Biases of Others*

David N. Danz† Kristóf Madarász‡ Stephanie W. Wang§

This version: April, 2019
First Online Draft 2014.

Abstract

When social beliefs are systematically biased, understanding how people think about the biases of others is essential since biased beliefs about the beliefs of others and the anticipation of the same biased tendency in the thinking of others are directly linked. We find that while people fail to engage in sufficient mental perspective taking and naively project their information onto differentially-informed others, they also anticipate that differentially-informed others mistakenly project onto them. In particular, we test a tight one-to-one relationship between the extent to which the typical person projects onto others, $\rho$, and the partial extent to which she anticipates but underestimates the projection of others onto her, $\rho^2$, as implied by projection equilibrium. We find that most people are partially biased and partially anticipate the biases of others and that the structure of this psychological phenomena is remarkably consistent with the proposed portable and tight idea of social beliefs arising from a coherent but fully egocentric belief hierarchy with a partial probabilistic adjustment to the truth. Applications to defensive agency and law and economics are discussed.

Keywords: social beliefs, metacognition, partial sophistication, projection, hindsight bias, tort law, defensive medicine, behavioral law and economics, behavioral organizations.

*We are grateful to Ned Augenblick, Douglas Bernheim, Arie Beresteanu, Colin Camerer, John Duffy, Emir Kamenica, Botond Koszegi, Dorothea Kübler, George Loewenstein, Muriel Niederle, Ted O’Donoghue, Al Roth, Andrew Schotter, Adam Szeidl, Lise Vesterlund, Axel Werwatz, Alistair Wilson, and seminar audiences at Amsterdam, Berkeley, CEU, Columbia, Cornell, CMU, UCL, USC, Stanford, Stanford SITE 2014, Utah, ESA Berlin 2018, and SEA 2018, for comments.

†University of Pittsburgh and WZB Berlin.
‡London School of Economics and Political Science and CEPR
§University of Pittsburgh.
“I found the concept of hindsight bias fascinating, and incredibly important to management. One of the toughest problems a CEO faces is convincing managers that they should take on risky projects if the expected gains are high enough. [...] Hindsight bias greatly exacerbates this problem, because the CEO will wrongly think that whatever was the cause of the failure, it should have been anticipated in advance. And, with the benefit of hindsight, he always knew this project was a poor risk. What makes the bias particularly pernicious is that we all recognize this bias in others but not in ourselves.”


1 Introduction

While there is growing interest in understanding how people’s perceptions systematically deviate from the truth (e.g., Tversky and Kahneman, 1974; Camerer et al., 2004; Gennaioli and Shleifer, 2010; Bénabou and Tirole, 2016; Augenblick and Rabin, 2018), there is much less direct evidence on the nature of metacognition about such tendencies. Do people explicitly think others form systematically wrong beliefs, how well-calibrated are they? Is there a relationship between their awareness of a mistaken tendency in others and the extent of the same mistake in their own judgement? Studying an individual misperception, such as people’s misprediction of their future preferences, one can proceed independently of how they think of the same misperception in others.¹ For a social bias, however, that is for a misperception of the beliefs of others, this no longer holds. What an agent thinks his principal thinks, or what a trader thinks his fellow trader thinks, entails his belief about what she thinks he thinks. For a social bias the nature of metacognition is thus essential, since here one must specify a person’s basic misperception and her perception of the same kind of misperception in others simultaneously.

¹In self-control problems, evidence suggests that people are too optimistic when predicting the time they will take to complete a task, but more pessimistic when predicting the time others will take, e.g., Buehler, Griffin, and Ross (1994). In the context of loss aversion, e.g., van Boven, Loewenstein, Dunning (2003) provide evidence consistent with the idea that people anticipate the fact that ownership changes preferences but underestimate the extent of this change both for themselves and for others. In level-k and cognitive hierarchy models, people form non-equilibrium beliefs about others’ beliefs of their reasoning about play. Instead, our paper considers people’s biased beliefs about others’ beliefs about the payoff relevant uncertainty.
Our paper then considers the fundamental domain, sometimes referred to as theory of mind or mentalizing capacity, of forming beliefs about the beliefs of others. In the presence of private information, classic evidence from ‘false-belief tasks’ (Piaget, 1953; Wimmer and Perner, 1983), hindsight tasks (Fischhoff, 1975), curse-of-knowledge tasks (Camerer et al., 1989), the illusion of transparency (Gilovich et al., 1998), or the outcome bias (Baron and Hershey, 1988) point to a robust mistake. People project their information onto others in that they too often act as if others have the same information as they do. In strategic settings, what matters, however, is not simply what information a person assigns to others; what information she thinks others assign to her is often at least as important. A key question is, thus, whether people anticipate others’ insufficient perspective taking and, if so, what is the relationship between appreciating this tendency in others and their own tendency toward insufficient perspective taking? At the same time, evidence on such a relationship appears to be lacking.

Anticipating the biases of others has direct economic implications. For example, in organizations, a principal (a board) evaluates the quality of an agent (a bureaucrat or a manager) by monitoring him with ex-post information. A naive principal projecting too often acts as if such ex-post information was readily available to the agent ex-ante. In turn, she misattributes the ex ante ex post information gap to the lack of sufficient skill or ex-ante care by the agent, on average. If the agent is sophisticated and anticipates the principal’s mistake, he will want to respond to such extensive monitoring by engaging in defensive practices (covering his ass). Specifically, he will want to distort the production of ex-ante information, overproduce ex-ante information that substitutes for, and underproduce ex ante information that is complemented by (can be understood with much greater clarity in hindsight) ex post information; undertake an ineffective selection of tasks; inefficiently delay decisions, or withdraw from a productive relationship (Madarasz, 2012). Such defensive agency is then predicated jointly on a principal naively projecting while the agent anticipating such projection.

---

2 Although hindsight bias is sometimes described as an intrapersonal phenomenon, the overwhelming evidence is predominantly from interpersonal settings.

3 An example of defensive agency is defensive medicine under medical malpractice liability. The radiologist Leonard Berlin, in his 2003 testimony on the regulation of mammography to the U.S. Senate Committee on Health, Education, Labor, and Pensions, describes explicitly how
More generally, but also practically, in the presence of transaction costs, tort law provides relevant economic incentives for internalizing externalities, e.g., two parties privately investing in ex ante efficient risk controls (e.g., Coase, 1960; Posner 2014). Liability judgements on whether the parties have adopted such risk controls given the ex-ante information, performed due care, as required by the commonly-adopted negligence standard of liability, Kaplow and Shavell (2002), occur, however, only ex-post once damages, along with new information, are realized.\(^4\) Such judgements are, then, considered to be heavily impacted by hindsight bias, information projection, e.g. Rachlinski (1998). A piece of data (a radiograph, a therapist’s note, a body cam footage, an engineer’s, or investment report) may have indicated that taking a certain costly precautionary action had a low probability of producing an appropriate benefit given the ex-ante information. The same piece of data will, however, suggest that such a precautionary action should have been taken when interpreted in light of the ex-post information leading to excessive liability rulings.\(^5\) Projection may, however, affect not only the compensatory aspect, but the perhaps more crucial incentive aspect of tort law. While the former depends on the extent of the basic bias, the latter on the anticipation of this bias in others. It is whether economic actors underestimate or exaggerate the impact of projection on liability rulings which determines their perceived incen-

\(^4\)The negligence standard is the predominant standard for ruling tort liability in the US, e.g., Kaplow and Shavell (2002). On the optimality of the negligence standard under unbiased neoclassical assumptions see, e.g., Shavell (1980).

\(^5\)For a case study in the context of radiology and torts, see, e.g., Berlin (2000).
tives and the departures from socially optimal investments. In turn, the economic content of the law is determined by the *relationship* between the extent of the basic bias and the extent to which people anticipate it in others.

The lack of evidence (and theory) on this relationship may have contributed to the fact that:

“despite the vast law and economics literature in the area of torts, no attention seems to have been paid to the potentially significant implications of hindsight bias for achieving optimal deterrence – the goal posited by that literature.” (Jolls, Sunstein, and Thaler, 1998, pp. 1525).

Understanding this relationship can then open up an avenue for incorporating these issues more carefully into the economic analysis of law and in many other strategic contexts from trade and trust formation to the impact of performance evaluations on risk taking or strategic communication. For example, it may help understand the ways tort reform affects physicians’ propensity to practice defensive medicine, trade-offs between ex ante regulation and judicial torts, Shleifer (2010), or how proposed changes to the evidentiary threshold for ruling negligence may help restore incentives or instead be counterproductive. We return to a more detailed discussion of this application in light of our results in Section 5.

Given the robust presence of the basic mistake, one may argue that anticipating it in others shall be uncommon. Being aware of it in others should prompt a person to recognize and correct this tendency in herself. This intuition would imply that a person is either naive, she is subject to this mistake and does not anticipate it in others, or sophisticated, she anticipates it in others and does not exhibit it herself. Under this dichotomy, the economic implications of this phenomenon may be limited. In domains where economic actors are naive, the potentially more severe consequence arising from anticipation may be absent. Conversely, one may prevent an anticipating radiologists’ incentive to practice defensive medicine by ensuring that the panel of ex-post malpractice evaluators consists of similar professionals. A basic test of whether such a dichotomy provides a good approximation of the perceptual heterogeneity in the population is, then, critical but also appears to be missing from the literature.
Clearly, if a person forms, on average, unbiased beliefs about the beliefs of others, thus fully anticipating their insufficient mental perspective taking, then she can not be systematically biased about their beliefs and can not engage in insufficient mental perspective taking herself. Beyond the above classic dichotomy, however, there is a bewildering variety of ways in which a relationship between the basic bias and its anticipation in others (and higher-order implications) could be specified.

In an already fairly restricted manner, one may suggest that while the typical person mistakenly assigns some probability $\rho \in [0,1]$ to differentially-informed others acting on her actual information about the payoff state, she thinks that her opponent mistakenly assigns some probability $\chi \in [0,1]$ to her acting on his differential information. One could adopt a sceptical stance and leave the relationship between $\rho$ and $\chi$ unspecified or propose some particular relationship. For example, one may propose that people commonly believe in a fictional event whereby others get the same information as they do with some probability $\rho$ beyond what is truly warranted. When, in reality, player $A$ has information $a$ and player $B$ information $b$ for sure, $A$ then believes both (i) that, with probability $\rho$, $B$ has information $a$ and, with probability $1 - \rho$, she has information $b$ and (ii) that $B$ mistakenly believes that, with probability $\rho$, $A$ has information $b$ and, with probability $1 - \rho$, he has information $a$. This suggestion of a common belief in such a fictional event would then amount to $\chi = \rho$ and people, on average, anticipating the full extent of the bias in others.

Alternatively, one may suggest that the typical person is naive and thinks that her opponent makes no systematic mistake vis-a-vis her information, $\chi = 0$, or that she exaggerates his bias relatively to the basic bias she exhibits, $\chi > \rho$, etc. In the absence of further guidance, economically key implications of this phenomenon are hard to determine. Yet, if the underlying psychology of this phenomenon implied a portable structural relationship, this would greatly facilitate the economic study of this phenomenon both theoretically and empirically.

The model of projection equilibrium, Madarasz (2014, revised 2016), offers a general yet parsimonious account of this phenomenon by fully specifying its higher-order implications for Bayesian games. Its key aspect is a model of partial projection, governed by a single scalar $\rho \in [0,1)$, which postulates a tight one-
to-one relationship between the partial extent to which a person projects onto others (\textit{first-degree projection}) and the partial extent to which she anticipates but underestimates the projection of others onto her (\textit{second-degree projection}).

