Projective Thinking: Model, Evidence, and Applications

Kristóf Madarász*, David Danz,†and Stephanie W. Wang†

September 2024

Abstract

We offer a parsimonious model of egocentric thinking by postulating a link between the extent to which people project their beliefs onto others and to which they anticipate others' projecting onto them. We provide evidence for this link in higher-order beliefs and derive predictions of such projective thinking. In torts, judges' excessive liability judgments are conjoint with agents' under-appreciation thereof. In dissent, people infer antagonistic preferences and the more costly dissent is, the more they conclude that the norm is genuinely popular. In trade, informed traders bluff too little, uninformed ones are cursed, and the predictions match the experimental evidence.

Keywords: Cognition, Perspective Taking, Torts, Free Speech and Pluralistic Ignorance, Trade, Winner's Curse, Under-Bluffing.

*LSE, Houghton Street, London, UK. Email: k.p.madarasz@lse.ac.uk. This paper is preceded by two distinct working papers, Madarasz (2016) and Danz, Madarasz, and Wang (2018). We thank seminar audiences at Arizona, UC Berkeley, Bonn, Columbia, Harvard, Princeton, UC San Diego, UCLA, Utah, Wash U, Yale, CEU, Essex, UCL, LSE, Royal Holloway, Stockholm, Southampton, ITAM, Berlin Behavioral Seminar 2011, European Behavioral Economics Meeting Berlin 2013, ESSET Gerzensee 2013, SITE 2015, Peter Bossaerts, Colin Camerer, Jeff Ely, Marina Halac, Philippe Jehiel, Navin Kartik, Drazen Prelec, Matthew Rabin, Luis Rayo, Larry Samuelson, Joel Sobel, Balazs Szentes, Adam Szeidl, Tomasz Strzalecki, and Jörgen Weibull, and especially Roland Bénabou and four anonymous referees for excellent comments. Financial support from the Deutsche Forschungsgemeinschaft (DFG) through CRC 649 Economic Risk" is gratefully acknowledged. Wang completed this work as a Fellow at the Center for Advanced Study in the Behavioral Sciences.

†Department of Economics, University of Pittsburgh, 230 South Bouquet Street, Pittsburgh, PA 15260, USA, Email: danz@pitt.edu and swwang@pitt.edu

1 Introduction

Beliefs about the beliefs of others are central to strategic behavior. The evidence shows that people project their information and often their ignorance onto others when engaged in this 'theory-of-mind' reasoning, e.g. Piaget and Inhelder (1948), Fischhoff (1975), Camerer et al. (1989), Madarasz (2012).[1] In strategic settings, however, it is not simply one's belief about what the other knows, but what she thinks he think she knows, which is often key. Her model of him here entails her view of his model of herself, and projection may affect such views as well.

To illustrate, consider a seller privately informed about the quality of an asset, as in Akerlof (1970). If she exaggerates the chance that the buyer also knows this quality, she may be less tempted to bluff. At the same time, if she also anticipates that the buyer may project his ignorance onto her, thus underestimating the scope for selection, she may now find bluffing more attractive. Similarly, judges in tort cases exaggerate how often an agent had the same outcome information they do (Rachlinski, 1998), but the consequences of such *hindsight bias*, in terms of deterrence, depend on the extent to which the agent anticipates the judge's biased belief about his information. Analogously, forecasting a common macroeconomic variable, such as inflation, the typical manager underestimates how often other managers will disagree with her expectations (Coibion et al., 2021). The pricing implications of such projection-induced *false consensus* will, however, depend on whether she also thinks others exhibit false-consensus regarding their potentially different information leading to greater (more competitive) price dispersion.

The consequences of biased social cognition, a departure from unbiased expectations about other's beliefs, will rest on whether a person thinks others exhibit such cognition and whether she thinks they anticipate the way she thinks about them. However, without additional guidance, the number of different ways such higher-order beliefs could be specified is truly vast. For a predictive theory, a portable relationship linking these higher-order beliefs is sorely needed.

Our paper introduces a general but very parsimonious model of egocentric social cognition that we term *projective thinking*. At the core of our model is a simple idea that when a person forms beliefs about the beliefs of the other, she combines a

[1] Evidence on such egocentricity dates back to Piaget and Inhelder (1948). In a classic study, Wimmer and Perner (1983) demonstrate such informational projection in young children while Birch and Bloom (2007) show that Yale undergraduates make the exact same mistake in slightly more complex tasks. See, e.g., Epley et al. (2004) for a review and further relevant evidence.

coherent but fully egocentric fantasy with a probabilistic adjustment to an unbiased view. Specifically, each player assigns probability ρ to a fictional version of her opponent who has her information and is also *omniscient* about her beliefs. He knows what information she has, he knows what she thinks about what his information may be, and so on. Each player then assigns the remaining probability to an unbiased estimate of her opponent's beliefs. This idea implies a tight relationship between the extent to which people project their beliefs onto others and to which they anticipate others' projecting their differential beliefs onto them. We provide a direct test of this relationship and also show that the model can help provide a more unified account of various puzzling empirical findings.

To illustrate, suppose again that the value of an asset is Judith's secret. She then falsely believes that, with probability ρ, Paul knows this secret (basic projection). Paul also projects and believes that, with probability ρ, Judith knows that he does not know her secret. In turn, he believes that she, on average, assigns only probability $\rho 0 + (1 - \rho)\rho$ to him knowing her secret, thus underestimating her basic projection by ρ^2. This parsimonious relationship emerges naturally when the basic bias and its anticipation in others are understood not as separate phenomena, but as the result of the repeated application of the same partial egocentricity posited by our model.

Our model departs both from an idea of a 'bias blind spot,' whereby people, despite having actively biased beliefs, are *sophisticated* about the same tendency in the beliefs of others (Pronin et al. 2002, Pronin, 2008), and of a 'logic of introspection,' whereby recognizing others' biased belief tendency eliminates the same tendency in oneself, thus those who project must be *naive* about others' projections. Rather, we propose a model where people partially anticipate, but by the very extent of their own egocentricity, partially underestimate the egocentricity in others' thinking. After demonstrating the economic relevance of this tight relationship in the context of tort liability, we turn to a direct test in a design that mirrors this context.

