidence. *Games and Economic Behavior*, 142:315–338, 2023.

J. M. Barrero. The micro and macro of managerial beliefs. *Journal of Financial Economics*, 143(2):640–667, 2022.

I. Ben-David, J. R. Graham, and C. R. Harvey. Managerial miscalibration. *The Quarterly Journal of Economics*, 128(4):1547–1584, 2013.

A. E. Bernardo and I. Welch. On the evolution of overconfidence and entrepreneurs. *Journal of Economics & Management Strategy*, 10(3):301–330, 2001.

E. S. Berner and M. L. Graber. Overconfidence as a cause of diagnostic error in medicine. *The American journal of medicine*, 121(5):S2–S23, 2008.

J. Blanes-i Vidal and M. Möller. When should leaders share information with their subordinates? *Journal of Economics & Management Strategy*, 16(2):251–283, 2007.

P. Bolton, M. K. Brunnermeier, and L. Veldkamp. Leadership, coordination, and corporate culture. *Review of Economic Studies*, 80(2):512–537, 2013.

Y.-K. Che and N. Kartik. Opinions as incentives. *Journal of Political Economy*, 117(5):815–860, 2009.

P. Croskerry and G. Norman. Overconfidence in clinical decision making. *The American journal of medicine*, 121(5):S24–S29, 2008.

K. Daniel and D. Hirshleifer. Overconfident investors, predictable returns, and excessive trading. *Journal of Economic Perspectives*, 29(4):61–88, 2015.

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

K. Eliaz and R. Spiegler. A model of competing narratives. *American Economic Review*, 110 (12):3786–3816, 2020.

F. Englmaier. Commitment in r&d tournaments via strategic delegation to overoptimistic managers. *Managerial and Decision Economics*, 32(1):63–69, 2011.

F. Englmaier and M. Reisinger. Biased managers as strategic commitment. *Managerial and Decision Economics*, 35(5):350–356, 2014.

L. G. Epstein and S. M. Tanny. Increasing generalized correlation: a definition and some economic consequences. *Canadian Journal of Economics*, pages 16–34, 1980.

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

S. Gervais and T. Odean. Learning to be overconfident. *The Review of Financial Studies*, 14 (1):1–27, 2001.

S. Gervais, J. B. Heaton, and T. Odean. Overconfidence, compensation contracts, and capital budgeting. *The Journal of Finance*, 66(5):1735–1777, 2011.

A. M. Goel and A. V. Thakor. Overconfidence, ceo selection, and corporate governance. *The Journal of Finance*, 63(6):2737–2784, 2008.

J. Goodman-Delahunty, P. A. Granhag, M. Hartwig, and E. F. Loftus. Insightful or wishful: Lawyers' ability to predict case outcomes. *Psychology, Public Policy, and Law*, 16(2):133, 2010.

A. T. Hansen, U. Hvidman, and H. H. Sievertsen. Grades and employer learning. *IZA Discussion Papers*, 2021.

U. Haran, D. A. Moore, and C. K. Morewedge. A simple remedy for overprecision in judgment. *Judgment and Decision Making*, 5(7):467, 2010.

Y. Heller. Overconfidence and diversification. *American Economic Journal: Microeconomics*, 6 (1):134–53, 2014.

N. Hestermann and Y. Le Yaouanq. Experimentation with self-serving attribution biases. *American Economic Journal: Microeconomics*, 2020.

B. Holmström. On incentives and control in organizations. *Ph.D. Thesis, Stanford University*, 1977.

B. Holmström. On the theory of delegation. In M. Boyer and R. Kihlstrom, editors, *Bayesian models in economic theory*. North-Holland, 1984.

P. Ilinov, A. Matveenko, M. Senkov, and E. Starkov. Optimally biased expertise. *arXiv preprint arXiv:2209.13689*, 2022.

A. Ispano. The perils of a coherent narrative. *Working Paper*, 2023.

A. Jain. Informing agents amidst biased narratives. *Working Paper*, 2023.

D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

E. Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11: 249–272, 2019.

E. Kamenica and M. Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011.

P. Koellinger, M. Minniti, and C. Schade. "I think I can, I think I can": Overconfidence and entrepreneurial behavior. *Journal of economic psychology*, 28(4):502–527, 2007.

D. Kreps. Some dimensions of behavior with wich economics should contend. *Nemmers Prize Lecture*, 2019.

A. S. Kyle and F. A. Wang. Speculation duopoly with agreement to disagree: Can overconfidence survive the market test? *The Journal of Finance*, 52(5):2073–2090, 1997.

F. Lange. The speed of employer learning. *Journal of Labor Economics*, 25(1):1–35, 2007.

G. Levy and R. Razin. Correlation neglect, voting behavior, and information aggregation. *American Economic Review*, 105(4):1634–45, 2015.

U. Malmendier and G. Tate. CEO overconfidence and corporate investment. *The Journal of Finance*, 60(6):2661–2700, 2005.

U. Malmendier and T. Taylor. On the verges of overconfidence. *Journal of Economic Perspectives*, 29(4):3–8, 2015.

N. L. Meikle, E. R. Tenney, and D. A. Moore. Overconfidence at work: Does overconfidence survive the checks and balances of organizational life? *Research in Organizational Behavior*, 36:121–134, 2016.

T. Mekonnen and R. Leal-Vizcaíno. Bayesian comparative statics. *Theoretical Economics*, 17 (1):219–251, 2022.

M. Meyer and B. Strulovici. Increasing interdependence of multivariate distributions. *Journal of Economic Theory*, 147(4):1460–1489, 2012.

D. A. Moore and P. J. Healy. The trouble with overconfidence. *Psychological review*, 115(2): 502, 2008.

D. A. Moore and D. Schatz. The three faces of overconfidence. *Social and Personality Psychology Compass*, 11(8):e12331, 2017.

M. Niu. Motivated misspecification. 2023.

F. Ostrizek. Vague by design: Performance evaluation and learning from wages. *Working Paper*, 2022.

F. Palomino and A. Sadrieh. Overconfidence and delegated portfolio management. *Journal of Financial Intermediation*, 20(2):159–177, 2011.

K. Phua, T. M. Tham, and C. Wei. Are overconfident ceos better leaders? evidence from stakeholder commitments. *Journal of Financial Economics*, 127(3):519–545, 2018.

C. Prendergast. The motivation and bias of bureaucrats. *American Economic Review*, 97(1): 180–196, 2007.

C. Prendergast. Intrinsic motivation and incentives. *American Economic Review: Papers & Proceedings*, 98(2):201–05, 2008.

C. Prendergast and L. Stole. Impetuous youngsters and jaded old-timers: Acquiring a reputation for learning. *Journal of political Economy*, 104(6):1105–1134, 1996.

L. Rayo and I. Segal. Optimal information disclosure. *Journal of political Economy*, 118(5): 949–987, 2010.

J. J. Rotemberg and G. Saloner. Visionaries, managers, and strategic direction. *RAND Journal of Economics*, pages 693–716, 2000.

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

J. F. Schulz and C. Thöni. Overconfidence and career choice. *PloS one*, 11(1):e0145126, 2016.

J. Schwartzstein and A. Sunderam. Using models to persuade. *American Economic Review*, 111 (1):276–323, 2021.

M. Shaked and J. G. Shanthikumar. *Stochastic orders*. Springer Science & Business Media, 2007.

A. Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. 1776.

A. H. Tchen. Inequalities for distributions with given marginals. *The Annals of Probability*, 8 (4):814–827, 1980.

E. Van den Steen. Organizational beliefs and managerial vision. *Journal of Law, Economics, and organization*, 21(1):256–283, 2005.