The model claims that each player’s belief hierarchy is based on a fully-egocentric projective fantasy of her opponent with a partial probabilistically adjustment towards an unbiased view of her opponent. This projective fantasy is fully-egocentric in that a person mistakenly believes that with probability $\rho$, her opponent not only has the same basic information about the payoff state as she does, but he also knows exactly what she thinks, that is, what information she has, what she believes about the information of others, etc., i.e., that he assigns probability one to her actual belief hierarchy. She then places the remaining weight on an unbiased estimate of the distribution of her opponent’s beliefs, as in the standard unbiased model, which now includes understanding that he projects to degree $\rho$. This way, the model implies a tight relationship whereby $\chi = (1 - \rho)\rho$ and each person underestimates the extent of others’ projection onto her exactly in proportion to her projection onto them and there is a non-monotone relationship between the size of the basic bias and the size of the anticipated bias. Our paper introduces an experimental design to study the key issue of anticipation, the underestimation of others’ projection, and the extent to which this idea may help organize key aspects of the data.

In our experiment, principals estimate the average success rate $\pi$ of reference agents in a real-effort task. While agents never receive the solution to the task, principals receive the solution to the task prior to the estimation in the informed treatment, as in the case of monitoring with ex-post information, but not in the uninformed treatment. Projection equilibrium predicts that a principal in the former, but not in the latter, treatment should systematically overestimate the success rate. We find that in the uninformed treatment, principals are very well-calibrated. In contrast, in the informed treatment, while the true success rate is 39\%, the principals’ estimate is 57\%, on average. This exaggeration allows us to identify the extent of first-degree projection in our data.

To first obtain a qualitative response regarding anticipation, strategically active agents choose between a sure payoff and an investment whose payoff is decreasing in the principal’s estimate of the success rate. We find strong evidence
of anticipation, in that only 32.7% of agents in the informed versus 60.8% in the uninformed treatment choose to invest. In the context of defensive agency, the agent’s willingness to take on ex ante risk drops by 30 percentage points when she is monitored by ex-post information.

We then turn to the main structural hypotheses of our paper. We elicit both strategically active agents’ first- and second-order estimates (their estimate of the principals’ estimates) of the success rate of the reference agents. In the uninformed treatment, projection equilibrium, just as the unbiased BNE, predicts that these two should be the same and that, on average, they should also equal the truth. Indeed, the data confirm all of these predictions. In contrast, in the informed treatment, while unbiased BNE makes the same prediction as in the uninformed treatment, projection equilibrium predicts two key departures. While the agent’s first-order estimate should still be well-calibrated, (i) her second-order estimate should be higher than her own first-order estimate but (ii) lower than the principal’s estimate, on average. We find exactly this pattern. The agents’ second-order estimate is 51%. As predicted, players explicitly anticipate that others are biased but underestimate the extent.

Next, we consider the link between the first- and second-degree projections. The model predicts that the extent to which the principal exaggerates the success rate in the informed treatment shall fully match the square-root extent to which the agent under-estimates the principal’s exaggeration. Our data are remarkably consistent with this prediction. The extent of projection calculated based on the principal’s exaggeration of the success rate is 0.326, while that calculated based on the square-root of the agent’s underestimation of this exaggeration is 0.334. We also perform an econometric estimation of projection equilibrium accounting for individual heterogeneity. We compare the specification in which the principal’s mistake and the agent’s mistake about the principal’s mistake can freely differ with the specification that constrains these to be the same in the above manner. We find that these two specifications deliver very similar parameter estimates and very similar log-likelihoods.

Finally, we turn to the realism of the proposed idea of partial sophistication at the individual level. We document three main facts. First, the majority of the principals are partially biased and partially exaggerate the success-rate of the
agents. Second, the majority of agents explicitly anticipate but underestimate the biases of the principals. Indeed, for the majority of our agents, we can reject the hypotheses that they are fully naive about the biases of others and that they fully anticipate the biases of others. Third, and most surprisingly, we find that (i) the distribution of the degree of projection inferred from the population of principals’ first-order estimates and (ii) the distribution of the degree of projection inferred from the population of agents’ second-order estimates are remarkably close to each other. The data thus provides strong support for the idea of social beliefs stemming from a coherent but fully-egocentric belief hierarchy with a partial probabilistic adjustment towards the truth.

To the best of our knowledge, our paper is the first to consider a direct and explicit test of the structure of a social belief bias while also demonstrating its impact on strategic behavior. The rest of the paper is organized as follows. Section 2 presents the design, Section 3 the predictions, and Section 4 the results. In Section 5, we discuss alternative hypotheses and turn to some economic implications of our results including the non-monotone relationship between the size of the basic bias and the size of the anticipated bias in others.

2 Experimental Design

2.1 Experimental task

All participants worked on the same series of 20 different change-detection tasks. In each basic task, the subjects had to find the difference between two otherwise identical images. Figure 1 shows an example. The change-detection task is a common visual stimulus (Rensink et al., 1997; Simons and Levin, 1997) and has already been studied in the context of the curse-of-knowledge (e.g., Loewenstein et al., 2006).

We presented each basic task in a short clip in which the two images were displayed alternately with short interruptions. Afterwards, subjects could submit

\footnote{Each image was displayed for one second, followed by a blank screen for 150 milliseconds. The total duration of each clip was 14 seconds.}
Figure 1: Example of an image pair. Image sequence in the experiment: A, B, A, B, ... 

an answer by indicating the location of the difference on a grid (see Instructions in the Appendix 8).

2.2 Principals

*Principals* had to estimate the performance of others in a series of basic tasks. Specifically, the principals were told that subjects in previous sessions had worked on each of these tasks and that these subjects (*reference agents*, henceforth) had been paid according to their performance. We took the performance data of 144 reference agents from Danz (2014), in which the subjects performed the tasks in winner-take-all tournaments and faced the tasks in the exact same way as the subjects in the current experiment.

In each of the 20 rounds, the principals were first exposed to a basic task. Afterwards, the principal stated her estimate \((b_P)\) of the fraction of reference agents who spotted the difference in that task (success rate \(\pi\) henceforth). After each principal stated his or her estimate, the next round, with a new and always different basic task, started.\(^7\)

For the principals, the two treatments differed as follows. In the *informed* (asymmetric) *treatment*, principals received the solution to each basic task before seeing the task. Specifically, during a countdown phase that announced the start

\(^7\)The principals first participated in three practice rounds to become familiar with the interface.
of each task, the screen showed one of the two images with the difference highlighted with a red circle. This mimics various motivating economic examples such as monitoring with ex post information after an accident, such as in a medical malpractice case or evaluating a realized portfolio allocation decision. In the uninformed (symmetric) treatment, instead, the principals were not given solutions to the basic tasks. Principals in both treatments then saw each task exactly as the reference agents had. The principals received no feedback during the experiment.

At the end of the sessions, the principals received €0.50 for each correct answer in the uninformed treatment and €0.30 in the informed treatment. In addition, they were paid based on the accuracy of their stated estimates in two of the 20 tasks (randomly chosen): for each of these two tasks, they received €12 if $b \in [\pi - 0.05, \pi + 0.05]$—that is, if the estimate was within five percentage points of the true success rate of the agents. We ran one session with informed principals, and one with uninformed principals, with 24 participants in each.

2.3 Agents

Agents in our experiment were also told that subjects in previous sessions had worked on each basic task and that these subjects (reference agents) had been paid according to their performance. Agents were also told that principals had estimated the average performance of these reference agents and that the principals had been paid according to the accuracy of their estimates.\(^8\) The agents were further told that they had been randomly matched with one of the principals at the outset of the experiment and that this match would remain the same for the duration of the experiment.

For the agents, the two treatments differed solely with respect to the kind of principal they were matched with: in the informed treatment, agents were randomly matched with one of the informed principals; in the uninformed treatment, agents were randomly matched with one of the uninformed principals. In both treatments, the agents where made fully aware of whether the principal had received the solution (both, of course, not of the existence of the other treatment).

\(^8\)There is no significant difference between the success rate of the reference agents and the strategic active agents.
In each of the 20 rounds, the agents in both treatments first performed the basic task in the same way as the reference agents had; that is, they went through the images and then submitted a solution. Following each of the first ten change-detection tasks, the agent then stated (i) his estimate of the fraction of reference agents who spotted the difference in that task (first-order estimate $b^I_A$ henceforth); and (ii) his estimate of the principal’s estimate of that success rate (second-order estimate $b^{II}_A$ henceforth).

Although our primary interest is in this direct belief data, we also collected more classical choice data. For the second ten change-detection tasks, the agent decided between two investment options, $A$ and $B$. Option $A$ provided a sure payoff of €4. Option $B$ was a lottery in which the agent received €10 if the principal’s estimate $b_P$ was not more than ten percentage points higher than the success rate $\pi$; otherwise, the agent received €0. This binary choice, implicitly, is also a function of the agent’s first- and second-order estimates of the success rate. Choosing option $B$ can be thought of as an investment whose perceived expected return is decreasing in the wedge between the agent’s second- and first-order estimates. Throughout the paper, we will refer to this choice as the agents’ investment decision. We also ran separate sessions without belief elicitation—that is, where the agents, following each of the 20 change-detection tasks, after solving this task, had to choose between option $A$ and option $B$ as described above. In the result section, we also present these data.

Agents also received feedback, in the exact same way in both treatments, regarding the solution to each task right after solving the task. Specifically, the screen showed one of the two images with the difference highlighted with a red circle; then, the images were shown again. Agents matched to informed principals were told that this feedback corresponded to what the principal had seen for that task. Agents matched with uninformed principals were told that the principal had not received this solution to the task. In neither treatment, did agents receive any information about the principals’ estimates.

Finally, agents received €0.50 for each correct answer to the change-detection tasks. In addition, at the end of the experiment, one round was randomly selected for additional payment. If this round involved belief elicitation, we randomly selected one of the agent’s stated estimates for payment—either her first- or second-
order estimate in that round.\textsuperscript{9} The subject received €12 if her stated estimate was within five percentage points of the actual value (the actual success rate in the case of a first-order estimate and the principal’s estimate of that success rate in the case of a second-order estimate), and nothing otherwise.\textsuperscript{10} If the round selected for payment involved an investment decision, the agent was paid according to the realized payoff of this investment decision.

2.4 Procedures

The experimental sessions were run at the Technische Universität Berlin in 2014. Subjects were recruited with ORSEE (Greiner, 2004). The experiment was programmed and conducted with z-Tree (Fischbacher, 2007). The average duration of the principals’ sessions was 67 minutes, and the average earning was €15.15. The agents’ sessions lasted one hour and 45 minutes, on average; the average payoff was €20.28.\textsuperscript{11} Participants received printed instructions that were also read aloud. They then had to answer a series of comprehension questions before they were allowed to begin the experiment.\textsuperscript{12} At the end of the experiment, but before receiving any feedback, the participants completed the four-question DOSE risk attitude assessment (Wang et al., 2010), a demographics questionnaire, the abbreviated Big-Five inventory (Rammstedt and John, 2007), and a personality survey (Davis, 1983).

3 Predictions

The predictions below are based directly on the model of projection equilibrium (Madarasz, 2016).\textsuperscript{13} All proofs for the claims below are in Appendix 6. To

\textsuperscript{9}This payment structure addresses hedging concerns (Blanco et al., 2010).
\textsuperscript{10}We chose this elicitation mechanism because of its simplicity and strong incentives. In comparison, the quadratic scoring rule is relatively flat incentive-wise over a range of beliefs, and its incentive compatibility is dependent on assumptions about risk preferences (Schotter and Trevino, 2014). The Becker-DeGroot-Marschak mechanism can be confusing and misperceived (Cason and Plott, 2014). The beliefs we elicited were coherent and sensible.
\textsuperscript{11}The average durations of the sessions (the average payoffs) in the treatments with and without belief elicitation were 115 and 96 minutes (€21.47 and €19.10), respectively.
\textsuperscript{12}Two participants did not complete the comprehension questions and were excluded from the experiment.
\textsuperscript{13}See https://works.bepress.com/kristof_madarasz/43/
describe these for our design, let there be a set of states $\Omega$, with generic element $\omega$, and a prior $\phi$ over it. Player $i$'s information about the state is generated by an information partition $P_i : \Omega \rightarrow 2^\Omega$, her action set is given by $A_i$, and her payoff function by $u_i(\omega, a)$ where $a$ is an action profile. The game is then summarized by $\Gamma = \{\Omega, \phi, P_i, A_i, u_i\}$. There are only two strategically active players in the game corresponding to our design: one agent and one principal. The reference agents are strategically passive; they perform only the basic task and have a dominant strategy of maximizing the probability of success. In what follows, subscript $A$ refers to the strategically active agent and subscript $P$ to the principal.