Direct Evidence. In our experiment, we develop independent and unconstrained measures of the basic bias and its anticipation in others in a design where rejecting the precise predictions of the model, as well as distinguishing it from key alternatives, is potentially very easy. All participants performed the same sequence of "spot-the-difference" tasks. After performing each task, participants in the role of principals stated their belief about the success rate ϕ of other participants on the same task. Participants in the role of agents, after performing each task, stated their belief

about the success rate ϕ, like the principals, but agents also stated their belief about the principal's belief. There are two treatments and they differ only in one aspect. Principals received the solution to each task in the *informed treatment*, but not in the *uninformed treatment*.

In the uninformed treatment, our model makes the same average prediction as unbiased Bayesian beliefs would: all stated beliefs should be equal to the true success rate ϕ, *on average*. Indeed, this is what we find. In the informed treatment, unbiased Bayesian beliefs would lead to the same prediction. Instead, our model here predicts that principals should exaggerate the success rate while agents should anticipate but underestimate the magnitude of the principals' exaggeration, *on average*. Furthermore, if principals exaggerate the success rate by $\rho(1-\phi)$, agents should underestimate this exaggeration by exactly $\rho^2(1-\phi)$. Our results confirm these very tight predictions. Informed principals overestimate the success rate, in line with our model and previous findings. Turning to what is novel in our setup, we find that while agents' estimates of the success rate is correct on average, as our model predicts, they anticipate but underestimate informed principals' exaggeration. Remarkably, the proportion to which informed principals exaggerate the success rate is roughly the square root of the proportion to which agents underestimate this exaggeration. This predicted functional relationship holds not only in the aggregate, but also at the distributional (and task) level.

Dissent and Free Speech. In many settings, such as adherence to norms, or obedience to authority, outcomes depend less on people's actual preferences and more on their inference about the preferences of others, (Hume, 1741). In settings where expressing dissent in front of a loyalist may be costly, the evidence indicates that by interacting with them, people come to systematically misinfer the preferences of others, e.g., Katz et al. (1931), O'Gorman (1975), Miller and McFarland (1987), Prentice and Miller (1993), Cantoni et al. (2016), Bursztyn et al. (2020). In particular, here the literature commonly documents the emergence of pluralistic ignorance: "the phenomenon that occurs when people erroneously infer that they feel differently from their peers, even though they are behaving similarly." (Prentice, 2007) Such systematic belief distortions are inconsistent with rational inference which, on average, should be neutral or with the heuristic that everyone else is just like me.

To fix ideas, suppose that Judith and Paul are members of an organization and each either agrees with or disagrees with a prevailing norm. Consider an entry game

where each can decide whether to speak up (dissent) or act loyal. If a member agrees, she acts loyal. If she opposes, she gains when dissenting in front of someone who also opposes. Dissenting in front of a loyalist, however, leads to some loss. A similar situation arises when initiating friendship while facing uncertainty about whether the other wants friendship.

Projecting information implies that a person will underestimate how much uncertainty others face about her preference. When expressing one's preferences is a dominant strategy (e.g., free speech), such misperceptions do not affect equilibrium inference. Instead, we show that when speech is not free, this initial misperception leads to false antagonism whereby after interacting with another person, people who oppose the norm infer that others support it, while those who support it infer that others oppose it, on average. A further non-Bayesian comparative static arises: the less free speech is, the more people come to believe that others genuinely support the norm (don't want to friendship with them) – irrespective of the direction of their own preference. In the context of repeated encounters, these effects jointly imply that as speech becomes potentially more free, provided it does so sufficiently gradually, people, on average, become more convinced that the norm is genuinely popular. Instead, if speech becomes more free sufficiently quickly, people, on average, update positively and are surprised to learn about the lack of support for the norm.

Trade. Finally, we apply the model to a canonical setting of trade with asymmetric information (Akerlof, 1970). When the privately informed party makes the offer, she bluffs too little, but under projective thinking, the efficiency of trade, as well as her payoff, can exceed their respective Bayesian upper bounds. When the uninformed party has the bargaining power, he underestimates selection and consequently falls prey to the winner's or loser's curse, and he is always hurt relative to the unbiased case. We compare the model's fit of the data of Samuelson and Bazerman (1984) and Holt and Sherman (1994) with that of BNE and cursed equilibrium (Eyster and Rabin, 2005). We find that not only does our model provide a better explanation of the data when the informed party makes the offer, but that it also does so when the uninformed party makes the offer.

Related Literature Our model is related to prior game-theoretic approaches where players form wrong theories of each other's behavior as a function of the true distribution of information. Jehiel (2005) and Jehiel and Koessler (2008) study

analogy-based expectations equilibria (ABEE), while Eyster and Rabin (2005) study cursed equilibrium (CE). The phenomenon postulated by the current model differs from these both in terms of its order and also often in its direction. Under both ABEE and CE, each player has correct beliefs about the information of her opponent and only a wrong theory of the link between her opponent's actions and his information (the mistake is zeroth-order). In addition, a cursed player thinks that her opponent's strategy is coarser than it actually is. In contrast, under projective thinking, each player has wrong beliefs about the beliefs of her opponent (the mistake is first-order), but decision-theoretically rational theories regarding the link between other players' beliefs and actions (there is no zeroth-order mistake). Furthermore, a player with private information often thinks that her opponent's strategy space is finer than it actually is.

Crucially, ABEE and CE are pinned down by the common identifying assumption that, while players may have wrong action expectations of others state- by-state, such expectations must be *correct* on average. In contrast, under projective thinking, such action expectations are often wrong on average. Indeed, the key qualitative predictions in this paper are based on such wrong average action expectations which are also perceived to be more fine-tuned that it is in reality.