E. Van den Steen. Culture clash: The costs and benefits of homogeneity. *Management Science*, 56(10):1718–1738, 2010.

A Proofs

Proof of Proposition 1

The belief design problem is as follows

$$\max_{\varepsilon\in[\underline{\varepsilon},\bar{\varepsilon}]} U\left(f+\varepsilon A\right).$$

We have

$$\begin{aligned} U(f+\varepsilon A) = \quad & -f(\theta_1,s_1)(\mathbb{E}_\varepsilon[y(\tilde{\theta})|s_1]-\theta_1)^2 - f(\theta_1,s_2)(\mathbb{E}_\varepsilon[y(\tilde{\theta})|s_2]-\theta_1)^2 \\ & -f(\theta_2,s_1)(\mathbb{E}_\varepsilon[y(\tilde{\theta})|s_1]-\theta_2)^2 - f(\theta_2,s_2)(\mathbb{E}_\varepsilon\left[y(\tilde{\theta})|s_2\right]-\theta_2)^2, \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}_\varepsilon[y(\tilde{\theta})|s_1] \quad &= \mathbb{E}_f[y(\tilde{\theta})|s_1] - \tfrac{\varepsilon}{f_S(s_1)}[y(\theta_2)-y(\theta_1)], \\ \mathbb{E}_\varepsilon[y(\tilde{\theta})|s_2] \quad &= \mathbb{E}_f[y(\tilde{\theta})|s_2] + \tfrac{\varepsilon}{f_S(s_2)}[y(\theta_2)-y(\theta_1)]. \end{aligned}$$

Thus,

$$\frac{\partial U(f+\varepsilon A)}{\partial \varepsilon} = \quad 2[y(\theta_2)-y(\theta_1)][\mathbb{E}_\varepsilon[y(\tilde{\theta})|s_1]-\mathbb{E}_f[\tilde{\theta}|s_1]-\mathbb{E}_\varepsilon[y(\tilde{\theta})|s_2]+\mathbb{E}_f[\tilde{\theta}|\tilde{s}=s_2]].$$

We can see that U is strictly concave in ε:

$$\frac{\partial^2 U\left(f+\varepsilon A\right)}{\partial \varepsilon^2} = -2\frac{[y(\theta_2)-y(\theta_1)]^2}{f_S(s_1)f_S(s_2)} < 0.$$

The first-order condition is given by

$$\mathbb{E}_\varepsilon[y(\tilde{\theta})|s_1]-\mathbb{E}_f[\tilde{\theta}|s_1] = \mathbb{E}_\varepsilon[y(\tilde{\theta})|s_2]-\mathbb{E}_f[\tilde{\theta}|s_2],$$

which is equivalent to

$$\mathbb{E}_f[c(\tilde{\theta})|s_1] - \frac{\varepsilon}{f_S(s_1)}[y(\theta_2)-y(\theta_1)] = \mathbb{E}_f[c(\tilde{\theta})|s_2] + \frac{\varepsilon}{f_S(s_2)}[y(\theta_2)-y(\theta_1)].$$

Note that

$$\mathbb{E}_f[c(\tilde{\theta})|s_2]-\mathbb{E}_f[c(\tilde{\theta})|s_1] = [c(\theta_2)-c(\theta_1)]\frac{f(\theta_2,s_2)f(\theta_1,s_1)-f(\theta_1,s_2)f(\theta_2,s_1)}{f_S(s_1)f_S(s_2)}.$$

Therefore, the interior solution (if exists) is given by

$$\begin{aligned} \varepsilon^* \quad &= \quad -\frac{c(\theta_2)-c(\theta_1)}{y(\theta_2)-y(\theta_1)}\left[f(\theta_2,s_2)f(\theta_1,s_1)-f(\theta_1,s_2)f(\theta_2,s_1)\right] \\ &= \quad -\frac{Cov_f(c(\tilde{\theta}),y(\tilde{\theta}))}{Var_f(y(\tilde{\theta}))}\left[f(\theta_2,s_2)f(\theta_1,s_1)-f(\theta_1,s_2)f(\theta_1,s_2)\right]. \end{aligned}$$

Since by assumption $f(\theta_2,s_2)f(\theta_1,s_1) > f(\theta_1,s_2)f(\theta_1,s_2)$, we have that the sign of ε^* equals the sign of $-Cov_f(c(\tilde{\theta}),y(\tilde{\theta}).)$. Finally, if $\varepsilon^* \notin [\underline{\varepsilon},\bar{\varepsilon}]$ then the solution is in a corner and will still have the same sign as ε^*. ■

Proof of Proposition 2

Optimality of the well-calibrated agent when $\mathbb{E}_f[c(\tilde{\theta})|\tilde{s}]$ is constant follows directly from the decomposition in subsection 3.2.

On the other hand, assume that $\mathbb{E}_f[c(\tilde{\theta})|\tilde{s}]$ is not constant. This is, there exist signal realizations s_l and $s_{l'}$ with $\mathbb{E}_f[c(\tilde{\theta})|s_{l'}] > \mathbb{E}_f[c(\tilde{\theta})|s_l]$. Choose any two states θ_k and $\theta_{k'}$, where $\theta_k < \theta_{k'}$, and define $g_\varepsilon \in \Delta(\Theta \times S)$ as follows

$$g_\varepsilon(\theta, s) := \begin{cases} f(\theta,s) + \varepsilon & \text{if} \quad (\theta,s) \in \{(\theta_{k'}, s_{l'}), (\theta_k, s_l)\} \\ f(\theta,s) - \varepsilon & \text{if} \quad (\theta,s) \in \{(\theta_k, s_{l'}), (\theta_{k'}, s_l)\} \\ f(\theta,s) + \varepsilon & \text{otherwise.} \end{cases}$$

Note that g_ε remains feasible whenever $\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon}]$, where

$$\begin{aligned} \underline{\varepsilon} &:= -\min\{f(\theta_{k'}, s_{l'}), 1 - f(\theta_k, s_{l'}), 1 - f(\theta_{k'}, s_l), f(\theta_k, s_l)\}, \\ \bar{\varepsilon} &:= \min\{1 - f(\theta_{k'}, s_{l'}), f(\theta_k, s_{l'}), f(\theta_{k'}, s_l), 1 - f(\theta_k, s_l)\}. \end{aligned}$$

The full support assumption implies that $\underline{\varepsilon} < 0 < \bar{\varepsilon}$.