All players solve the basic task which amounts to picking a cell $x \in D$ from the finite grid on the visual image. The action set of the principal includes her estimation task and is $A_P = D \times [0, 1]$. The action set of the strategically-active agent is $A_A = D \times [0, 1] \times [0, 1]$. Since for no player $i$ does the payoff from choosing $x_i$ directly interact with the payoff from any other decision, we denote this payoff by $f(x_i, \omega)$ and normalize it to be one if the solution is a success and zero otherwise.

**Projection Equilibrium.** Under projection equilibrium, each real player $i$, real Aaron, best responds to a biased perception of his opponent $j$'s, Pepi's, play. Real Aaron assigns probability $\rho_A$ to a fictional projected version of Pepi. Such projection is fully egocentric or all-encompassing in the following sense. The projected version of Pepi, who exists only in real Aaron’s imagination, not only has the exact same information about the payoff state as real Aaron does, but also knows exactly the way he thinks — i.e., in each state, she assigns probability one to his actual belief hierarchy.\(^{14}\) Given such beliefs, the projected version of Pepi is then thought to best respond to real Aaron’s strategy. Real Aaron then assigns the remaining probability $1 - \rho_A$ to the real version of Pepi, that is, to an, on average, unbiased estimate to how she thinks and behaves in reality just as in the standard case under unbiased assumptions. Analogously, real Pepi assigns probability $\rho_P$ to a fully egocentric projective fantasy of Aaron who has the same information as

\(^{14}\)In a projection equilibrium each player projects both his information and his ignorance in the sense that he believes that the projected version of his opponent has the exact same information about the payoff state as he does. For details, see Section 5 and the Appendix for the general $N$-player case in Madarasz (2016). The relevant aspect of the $N$–player case employed here is described in the footnote following Claim 2.
real Pepi does and knows exactly what she thinks, and the remaining probability to the real version of Aaron.\footnote{The predictions of projection equilibrium nest the predictions of unbiased BNE; these correspond to the unbiased case of $\rho_i = 0$ for $i \in \{A, P\}$.}

In the context of our design, all-encompassing projection means that real Aaron thinks that the projected Pepi (i) has the same belief about the solution to the basic task as he does, (ii) knows Aaron’s belief about the solution to this basic task (hence, she is unbiased about his belief of the solution), (iii) knows Aaron’s belief about the distribution of others’ beliefs, etc.

Projection equilibrium, via the above stated assumption that beliefs are generated from a coherent and fully-egocentric belief hierarchy with a partial probabilistic adjustment to the truth, pins down the meaning of informational projection for strategic settings. It implies a polynomially vanishing bias structure along each player’s belief hierarchy. In particular, it postulates a parsimonious one-to-one relationship between the extent to which a player projects onto her differentially-informed opponent and the extent to which she simultaneously anticipates but underestimates the projection of her differentially-informed opponent onto her. The predictions below will highlight this relationship.

Before turning to the predictions, some additional remarks are in order. Below, do not need to assume that people make the same inference from watching the images per se. Instead, we allow players to obtain different private signal realizations about the task. We assume only that, from the relevant ex-ante perspective — i.e., before the identity of each player is randomly determined — the distribution of these signal realizations is the same for each player. In turn, we focus on the average estimate within each treatment, and the predicted differences across the treatments hold in the ex ante expected sense. Focusing on the average estimate within each treatment then allows us to pin down the predictions of the model without having to impose further, potentially harder-to-justify, structural assumptions on the data generating process.

Consider first the ex-ante probability with which a randomly chosen player $i$ who sees only the images but not the solution can solve the basic task. Denote this success rate by $\pi$. Formally, where $E$ denotes the standard expectations operator,
it follows that:
\[ \pi \equiv E_\omega [\max_{x \in D} E [f(x, \omega) \mid P_i(\omega)]] . \]

Consider now the ex-ante expected difference between the probabilities with which a randomly chosen principal and and a randomly chosen agent can solve the basic task respectively. Let
\[ d \equiv E_\omega [\max_{x \in D} E [f(x, \omega) \mid P_P(\omega)] - \max_{x \in D} E [f(x, \omega) \mid P_A(\omega)]] , \]
denote this difference. In the uninformed treatment, neither the agent nor the principal is given the solution. Hence, \( d = 0 \) by construction. In the informed treatment, the principal also has access to the solution. Since the solution always helps in expectation, here, \( d > 0 \) must hold. Consider first the principal.

**Claim 1.** Under projection equilibrium, the principal’s ex-ante expected mean estimate of \( \pi (b_{IP}^I) \) is \( \pi + \rho_P d \).

In the uninformed treatment, the principal’s estimate is unbiased. While her estimate, conditional on her own success or failure on the basic task, may well be affected by projection, e.g., inflated following success and deflated following failure — such distortions must cancel out, on average. This follows from the facts that the players here have the same ex-ante probability of success on the basic task, and that roles are determined randomly. Hence, even if the principal believes that with probability \( \rho_P \) the agent has the exact same belief about the task as she does, this leads to no distortion on average.

In the informed treatment, in contrast, since the principal always knows the solution, she does so more often than the agent. A \( \rho_P \)-biased principal always thinks that, with probability \( \rho_P \), the agent must also know the solution. A projecting principal then exaggerates the success rate, on average, and the extent of this is \( \rho_P d \). Consider now the agent.

---

16In our design, strategically active players always estimate the success rate of the strategically passive (reference) agents. This feature ensures that there cannot be an equilibrium where the strategically active agent and the principal may coordinate on suboptimal performance on the basic task to achieve higher earnings on the estimation tasks. Our predictions then already follow from interim iterative dominance, e.g., Fudenberg and Tirole (1983), given the structure of biased beliefs implied by the model.
Claim 2. Under projection equilibrium, the agent’s ex-ante expected mean

1. estimate of $\pi$ (first-order estimate $b^I_A$) is $\pi$;

2. estimate of the principal’s estimate of $\pi$ (second-order estimate $b^{II}_A$) is $\pi + (1 - \rho)\rho P d$.

In the uninformed treatment, the agent’s first-order and second-order estimates are again unbiased. Just as under the predictions of unbiased BNE, all estimates are predicted to equal $\pi$, that is, $b^I_P = b^{II}_A = b^I_A = \pi$ must hold for any $\rho_A, \rho_P \geq 0$.

In the informed treatment, while the same equivalence holds under the unbiased BNE, projection equilibrium predicts two departures. First, the agent’s second-order estimate is predicted to be systematically higher than his own first-order estimate, i.e., he thinks that she is biased. Second, the agent’s second-order estimate is systematically lower than the principal’s first-order estimate. The agent anticipates but, in proportion to his own projection, underestimates the principal’s projection. The inequality $b^I_P > b^{II}_A > b^I_A = \pi$ holds if and only if $\rho_A, \rho_P \in (0, 1)$.\(^{17}\)

To describe the logic, consider, first, an unbiased agent, $\rho_A = 0$. Such an agent assigns probability one to the real version of the principal thus has correct beliefs about the principal’s beliefs, on average. In turn, he fully anticipates the principal’s bias and there is no wedge between his second-order estimate and the principal’s first-order estimate, on average, but there is a wedge between his second-order and first-order estimates, i.e., $b^I_P = b^{II}_A > b^I_A$. In contrast, consider a fully biased agent, $\rho_A = 1$. Such an agent thinks that the principal (and the reference agents) always have the same beliefs about the solution to the task as he does and, by the fully-egocentric nature of projection, that the principal knows whether or not he knows the solution. In turn, there is no wedge between his own first- and second-order estimates on average, no anticipation, i.e., $b^I_P > b^{II}_A = b^I_A$.

Given a partially biased agent, however, all inequalities are strict. Since the agent attaches probability $1 - \rho_A$ to the real version of the principal, he anticipates

\(^{17}\)In the projection equilibrium of an $N$-player game, each player $i$ believes that the projected versions of her opponents occur in a perfectly correlated manner, that is, the agent believes that with probability $\rho_A$ the reference agents and the principal all have the same belief about the solution to the basic task as he does and that they know this about each other. See Madarasz (2016).
her projection, on average. Since the agent places probability $\rho_A$ on the projected version of the principal and such a version knows exactly what the agent thinks, thus also whether he does or does not figure out the solution. In turn, the agent understands that the principal exaggerates the success rates, but in proportion to his own projection onto her, he underestimates its extent on average, i.e., $b_A^{II} = \pi + (1 - \rho_A)\rho蒲$.

**Homogeneous Projection.** The above predictions are formulated allowing the principal and the agent to project to differing degrees, i.e., $\rho_P \neq \rho_A$. Under homogeneous projection, i.e., $\rho_A = \rho_P$, the predictions are further refined. Here, the extent of the principal’s exaggeration fully pins down the extent to which the agent underestimates this exaggeration and vice versa. Since we infer the basic mistake (*first-degree projection*) from the estimate of the principal and the misperception of the mistake in others (*second-degree projection*) from the estimate of the agent, this case is the one that is of particular interest for our study. Imposing homogeneous projection allows us to directly test the key relationship between the extent of the basic mistake and the extent of its anticipation in others predicted by the model.

The predictions under homogeneous projection are then summarized below:

<table>
<thead>
<tr>
<th>Ex-ante expected bias</th>
<th>Uninfo Treatment</th>
<th>Info Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>principal’s first-order estimate</td>
<td>0</td>
<td>$\rho(1 - \pi)$</td>
</tr>
<tr>
<td>agent’s first-order estimate</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>agent’s second-order estimate</td>
<td>0</td>
<td>$-\rho^2(1 - \pi)$</td>
</tr>
</tbody>
</table>

Given the notation introduced in the Introduction, out of the continuum of possible relationships between these probabilities — e.g., the agent could think that the principal has unbiased beliefs about his information, $\chi = 0$, or that the agent and the principal could simply have a common belief in common information, $\chi = \rho$, etc. — projection equilibrium implies that $\chi = \rho(1 - \rho) = \rho - \rho^2$. This follows directly from projection being all-encompassing and social beliefs taking the form of a fully-egocentric projective fantasy with a partial probabilistic adjustment to the truth.
Finally, we also consider a less structured setting in which the principal’s action set is unchanged, but where the active agent’s action set is $A_A = D \times \{\text{Invest, Not Invest}\}$.