We incorporate our model of social (higher-order) belief formation into the standard equilibrium framework, and describe a special case of equilibrium with non-truthful or heterogeneous priors. Since such perceptions can depart from being truthful in arbitrary ways, our contribution here is in specifying a parsimonious but portable way on how higher-order perceptions about the distribution of information are systematically distorted as a function of the truth. At the same time, since our model of social belief formation is separate from a alternative model of play, it could also be incorporated into non-equilibrium models of play such as cognitive hierarchy or level-k, e.g., Crawford and Iriberri (2007).

2 Projective Thinking

This section develops the model. For ease of exposition, we restrict attention to two-player games and present the extension to N players in Appendix A. Consider a Bayesian game Γ. Let there be a finite set of states Ω and a strictly positive prior associated with it, $\pi \in \Delta\Omega$. Player i's information about the state ω is given by a standard information partition $P_i : \Omega \to 2^{\Omega}$; her finite action set is A_i; and her

payoff is $u_i(a, \omega) : A \times \Omega \to \mathbb{R}$, where $a \in A = \times_i A_i$ is an action profile. The game is then summarized by the tuple $\Gamma = \{\Omega, \pi, P_i, A_i, u_i\}$.

We first distinguish between the real and the fictional projected versions of each player i. The *real* version of i conditions her strategy on her true information, that is, she chooses a strategy from the set:

$$S_i = \{\sigma_i(\omega) \mid \sigma_i(\omega) : \Omega \to \Delta A_i \text{ measurable with respect to } P_i\}.$$

The fictional *projected* version of player i—who is real only in the imagination of player j—conditions her strategy on j's information, that is, she chooses a strategy from the set:

$$S_i^j = \{\sigma_i(\omega) \mid \sigma_i(\omega) : \Omega \to \Delta A_i \text{ measurable with respect to } P_j\}.$$

Below, we first state our formal definition, which implicitly describes players' belief hierarchy about the distribution of information and then make the logic and our assumptions explicit. Let the operator + denote the mixture of two lotteries, and let BR be the standard best-response operator. Its subscript refers to the set of strategies over which the indexed player maximizes; its argument is this player's belief about her opponent's strategy.

Definition 1 *A strategy profile $\sigma^\rho \in S_i \times S_j$ is a ρ projection equilibrium of Γ, where $\rho \in [0, 1)$, if there exists $\sigma^+ \in S_i^j \times S_j^i$ such that for each i and j,*

$$\sigma_i^\rho \in BR_{S_i}\{(1-\rho)\sigma_j^\rho + \rho\sigma_j^+\}, \tag{1}$$

and

$$\sigma_j^+ \in BR_{S_j^i}\{\sigma_i^\rho\}, \tag{2}$$

If $\rho = 0$, players form unbiased beliefs about each other's strategies and the predictions collapse to those of BNE in any given game Γ. If $\rho > 0$, each player i mistakenly assigns positive probability to her opponent best responding to her *true* strategy conditioning his best response on her true information in the game. She assigns the remaining probability to him playing the strategy he truly plays.

In our model, each real player i assigns probability ρ to the projected version of her opponent (who has the same information she does) and probability $1 - \rho$ to the real version of her opponent. Furthermore, each projected version assigns probability

one to her opponent being real. Implicit in this definition is a belief hierarchy about the distribution of information in the game which describes the players' higher-order misperceptions. Since it is this hierarchy which is at the core of our model, to make it explicit, it is useful to *decouple* basic projection, describing how one thinks about her opponent's information, from the way she thinks he thinks about her beliefs.

Suppose then that ρ only described the probability that each player wrongly assigned to her opponent having the same information she does (the basic bias). Even if it was clear that the real and projected versions of each player are the only two versions that enter into players' theories of each other, this assumption does not determine how a player thinks these versions of her opponent think about her. Consider then real Judith's possible perceptions of the probabilities that these two versions of Paul may assign to her two versions:

	projected Judith	real Judith
projected Paul's belief:	α	$1-\alpha$;
real Paul's belief:	β	$1-\beta$.

Table 1: Higher-order Beliefs

where *any* $\alpha, \beta \in [0,1]$ is possible. The following two assumptions, implicit in the definition above, fully pin down the hierarchy of misperceptions about the distribution of information.

1. All-encompassing Projection: $\boldsymbol{\alpha = 0}$. Real Judith thinks that projected Paul knows that she is real (Eq.2). He is thus omniscient about her; he knows what information Judith has (knows her first-order belief), what information she believes Paul may have (knows her second-order belief), and so on. He assigns probability one to real Judith's actual belief hierarchy. Since a player always knows what she herself believes, this property is in-line with the very idea of projective thinking.

2. Consistency with Feedback: $\boldsymbol{\beta = \rho}$. Real Judith correctly thinks that real Paul assigns probability ρ to her being real (Eq.1). She understands the extent to which he may have different information than she does as well as the extent to which he wrongly projects such information onto her. This means that any outcome that may occur in equilibrium remains within the support of each player's expectation about what may happen. Nothing that a player learns, be it about realized actions or payoffs, contradicts what this player believes may happen. Simply, each player keeps expecting things to happen that may never happen, or happen with a different

probability than expected.

The Biases of Others. A key aspect of our model, the source of its potential predictive power, is the parsimonious structure concerning higher-order beliefs. The extent to which a player has biased belief about her opponent's information fully determines the parameter ρ, and this same parameter then fully pins down the misperception along the entire belief hierarchy.[2]

A consequence is that each player anticipates but, precisely in proportion to her basic projection, *underestimates* her opponent's basic projection onto her. Formally, while real Paul assigns probability ρ to Judith's projected version (thus having his information), real Judith believes that, on average, Paul assigns only probability $\rho\alpha + (1-\rho)\beta = (1-\rho)\rho$ to her projected version. She thus underestimates this probability by ρ^2. For example, suppose that, in reality, Paul has a secret and let $\rho = 2/3$. He then falsely believes that Judith knows his secret with probability $2/3$. Instead, in reality, Judith believes that, on average, Paul falsely believes that she knows his secret with probability $2/9$.