Note that when agent's beliefs are given by g_ε, he takes the following action after a given signal realization s

$$\mathbb{E}_{g_\varepsilon}[y(\tilde{\theta})|s] = \begin{cases} \mathbb{E}_f[y(\tilde{\theta})|s] - \varepsilon \frac{y(\theta_{k'}) - y(\theta_k)}{f_S(s_l)} & \text{if} \quad s = s_l \\ \mathbb{E}_f[y(\tilde{\theta})|s] + \varepsilon \frac{y(\theta_{k'}) - y(\theta_k)}{f_S(s_{l'})} & \text{if} \quad s = s_{l'} \\ \mathbb{E}_f[y(\tilde{\theta})|s] & \text{otherwise.} \end{cases}$$

Therefore, the principal's expected payoff from choosing g_ε is given by

$$\begin{aligned} U(g_\varepsilon) = U(f) &- \varepsilon^2 [y(\theta_{k'}) - y(\theta_k)]^2 \left(\frac{1}{f_S(s_{l'})} + \frac{1}{f_S(s_l)} \right) \\ &-2\varepsilon [y(\theta_{k'}) - y(\theta_k)](\mathbb{E}_f[c(\tilde{\theta})|s_{l'}] - \mathbb{E}_f[c(\tilde{\theta})|s_l]). \end{aligned}$$

Differentiating this expression with respect to ε we get

$$\begin{aligned} \frac{\partial U(g_\varepsilon)}{\partial \varepsilon} = &- 2[y(\theta_{k'}) - y(\theta_k)]^2 \varepsilon \left(\frac{1}{f_S(s_{l'})} + \frac{1}{f_S(s_l)} \right) \\ &- 2[y(\theta_{k'}) - y(\theta_k)](\mathbb{E}_f\left[c(\tilde{\theta})|\tilde{s} = s_{l'}\right] - \mathbb{E}_f[c(\tilde{\theta})|\tilde{s} = s_l]). \end{aligned}$$

Finally, $y(\theta_{k'}) \neq y(\theta_k)$ and $\mathbb{E}_f[c(\tilde{\theta})|s_{l'}] > \mathbb{E}_f[c(\tilde{\theta})|s_l]$ imply that this derivative never equals zero at $\varepsilon = 0$:

$$\frac{\partial U(g_\varepsilon)}{\partial \varepsilon}\Big|_{\varepsilon=0} = -2[y(\theta_{k'}) - y(\theta_k)](\mathbb{E}_f[c(\tilde{\theta})|s_{l'}] - \mathbb{E}_f[c(\tilde{\theta})|s_l]) \neq 0.$$

Therefore, f cannot solve the belief design problem.■

Proof of Proposition 3

We start with the observation that any $g \in \mathcal{G}$ can be equivalently expressed as $f + CtB$ where $t \in \mathbb{R}^{(n-1)\times(m-1)}$ is a matrix of elementary transformations and C and B are matrices of constants. The decomposition is as follows:

$$t := \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1,m-1} \\ t_{21} & t_{22} & \cdots & t_{2,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n-1,1} & t_{n-1,2} & \cdots & t_{n-1,m-1} \end{bmatrix} \in \mathbb{R}^{(n-1)\times(m-1)},$$

$$C := \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & \cdots & 0 & 0 \\ 0 & -1 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & \cdots & 0 & -1 \end{bmatrix} \in \mathbb{R}^{n\times(n-1)},$$

$$B := \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \in \mathbb{R}^{(m-1)\times m}.$$

Note that each t_{kl} affects posterior beliefs only after signal realizations s_l and s_{l+1}. Moreover, $t_{kl} > 0$ moves probability mass from θ_{k+1} to θ_k after s_l, while the opposite happens after s_{l+1}.

Let $\delta(t, s_j)$ denote the difference between the actions of an agent that interprets signals according to t and those of the well-calibrated agent, i.e.

$$\delta(t, s_j) := \mathbb{E}_g[y(\tilde{\theta})|s_j] - \mathbb{E}_f[y(\tilde{\theta})|\tilde{s} = s_j] = \frac{1}{f_S(s_j)} \sum_{i=1}^{n-1} [y(\theta_{i+1}) - y(\theta_i)][t_{i,j-1} - t_{ij}].$$

In particular, since $t_{i0} = t_{im} = 0$ we have that

$$\begin{aligned} \delta(t, s_1) &= -\frac{1}{f_S(s_1)} \sum_{i=1}^{n-1} [y(\theta_{i+1}) - y(\theta_i)] t_{i1}, \\ \delta(t, s_m) &= \frac{1}{f_S(s_m)} \sum_{i=1}^{n-1} [y(\theta_{i+1}) - y(\theta_i)] t_{i,m-1}. \end{aligned}$$

As a result

$$\frac{\partial \delta(t, s_j)}{\partial t_{kl}} = \begin{cases} -\frac{y(\theta_{k+1}) - y(\theta_k)}{f_S(s_l)} & \text{if} \quad j = l \\ \frac{y(\theta_{k+1}) - y(\theta_k)}{f_S(s_{l+1})} & \text{if} \quad j = l + 1 \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand, $g \in \mathcal{G}$ if and only if

$$t_{i-1,j-1} - t_{i-1,j} - t_{i,j-1} + t_{ij} \in [-f(\theta_i, s_j), 1 - f(\theta_i, s_j)]$$

for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$, where $t_{0j} = t_{i0} = t_{nj} = t_{im} = 0$. Let $\mathcal{T}$ denote the set of feasible transformations. The full-support assumption guarantees that the interior of $\mathcal{T}$, denoted $\mathcal{T}^\circ$, is non-empty (in particular the $(n-1)\times(m-1)$-matrix with all entries equal to zero belongs to $\mathcal{T}^\circ$).

Therefore, the belief design problem corresponds to choosing $t \in \mathcal{T}$ to maximize $U(f + AtB)$. Since $U.(f + AtB.)$ is continuous in t and $\mathcal{T}$ is a compact set, it follows that this problem has a solution.

Note that $g \succeq f$ if and only if $t \geq 0$, i.e. $t_{ij} \geq 0$ for all $i \in \{1, \ldots, n-1\}$ and $j \in \{1, \ldots, m-1\}$.[13] Therefore, if $t^* \gneqq 0$ solves this problem, we would say that the optimal agent is *overconfident*.

We can decompose the objective function as follows

$$U(g) = -\mathbb{E}_f\left[\delta(t, \tilde{s})^2\right] - 2\mathbb{E}_f\left[\delta(t, \tilde{s})\mathbb{E}_f\left[c(\tilde{\theta})|\tilde{s}\right]\right] + U(f).$$

Thus,

$$\frac{\partial U(f + CtB)}{\partial t_{kl}} = 2[y(\theta_{k+1}) - y(\theta_k)][\delta(t, s_l) + \mathbb{E}_f[c(\tilde{\theta})|s_l] - \delta(t, s_{l+1}) - \mathbb{E}_f[c(\tilde{\theta})|s_{l+1}]].$$

Since we assumed that $y(\theta_{k+1}) > y(\theta_k)$, we have that in any optimum $t^* \in \mathcal{T}^\circ$ the following expression must hold for each $l \in \{1, \ldots, m-1\}$

$$\delta(t^*, s_l) + \mathbb{E}_f[c(\tilde{\theta})|s_l] = \Delta(t^*, s_{l+1}) + \mathbb{E}_f[c(\tilde{\theta})|s_{l+1}]. \tag{1}$$

Expression (1) requires $\delta(t^*, s_l) + \mathbb{E}_f[c(\tilde{\theta})|s_l]$ to be constant in l. Moreover, $\mathbb{E}_f[\delta(t^*, s_l)] = 0$ implies that this constant must equal $\mathbb{E}_f[c(\tilde{\theta})]$. Therefore, in any optimum $t^* \in \mathcal{T}^\circ$ we must have $\delta(t^*, s_l) = \mathbb{E}_f[c(\tilde{\theta})] - \mathbb{E}_f[c(\tilde{\theta})|s_l]$ for all $l \in \{1, \ldots, m\}$. Using the definition of $\delta(t^*, s_l)$, we conclude that $\mathbb{E}_{g^*}[\tilde{y}|s_l] = \mathbb{E}_f[c(\tilde{\theta})] + \mathbb{E}_f[\tilde{\theta}|s_l]$, i.e. the optimal agent behave as if the conflict of interest was constant.