**Claim 3.** *Under projection equilibrium the agent’s propensity to invest is lower in the informed than in the uninformed treatment on average iff $\rho_A, \rho_P > 0$.***

### 4 Results

#### 4.1 Principals

The data on the principals’ estimates confirm Claim 1. Principals in the uninformed treatment are, on average, very well calibrated: there is virtually no difference between their average estimate of 39.76% and the true success rate of 39.25% ($p = 0.824$).\(^{18}\) In contrast, principals in the informed treatment grossly overestimate the success rate for the vast majority of tasks.\(^{19}\) Their average estimate of the success rate amounts to 57.45%, which is significantly higher than both the true success rate ($p < 0.001$) and the average estimate of the principals in the uninformed treatment ($p < 0.001$). Accordingly, principals in the informed treatment had significantly lower expected earnings ($\text{€} 2.40$) than principals in the uninformed treatment ($\text{€} 3.65$; one-sided $t$-test: $p = 0.034$).\(^{20}\)

**Result 1.** *Principals in the informed, but not in the uninformed, treatment overestimate the true success rate, on average.*

---

\(^{18}\)We employed a $t$-test of the average estimate per principal against the average success rate (over all tasks). Figure 5 in the Appendix shows the distribution of individual performance estimates by informed and uninformed principals, together with the actual performance of the reference agents. A Kolmogorov-Smirnov test of the distributions of average individual estimates between treatments yields $p = 0.001$. Unless stated otherwise, $p$-values throughout the Results section refer to two-sided $t$-tests that are based on average values per subject.

\(^{19}\)Figure 6 in the Appendix provides a plot of the principals’ average first-order estimate for each task.

\(^{20}\)The average payoffs in the rounds randomly selected for payment were $\text{€} 1.50$ and $\text{€} 2.50$ in the informed and the uninformed treatment, respectively.
The above result and the size of the exaggeration is not surprising given the previous findings in the literature such as in Loewenstein, Moore and Weber (2006). Reproducing this reinforces the validity of our experiment.\footnote{Following Moore and Healy (2008), we can also examine the extent to which task difficulty per se plays a role here. If we divide the tasks into hard and easy ones by the median one (yielding ten hard tasks with success rates of 0.42 and below and ten easy tasks with success rates of 0.43 and above), we find that principals in the uninformed treatment, on average, overestimate the success rate for hard tasks by seven percentage points ($p = 0.047$, sign test) but underestimate the true success rate for easy tasks by six percentage points ($p = 0.059$). This reversal is well known as the Bayesian hard-easy effect as described by Moore and Healy (2008). This reversal, however, is not observed for principals in the informed treatment. The difference between the informed and uninformed treatments is significant for both easy and hard tasks ($p < 0.01$ for each difficulty level). Principals in the informed treatment significantly overestimate the success rate by 21 percentage points and 16 percentage points for hard and easy tasks, respectively.}

### 4.2 Agents

**Investment decisions**

We now move to the agent’s investment decision. The data clearly support Claim 3.\footnote{We pool the sessions with belief elicitation and those without. Within the informed [uninformed] treatments, the average investment rates per agent in sessions with belief elicitation do not differ from the average investment rates in sessions without belief elicitation ($t$-test, $p = 0.76 [p = 0.70]$). There are also no significant time trends in the investment rates (see Figure 8 in the Appendix).} Agents matched to informed principals invest at a significantly lower rate than agents matched to uninformed principals.\footnote{Figure 7 in the Appendix shows the distribution of individual investment rates in the informed and the uninformed treatment.} The average investment rate of agents matched to uninformed principals is 67.3\%, whereas the average investment rate of agents matched to informed principals is only 39.2\% ($p < 0.001$).

**Result 2.** Agents invest significantly less often in the informed than in the uninformed treatment.

The agents in the informed treatment, relative to agents in the uninformed treatment, shy away from choosing an option whose payoff decreases—in the sense of first-order stochastic dominance—in the principal’s belief. Result 2 is consistent with agents anticipating the projection of the principals. Recall that for the agents, the only difference between the two treatments is that the agent in the informed treatment was told that his principal had access to the solution, while the agent...
in the uninformed treatment was told that his principal had not been given the solution. Hence, the difference in the propensity to invest has to do with the difference between the agent’s first-order and second-order estimates.\textsuperscript{24}

Stated Estimates

We now turn to the key structured hypothesis of our study. Figure 2 summarizes our first key findings. It shows a bar chart which collects the average stated estimates of the agents in each treatment, together with the true success rate and the corresponding estimates of the principals.

![Figure 2: Agents’ first-order estimates and second-order estimates, conditional on being matched with informed or uninformed principals. Capped spikes represent 95% confidence intervals.](image)

We find no significant treatment difference in the performance of agents (their success rate is 41.35\% in the informed treatment and 39.89\% in the uninformed treatment; \(p = 0.573\)). Thus, any treatment differences in the agents’ investment decision or the agents’ beliefs cannot be attributed to differences in task performance.\textsuperscript{24}
The left panel shows the data from the uninformed treatment. Under projection equilibrium, all beliefs should be correct, on average, and this is indeed what we find. None of the elicited estimates are significantly different from the true success rate: not the agents’ first-order estimates ($p = 0.917$), not their second-order estimates ($p = 0.140$), and not the principals’ first-order estimates ($p = 0.337$).\textsuperscript{25}

The right panel shows the data from the informed treatments. The principals vastly overestimate the true success rate, while the agents’ first-order estimates are well calibrated ($p = 0.967$), as predicted. The agents’ second-order estimates, as predicted, are significantly higher than their own first-order estimates ($p < 0.001$). Finally, as predicted, the agents’ second-order estimates are also significantly lower than the principals’ estimates (one-sided $t$-test: $p = 0.047$). Agents both anticipate and underestimate the principals’ mistaken exaggeration.

When comparing treatments, the agents’ first-order estimates are not significantly different ($p = 0.956$) across treatments. The agents’ second-order estimate in the informed treatment is significantly higher than that in the uninformed treatment ($p = 0.0314$ for one-sided $t$-test). Finally, the difference between the agents’ second- and own first-order estimates is also significantly larger in the informed than in the uninformed treatment ($p < 0.001$).\textsuperscript{26} In sum, the following results hold for the agent’s beliefs.

**Result 3** (Partial anticipation of information projection). *The following results hold:*

1. *The agent’s first-order estimate of the success rate is correct in both treatments.*

2. *The difference between the agent’s second-order estimate and her own first-order estimate is significantly larger in the informed than in the uninformed treatment.*

\textsuperscript{25}In the uninformed treatment agents’ second-order estimates are somewhat higher than the principals’ first-order estimates, but this difference is not significant either ($p = 0.080$).

\textsuperscript{26}The size of the treatment effect on the wedge between the agents’ first- and second-order estimates and also on the agents’ investment decisions is unchanged when controlling for successful task performance; the treatment difference is also robust to controlling for individual characteristics (see Tables 2, 3, 4, and 5 in the Appendix, respectively).
3. In the informed treatment, the agent’s estimate of the principal’s estimate is significantly higher than the agent’s own estimate and significantly lower than the principal’s estimate.

The evidence clearly violates the predictions of the unbiased beliefs, but confirms all predictions on the structure of biased beliefs postulated under projection equilibrium.

4.3 Estimation of Projection Equilibrium

The analysis above confirmed all three predictions on the structure of biased beliefs under heterogeneous projection, the presence of the basic bias, the anticipation of this bias, and the underestimation of this bias. We now turn to the postulated link between the extent of the basic bias (first-degree projection) and the extent of its underanticipation in others (second-degree projection). Projection equilibrium provides a joint account of the basic mistake and the mistake in the anticipation of this basic mistake in others; people underestimate others projection exactly in proportion to their own projection onto them. Since, in our design, the former is recoverable from the mispredictions of the principal and the latter from the agent’s misprediction of the principal’s misprediction, to do so, we now turn to homogeneous projection—i.e., \( \rho = \rho_A = \rho_P \).

We consider, first, the aggregate data. We can directly solve for the two equations in Claims 1 and 2 by allowing for heterogeneous projection. This yields the unique solution of \( \hat{\rho}_P = 0.3 \) and \( \hat{\rho}_A = 0.33 \). Remarkably, the degree of projection inferred from the principals’ proportional exaggeration and that inferred from the square-root from the agents’ proportional underestimation of this exaggeration are very close to being homogeneous. At the aggregate level, the ratio of the agent’s underestimation of the principals’ exaggeration of the success rate over the principals’ exaggeration of the success rate is very close to the degree of the principals’ projection measured directly from their exaggeration (first-degree projection).

An Econometric Analysis

We now turn to a structured econometric test of the hypothesis of homogenous projection. We employ a Maximum Likelihood estimation of a random-coefficient
model and start with a flexible specification that allows for different degrees of projection both across roles and within roles. The parameters of the unrestricted model are $\Theta_{UR} = \{\rho_P, \rho_A, \phi_\rho, \phi_b\}$, where $\rho_P$ and $\rho_A$ denote the average degree of projection in the principal and the agent populations, respectively, and $\phi_\rho$ and $\phi_b$ are precision parameters governing variance in individual projection and noise in response, as we explain below. We then estimate the model under homogeneous projection—i.e., with restricted parameters $\Theta_R = \{\rho_P = \rho_A, \phi_\rho, \phi_b\}$. A comparison of the restricted and the unrestricted specification provides a test of homogeneous projection.

Our econometric model makes repeated use of the beta distribution. We will use a convenient alternative parameterization of the beta distribution $x \sim \text{Beta}(\mu, \phi)$ with density

$$f(x; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\phi\mu)\Gamma(\phi(1-\mu))} x^{\phi\mu-1} (1-x)^{\phi(1-\mu)-1},$$

where the first parameter $\mu$ is the expected value of $x$, and the second parameter $\phi$ is a precision parameter that is inversely related to the variance of $x$, $\text{var}(x) = \mu(1-\mu)/(1+\phi)$ (Ferrari and Cribari-Neto, 2004). That is, conditional on the mean $\mu$, higher values of $\phi$ translate into a lower variance.

Note that, by virtue of our design involving a real-effort task, without further assumptions, we can not pin down the full distribution of conditional estimates, that is the distribution of a player’s estimate conditional on her realized signal and performance even under the unbiased BNE. We only know, by virtue of the martingale property of beliefs, that the unbiased ex ante expected estimate must equal the truth. This then allows us to derive also only the ex ante expected biased estimate. We, however, know that the difference between the conditional estimates and the ex ante expected estimates are always mean zero irrespective of the degree of projection. Since probabilities are always between zero and one, our first structural assumption is that the subjects’ stated estimates are beta distributed centered around the task- and individual-specific mean estimates predicted by Claim 1 and Claim 2 nesting the unbiased BNE predictions. Specifically,
the estimates of principal $i$ and agent $j$ for task $t$ are:

\[
\begin{align*}
    b_{P,t}^I &\sim \text{Beta}(\mu_{P,t}, \phi_b), \\
    b_{A,t}^H &\sim \text{Beta}(\mu_{A,t}, \phi_b),
\end{align*}
\]

where

\[
\begin{align*}
    \mu_{P,t} &= \rho_{P_i} + (1 - \rho_{P_i}) \pi_t, \quad \text{(see Claim 1)} \\
    \mu_{A,t} &= \pi_t + (1 - \rho_{A_j}) \rho_P (1 - \pi_t). \quad \text{(see Claim 2)}
\end{align*}
\]

Our second structural assumption captures unobserved individual heterogeneity in the degree of informational projection and accounts for repeated observations on the individual level. We employ a random-coefficient specification in which individual degrees projection in the principal and the agent populations follow beta distributions with

\[
\begin{align*}
    \rho_{P_i} &\sim \text{Beta}(\rho_P, \phi_{\rho}), \\
    \rho_{A_j} &\sim \text{Beta}(\rho_A, \phi_{\rho}).
\end{align*}
\]

We impose tight restrictions on the distributions of the degree of projection in the agent and the principal populations by allowing them to differ only with respect to their location parameter. This greatly facilitates our test of equality of the average degree of projection across roles, which is the focus of this section. We take a closer look at heterogeneity and also provide a test of this structural assumption in Section 4.4.