This structure of underestimation holds along higher-order beliefs as well. One can iteratively derive the real players' higher-order beliefs about Judith being the projected version. Consider the sequence $\sum_{s=1}^{k}(-1)^{s-1}\rho^s$. Odd elements of this sequence describe real Paul's subsequent beliefs about the probability that Judith is the projected version.[3] Even elements describe real Judith's subsequent beliefs about the same. The gap between the subsequent beliefs of Paul and Judith is always ρ^k. Thus the same diminishing polynomial structure of underestimation continues to hold in higher-order beliefs and both the sub-sequence of odd and sub-sequence of even elements converge to $\rho/(1+\rho)$.

Alternative Hierarchies: Sophistication and Naivite Our main psychological assumption is all-encompassing projection, $\alpha = 0$. If $\alpha > 0$, projected Paul would no longer be omniscient about real Judith, but instead would be uncertain about Judith's information and may also think that she could know things he does not. When relaxing all-encompassing projection, a salient alternative is *full sophis-*

[2]To further underscore this parsimony, note that one could extend the model by adding more than one fictional version for each player to describe further mis-coordination in people's theories of each other.

[3]For instance, $k = 3$ refers to Paul's belief that, on average, Judith believes that he assigns probability $\rho - \rho^2 + \rho^3$ to Judith being the projected version ($14/27$ to her knowing his secret).

tication about the basic bias of the other: $\alpha = \beta = \rho$. Here Judith would think that the projected and the real version of Paul had the same false beliefs about her information. Hence, she would have an unbiased belief about Paul's belief of her information. Such sophistication corresponds to a bias blind spot (Pronin 2008) where each person correctly recognizes the biased belief tendency in others despite exhibiting the same tendency herself. More generally, if $\alpha < \rho$, each player would still underestimate the other's basic projection, but the extent of such underestimation would be smaller than in our model. Instead, if $\alpha > \rho$, each player would overestimate it. Our experimental study in Section 3 allows us to directly test for many of these alternatives.

The consistency with feedback assumption could also be relaxed. If $\beta \neq \rho$, then a player may no longer assign positive probability to her opponent's actual beliefs (type), hence her view of her opponent's beliefs may now be completely misspecified. A salient case here is *full naivete*: $\alpha = \beta = 0$. Here Judith is simply unaware of Paul projecting onto her. A more nuanced version of such naivete is the *logic of introspection*, whereby people who anticipate others' basic projection, must not project themselves, thus biased people must be naive about the projection of others, if $\rho > 0$, then $\alpha = \beta = 0$. Our experimental study in Section 3 is also able to test for both of these alternatives.

Heterogeneous Projection. The model immediately extends to heterogeneous projection. Here, a different ρ_i replaces ρ for each i in Eq.(1). The *same* all-encompassing projection, $\alpha_i = 0$, and consistency with feedback, $\beta_i = \rho_j$, properties continue to hold for each i and j. If $\rho_i = 0$, player i is sophisticated; she does not project and fully anticipates her opponent's basic projection. Otherwise, she estimates it to be $(1 - \rho_i)\rho_j$, underestimating it directly in proportion to her own basic projection. As $\rho_i \to 1$, i becomes fully naive; she believes that her opponent knows her beliefs perfectly and, hence, does not project onto her.

Projecting Information Only. Maintaining the structure of our model, all-encompassing projection, $\alpha = 0$, and consistency with feedback, $\beta = \rho$, we introduce an alternative specification where people project only their information, but *not* their ignorance. Projected Paul is still omniscient about Judith, the only difference is that the he now not only has the information of real Judith, but also that of real Paul. For example, in poker, Judith exaggerates the chance that Paul knows her hand, but

does not underestimate the chance that he knows his own hand. Formally, consider the following correspondence:

$$P^+(\omega) = \{\widehat{\omega} \in \Omega \mid \widehat{\omega} \in P_i(\omega) \cap P_j(\omega)\}$$

describing the coarsest common refinement of the two players' partitions. If an event is known at a state ω by either of the players, it is also known at that state under P^+. Conversely, any event known at a state under P^+ is also known at that state given the pooled information of the two players. We can then define the strategy set of the projected version of i to be

$$S_i^+ = \{\sigma_i(\omega) \mid \sigma_i(\omega) : \Omega \to \Delta A_i \text{ measurable with respect to } P^+\}.$$

The definition of ρ *information-projection equilibrium* (IPE) is then obtained simply by replacing S_i^j with S_i^+ for all i in Definition 1. All other parts of the model, as described below, will remain exactly the same.

Standard fixed-point results imply the existence of both solutions. Some simple but general observations are in order. First, in symmetric information games, projective thinking is inconsequential. Second, if i is better informed than j (one-sided private information) and i's best responses are the same whether she conditions on her own or only on the lesser-informed party's information, then PE and IPE are equivalent. Third, if a BNE is also an ex-post equilibrium, then it is information-projection proof, but may not be projection-proof. The reason for this last observation is that while in an IPE, each player perceives the interim strategy of her opponent to be at least as fine as it actually is, in a PE, a player may believe that her opponent's interim strategy must be coarser than it actually is.[4]

Corollary 1 *1.If $P_i = P_j$, the sets of PE and IPE are constant in ρ and equal to the set of BNE. 2. If P_i is a refinement of P_j and $BR_{S_i} = BR_{S_i^j}$ then, for any given ρ, the sets of PE and IPE are equivalent. 3. If σ is a BNE that is also an ex-post equilibrium, then, for any given $\rho > 0$, it is also an IPE, but may not be a PE.*

While we offer a very parsimonious approach, our paper does not pin down the contextual factors that determine in which games people project only their informa-

[4]We provide a simple example of a BNE that is also an ex post equilibrium, but is not a PE in Appendix B. See Madarasz (2016) for a nested version of these two specifications.

tion and in which also their basic ignorance. Instead, it encourages future empirical research to develop a better understanding of these factors. We conjecture that such factors may relate to the relation between the people who interact, the kind of private information present, and various aspects of salience. In settings where different pieces of private information relate to the same aspect of a situation or attribute of an object, such as the quality of an asset, the solution to a puzzle, or some physical property, or when interactions are less direct, the simultaneous projection of information and ignorance is likely to be the norm. In settings where different pieces of information relate to different things and each player has independent private information about her own attitude, or when interactions are more direct, people are likely to project only their information but not their ignorance.