Using the expression for $\delta(t, s_l)$ in $\delta(t^*, s_l) = \mathbb{E}_f[c(\tilde{\theta})] - \mathbb{E}_f[c(\tilde{\theta})|\tilde{s} = s_l]$ we get

$$\frac{1}{f_S(s_l)} \sum_{i=1}^{n-1} [y(\theta_{i+1}) - y(\theta_i)][t^*_{i,l-1} - t^*_{il}] = \mathbb{E}_f[c(\tilde{\theta})] - \mathbb{E}_f[c(\tilde{\theta})|s_l]. \tag{2}$$

Iterating (2) we get the following expression for all $l \in \{1, \ldots, m-1\}$

$$\sum_{i=1}^{n-1} [y(\theta_{i+1}) - y(\theta_i)]t^*_{il} = \sum_{j=1}^{l} f_S(s_j)(\mathbb{E}_f[c(\tilde{\theta})|s_j] - \mathbb{E}_f[c(\tilde{\theta})]). \tag{3}$$

[13] For a more general version and proof of this statement see Tchen (1980) or Epstein and Tanny (1980).

Now, consider the matrix of transformations $t^\phi \in \mathbb{R}^{(n-1)\times(m-1)}$ with typical entry defined as follows

$$t^\phi_{kl} = \frac{\sum_{j=1}^{l} f_S(s_j)(\mathbb{E}_f[c(\tilde{\theta})|s_j] - \mathbb{E}_f[c(\tilde{\theta})])}{y(\theta_{k+1}) - y(\theta_k)}\phi_k$$

where $\phi = (\phi_1, \ldots, \phi_{n-1}) \in \Delta^{n-1}$.[14]

By construction t^ϕ satisfies (3). Under the assumptions that $\mathbb{E}_f[c(\tilde{\theta})|s]$ is strictly decreasing in s and that $y(\cdot)$ is strictly increasing we have that $t^\phi_{kl} \geq 0$. Thus, if $t^\phi \in \mathcal{T}$ for some $\phi \in \Delta^{n-1}$ then it solves the belief design problem and we can conclude that there is an overconfident optimal agent.

Let $\alpha := \max_{s\in S} |\mathbb{E}_f[c(\tilde{\theta})|s] - \mathbb{E}_f[c(\tilde{\theta})]|$. Therefore,

$$t^\phi_{kl} = |t^\phi_{kl}| = \left|\frac{\sum_{j=1}^{l} f_S(s_j)(\mathbb{E}_f[c(\tilde{\theta})|\tilde{s} = s_j] - \mathbb{E}_f[c(\tilde{\theta})])}{y(\theta_{k+1}) - y(\theta_k)}\phi_k\right| \leq \alpha\frac{\phi_k}{y(\theta_{k+1}) - y(\theta_k)}.$$

This implies that

$$t^\phi_{i-1,j-1} - t^\phi_{i-1,j} - t^\phi_{i,j-1} + t^\phi_{ij} \geq -t^\phi_{i-1,j} - t^\phi_{i,j-1} \geq -\alpha\left[\frac{\phi_i}{y(\theta_{i+1}) - y(\theta_i)} + \frac{\phi_{i-1}}{y(\theta_i) - y(\theta_{i-1})}\right]$$

and

$$t^\phi_{i-1,j-1} - t^\phi_{i-1,j} - t^\phi_{i,j-1} + t^\phi_{ij} \leq t^\phi_{i-1,j-1} + t^\phi_{ij} \leq \alpha\left[\frac{\phi_i}{y(\theta_{i+1}) - y(\theta_i)} + \frac{\phi_{i-1}}{y(\theta_i) - y(\theta_{i-1})}\right].$$

Therefore, there exist a $\bar{\alpha}(\phi) > 0$ such that $\alpha < \bar{\alpha}(\phi)$ implies that t^ϕ is feasible. Finally, let $\bar{\alpha} := \sup_{\phi\in\Delta^{n-1}} \bar{\alpha}(\phi) > 0$. Thus, whenever $\alpha < \bar{\alpha}$ there exists some $\phi \in \Delta^{n-1}$ such that $t^\phi \in \mathcal{T}$, which gives the desired result. ■

Proof of Proposition 4

The four incentive compatibility constraints are given by

$$\begin{aligned}
-\mathbb{E}_\varepsilon[(x_1 - y(\tilde{\theta}))^2|s_1] + w_1 &\geq -\mathbb{E}_\varepsilon[(x_2 - y(\tilde{\theta}))^2|s_1] + w_2,\\
-\mathbb{E}_\varepsilon[(x_1 - y(\tilde{\theta}))^2|s_1] + w_1 &\geq -Var_\varepsilon(y(\tilde{\theta})|s_1),\\
-\mathbb{E}_\varepsilon[(x_2 - y(\tilde{\theta}))^2|s_2] + w_2 &\geq -\mathbb{E}_\varepsilon[(x_1 - y(\tilde{\theta}))^2|s_2] + w_1,\\
-\mathbb{E}_\varepsilon[(x_2 - y(\tilde{\theta}))^2|s_2] + w_2 &\geq -Var_\varepsilon(y(\tilde{\theta})|s_2).
\end{aligned}$$

Thus, the principal's problem is

$$\max_{\{x_1,x_2,w_1,w_2,\varepsilon\}} \quad -f_S(s_1)(\mathbb{E}_f[(x_1 - \tilde{\theta})^2|s_1] + w_1) - f_S(s_2)(\mathbb{E}_f[(x_2 - \tilde{\theta})^2|s_2] + w_2)$$

[14] Δ^a denotes the a-dimensional simplex, i.e., the set of a-vectors satisfying $\phi_i \geq 0$ and $\sum_{i=1}^{a} \phi_i = 1$.

subject to the incentive compatibility constraints that can be expressed as

$$\begin{array}{lrcl} [1] & w_2 - w_1 & \leq & x_2^2 - x_1^2 - 2(x_2 - x_1)\mathbb{E}_\varepsilon[y(\tilde{\theta})|s_1], \\ [2] & w_1 & \geq & (x_1 - \mathbb{E}_\varepsilon[y(\tilde{\theta})|s_1])^2, \\ [3] & w_2 - w_1 & \geq & x_2^2 - x_1^2 - 2(x_2 - x_1)\mathbb{E}_\varepsilon[y(\tilde{\theta})|s_2], \\ [4] & w_2 & \geq & (x_2 - \mathbb{E}_\varepsilon[y(\tilde{\theta})|s_2])^2. \end{array}$$

A necessary condition for [1] and [3] is

$$(x_2 - x_1)(\mathbb{E}_\varepsilon[y(\tilde{\theta})|s_2] - \mathbb{E}_\varepsilon[y(\tilde{\theta})|s_1]) \geq 0.$$

In any optimum we must have $x_2 \geq x_1$. Otherwise, the principal is strictly better off asking for the average action regardless of the signal; such a change can only decrease the transfers to the agent. Therefore, we also need that $\mathbb{E}_\varepsilon[y(\tilde{\theta})|s_2] \geq \mathbb{E}_\varepsilon[y(\tilde{\theta})|s_1]$.

Additionally, in the optimum, either [2] or the [4] bind. Otherwise the principal can reduce both payments by the same small amount without affecting the other two constraints. Moreover, at least two constraints must bind. Suppose [2] binds. If [3] and [4] are slack, the principal can decrease w_2. Likewise, if [4] binds, either [1] or [2] must also bind. The binding constraints determine the optimal wages as a function of the recommended actions and the level of confidence.