We now formulate the log-likelihood function. Conditional on $\rho_{k_i}$ and $\phi_{\rho}$, the likelihood of observing the sequence of stated estimates $(b_{k,t})_t$ of subject $i$ in role $k \in \{A, P\}$ is given by

\[
L_{k_i}(\rho_{k_i}, \phi_b) = \prod_t f_b(b_{k,t}; \mu_{k,t}(\rho_{k_i}), \phi_b).
\]

Hence, the unconditional probability amounts to

\[
L_{k_i}(\rho_k, \phi_{\rho}, \phi_b) = \int \left[ \prod_t f_b(b_{k,t}; \mu_{k,t}(\rho_{k_i}), \phi_b) \right] f_{\rho}(\rho_{k_i}; \rho_{\rho}, \phi_{\rho}) \, d\rho_{k_i}. \quad (1)
\]
Table 1: Maximum likelihood estimates of projection bias $\rho$ based on Claim 1 and 2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unrestricted model with heterogeneous $\rho$ ($\rho_P \neq \rho_A$)</th>
<th>Restricted model with homogeneous $\rho$ ($\rho_P = \rho_A$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_P$</td>
<td>0.326*** [0.247, 0.405]</td>
<td>0.324*** [0.252, 0.397]</td>
</tr>
<tr>
<td>$\rho_A$</td>
<td>0.334** [0.110, 0.558]</td>
<td></td>
</tr>
<tr>
<td>$\phi_{\rho}$</td>
<td>3.103*** [1.169, 5.036]</td>
<td>3.092*** [1.177, 5.007]</td>
</tr>
<tr>
<td>$N$</td>
<td>720</td>
<td>720</td>
</tr>
<tr>
<td>$\ln L$</td>
<td>123.597</td>
<td>123.593</td>
</tr>
</tbody>
</table>

Note: Values in square brackets represent 95% confidence intervals. Asterisks represent $p$-values: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$ Testing $H_0: \rho_P = \rho_A$ in column (1) yields $p = 0.8389$.

The joint log likelihood function of the principals’ and the agents’ responses can then be written as

$$l(\rho_P, \rho_A, \phi_{\rho}, \phi_b) = \sum_k \sum_i \log L_k(\rho_k, \phi_{\rho}, \phi_b).$$

(2)

We estimate the parameters in (2) by maximum simulated likelihood (Train, 2009; Wooldridge, 2010). Table 1 shows the estimation results for the unrestricted model with $\rho_P \neq \rho_A$ in the left column and the restricted model with $\rho_P = \rho_A$ in the right column. Focussing on the unrestricted model first, we make three observations.

First, the principals’ average degree of projection is estimated to be $\hat{\rho}_P = 0.326$, with a confidence interval of [0.247, 0.405]. This estimate indicates the relevance

---

27. The estimation is conducted with GAUSS. We use Halton sequences of length $R = 100,000$ for each individual, with different primes as the basis for the sequences for the principals and the agents (see Train, 2009, p.221ff).

28. The results are robust with respect to alternative starting values for the estimation procedure. All regressions for a uniform grid of starting values converge to the same estimates (for both the restricted and the unrestricted models). Thus, the likelihood function in (2) appears to assume a global (and unique) maximum at the estimated parameters.

25
of informational projection: the unbiased BNE—which is the special case in which \( \rho_P = 0 \)—is clearly rejected. Second, the agents’ average degree of informational projection is estimated to be 0.334, with a confidence interval of \([0.110, 0.558]\). The \( \tilde{\rho}_A = 0.334 \) estimate—which is significantly different from 0 and 1—gives structure to our observation that agents do anticipate the projection of the principals, but, due to their own projection onto them, underanticipate the principals’ level of projection.

Crucially, the estimated parameters of the degree of projection are not significantly different between the principals and the agents \((p = 0.914)\). Furthermore, the log likelihood of the unrestricted model is very close to that of the restricted model (right column of Table 1), and standard model selection criteria (e.g., BIC) clearly favor the single-parameter model of homogeneous projection over the unrestricted model with two parameters. In short, the data are remarkably consistent with the structure of biased beliefs hypothesized.

4.4 Partial Bias and Anticipation

The aggregate estimates are consistent with homogeneous projection and the idea of a partial bias proposed. So far nothing indicates, however, that this idea has descriptive accuracy at the individual level. An account of metacognition may still hold whereby people who anticipate projection should not project themselves. In this case the fraction of agents who anticipate projection should match the fraction of principals who do not project. It could even be the case that those who anticipate the mistakes of others, anticipate it correctly. If any of this was the case, our hypothesis on the nature of metacognition and the proposed idea of a partial social bias may have little actual realism. As mentioned in the Introduction and discussed in Section 5, the economic implications of this phenomenon may then also be very different. The last, but key part of the analysis is then devoted to exploring the degree of informational projection at the individual level.

Below we then compare the distribution of principals’ degree of projection as inferred from the extent of their exaggeration of the success rate with the distribution of the agent’s degree of projection as inferred from the square-root of the extent of their underestimation of this exaggeration given the structure
of projection equilibrium. This then allows us to explore whether the estimated mean degree of projection $\hat{\rho} = 0.324$ is the result of a mixture of some people being naive and not anticipating the mistakes of others at all while others being sophisticated and anticipating it (correctly) or if instead partial anticipation is the norm. Furthermore, it allows us to perform a very conservative test implied by our hypothesis namely that these two distributions shall be the same.

We base our specification test on non-parametric density estimates of individual degrees of projection in the principal and agent populations. To this end, we first obtain individual estimates of $\rho$ for each principal and each agent from the informed treatment using simple linear regressions, without imposing any restrictions either on the size or on the sign of the parameters. In contrast to the previous subsection, we now adopt a simple OLS framework. Specifically, for each principal $i$ in the informed treatment, we estimate her degree of projection $\rho_{P_i}$ from Claim 1 via:

$$b_{P_i t} = \pi_t + \rho_{P_i} (1 - \pi_t) + \epsilon_{it}, \quad (3)$$

where $\epsilon_{it}$ denotes an error term with mean zero and variance $\sigma^2_{i}$. Analogously, for each agent $j$ in the informed treatment, we estimate his degree of projection $\rho_{A_j}$ from Claim 2 via

$$b_{A_j t} = b_{A_j t} + (1 - \rho_{A_j}) \rho_{P} (1 - \pi_t) + \epsilon_{jt}, \quad (4)$$

where $\rho_{P}$ is the mean projection bias in the principal population, and $\epsilon_{jt}$ denotes an error term with mean zero and variance $\sigma^2_{j}$. For the derivation of these equations, which now also use the information in the agent’s first-order estimates directly, see Appendix 6.2. We estimate the parameters in (3) and (4) by OLS, where we replace $\rho_{P}$ in (4) with the average estimate of $\rho_{P_i}$ obtained from the regressions in (3).\textsuperscript{29}

\textsuperscript{29}The results are qualitatively the same when using the true success rate instead of the agents’ first-order estimates in Eq. (4), that is replacing $b_{A_j t}$ with $\pi$. Figure 9 in the Appendix shows the corresponding distribution of individual $\rho$ estimates for this alternative specification for a comparison with Figure 3.

\textsuperscript{30}We base all inference on the individual level on heteroskedasticity-robust standard errors. Unlike in the simultaneous estimation of the agents’ and the principals’ projection bias from (2), the simple estimation approach applied here ensures that the individual estimates of the
Figure 3: Empirical cumulative distribution functions of principals’ (solid) and agents’ (dashed) projection bias $\rho$ in the informed treatment. The smooth line shows the estimated beta distribution from the model with homogeneous projection (right column in Table 1).

Figure 3 shows the empirical CDFs of the individual degrees of projection in the principal and agent populations. Two important observations follow.

First, the figure already suggests that the empirical CDFs of the principals’ and the agents’ estimated $\rho$s are remarkably similar. In fact, a Kolmogorov-Smirnov test does not reveal any significant difference between the distributions ($p = 0.441$). Remarkably, not only are the average degrees of projection exhibited by the principals, inferred from the extent to which the principals exaggerate the success rate, and those of the agents, inferred from the square-root of their underappreciation of the principals’ exaggeration of the success rate, are the same, principals’ projection bias are not informed by the data of the agents’ choices, a feature that is desirable for our specification test below.
but the two distributions describing the underlying heterogeneity in partialness are also not significantly different.

Second, as Figure 3 shows, partial projection is the norm when it comes to both first-degree and second-degree projection. The majority of the principals (70.8%) exhibit an estimated $\rho_P$ that is significantly larger than zero and significantly smaller than one. Similarly, the majority of the agents (50%) have an estimated $\rho_A$ that is significantly larger than zero, but significantly smaller than one. The majority of people exhibit the basic bias, but also believe that others are biased while, at the same time, partially underestimating its extent.

In sum, the findings go against the descriptive validity of a dichotomous model of metacognition. Instead, it data lends strong support to the realism of underlying logic a partial belief bias stemming from a fully egocentric belief hierarchy with a partial adjustment to the truth.

Figure 3 conveys an additional result regarding the econometric specification in (2). The pooled empirical CDF of the principals’ and the agents’ projection is not significantly different from the beta distribution $f(\hat{\rho} = 0.324, \hat{\phi}_\rho = 3.092)$ obtained from model (2) with homogeneous projection (smooth line in Figure 3; Kolmogorov-Smirnov test: $p = 0.529$).32

**Descriptive Accuracy at the Task Level**

We conclude the analysis with an assessment of the descriptive accuracy of the estimates of the agents’ anticipation and second-degree projection also at the task level. Specifically, we ask how well do their individual $\rho$ estimates, inferred from all tasks, capture the wedge between the agents’ first- and second-order estimates

---

31 The second most common category (33.3%) in the agent population is $\rho$ being not significantly different from 0 but significantly different from 1—i.e., full anticipation of others’ information projection. Further, 12.5% of the agents have an estimated $\rho$ that is not significantly different from 1 but significantly different from 0—i.e., no anticipation. Only one agent has an estimated $\rho$ that is significantly larger than one. The corresponding fractions for the principals are similar, and there is no significant difference between the agents’ and the principals’ categorized distribution of projection (Fisher’s exact test: $p = 0.179$).

32 The separate empirical CDFs principals’ and the agents’ degrees of informational projection are also well described by beta distributions, Kolmogorov-Smirnov tests of the empirical CDF against the best-fitting beta distribution yields $p = 0.941$ for the principals and $p = 0.974$ for the agents.
in any given task separately? If the predictions hold at the task level, not only over all tasks, this can then help further confirm the validity of our results.

Figure 4 compares the *empirical* average second-order estimates of the agents to their *predicted* average second-order estimates, given the model using the individual $\rho$-estimates from equation (4) estimated over all tasks, at the level of each task.\footnote{Formally, for any given basic task $t$, we first take each agent’s estimated $\rho_{A_j}$ and calculate that agent’s predicted second-order estimate given Eq.(4) for that task $t$, then average this across all agents for that task $t$.} It is apparent that the predicted second-order estimates fit the empirical data well. The correlation between the predicted and actual second-order estimates across tasks amounts to 0.88, and the share of unexplained variance in the empirical estimates is only 26.8%. This further confirms that the postulated tight relationship holds not only over all tasks, which could then potentially be due to some composition of the tasks, but at the task level as well.
5 Discussion

We are unaware of any other existing model of strategic behavior that would provide a tight explanation of the data. Below we describe the implications of some leading candidates from the literature.

5.1 Alternative Models and Mechanisms

Coarse Thinking. Unlike a number of other prominent behavioral models of play in games with private information, projection equilibrium focuses directly and explicitly on players systematically misperceiving each others’ beliefs rather than misperceiving the relationship between other players’ information and their actions. In particular, the models of ABEE (Jehiel, 2005), and cursed equilibrium
(Eyster and Rabin, 2005), assume that people have correct expectations about
the information of others, but have coarse or misspecified expectations about the
link between others’ actions and their information. Crucially, these models are
closed by the identifying assumption that those expectations about actions are
nevertheless correct, on average; that is, each player has correct expectations
about the distribution of his or her opponent’s actions.