2.1 Example: Negligence and Deterrence

> "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." (Sunstein, Jolls, and Thaler, 1998).

To illustrate the logic of the model, consider an example of liability for accidents or medical malpractice. This institution is often described as aiming to provide efficient incentives for agents to internalize negative externalities (Coase, 1960). The dominant liability rule is the negligence rule where a judge assigns liability based on her assessment of the acting agent's information, that is, based on her assessment of the foreseeability of the accident for him (Cooter 1991, Kaplow and Shavell, 2002).

Game. If the agent engages, he receives benefit $b > 0$ (saves on precaution), but may cause an accident; if he abstains (takes precaution), his payoff is 0. The state determines whether engagement leads to an accident. The agent observes only signal $s \in [0, 1]$ describing the probability of such an accident, then decides. Simultaneously, the principal (judge) learns the state and forms an expectation $a_p \in [0, 1]$ about the probability the agent should assign to such an accident.[5] If engagement causes an accident and the principal's expectation a_p exceeds the evidentiary threshold $z \in [0, 1]$, negligence is met and the agent incurs a normalized payoff loss of 1.

[5]Formally, for each $\omega \in \Omega$, $P_{Principal}(\omega) = \omega$, and the agent's coarser partition $P_{Agent}(\omega)$ implies, given the prior π, some probability $s[P_{Agent}(\omega)]$ of an accident. The analysis is the same whether b is public or is the agent's private information in reality.

Here, z is a policy variable and $z = 0$ corresponds to strict liability and $z = 1$ to no liability. Consider now the unique projection equilibrium (by Corollary 1, the unique information projection equilibrium is equivalent).

1. If $\rho = 0$ (BNE), the principal forms $a_p^0 = s$, the agent is liable iff $s > z$, and he abstains iff $s > max\{z, b\}$. We say that *deterrence is optimal*, given policy target z, iff the agent follows this strategy for any realization of b and s.

2. If $\rho > 0$, and there is an accident, the principal forms $a_p^\rho = (1-\rho)s+\rho$, and the agent is liable iff $a_p^\rho > z$. She overestimates the probability that the negligence threshold is met and her rulings are excessive, increasingly so in ρ.

3. If $\rho > 0$, the agent believes that, with probability ρ, the principal correctly believes that his information is s thus forms $a_p = s$, while with probability $1 - \rho$, she forms a_p^ρ. He abstains (weakly) more often than optimal, but his underestimation of excess liability also increases in ρ.[6]

Tort Reform. Consistent with the above, negligence rulings are considered to be excessive in practice due to judges projecting superior information (Rachlinski, 1998, Harley 2007). This led to a broader policy discussion on easing liability for defendants, and by how much, to counteract the effect of such excessive liability on optimal deterrence. Specifically, Sunstein et al. (1998) propose easing liability whereby "the overestimation of the likelihood that the negligence threshold is met could in theory be precisely offset by a change in the evidentiary threshold."

Their proposal directly translates to raising the threshold to $z^\rho = (1 - \rho)z + \rho$ in our setup. This reform requiring greater foreseeability indeed eliminates excessive liability rulings since $a_p^\rho > z^\rho$ is equivalently to $s > z$. Deterrence, however, depends on the agent's belief about the principal's belief of the agent's information. Our model predicts that this reform also achieves optimal deterrence when projection is smaller, but backfires precisely when it is larger. Let z^* be the de jure threshold under which deterrence is optimal given policy target z.

Proposition 1 *1. If $\rho \geq 1 - b/z$, deterrence is optimal without reform and the unique $z^* = z$. 2. If $\rho < 1 - b/z$, there is too much abstention without reform and the unique $z^* = z^\rho$.*

[6] If $b \geq s$, the agent engages. If $b < s$, (i) when $s \geq z$, he abstains; (ii) when $s \in [(z-\rho)/(1-\rho), z)$ he abstain iff $b < (1 - \rho)s$; (iii) otherwise he engages.

The optimal easing of liability for deterrence is non-monotonic in ρ. As the bias tends to full, the judge always holds the agent liable for an accident (de jure negligence becomes de facto strict liability), but the agent does not anticipate any excess liability. Deterrence remains optimal and easing liability induces too little precaution. More generally, as ρ increases, both excess liability and the agent's underestimation thereof are increasing. If ρ is smaller, the first effect dominates, the agent takes too much precaution without reform, and easing liability, in proportion to ρ, achieves optimal deterrence. If ρ is larger, the agent's underestimation dominates, and easing liability leads to too little precaution.

Two comparative static follow: as the agent's relative benefit b, or distribution thereof, increases, the more likely it is that easing liability only backfires. Similarly, the closer the desired policy is to strict liability, the lower z, the more likely that easing liability, which de facto corrects for excessive liability rulings , will backfire. Finally, the above directly rests on projection being all-encompassing. If it was characterized by full sophistication ($\alpha = \beta = \rho$), the proposed reform would be optimal, while under full naivete ($\alpha = \beta = 0$), optimal reform would always be null.

3 Evidence

Our model implies a tight relationship between basic projection and its misperception in others by linking both to the same process of egocentricity. We now introduce an experimental design that develops independent and unconstrained measures of these two to investigate their relationship and offer a sharp test of our model.