Let $\mu_i := \mathbb{E}_\varepsilon[y(\tilde{\theta})|s_i]$ and $\bar{\mu} := (\mu_1 + \mu_2)/2$. The three possible cases are:

Case	Binding constraints	$\mathbf{w_1}$	$\mathbf{w_2}$	Restriction
1	[2] and [3]	$(x_1 - \mu_1)^2$	$(x_2 - \mu_2)^2$ $+2(\mu_2 - \mu_1)(x_1 - \bar{\mu})$	$x_2 \geq x_1 > \bar{\mu}$
2	[2] and [4]	$(x_1 - \mu_1)^2$	$(x_2 - \mu_2)^2$	$x_1 \leq \bar{\mu} \leq x_2$
3	[1] and [4]	$(x_1 - \mu_1)^2$ $+2(\mu_2 - \mu_1)(\bar{\mu} - x_2)$	$(x_2 - \mu_2)^2$	$x_1 \leq x_2 < \bar{\mu}$

The subsequent analysis will be divided in the previous cases. Since beliefs enter principal's objective only through wages, as a first step we will focus on choosing the beliefs that minimize expected transfers, defined as $W(x_1, x_2, \mu_1) := f_S(s_1)w_1 + f_S(s_2)w_2$, to implement a given pair of actions (x_1, x_2). After that, we will solve for the optimal recommended actions. Finally, we will be able to conclude that under the assumption that $|\mathbb{E}_f[c(\tilde{\theta})]| \leq \mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]$ the solutions corresponds to the one in case 2.

Case 1: $x_2 \geq x_1 > \bar{\mu}$.

$$W(x_1, x_2, \mu_1) = (x_1 - \mu_1)^2 + f_S(s_2)(x_2^2 - x_1^2) - 2f_S(s_2)\mu_2(x_2 - x_1).$$

Thus,

$$\begin{aligned}\frac{\partial W(x_1,x_2,\mu_1)}{\partial \mu_1} &= -2\,(x_1-\mu_1)+2f_S(s_1)(x_2-x_1),\\ \frac{\partial^2 W(x_1,x_2,\mu_1)}{\partial \mu_1^2} &= 2.\end{aligned}$$

As a result, $\mu_1^* \coloneqq [1+f_S(s_1)]x_1 - f_S(s_1)x_2$ minimizes the expected wage. Note that $\mu_1^* \le x_1$ with equality only if $x_1 = x_2$.

Additionally, $\bar{\mu}^* < x_1$ requires

$$\mathbb{E}_f[y(\tilde{\theta})] < x_1 + f_S(s_1)[f_S(s_2)-f_S(s_1)](x_2-x_1).$$

By the envelope theorem

$$\begin{aligned}\frac{\partial W(x_1,x_2,\mu_1^*)}{\partial x_1} &= 2\,(x_1-\mu_1^*)+2f_S(s_2)(\mu_2^*-x_1),\\ \frac{\partial W(x_1,x_2,\mu_1^*)}{\partial x_2} &= 2f_S(s_2)\,(x_2-\mu_2^*)\,.\end{aligned}$$

Now, we focus on optimal actions. The principal's problem becomes

$$\max_{\{x_1,x_2\}} -f_S(s_1)\mathbb{E}_f[(x_1-\tilde{\theta})^2|s_1] - f_S(s_2)\mathbb{E}_f[(x_2-\tilde{\theta})^2|s_2] - W(x_1,x_2,\mu_1^*).$$

First order conditions are given by

$$\begin{aligned}-2f_S(s_1)(x_1-\mathbb{E}_f[\tilde{\theta}|s_1]) &= \frac{\partial W(x_1,x_2,\mu_1^*)}{\partial x_1},\\ -2f_S(s_2)(x_2-\mathbb{E}_f[\tilde{\theta}|s_2]) &= \frac{\partial W(x_1,x_2,\mu_1^*)}{\partial x_2}.\end{aligned}$$

Equivalently,

$$\begin{aligned}-f_S(s_1)(x_1-\mathbb{E}_f[\tilde{\theta}|s_1]) &= x_1-\mu_1^*+f_S(s_2)(\mu_2^*-x_1),\\ -(x_2-\mathbb{E}_f[\tilde{\theta}|s_2]) &= x_2-\mu_2^*.\end{aligned}$$

Thus, the following system of linear equations characterizes the optimum:

$$\begin{aligned}x_1^* &= \frac{\mu_1^*+\mathbb{E}_f[\tilde{\theta}|s_1]}{2} - \frac{f_S(s_2)}{2f_S(s_1)}(\mu_2^*-\mu_1^*),\\ x_2^* &= \frac{\mu_2^*+\mathbb{E}_f[\tilde{\theta}|s_2]}{2},\\ \mu_1^* &= [1+f_S(s_1)]x_1^* - f_S(s_1)x_2^*,\\ \mu_2^* &= \frac{\mathbb{E}_f[y(\tilde{\theta})]-f_S(s_1)\mu_1^*}{f_S(s_2)}.\end{aligned}$$

The solution is given by

$$\begin{bmatrix} x_1^* \\ x_2^* \\ \mu_1^* \\ \mu_2^* \end{bmatrix} = \begin{bmatrix} \frac{1+f_S(s_2)}{2[1-f_S(s_1)f_S(s_2)]} & \frac{f_S(s_1)^3-2f_S(s_1)f_S(s_2)}{2[1-f_S(s_1)f_S(s_2)]} & \frac{f_S(s_1)f_S(s_2)[1+f_S(s_1)]}{2[1-f_S(s_1)f_S(s_2)]} \\ \frac{f_S(s_2)}{2[1-f_S(s_1)f_S(s_2)]} & \frac{f_S(s_1)^2[1+f_S(s_1)]}{2[1-f_S(s_1)f_S(s_2)]} & \frac{[1+f_S(s_1)^2]f_S(s_2)}{2[1-f_S(s_1)f_S(s_2)]} \\ \frac{1}{1-f_S(s_1)f_S(s_2)} & -\frac{f_S(s_1)f_S(s_2)[1+f_S(s_1)]}{1-f_S(s_1)f_S(s_2)} & \frac{f_S(s_1)^2f_S(s_2)}{1-f_S(s_1)f_S(s_2)} \\ \frac{f_S(s_2)}{1-f_S(s_1)f_S(s_2)} & \frac{f_S(s_1)^2[1+f_S(s_1)]}{1-f_S(s_1)f_S(s_2)} & -\frac{f_S(s_1)^3}{1-f_S(s_1)f_S(s_2)} \end{bmatrix} \begin{bmatrix} \mathbb{E}_f[y(\tilde{\theta})] \\ \mathbb{E}_f[\tilde{\theta}|s_1] \\ \mathbb{E}_f[\tilde{\theta}|s_2] \end{bmatrix}$$

Note that

$$(\mu_2^* - \mu_1^*)\frac{1 - f_S(s_1)f_S(s_2)}{f_S(s_1)} = -\mathbb{E}_f[c(\tilde{\theta})] - [\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]].$$

therefore $\mu_2^* \geq \mu_1^*$ if and only if $\mathbb{E}_f[c(\tilde{\theta})] \leq -[\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]]$.

Case 2: $x_1 \leq \bar{\mu} \leq x_2$.