Their identifying assumption directly implies that in our design both models
predict a null treatment effect. They have the same overall predictions as the un-
biased BNE. A cursed principal should never exaggerate the agent’s performance,
on average, and a cursed agent should never anticipate any systematic mispre-
diction by the principal, on average. Instead, we find both patterns and do so
explicitly by eliciting beliefs directly.\footnote{The same applies to the model of Esponda (2008). Note, also, that QRE also predicts no
treatment difference since the principal’s incentives in the two treatments are exactly the same.
The same is true for level-k models that hold the level zero play constant across treatments.
Furthermore, in contrast to the defining feature of our hypothesis, under cursed equilibrium
players’ behavior need not be consistent with a coherent belief hierarchy.}

**Risk Aversion.** We find no evidence that risk aversion matters for the sub-
jects’ choices. (See Tables 3 and 5 in the Appendix). Note, also, that if more
information should help an unbiased principal to make more accurate forecasts,
on average, as it should be the case, a risk-averse agent should choose the risky
option over the safe option more often in the informed than in the uninformed
treatment. Instead, we find the exact opposite pattern.

**Overconfidence.** Note that overconfidence cannot explain the subjects’
choices either. If an agent believes that he is better than average, he might under-
estimate the reference agents’ performance relative to his own, but this will not
differ across treatments. Similarly, a principal may be over- or under-confident
when inferring others’ performance on a given task, but there is no reason for this
to systematically interact with the treatment. As the data show, however, agents,
as well as principals in the uninformed treatment, are very well calibrated about
the success rate of others showing no sign of systematic under- or over-confidence
on average.

**Everybody is just like me.** Finally, one may propose a general heuristic
whereby people simply think that others are just like them. While the exact
meaning of such a heuristic may be unclear, note that if people just believe that others have the same beliefs as they do, then we cannot account for our key finding; the systematic wedge between the agents’ own first-order and second-order estimates. Such a heuristic cannot account for the fact that the typical subject explicitly thinks that others form systematically wrong (hence differing from her) beliefs about his or her true beliefs.

5.2 Conditional Estimates

As mentioned, without imposing further assumptions we cannot pin down what unbiased or the biased (unless the bias is full) conditional estimates of a player. We could only state that whatever is the right difference between these expected conditional estimates in the unbiased case projection may well inflate this difference.35 At the same time, we can meaningfully compare the estimates of the principals in the informed treatment who were given the solution with the estimates of the principals in the uninformed treatment who figured out the solution themselves. If any systematic distortion in conditional estimates is due to informational projection only, then these two sets of conditional estimates should be the same on average.

We do indeed find that the estimates of the principals in the uninformed treatment who figured out the difference themselves (60.93%) is very close to the estimates of the principals in the informed treatment who were given the solution (57.45%). This finding further supports the fact that the systematic distortion in the principals’ estimates is due to informational projection as opposed to an alternative psychological mechanism whose implications would greatly differ in the way information is acquired, e.g., problems that one solves may just appear more difficult, while problems for which one is exogenously given the solution appear too easy.

35 Consistent with this, in the uninformed treatment the principal’s estimate of the success rate conditional on spotting the difference was 60.93% while conditional on not figuring it out was 29.93%.
5.3 Some Economic Implications and Conclusion

Social cognition is a key determinant of behavior in many economic, legal, and organizational settings. Our paper provides careful evidence that, while the typical person naively engages in insufficient informational perspective-taking vis-a-vis differentially-informed others, she also anticipates that differentially-informed others engage in insufficient informational perspective-taking vis-a-vis her. The data do not support an account of metacognition whereby being aware of others’ projection implies not exhibiting the very same mistake. In that case the fraction of principals who do not project should match the fraction of agents who anticipate projection. Instead, we find that a partial basic bias and a partial anticipation of this bias in others is the norm and the data is remarkably consistent with the portable idea, proposed by projection equilibrium, of social beliefs arising from a coherent, but fully egocentric belief hierarchy with a partial probabilistic adjustment to the truth which implies a tight link between these two, \( \chi = (1 - \rho)\rho \).

There are a number of potentially important economic applications of our main findings. Principals in our study exaggerate the predictability of an initially uncertain outcome, underestimate the ex-ante noisiness of the data, by 32 % of the ex-ante ex-post information gap. Agents believe that principals exaggerate such predictability by 21 % of this gap. Since this gap is often quite substantial, these magnitudes imply a significant perceived incentive for defensive agency, but one that falls short of the true incentive.

Our empirical findings, and their support for a portable and tight model, which includes a non-monotonic relationship between the size of the basic bias and the size of the anticipated bias, may help shed novel light on a number of phenomena such as the propensity for excessive risk avoidance by managers as a function of how extensively they are monitored, or the impact of tort reform on defensive medicine. It may be incorporated into the extensive law and economics literature on deterrence, where it can inform classic results on the optimality and the implications of various commonly adopted liability standards for incentive provision, e.g., Shavell (1980), or the size of punitive damages, further qualify relative advantages of judicial tort versus more rigid direct ex ante regulation, e.g., Shleifer (2010), or classic trade-offs between information and incentives in
organizations, e.g., Holmström (1979). In many contexts it is the joint presence of our three qualitative findings, the basic mistake and the awareness of, but limited learning about, others’ mistakes that may have key implications.

**Tort Law and the Burden of Proof.** To illustrate, let’s return to the example of the economics of tort and precautions. The predominant standard for assessing tort liability in the United States is the negligence rule (Kaplow and Shavell, 2002). Under this rule, jurors and judges are required to determine the likelihood that the ex-ante information suggested that a costly action by the defendant (or the plaintiff) would have produced an appropriate expected benefit, conformed to due care, and avoided a harmful outcome, i.e., the likelihood that the negligence standard is met. If the ex-post information suggests that this action would have avoided harm, this likelihood will, however, be greatly overestimated due to the jury or judge projecting information thus being hindsight biased.

Recognizing this, Jolls, Sunstein, and Thaler (1998) suggest to offset the distortionary impact of hindsight bias on incentive provision: raising the evidentiary standard, the burden of proof, for ruling liability. They propose to exactly counterbalance the degree of ex-post overestimation (“the overestimation of the likelihood that the negligence threshold is met could in theory be precisely offset by a change in the evidentiary threshold”). Specifically, they argue that the common “preponderance of the evidence” standard of civil cases, in which liability is ruled when the evidence suggests that there is an at least a 50% chance that the negligence standard is met, could be replaced by the “clear and convincing evidence” standard corresponding to a higher, roughly 2/3 required chance that the negligence standard is met.

Our estimate of $\rho$ gives structure to the above proposal. It implies that if the evidence, in reality, suggests that there is a 50% chance that the ex-ante information indicated that taking a certain action would have been warranted, but ex-post following the accident this action turns out to have been clearly warranted,

---

36 See p. 1530.

37 On the preponderance of evidence standard corresponding to a 50% threshold, see, e.g., Lord Denning in Miller v. Minister of Pensions [1947] 2 All ER 372. This standard is also referred to as “the greater weight of all the evidence” standard. On the clear and convincing evidence standard corresponding to such a 65 – 70% threshold, see, e.g., United States v. Fatico, 458 F. Supp. 388, 410 (E.D.N.Y. 1978), or Jolls, Thaler, and Sunstein (1998).
then people will perceive that the evidence suggests that there is a \((1 - \rho)0.5 + \rho 1 \simeq 2/3\) chance that the ex-ante information indicated taking this action. Our first result may then suggest that such a raise in the burden of proof is warranted.

It is, however, not the basic bias but the anticipation of such biased liability assessments which distorts economic incentives, e.g., induces excessive precautions by potential defendants or underinvestment in care by potential plaintiffs. The logic of Jolls, Sunstein, and Thaler (1998) then implicitly presupposes that economic agents anticipate the full extent of the bias in others (they are unbiased). If economic actors were fully naive and did not anticipate the high probability of being mistakenly found liable, such a rise in the evidentiary standard would actually be counterproductive. Instead, keeping it unchanged would induce the same incentives as in the unbiased case irrespective of the bias of those judging ex post. Indeed our findings indicate that, by virtue of they themselves projecting too, thus believing that with probability \(\rho\) those judging in hindsight would surely realize what their ex ante information truly suggested, they will underestimate the bias in ex-post judgements, and do so by \(\rho^2\). While economic actors will not be fully naive, they will appear overly optimistic vis-à-vis the true probability of being found liable. Under the maintained hypothesis of projection equilibrium, the more so the greater is the basic bias. In turn, a raise, but a more modest raise than suggested above, may help achieve the desired goal.

**Punitive Damages.** A similar logic may be incorporated into the neoclassical determination of punitive damages (Polinsky and Shavell, 1998). Punitive damages are defined as the damages in excess of compensatory damages to ensure that sufficient precautions are taken ex ante when the injurers sometimes escape liability. The optimal level of punitive damages for this deterrence objective should be some multiple of the harm. Specifically, Polinsky and Shavell (1998) argue that the punitive damages multiplier is the ratio of the probability of escaping detection to the probability of being detected. At the same time, if injurers will too often be found liable due to the jury or the judge projecting information and economic actors anticipate this, then such a determination of punitive damages may well be excessive theoretically.\(^{38}\)

\^{38}See, e.g., Hastie and Viscusi (1998) on a discussion of the potential role of hindsight bias in the determination of punitive damages in the context of the Exxon Valdez oilspill disaster.
Under the maintained hypothesis of projection equilibrium, however, either if projection is full, \( \rho \rightarrow 1 \), or if it is null, \( \rho = 0 \), this provides no rationale for curbing punitive damages. Instead, given the non-monotone relationship between the size of the basic bias and the size of the anticipated bias, it is when the degree of projection \( \rho \) is moderate — \((1 - \rho)\rho\), is maximal at \( \rho = \frac{1}{2} \) — that the logic may imply the greatest rationale for curbing punitive damages. Under our maintained hypothesis, estimating the extent of the basic bias may already give economically meaningful guidance for legal policy, but only if adjusted in this fashion.

Debiasing. Our findings may also shed light on the robustness of the basic mistake to debiasing methods, e.g., Wu et al. (2010). People may correctly be aware of a mistaken tendency, yet consistently suffer from it themselves. The structure of biased social beliefs hypothesized renders this into a consistent proposition. Our results suggest that organizations that try to address the adverse consequences of information projection may be less likely to succeed if they focus on a simple sorting mechanism of naifs and sophisticates or trying to debias first-degree projection by calling attention to it.

Finally, our results illustrate the potential of obtaining economically key insights about the structure of psychological phenomena by directly eliciting higher-order beliefs. While multiple factors might capture observed departures from classic rational choice outcomes in strategic settings, it is the pattern of beliefs about those misperceptions, that may be especially helpful to offer more realistic accounts. The support for this portable idea of a coherent but fully egocentric belief hierarchy with a partial probabilistic adjustment to the truth may help understand a number of phenomena. For example, it may be key in generating systematic social biases such as pluralistic ignorance in norm formation whereby the majority systematically misperceives itself as the minority — for recent evidence on this phenomenon in the context of women’s rights, see, e.g., Bursztyn et al. (2018) — or its seeming opposite, a (truly) false consensus effect. Both of these are predicted by projection equilibrium (Madarasz, 2016). The logic of this idea may then help understand the type of interventions that may help alleviate these or the type of social interactions that reinforce these. This structure of egocentric thinking with a grain of truth may also be relevant in understanding features of deception and the winner’s curse in many classic trade settings. Indeed, in common-value
lemons problems (Akerlof, 1970) projection equilibrium provides a very good fit of the observed deviations from the unbiased predictions in the aggregate experimental choice data both for uninformed buyers and privately informed sellers.\footnote{For example, Samuelson and Bazerman (1985) and Holt and Sherman (1994). See, Madarasz (2016) for details.}

A natural limitation of our study is that it focuses on tasks in one domain. Future research can then help understand the extent to which these findings and the structural link observed may hold in the other domains where the basic bias is observed. One advantage of our design, however, is that it allowed us to identify a functional relationship without the need to impose too many assumptions on the data generating process. It could also be fruitful to understand the extent to which a similar link may be detected for some other motivated social misperceptions, e.g., Mobius et al. (2015), when people reason about each other’s relative overconfidence.