In our experiment, all participants perform a series of guessing tasks. After performing a given guessing task, principals estimate the average success rate of reference agents who performed this task previously. After performing a given guessing task, current agents also estimate the same success rate as well as the principals' estimates thereof. In the *Informed treatment*, principals, but not agents, receive the solution to each guessing task. In the *Uninformed treatment* no one receives the solution prior to performing the guessing task. The two treatments are identical in every other way.

3.1 Experimental Design

Guessing Task. All participants worked on the same series of 10 difference-detection tasks. In each task, subjects had to find the single difference between two otherwise

identical images presented via a video clip (see Instructions in the Online Appendix). Each participant saw each task in the exact same way as all other participants performing the task.[7]

Principals For each of the 10 tasks, the principal first performed the task. The principal knew that participants in previous sessions (reference agents) had been paid according to their performance on the same tasks.[8] After performing each task, the principal was asked to state their belief (first-order belief b_P^I) about the share of reference agents who spotted the difference in that task (success rate ϕ henceforth). After that, they moved onto the next round with a new and different change-detection task.

For the principals, the two treatments differed as follows. In the *Informed treatment*, they received the solution to each guessing task before solving it.[9] In the *Uninformed treatment*, instead, the principals were not given solutions to the guessing tasks. We ran one session with informed principals, and one with uninformed principals, with 24 participants in each (N=48). In both treatments principals were told that reference agents performing the guessing task only saw the video clip (did not receive the solution).[10]

Agents For each of the 10 tasks, the active agent first performed the task.[11] The agents were also told, in the same way as the principal, that participants in previous sessions (reference agents) had been paid according to their performance on the same tasks. In addition, they were told that the principal had estimated the average

[7]The detection task is a common visual stimulus (Rensink et al., 1997; Simons and Levin, 1997) and has already been studied in the context of documenting the basic bias (e.g., Loewenstein et al., 2006).

[8]We took the performance data of 144 participants from Danz (2020), in which participants performed the tasks in winner-take-all tournaments and faced the tasks in the exact same way as the participants in the current experiment.

[9]Specifically, during a countdown phase that announced the start of each task, the screen showed one of the two images with the difference highlighted with a red circle.

[10]At the end of the sessions, the principals received EUR 0.50 for each correct answer in the uninformed treatment and EUR 0.30 in the informed treatment. In addition, they were paid based on the accuracy of their stated estimates in two randomly chosen tasks: for each of these two tasks, they received EUR 12 if their guess was within 5 percentage points of the truth. We chose this elicitation mechanism because the strong incentives are simple and transparent which is important for behavioral incentive compatibility (Danz et al., 2022). The beliefs we elicited were coherent and sensible.

[11]There is no significant difference between the success rate of the reference agents and the active agents.

performance for the task (ϕ) and that the principal had also been paid based on the accuracy of her estimate. Each agent was randomly matched with a principal for the duration of the experiment. Then, the agent was asked to state their belief (i) about the share of reference agents who spotted the difference in that task (first-order belief b_A^I), and (ii) about the principal's estimate of this success rate (second-order belief b_A^{II}). [12]

For the agents, the two treatments differed *solely* with respect to the kind of principal they were told they are 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 were made fully aware of whether the principal had received the solution. After performing each task, agents in both treatments received the solution to the task in the exact same way. The only difference across the treatments was that agents matched to informed principals were told that this feedback corresponded to what the principal had seen for that task prior to performing it while agents matched with uninformed principals were told that the principal had not received this solution. In neither treatment did agents (or principals) receive any other feedback, e.g., about the principals' estimates or about the success rate. We ran one session with 24 agent matched to informed principals, and one with 23 agents matched to uninformed principals (N=47).

3.2 Predictions

We now present the predictions of projection equilibrium (the extension to N-players and the proofs are in the Appendix). Let d denote the ex-ante expected difference between the probabilities with which a randomly chosen principal and a randomly chosen agent can solve the task respectively. In the *Uninformed treatment*, since neither the agent nor the principal is given the solution before performing the task and roles are determined randomly, $d = 0$ by construction. In the *Informed treatment*, since the solution always helps, instead $d > 0$ ($d = 1 - \phi$). Below, we do not

[12] Agents received EUR 0.50 for each correct answer to the detection tasks; at the end of the experiment, one round was randomly selected for payment. We randomly selected one of the agent's stated estimates for payment—either her first- or second-order belief in that round. This payment structure addresses hedging concerns (Blanco et al., 2010). We used the same belief elicitation method as for principals. Each agent received EUR 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 belief and the principal's estimate of that success rate in the case of a second-order belief), and nothing otherwise.

assume that people need to make the same inference from watching the video per se. Instead, we allow players to obtain different private signal realizations about the solution to 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. We can then pin down the differences across the treatments in the ex-ante expected sense. Below, we first describe the *unique* predictions of projection equilibrium allowing for role-dependent degrees of projective thinking (ρ_P for the principal, ρ_A for the agent). In this Section we use elicited estimates and elicited beliefs interchangeably.[13]

Claim 1 *The principal's ex-ante expected mean estimate of ϕ is ${b_P}^I = \phi + \rho_P d$.*

In the *Uninformed treatment*, the principal's estimate of ϕ is unbiased. While her estimate *conditional* on her own success or failure on the guessing task shall be affected by projection,—e.g., inflated following success and deflated following failure— since the principal and the agent have the same probability of success, such distortions must cancel out on average. In the *Informed treatment*, the principal instead always knows the solution, and thus always thinks that the projected version of an agent must also succeed leading to an exaggeration of the agents' success rate.

Claim 2 *The agent's ex-ante expected first-order and second-order mean estimates are $b_A^I = \phi$ and $b_A^{II} = \phi + (1 - \rho_A)\rho_P d$.*

In the *Uninformed treatment* the agent's first- and second-order estimates are equal to ϕ for the same reason as the principal's estimate. In the *Informed treatment*, the agent's first-order estimate is again unbiased for the same reason. His second-order estimate is instead predicted to be systematically higher than his *own* first-order one, but systematically lower than the principal's. This is because the agent believes that with probability ρ_A the principal has a correct belief about the agent's information. The agent is predicted to anticipate, but underestimate, the principal's exaggeration on average.