We have

$$W(x_1, x_2, \mu_1) \quad = f_S(s_1)\,(x_1 - \mu_1)^2 + f_S(s_2)\,(x_2 - \mu_2)^2\,.$$

Thus,

$$\begin{aligned} \frac{\partial W(x_1, x_2, \mu_1)}{\partial \mu_1} &= 2f_S(s_1)\,[x_2 - x_1 - (\mu_2 - \mu_1)]\,, \\ \frac{\partial^2 W(x_1, x_2, \mu_1)}{\partial \mu_1^2} &= 2\frac{f_S(s_1)}{f_S(s_2)}. \end{aligned}$$

As a result, $\mu_2^* - \mu_1^* = x_2 - x_1$, equivalently $\mu_1^* := \mathbb{E}_f[y(\tilde{\theta})] - f_S(s_2)(x_2 - x_1)$, minimizes the expected wage.

Additionally, $x_2 \geq \bar{\mu}^* \geq x_1$ requires

$$\mathbb{E}_f[y(\tilde{\theta})] \in [x_1 + \frac{f_S(s_2) - f_S(s_1)}{2}(x_2 - x_1), x_2 + \frac{f_S(s_2) - f_S(s_1)}{2}(x_2 - x_1)].$$

By the envelope theorem

$$\begin{aligned} \frac{\partial W(x_1, x_2, \mu_1^*)}{\partial x_1} &= 2f_S(s_1)(x_1 - \mu_1^*), \\ \frac{\partial W(x_1, x_2, \mu_1^*)}{\partial x_2} &= 2f_S(s_2)(x_2 - \mu_2^*). \end{aligned}$$

Now, we focus on optimal actions. The first order conditions of the principal's problem are given by

$$\begin{aligned} -(x_1 - \mathbb{E}_f[\tilde{\theta}|s_1]) &= x_1 - \mu_1^*, \\ -(x_2 - \mathbb{E}_f[\tilde{\theta}|s_2]) &= x_2 - \mu_2^*. \end{aligned}$$

Thus, the following system of linear equations characterizes the optimum:

$$\begin{aligned} x_1^* &= \frac{\mu_1^* + \mathbb{E}_f[\tilde{\theta}|s_1]}{2}, \\ x_2^* &= \frac{\mu_2^* + \mathbb{E}_f[\tilde{\theta}|s_2]}{2}, \\ \mu_1^* &= \mathbb{E}_f[y(\tilde{\theta})] - f_S(s_2)(x_2^* - x_1^*), \\ \mu_2^* &= \frac{\mathbb{E}_f[y(\tilde{\theta})] - f_S(s_1)\mu_1^*}{f_S(s_2)}. \end{aligned}$$

The solution is given by

$$\begin{aligned} x_i^* &= \tfrac{\mathbb{E}_f[c(\tilde{\theta})]}{2} + \mathbb{E}_f[\tilde{\theta}|s_i], \\ \mu_i^* &= \mathbb{E}_f[c(\tilde{\theta})] + \mathbb{E}_f[\tilde{\theta}|s_i]. \end{aligned}$$

Note that $x_1^* \le \bar{\mu}^* \le x_2^*$ is equivalent to $|\mathbb{E}_f[c(\tilde{\theta})]| \le \mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]$.

Case 3: $x_2 < \bar{\mu}$.

We have

$$W(x_1, x_2, \mu_1) = f_S(s_1)x_1^2 + f_S(s_2)x_2^2 + \mu_2^2 - 2x_2\mu_2 + 2f_S(s_1)\mu_1(x_2 - x_1).$$

Thus,

$$\begin{aligned} \frac{\partial W(x_1, x_2, \mu_1)}{\partial \mu_1} &= -2\mu_2\frac{f_S(s_1)}{f_S(s_2)} + 2x_2\frac{f_S(s_1)}{f_S(s_2)} + 2f_S(s_1)(x_2 - x_1), \\ \frac{\partial^2 W(x_1, x_2, \mu_1)}{\partial \mu_1^2} &= 2\frac{f_S(s_1)^2}{f_S(s_2)^2}. \end{aligned}$$

As a result, $\mu_2^* := [1 + f_S(s_2)]x_2 - f_S(s_2)x_1$ minimizes the expected wage. Note that $\mu_2^* \ge x_2$ with equality only if $x_1 = x_2$. Additionally, $\bar{\mu}^* > x_2$ requires

$$\mathbb{E}_f[y(\tilde{\theta})] > x_2 + f_S(s_2)[f_S(s_2) - f_S(s_1)](x_2 - x_1).$$

By the envelope theorem

$$\begin{aligned} \frac{\partial W(x_1, x_2, \mu_1^*)}{\partial x_1} &= 2f_S(s_1)(x_1 - \mu_1^*), \\ \frac{\partial W(x_1, x_2, \mu_1^*)}{\partial x_2} &= 2(x_2 - \mu_2^*) - 2f_S(s_1)(x_2 - \mu_1^*). \end{aligned}$$

Now, we focus on optimal actions. The first order conditions of the principal's problem are given by

$$\begin{aligned} -(x_1 - \mathbb{E}_f[\tilde{\theta}|s_1]) &= x_1 - \mu_1^*, \\ -f_S(s_2)(x_2 - \mathbb{E}_f[\tilde{\theta}|s_2]) &= x_2 - \mu_2^* - f_S(s_1)(x_2 - \mu_1^*). \end{aligned}$$

Thus, the following system of linear equations characterizes the optimum:

$$\begin{aligned} x_1^* &= \frac{\mu_1^* + \mathbb{E}_f[\tilde{\theta}|s_1]}{2}, \\ x_2^* &= \frac{\mu_2^* + \mathbb{E}_f[\tilde{\theta}|s_2]}{2} + \frac{f_S(s_1)}{2f_S(s_2)}(\mu_2^* - \mu_1^*), \\ \mu_1^* &= \frac{\mathbb{E}_f[y(\tilde{\theta})] - f_S(s_2)\mu_2^*}{f_S(s_1)}, \\ \mu_2^* &= [1 + f_S(s_2)]x_2^* - f_S(s_2)x_1^*. \end{aligned}$$

The solution is given by

$$\begin{bmatrix} x_1^* \\ x_2^* \\ \mu_1^* \\ \mu_2^* \end{bmatrix} = \begin{bmatrix} \frac{f_S(s_1)}{2[1-f_S(s_1)f_S(s_2)]} & \frac{f_S(s_1)[1+f_S(s_2)^2]}{2[1-f_S(s_1)f_S(s_2)]} & \frac{f_S(s_2)^2[1+f_S(s_2)]}{2[1-f_S(s_1)f_S(s_2)]} \\ \frac{1+f_S(s_1)}{2[1-f_S(s_1)f_S(s_2)]} & \frac{f_S(s_1)f_S(s_2)[1+f_S(s_2)]}{2[1-f_S(s_1)f_S(s_2)]} & \frac{f_S(s_2)^3-2f_S(s_1)f_S(s_2)}{2[1-f_S(s_1)f_S(s_2)]} \\ \frac{f_S(s_1)}{1-f_S(s_1)f_S(s_2)} & -\frac{f_S(s_2)^3}{1-f_S(s_1)f_S(s_2)} & \frac{f_S(s_2)^2[1+f_S(s_2)]}{1-f_S(s_1)f_S(s_2)} \\ \frac{1}{1-f_S(s_1)f_S(s_2)} & \frac{f_S(s_1)f_S(s_2)^2}{1-f_S(s_1)f_S(s_2)} & -\frac{f_S(s_1)f_S(s_2)[1+f_S(s_2)]}{1-f_S(s_1)f_S(s_2)} \end{bmatrix} \begin{bmatrix} \mathbb{E}_f[y(\tilde{\theta})] \\ \mathbb{E}_f[\tilde{\theta}|s_1] \\ \mathbb{E}_f[\tilde{\theta}|s_2] \end{bmatrix}$$

Note that

$$(\mu_2^* - \mu_1^*)\frac{1 - f_S(s_1)f_S(s_2)}{f_S(s_2)} = \mathbb{E}_f[c(\tilde{\theta})] - [\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]].$$

therefore $\mu_2^* \geq \mu_1^*$ if and only if $\mathbb{E}_f[c(\tilde{\theta})] \geq \mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]$.