References


6 Proofs

6.1 Appendix for Section 3

We now formally present the predictions based on the model of projection equilibrium stated in Madarasz (2016). Throughout the analysis, we assume that the strategically active players form their reported estimates at the time of solving the basic task. Below, $E$ refers to the expectations operator over $\omega$ with respect to the true distribution of actions and signals in the game. Since the reference agents solving the basic task have no relevant strategic interactions and are ex ante equivalent, we can introduce a representative reference agent and denote it by $\overline{A}$. With a slight abuse of notation, we can then represent the average success rate of the representative agent in state $\omega$ by

$$E \left[ E \left[ f^*(\omega, x) \mid P_{\overline{A}}(\omega) \right] \mid P_A(\omega) \right] = \pi$$

must then hold.

Let $E^{\rho_P}$ denote the expectations of a $\rho_P$-biased principal. The ex-ante expected estimate of $\pi$ by a $\rho_P$-biased principal is then:

$$E_{\omega} \left[ E^{\rho_P} \left[ E \left[ f^*(\omega, x) \mid P_{\overline{A}}(\omega) \right] \mid P_{P}(\omega) \right] \mid P_P(\omega) \right].$$

Using the definition of projection equilibrium, we obtain that the above equals to:

$$E_{\omega} \left[ \rho_P \max_{x \in D} E \left[ f(\omega, x) \mid P_P(\omega) \right] + (1 - \rho_P) E \left[ E \left[ f^*(\omega, x) \mid P_{\overline{A}}(\omega) \right] \mid P_P(\omega) \right] \right],$$

thus, given the law of iterated expectations, this equals $\rho_P (d + \pi) + (1 - \rho_P) \pi = \pi + \rho_P d$, as stated in Claim 1.

Let $E^{\rho_A}$ denote the expectations of a $\rho_A$-biased agent. The ex-ante expected first-order estimate of $\pi$ by the agent is analogous to the above and is then given by:

$$E_{\omega} \left[ \rho_A \max_{x \in D} E \left[ f(\omega, x) \mid P_A(\omega) \right] + (1 - \rho_A) E \left[ E \left[ f^*(\omega, x) \mid P_{\overline{A}}(\omega) \right] \mid P_A(\omega) \right] \right].$$

\[40\] See Section 5 Madarasz (2016). Projection equilibrium involves each player projecting both her information and her ignorance simultaneously. Note that formally our game is an $N$-player game given the presence of the reference agents, hence, we invoke the $N$-player definition we describing the predictions, See the Appendix of Madarasz (2016).
which then becomes $\rho_A \pi + (1 - \rho_A)\pi = \pi$, establishing the first part of Claim 2.

Now consider the agent’s ex-ante expected second-order estimate, her estimate of the principal’s estimate of $\pi$. This equals:

$$E_\omega \left[ E^{A'} \left[ E^{P'} \left[ f^*(\omega, x) \mid P_A(\omega) \right] \mid P_P(\omega) \right] \mid P_A(\omega) \right].$$

Using the definition of projection equilibrium, we can rewrite this as:

$$E_\omega \left[ \rho_A \max_{x \in D} E[f(\omega, x) \mid P_A(\omega)] + (1 - \rho_A)E[\rho_P E[f^*(\omega, x) \mid P_A(\omega)] \mid P_P(\omega)] \mid P_A(\omega) \right].$$

The first part of the above expression is based on the feature of projection equilibrium, when applied to a $N$-player setting, that the agent projects both on the principal and the reference agents. Specifically, he believes (i) that the projected versions of the reference agents and the principal have the same estimates of the solution to the basic task as the he does and (ii) that such versions occur in a perfectly correlated manner which they know. Rearranging the above expression, and given Claim 1, the above then equals $\rho_A \pi + (1 - \rho_A)(\pi + \rho_P d) = \pi + (1 - \rho_A)\rho_P d$, as stated in Claim 2.

Finally, to prove Claim 3, note the following. Since the agent attaches weight $\rho_A$ to the projected version of the principal the agent’s belief about the belief of the projected version of the principal is constant across the treatments – conditional on any given performance and signal realization of the agent on the basic task. Since the agent places weight $1 - \rho_A$ on the real version of the principal, and the agent’s belief about the real principal’s belief — again, conditional on any given own performance and signal realization by agent on the task — is strictly higher, in the sense of first-order stochastic dominance, in the informed than in the uninformed treatment, the agent is predicted to choose the sure payment more often in the former than in the latter treatment.

### 6.2 Appendix for Section 4.4

The specification for the principal follows from the above derivation of the mean estimate allowing now for an additive noise term to capture the difference between the conditional and the ex ante expected estimates. Consider, now, the agent’s
estimate conditional on the agent’s own success or failure and signal realization. If the agent figures out the solution, her success rate on the basic task is 1. If the agent does not figure it out, it is some number weakly higher than 0, — e.g., random clicking on a 7x7 grid does allow for a positive chance of success. For short, we denote the agent’s estimate of her own success probability by the variable [own success]. Under projection equilibrium, agent j’s stated second-order estimate in task t is then given by:

\[ b^{II}_{A,t} = \rho_{A_j}[\text{own success}] + (1 - \rho_{A_j}) E[b^{I}_{P,t} | \text{agent } j\text{'s info}] \]  

(5)

Note also, given the logic described in the Appendix for Section 3, that

\[ b^{I}_{A,t} = \rho_{A_j}[\text{own success}] + (1 - \rho_{A_j}) E[\pi_t | \text{agent } j\text{'s info}] \]  

(6)

By substituting Eq. (6) into Eq. (5), we get that \[ b^{II}_{A,t} = b^{I}_{A,t} - (1 - \rho_{A_j}) E[\pi_t | \text{agent } j\text{'s info}] + (1 - \rho_{A_j}) E[b^{I}_{P,t} | \text{agent } j\text{'s info}] \]. Given the linearity of the expectation operator, we can allow for an additive error term to describe the difference between the ex ante expected mean of a random variable and its realization. Hence, we can write this as:

\[ b^{II}_{A,t} = b^{I}_{A,t} - (1 - \rho_{A_j})\pi_t + (1 - \rho_{A_j})(\rho_P + (1 - \rho_P)\pi_t + \epsilon_{j,t}) = b^{I}_{A,t} + (1 - \rho_{A_j})\rho_P(1 - \pi_t) + \epsilon_{j,t}. \]
7 Supplementary analysis

7.1 Stated estimates of the principals

Figure 5: Distribution of average first-order estimates per principal in the informed and the uninformed treatment.

Figure 6: Average performance estimates of principals and actual success rate of the reference agents per task.
7.2 Investment decisions of the agents

Figure 7: Distribution of individual investment rates in the informed and the uninformed treatment.

Figure 8: Investment rates per task by session.
Table 2: Propensity to invest conditional on treatment and successful task completion.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Investment decision (Probit)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>−0.727***</td>
<td>−0.754***</td>
<td>−0.738***</td>
<td></td>
</tr>
<tr>
<td>(1-informed)</td>
<td>(0.205)</td>
<td>(0.211)</td>
<td>(0.221)</td>
<td></td>
</tr>
<tr>
<td>Success</td>
<td>0.429***</td>
<td>0.451***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1-task solved)</td>
<td>(0.096)</td>
<td>(0.126)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment × Success</td>
<td>−0.040</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.467***</td>
<td>0.299**</td>
<td>0.291**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.146)</td>
<td>(0.149)</td>
<td>(0.147)</td>
<td></td>
</tr>
</tbody>
</table>

N = 1410
R² = −916.436
F = 12.575

Note: Values in parentheses represent standard errors corrected for clusters on the individual level. Asterisks represent p-values: *p < 0.1, **p < 0.05, ***p < 0.01.

Table 3: Regressions of individual investment rates on treatment, gender, and risk attitude.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Individual investment rate (OLS)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>−0.281***</td>
<td>−0.279***</td>
<td>−0.255***</td>
<td>−0.259***</td>
<td>−0.254**</td>
<td></td>
</tr>
<tr>
<td>(1-informed)</td>
<td>(0.075)</td>
<td>(0.075)</td>
<td>(0.102)</td>
<td>(0.077)</td>
<td>(0.102)</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>−0.059</td>
<td>−0.032</td>
<td>−0.048</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1-female)</td>
<td>(0.075)</td>
<td>(0.108)</td>
<td>(0.109)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment × Gender</td>
<td>−0.053</td>
<td>−0.009</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.151)</td>
<td>(0.157)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coef. risk aversion</td>
<td></td>
<td>−0.026</td>
<td>−0.024</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(DOSE)</td>
<td></td>
<td>(0.022)</td>
<td>(0.023)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.673***</td>
<td>0.698***</td>
<td>0.687***</td>
<td>0.695***</td>
<td>0.715***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.053)</td>
<td>(0.063)</td>
<td>(0.071)</td>
<td>(0.057)</td>
<td>(0.076)</td>
<td></td>
</tr>
</tbody>
</table>

N = 94
R² = 0.134
F = 14.230

Note: Values in parentheses represent standard errors. Asterisks represent p-values: *p < 0.1, **p < 0.05, ***p < 0.01.
7.3 Stated estimates of the agents

Table 4: Individual differences between second-order estimates and first-order estimates conditional on treatment and successful task completion.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>((b^{II}<em>{A,t} - b^{I}</em>{A,t}))</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(OLS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>0.068***</td>
<td>0.068***</td>
<td>0.067***</td>
<td></td>
</tr>
<tr>
<td>(1-informed)</td>
<td>(0.019)</td>
<td>(0.019)</td>
<td>(0.024)</td>
<td></td>
</tr>
<tr>
<td>Success</td>
<td>-0.039***</td>
<td>-0.041**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1-task solved)</td>
<td>(0.011)</td>
<td>(0.017)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment×Success</td>
<td></td>
<td>0.003</td>
<td></td>
<td>(0.022)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.044***</td>
<td>0.058***</td>
<td>0.058***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.017)</td>
<td>(0.019)</td>
<td></td>
</tr>
<tr>
<td>(N)</td>
<td>470</td>
<td>470</td>
<td>470</td>
<td></td>
</tr>
<tr>
<td>(R^2)</td>
<td>0.090</td>
<td>0.117</td>
<td>0.117</td>
<td></td>
</tr>
<tr>
<td>(F)</td>
<td>12.828</td>
<td>17.767</td>
<td>13.296</td>
<td></td>
</tr>
</tbody>
</table>

Note: Values in parentheses represent standard errors corrected for clusters on the individual level. Asterisks represent \(p\)-values: ∗\(p < 0.1\), ∗∗\(p < 0.05\), ∗∗∗\(p < 0.01\).
Table 5: Mean individual differences in second-order estimates (estimate of the principal’s estimate) and first-order estimates $b_{1,t}^H$ (estimate of success rate) by treatment and further controls.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>(OLS)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment (1-informed)</td>
<td>0.068***</td>
<td>0.067***</td>
<td>0.080***</td>
<td>0.073***</td>
<td>0.090***</td>
<td></td>
</tr>
<tr>
<td>Gender (1-female)</td>
<td>0.013</td>
<td>0.047</td>
<td>0.045</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment × Gender</td>
<td>-0.062</td>
<td>-0.056</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coef. risk aversion (DOSE)</td>
<td>-0.006</td>
<td>-0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.044***</td>
<td>0.040**</td>
<td>0.030*</td>
<td>0.048***</td>
<td>0.034*</td>
<td></td>
</tr>
</tbody>
</table>

| N | 47 | 47 | 47 | 47 | 47 |
| R² | 0.220 | 0.228 | 0.270 | 0.236 | 0.278 |

Note: Values in parentheses represent standard errors. Asterisk represent p-values: *p < 0.1, **p < 0.05, ***p < 0.01.
7.4 Estimating projection equilibrium parameters

Figure 9: CDFs of principals’ (solid) and agents’ (dashed) projection bias $\rho$ in the informed treatment with alternative specification replacing the agents’ first-order estimates $b_{A,t}^I$ with the success rates $\pi_t$ in (4). Black lines represent empirical CDFs; gray lines represent best-fitting beta CDFs.
8 Instructions

8.1 Instructions for principals (translated from German)

Welcome to the experiment! The experiment that you will be participating in is part of a project funded by the German Research Foundation (DFG). It aims to analyze economic decision making.