The above are formulated allowing for role-specific degrees of projection ($\rho_A \neq \rho_P$). Since we independently infer (i) the extent of basic projection from the estimates

[13] In our design, active players always estimate the success rate of the strategically passive (reference) agents. This feature ensures that there cannot be another equilibrium where the active agent and the principal may coordinate on detection task performance to achieve higher earnings on the estimation tasks.

Ex-ante expected bias in	Uninformed	Informed
principal's first-order belief b_P^I	0	$\rho(1-\phi)$
agent's first-order belief b_A^I	0	0
agent's second-order belief b_A^{II}	0	$-\rho^2(1-\phi)$

Table 2: Predictions by Treatment

of the principals and (ii) the extent of its projection-based misperception in others from the estimates of the agents, to directly test the model, we need to impose their equivalence, i.e., $\rho_A = \rho_P = \rho$. The predictions, in terms of the ex-ante expected difference between an estimate and the corresponding estimand are then summarized in Table 2. For a summary of the predictions of leading alternative models and why they don't predict our qualitative hypothesis, see Appendix C.

3.3 Results

Figure 1 summarizes our main aggregate findings. It shows the agents' and principals' average elicited beliefs in each treatment and the true success rate. The beliefs elicited from the principals confirm Claim 1. In the uninformed treatment, there is no significant difference between the principals' average belief of 37.5% and the true success rate of 40%; ($p = 0.337$).[14] In the informed treatment, principals significantly overestimate the true success rate with an average belief of 58.7% ($p < 0.001$).[15] These results are quantitatively similar to the previous findings of Loewenstein et al. (2006).[16]

The agents' elicited beliefs confirm Claim 2. In the uninformed treatment, the agents' first-order belief 39.7% ($p = 0.917$) and the agents' second-order belief 44.1% ($p = 0.140$) are statistically indistinguishable from the true success rate.[17] In the informed treatment, the agents' first-order belief 39.9% is again the same as the

[14] We run a t-test of the average belief per principal against the average success rate (over all tasks). Figure D.1 in Appendix C shows the distribution of individual beliefs by informed and uninformed principals. Unless stated otherwise, throughout the analysis, p-values refer to two-sided t-tests that are based on average values per participant (paired t-tests whenever applicable).

[15] Accordingly, principals in the informed treatment had lower expected earnings (EUR 2.40) than principals in the uninformed treatment (EUR 3.65; one-sided t-test: $p = 0.034$).

[16] Moreover, while these statistics are possibly affected by skill-based selection, consistent with our logic, the average estimate of uninformed principals who found the solution is statistically indistinguishable from the average estimate of informed principals who received the solution upfront (60% and 58.7% average estimate, respectively; $p = 0.542$)

[17] In the uninformed treatment, agents' second-order beliefs are somewhat higher than the principals' first-order beliefs, but this difference is not significant ($p = 0.080$).

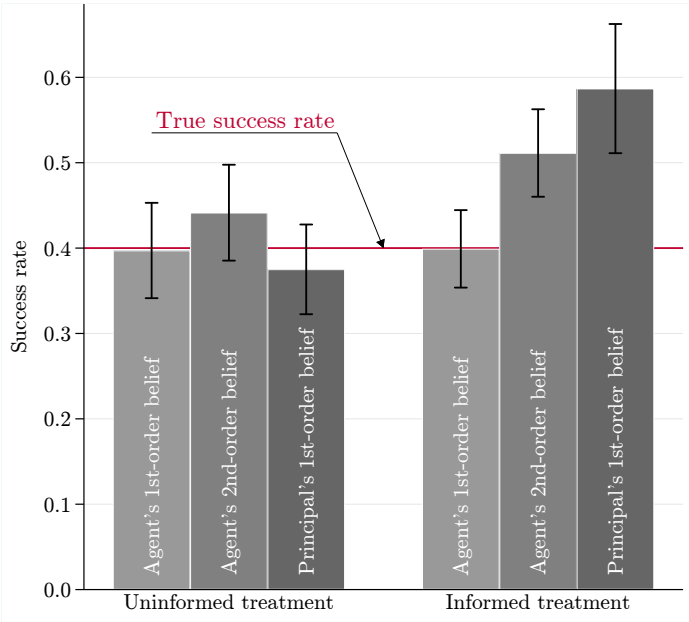

Figure 1: **Aggregate Beliefs** (capped spikes represent 95% confidence intervals).

true success rate ($p = 0.967$). However, their average second-order belief of 51.1% is significantly higher than the true success rate and also than their own first-order belief ($p < 0.001$). It is also significantly lower than the principals' first-order belief 58.7% (one-sided t-test: $p = 0.047$). That is, the data in the informed treatment confirms $b_P^I > b_A^{II} > b_A^I$. When comparing treatments, as predicted, we obtain the following result.

Result 1 *(i) b_A^I is the same across treatments ($p = 0.956$); (ii) b_A^{II} is significantly larger in the informed than in the uninformed treatment (one-sided t-test: $p = 0.031$); (iii) $b_A^{II} - b_A^I$ is larger in the informed than in the uninformed treatment ($p = 0.001$); (iv) $b_P^I - b_A^{II}$ is significantly larger in the informed than in the uninformed treatment ($p = 0.016$).*

3.4 Equivalence and Partial Projection

The above findings are consistent with the qualitative predictions of our model. We now turn to a more direct quantitative test of the hypothesis of $\rho_P = \rho_A$ implied by the model.