In conclusion. If $|\mathbb{E}_f[c(\tilde{\theta})]| < \mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]$ only the first order conditions in case 2 can hold. Moreover, it is a maximum because in that case the objective function is strictly concave (see below).

After beliefs have been chosen to minimize expected wages, the principal's problem is

$$\max_{\{x_1,x_2\}} \{-f_S(s_1)\mathbb{E}_f[(x_1 - \tilde{\theta})^2|s_1] - f_S(s_2)\mathbb{E}_f[(x_2 - \tilde{\theta})^2|s_2] - W(x_1, x_2, \mu_1^*)\}.$$

The Hessian of the objective function is given by

$$\mathcal{H} = \begin{bmatrix} -2f_S(s_1) - \frac{\partial^2 W(x_1,x_2,\mu_1^*)}{\partial x_1^2} & -\frac{\partial^2 W(x_1,x_2,\mu_1^*)}{\partial x_1 \partial x_2} \\ -\frac{\partial^2 W(x_1,x_2,\mu_1^*)}{\partial x_1 \partial x_2} & -2f_S(s_2) - \frac{\partial W(x_1,x_2,\mu_1^*)}{\partial x_2^2} \end{bmatrix}.$$

For the second case we have

$$\begin{aligned} \frac{\partial^2 W(x_1, x_2, \mu_1^*)}{\partial x_1^2} &= 2f_S(s_1)\left(1 - f_S(s_2)\right) = 2f_S(s_1)^2, \\ \frac{\partial W(x_1, x_2, \mu_1^*)}{\partial x_2^2} &= 2f_S(s_2)\left(1 - f_S(s_1)\right) = 2f_S(s_2)^2, \\ \frac{\partial^2 W(x_1, x_2, \mu_1^*)}{\partial x_1 \partial x_2} &= 2f_S(s_1)f_S(s_2). \end{aligned}$$

It follows that in that case the Hessian is negative-definite and the objective function is concave.■

Proof of Proposition 5

We start with a lemma.

Lemma 1 $Var_f(\tilde{s}) = \frac{|f|^2}{f_S(s_1)f_S(s_2)}(\theta_2 - \theta_1)^2$.

Proof: Using the following equalities

- $\mathbb{E}_f[\mathbb{E}_f[\tilde{\theta}|\tilde{s}]^2] = \mathbb{E}_f[\tilde{\theta}|s_2]^2 - f_S(s_1)[\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]][\mathbb{E}_f[\tilde{\theta}|s_2] + \mathbb{E}_f[\tilde{\theta}|s_1]]$.
- $\mathbb{E}_f[\tilde{\theta}]^2 = \mathbb{E}_f[\tilde{\theta}|\tilde{s}]^2 - 2f_S(s_1)\mathbb{E}_f[\tilde{\theta}|s_2][\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]] + f_S(s_1)^2[\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]]^2$.
- $\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1] = \frac{1}{f_S(s_2)}[f(\theta_1, s_2)\theta_1 + f(\theta_2, s_2)\theta_2] - \frac{1}{f_S(s_1)}[f(\theta_1, s_1)\theta_1 + f(\theta_2, s_1)\theta_2] = \frac{|f|}{f_S(s_1)f_S(s_2)}(\theta_2 - \theta_1)$.

we get

$$\begin{aligned}
Var_f(\tilde{s}) &= Var_f(\mathbb{E}_f[\tilde{\theta}|\tilde{s}]) = \mathbb{E}_f[\mathbb{E}_f[\tilde{\theta}|\tilde{s}]^2] - \mathbb{E}_f[\tilde{\theta}]^2 \\
&= -f_S(s_1)[\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]][\mathbb{E}_f[\tilde{\theta}|s_2] + \mathbb{E}_f[\tilde{\theta}|s_1]] \\
&+ 2f_S(s_1)\mathbb{E}_f[\tilde{\theta}|s_2][\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]] - f_S(s_1)^2[\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]]^2 \\
&= f_S(s_1)[\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]] \\
&\quad (-[\mathbb{E}_f[\tilde{\theta}|s_2] + \mathbb{E}_f[\tilde{\theta}|s_1]] + 2\mathbb{E}_f[\tilde{\theta}|s_2] - f_S(s_1)[\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]]) \\
&= f_S(s_1)f_S(s_2)[\mathbb{E}_f[\tilde{\theta}|s_2] - \mathbb{E}_f[\tilde{\theta}|s_1]]^2 \\
&= \frac{|f|^2}{f_S(s_1)f_S(s_2)}(\theta_2 - \theta_1)^2.\blacksquare
\end{aligned}$$

Delegating to an agent with beliefs g^* is optimal if and only if

$$U(g^*) = -Var_f(\mathbb{E}_{g^*}[y(\tilde{\theta})|\tilde{s}] - \mathbb{E}_f[\tilde{\theta}|\tilde{s}]) - \mathbb{E}_f[c(\tilde{\theta})]^2 - \mathbb{E}_f[Var_f(\tilde{\theta}|\tilde{s})] \geq -Var_f(\tilde{\theta}).$$

In any interior optimum we have that $Var_f(\mathbb{E}_{g^*}[y(\tilde{\theta})|\tilde{s}] - \mathbb{E}_f[\tilde{\theta}|\tilde{s}]) = 0$. Therefore, this condition is equivalent to

$$\mathbb{E}_f[c(\tilde{\theta})]^2 \leq Var_f(\tilde{\theta}) - \mathbb{E}_f[Var_f(\tilde{\theta}|\tilde{s})] = Var_f(\tilde{s}).$$

Using lemma 1 this becomes

$$\mathbb{E}_f[c(\tilde{\theta})]^2 \leq \frac{|f|^2}{f_S(s_1)f_S(s_2)}(\theta_2 - \theta_1)^2$$

which is equivalent to

$$f_S(s_1)f_S(s_2)\mathbb{E}_f[c(\tilde{\theta})]^2 - |f|^2(y(\theta_2) - y(\theta_1))(\theta_2 - \theta_1) \leq$$
$$|f|^2(\theta_2 - \theta_1)^2 - |f|^2(y(\theta_2) - y(\theta_1))(\theta_2 - \theta_1).$$

Therefore

$$\frac{f_S(s_1)f_S(s_2)\mathbb{E}_f[c(\tilde{\theta})]^2}{|f|(y(\theta_2) - y(\theta_1))(\theta_2 - \theta_1)} - |f| \leq [\frac{(\theta_2 - \theta_1)}{y(\theta_2) - y(\theta_1)} - 1]|f| = \varepsilon^*$$

where the definition of ε^* follows from the proof of proposition 1.■

B Truth-or-noise information structure

In this appendix I analyze the baseline model described in section 2 under alternative assumptions on the state space and the information structure. Specifically, I assume that the state space is a compact interval, i.e. $\Theta = [\underline{\theta}, \bar{\theta}]$, and that the information structure takes the particular form I describe below.

The agent's information is represented by an information structure $\langle S, F\rangle$, where $S \subseteq \mathbb{R}$ is the signal space and F is a *c.d.f.* over states and signals. Let F_Θ be marginal c.d.f. of the state under F. I assume that the signal equals the state with probability $\rho \in (0,1)$, otherwise the signal equals an independent draw from F_Θ. Moreover, F_Θ is assumed to have strictly positive density f_Θ.