You are not allowed to use any electronic devices or to communicate with other participants during the experiment. You may use only programs and features that are allocated to the experiment. Do not talk to any other participant. Please raise your hand if you have any questions. We will then approach you and answer your question in private. Please do not under any circumstances raise your voice while asking a question. Should the question be relevant for everyone, we will repeat it aloud and answer it. If you break these rules, we may have to exclude you from the experiment and from receiving payment.

You will receive a show-up fee of 5 Euros for your attendance. You can earn additional money through the experiment. The level of your earnings depends on your decisions, on the decisions of participants in former experiments, and on chance. The instructions are the same for everyone. A detailed plan of procedures and the conditions of payment will be explained below.

Tasks

You will face 20 tasks in the course of the experiment. For each task, you will be shown a short video. Each video consists of two images that are shown alternately. Your task is to spot the difference between the images.

The duration of each video is 14 seconds. After viewing each video, you will have 40 seconds to submit an answer. The interface will show a numbered grid, together with the image containing the difference. To solve the task, enter one of the numbers corresponding to a field containing the difference. If the difference is covered by more than one field, each field containing the difference will be evaluated as a correct answer.

You will receive [informed treatment: 0.30 Euros] [uninformed treatment: 0.50 Euros] for each task you solve correctly.
Estimates

Participants in previous experiments performed all of the tasks you will face in this experiment. Like you, these previous performers also had to spot the difference between the two images.

After performing each task, you will have the opportunity to estimate the percentage of previous performers that spotted the difference. Therefore, you will watch exactly the same videos as the previous performers watched. Like you, the previous performers also had 40 seconds to submit an answer and were paid according to their performance. [INFORMED TREATMENT ONLY: Before each video, you will receive a guide to the solution of the task. Please note that the previous performers did not receive solution guides.]

At the end of the experiment, the computer will randomly select two videos. Your estimates for these videos will be relevant to your payoff. For each of the payoff-relevant videos, the following holds: If your estimate is within the interval +/- 5 percentage points around the true percentage of previous performers that correctly identified the difference, you will receive 12 Euros.

Consider the following example. Assume that for one of the two payoff-relevant videos, 50% of the previous performers correctly identified the difference in this video. If you estimated that 53% of the previous performers spotted the difference, then you will receive 12 Euros. However, if you estimated that 57% of the previous performers spotted the difference, then you will receive 0 Euros.

You will start with three practice videos to familiarize yourself with the procedure. The practice rounds are not payoff-relevant. The 20 payoff-relevant videos will then follow.

Further procedure

After you submit your estimates for the 20 videos, you can earn money with additional decision-making problems. Further details will be given during the experiment.

At the end of the experiment, we will ask you to fill out a questionnaire. Even though your answers will not affect your payoff, we kindly ask you to answer the questions carefully.

After you have completed the questionnaire, you will be informed about your payoff for performing the tasks, your payoff for your estimates, your payoff for the additional decision-making problems, as well as your total payoff in this experiment. Please remain seated until the experimenter lets you know that you may collect your payment.

Do you have questions? If yes, please raise your hand. We will answer your questions in private.

Thank you for participating in this experiment!
Comprehension questions

1. How many (potentially payoff-relevant) videos will you evaluate in total?

2. How many of these videos will be selected for payment regarding your estimate of the previous performers?

3. What are the components of your total payoff?

4. Assume that the computer randomly selected the following exemplary videos for payment regarding your estimates:

   Video a): 62% of the previous performers found the solution.
   Video b): 35% of the previous performers found the solution.

   What is your payoff from your estimates if you estimated that 57% and 36% of the previous performers spotted the difference in video a and b, respectively?

   What is your payoff from your estimates if you estimated that 87% and 29% of the previous performers spotted the difference in video a and b, respectively?
8.2 Instructions for agents: Insurance only (translated from German)

Welcome to the experiment!

The experiment that you will be participating in is part of a project funded by the German Research Foundation (DFG). It aims to analyze economic decision making.

You are not allowed to use any electronic devices or to communicate with other participants during the experiment. You may use only programs and features that are allocated to the experiment. Do not talk to any other participant. Please raise your hand if you have any questions. We will then approach you and answer your question in private. Please do not under any circumstances raise your voice while asking a question. Should the question be relevant for everyone, we will repeat it aloud and answer it. If you break these rules, we may have to exclude you from the experiment and from receiving payment.

You will receive a show-up fee of 8 Euros for your attendance. You can earn additional money through the experiment. The level of your earnings depends on your decisions, on the decisions of participants in former experiments, and on chance. The instructions are the same for everyone. A detailed plan of procedures and the conditions of payment will be explained below.

Tasks

You will face 20 tasks in the course of the experiment. For each task, you will be shown a short video. Each video consists of two images that are shown alternately. Your task is to spot the difference between the images.

Figure 1 shows an example. Image A of the example shows a kayaker on the left side. In image B, the kayaker is not present.

The duration of each video is 14 seconds. After viewing each video, you will have 40 seconds to submit an answer. The interface will show a numbered grid, together with the image containing the difference (see Figure 2).

In each video, the difference between the images covers at least two fields. To solve the task, enter one of the grid numbers corresponding to a field containing the difference. That is, the number of any field containing the difference will be evaluated as a correct answer.

The experiment consists of 20 such tasks. The order of the tasks was randomly determined by the computer and is the same for all participants. You will receive 0.50 Euros for each task you solve correctly.
Previous participants

Performers

Participants in previous experiments performed all of the tasks you will face in this experiment. Like you, these previous performers had to spot the difference between the two images in each video. They were also paid according to their performance.

Evaluators

The degree of difficulty of the tasks that the previous participants performed (and that you will perform in this experiment) has been evaluated by further participants of previous experiments.

These evaluators were shown the tasks in the same way as the previous performers (and you), including the 40-second response time. After watching a video, the evaluators estimated the fraction of previous performers that solved this task correctly. The evaluators were paid according to the accuracy of their estimates.
[INFORMED TREATMENT ONLY: In contrast to the previous performers (and you), the evaluators received guides to the solution before each task. The evaluators were informed that the previous performers did not receive solution guides.]

At the beginning of the experiment, one of the evaluators will be randomly matched to you.

**Insurance decision**

During the experiment, you can earn additional money through insurance decisions. You will make one insurance decision after each task. At the end of the experiment, one of your insurance decisions will be randomly selected for payment.

Your endowment for each insurance decision is 10 Euros.

**Not buying insurance**

If you do not buy the insurance, your payoff depends on the following factors:

1. the number of previous performers (percent) who solved the current task + 10 (percent),

2. the number of previous performers (percent) who solved the current task in the evaluator’s estimation.

If (1) is at least as high as (2), you will keep your endowment of 10 Euros.

If (1) is smaller than (2), you will lose your endowment; that is, you will receive 0 Euros.

In other words, if you do not buy the insurance, your payoff will be determined as follows: You will keep your endowment of 10 Euros if the evaluator’s estimate of the previous participants’ performance is correct, an underestimation, or an overestimation by not more than 10 percentage points. Otherwise, you will lose your endowment.

**Buying insurance**

You have the opportunity to insure against this risk. The insurance costs 6 Euros. If you buy the insurance, you will receive your endowment minus the cost of insurance—that is, 4 Euros.

**Example 1**

Assume that 50% of the previous performers actually solved the task. The evaluator’s estimation is that 60% of the previous performers solved the task. Of course, you will not learn these values during the experiment.

If you did not buy the insurance, you will keep your endowment, because 50% + 10% ≥ 60%. The payoff from your insurance decision will, therefore, be 10 Euros.
If you bought the insurance, you will receive your endowment minus the cost of insurance—that is, 4 Euros.

**Example 2**

Assume again that 50% of the previous performers actually solved the task. In the evaluator’s estimation, 20% of the previous performers solved the task. As in the previous example, you will receive 10 Euros if you did not buy the insurance and 4 Euros if you bought the insurance.

**Example 3**

Assume again that 50% of the previous performers actually solved the task. In the evaluator’s estimation, 70% of the previous performers solved the task. In this example, you will receive 0 Euros if you did not buy the insurance and 4 Euros if you bought the insurance.

**Summary and further procedure**

The experiment consists of 20 rounds in total. At the beginning of each round, you will work on the task; that is, you will receive 0.50 Euros if you spot the difference between the two images in the video.

After each task, you can earn additional money through an insurance decision. At the end of the experiment, one of your insurance decisions will be randomly selected for payment. Because you do not know which insurance decision will be selected for payment, you should treat each insurance decision as if it were payoff-relevant.

After you make your insurance decision, you will receive feedback with a guide to the solution to the current task. [**Informed treatment:** The evaluators received the same guide before watching the task.] [**Uninformed treatment:** The evaluators did not receive solution guides.]

You will start with five practice rounds to familiarize yourself with the procedure. The practice rounds are not payoff-relevant. Afterwards, the 20 payoff-relevant rounds will follow.

After the 20 rounds, you can earn money through additional decision-making problems. Further details will be given during the experiment.

At the end of the experiment, we will ask you to fill out a questionnaire. Even though your answers will not affect your payoff, we kindly ask you to answer the questions carefully.

Do you have questions? If yes, please raise your hand. We will answer your questions in private.

Thank you for participating in this experiment!
Comprehension questions

1. When do you make your insurance decision—before or after you viewed the task?

2. Are the following statements true or false?

   (a) The evaluators received a guide to the solution to each task.
   (b) The evaluators did not receive guides to the solution to the tasks.
   (c) The evaluators received a guide to the solution before watching and evaluating the task.
   (d) At the end of the experiment, one of my insurance decisions will be randomly selected by the computer for payment.
   (e) My payment at the end of the experiment consists of the show-up fee (8 Euros), the payoff from performing the tasks (0.50 Euros per task solved), the payoff from one insurance decision, and the payoff from further decision-making problems.

3. Assume that the computer randomly selected videos for payment with the following characteristics. Provide the payoff from your insurance decision for each example.

   Example a)
   • 62% of the previous performers found the solution.
   • The evaluator randomly matched to you estimated that 72% of the previous performers spotted the difference.
   • You did not buy the insurance.

   Example b)
   • 62% of the previous performers found the solution.
   • The evaluator randomly matched to you estimated that 75% of the previous performers spotted the difference.
   • You did not buy the insurance.

   Example c)
   • 34% of the previous performers found the solution.
   • The evaluator randomly matched to you estimated that 41% of the previous performers spotted the difference.
   • You did not buy the insurance.

   Example d)
   • 34% of the previous performers found the solution.
   • The evaluator randomly matched to you estimated that 30% of the previous performers spotted the difference.
• You did not buy the insurance.

4. Consider again the examples in 3. For each example, assume that you did buy the insurance. What is your payoff from your insurance decisions in each of the examples?