In this context, ρ represents the *precision* of the agent's information (or, for short, the agent's precision). Additionally, I consider the case in which the agent can misperceive his precision. Specifically, the agent thinks that his precision equals $\rho + \kappa$ where $\kappa \in [-\rho, 1-\rho]$. I refer to κ as the agent's *level of confidence*. I say that the agent is *underconfident* if $\kappa < 0$, *well-calibrated* if $\kappa = 0$, and *overconfident* if $\kappa > 0$.

Finally, assume that the principal can select the agent based on his confidence level. The timing of events is as follows

1. Belief design: principal chooses $\kappa \in [-\rho, 1-\rho]$.
2. Nature draws (θ, s) according to F.
3. Agent observes s, interpreting its precision as $\rho + \kappa$, and chooses $x \in \mathbb{R}$.

Payoffs are given by $-(x-\theta)^2$ for the principal and by $-(x - y(\theta))^2$ for the agent. Let $c(\theta) := y(\theta) - \theta$.

The main result in this setting is as follows.

Proposition 6 *The unique optimal agent is*

(i) underconfident if and only if $y(\tilde{\theta})$ and $c(\tilde{\theta})$ are positively correlated.

(ii) well-calibrated if and only if $y(\tilde{\theta})$ and $c(\tilde{\theta})$ are uncorrelated.

(iii) overconfident if and only if $y(\tilde{\theta})$ and $c(\tilde{\theta})$ are negatively correlated.

Let

$$\beta := \frac{Cov(y(\tilde{\theta}), c(\tilde{\theta}))}{Var(y(\tilde{\theta}))}.$$

We have that β represent the slope of the best affine predictor of $c(\tilde{\theta})$ given $y(\tilde{\theta})$. The basic intuition in proposition 6 is the same as for proposition 1: overconfidence helps when the agent's preferences are such that he under-responds to his private information (relative to what the principal would do). The proposition and its proof shows is that β captures the precise form of under-responsiveness that overconfidence solves. Moreover, the optimal confidence level is given by

$$\kappa^*(\beta) = \begin{cases} -\rho & \text{if} \quad \beta \geq 1 \\ -\rho\beta & \text{if} \quad \beta \in (-(1-\rho)/\rho, 1) \\ 1-\rho & \text{if} \quad \beta \leq -(1-\rho)/\rho \end{cases}$$

Proof of proposition 6

Given signal realization s the optimal action for the agent equals $x_\kappa(s) = \mathbb{E}_\kappa[y(\tilde{\theta})|s] = (\rho+\kappa)y(s) + (1-\rho-\kappa)\mathbb{E}_F[y(\tilde{\theta})]$. Therefore, the principal's expected payoff is given by

$$U(\kappa) := -\mathbb{E}_F[(x_\kappa(\tilde{s}) - \tilde{\theta})^2] = -\mathbb{E}_F[x_\kappa(\tilde{s})^2] + 2\mathbb{E}_F[x_\kappa(\tilde{s})\tilde{\theta}] - \mathbb{E}_F[\tilde{\theta}^2].$$

Note that

$$\begin{aligned} \mathbb{E}_F[x_\kappa(\tilde{s})^2] &= (\rho+\kappa)^2\mathbb{E}_F[y(\tilde{s})^2] + 2(\rho+\kappa)(1-\rho-\kappa)\mathbb{E}_F[y(\tilde{s})]\mathbb{E}_F[y(\tilde{\theta})] \\ &+ (1-\rho-\kappa)^2\mathbb{E}_F[y(\tilde{\theta})]^2. \end{aligned}$$

Since $\tilde{s} \sim F_\Theta$, we have that $\mathbb{E}_F[y(\tilde{s})^2] = \mathbb{E}_F[y(\tilde{\theta})^2]$ and $\mathbb{E}_F[y(\tilde{s})] = \mathbb{E}_F[y(\tilde{\theta})]$. As a result,

$$\mathbb{E}_F[x_\kappa(\tilde{s})^2] = (\rho+\kappa)^2 Var_F(y(\tilde{\theta})) + \mathbb{E}_F[y(\tilde{\theta})]^2.$$

Similarly,

$$\mathbb{E}_F[x_\kappa(\tilde{s})\tilde{\theta}] = (\rho+\kappa)\mathbb{E}_F[y(\tilde{s})\tilde{\theta}] + (1-\rho-\kappa)\mathbb{E}_F[y(\tilde{\theta})]\mathbb{E}_F[\tilde{\theta}]$$

By the law of iterated expectations we know that

$$\begin{aligned}\mathbb{E}_F[y(\tilde{s})\tilde{\theta}] &= \mathbb{E}_F[y(\tilde{s})\mathbb{E}_F[\tilde{\theta}|\tilde{s}]] \\ &= \mathbb{E}_F[y(\tilde{s})[\rho\tilde{s} + (1-\rho)\mathbb{E}_F[\tilde{\theta}]]] \\ &= \rho\mathbb{E}_F[y(\tilde{s})\tilde{s}] + (1-\rho)\mathbb{E}_F[y(\tilde{s})]\mathbb{E}_F[\tilde{\theta}].\end{aligned}$$

Using the fact that $\tilde{\theta}$ and $\tilde{s}$ have the same marginal distributions we conclude that

$$\mathbb{E}_F[y(\tilde{s})\tilde{\theta}] = \rho\mathbb{E}_F[y(\tilde{\theta})\tilde{\theta}] + (1-\rho)\mathbb{E}_F[y(\tilde{\theta})]\mathbb{E}_F[\tilde{\theta}] = \rho Cov_F(y(\tilde{\theta}), \tilde{\theta}) + \mathbb{E}_F[y(\tilde{\theta})]\mathbb{E}_F[\tilde{\theta}].$$

Taking the derivative of the principal's expected payoff with respect to the confidence level we get

$$\frac{\partial U(\kappa)}{\partial \kappa} = -2\kappa Var_F(y(\tilde{\theta})) + 2\rho[Cov_F(y(\tilde{\theta}), \tilde{\theta}) - Var_F(y(\tilde{\theta}))].$$

It follows that the principal's expected payoff is strictly concave in the agent's level of confidence:

$$\frac{\partial^2 U(\kappa)}{\partial \kappa^2} = -2Var_F(y(\tilde{\theta})) < 0.$$

As a result, there exists a unique level of confidence $\kappa^*(\beta)$ that maximizes the principal's expected payoff, which is given by the following expression

$$\kappa^*(\beta) = \rho\left(\frac{Cov_F(y(\tilde{\theta}), \tilde{\theta})}{Var_F(y(\tilde{\theta}))} - 1\right) = -\rho\frac{Cov_F(y(\tilde{\theta}), c(\tilde{\theta}))}{Var_F(y(\tilde{\theta}))} = -\rho\beta.$$

Finally, $\kappa^*(\beta) > -\rho$ is equivalent to $Cov_F(y(\tilde{\theta}), \tilde{\theta})/Var_F(y(\tilde{\theta})) > 0$ and $\kappa^*(\beta) < 1-\rho$ is equivalent to $Cov_F(y(\tilde{\theta}), \tilde{\theta})/Var_F(y(\tilde{\theta})) < 1/\rho$. Therefore, $\kappa^*(\beta) \in (-\rho, 1-\rho)$ is equivalent to $\beta \in (-(1-\rho)/\rho, 1)$. ■