First, if we simply use the aggregate data from the Informed Treatment to directly solve for these two parameters using Claims 1 and 2, we obtain $\rho_P = 0.32$ and

$\rho_A = 0.39$ which are fairly close. To measure these using individual data and to actually *test* for equivalence, we use a structural Maximum-Likelihood model based on Claims 1 and 2 with random coefficients to capture individual heterogeneity in the degree of projection (see Appendix D3 for details). In the unrestricted model, the average degree of projection can freely differ between principals and agents, any $\rho_P \in [0,1]$ and, independently, any $\rho_A \in [0,1]$ is allowed. In the restricted model the average degree of projection is constrained to be the *same* across roles, $\rho_P = \rho_A = \rho \in [0,1]$.

In the unrestricted model the average $\hat{\rho}_P = 0.340$ and the average $\hat{\rho}_A = 0.354$ are *not* statistically different ($p = 0.869$). See Table D.2 in the Appendix. Furthermore, in the restricted model we find that the average $\hat{\rho} = 0.337$ is very close to these parameter estimates. The value of the maximum likelihood function of the restricted model is also very close to that of the unrestricted model such that both the Akaike and Bayesian information criteria select the former, parsimonious model.

Individual Estimates. The above estimates are consistent with the model, but do not establish that it has descriptive accuracy at the individual level. While full sophistication, $\rho_A = 0$, or full naivete, $\rho_A \to 1$, about the basic projection of others clearly does not hold in the overall data, these may still be common at the individual level. Furthermore, the logic of introspection, as described in Section 2, may still hold and the fraction of agents who anticipate the principals' exaggeration may match the fraction of principals who do not exaggerate. Finally, even if partial anticipation were common at the individual level, the distributions of ρ_P and ρ_A may be very different.

We now compare the distribution of individual principals' degree of projection as inferred from (Claim 1) and the distribution of individual agents' degree of projection as inferred from (Claim 2). To this end, we first obtain individual estimates of ρ_i for each principal and ρ_j for each agent from the informed treatment using simple linear regressions. We do *not* impose any restrictions on the *size* or *sign* of these parameters. For each principal i in the informed treatment, we estimate ρ_{P_i} from Claim 1 via:

$$b^I_{P_i l} = \phi_l + \rho_{P_i}(1 - \phi_l) + \epsilon_{il}, \tag{3}$$

where ϵ_{il} denotes a mean-zero error term with variance σ_i^2 and the index $l \in \{1, 2, ...10\}$ refers to a given guessing task in the sequence of tasks. Analogously,

for each agent j in the informed treatment we estimate ρ_{A_j} from Claim 2 via:

$$b^{II}_{A_j l} = b^{I}_{A_j l} + (1 - \rho_{A_j})\rho_P(1 - \phi_l) + \epsilon_{jl}, \tag{4}$$

where ϵ_{jl} denotes a mean-zero error term with variance σ^2_j. We estimate the parameters in (3) and (4) by OLS, where ρ_P in (4) is the average estimate of ρ_{P_i} from the regressions in (3).

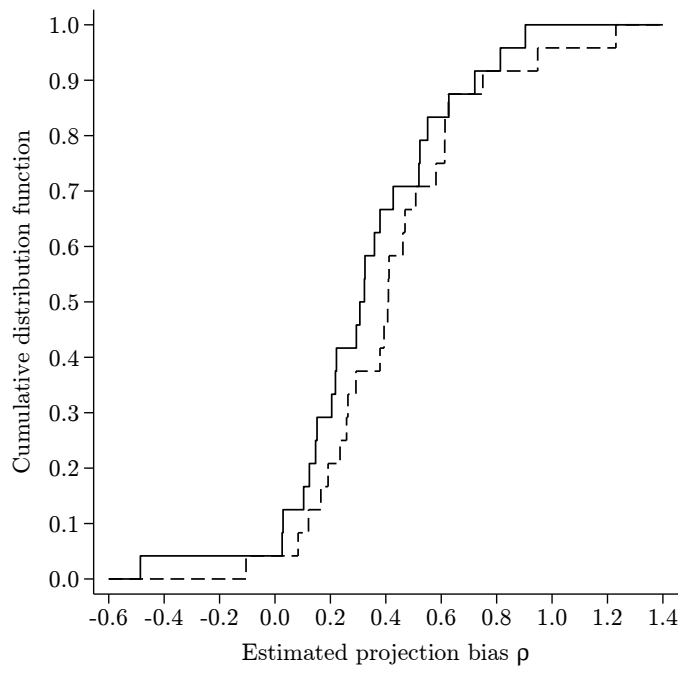

Figure 2: **Individual-level estimates.** Empirical cumulative distribution functions of principals' (solid) and agents' (dashed) ρ in the informed treatment.

Figure 2 shows the empirical cdfs of the individual degrees of projection for the principals and the agents respectively. The two distributions are not significantly different from each other ($p = 0.580$ from bootstrapped Kolmogorov-Smirnov test; see Davison and Hinkley, 1997).[18] Furthermore, partial projection is the most preva-

[18]We applied a two-level bootstrap procedure (with 10,000 iterations) to account for variation in the individual ρ estimates. At the first level, we resampled participant IDs (stratified by participant role) to account for variation in the average of the principals' estimates as an estimate for ρ_P in equation (4). At the second level, we resampled individual observations for each participant ID from the first stage, to account for variation in the individual estimates. The result is qualitatively the same for the test without bootstrapped test statistic ($p = 0.441$).

lent case for both principals and agents. For 95.8% of the principals and 91.7% of the agents, the value of the estimated parameter is larger than zero and smaller than one. Although our power on the individual level is limited, for the majority of participants the estimated parameter is both significantly larger than zero and significantly smaller than one (50% of the principals and 54.2% of the agents). The data is thus inconsistent with the hypotheses of full naivete or full sophistication at the individual level. Given the individual estimates, we can also test the alternative hypothesis of the logic of introspection that the fraction of agents who anticipate the principal's projection at least to some extent ($\rho_{A_j} < 1$) matches the share of unbiased principals (ρ_{P_i} not significantly different from 0). The data rejects this alternative as well ($p = 0.030$; Fisher's exact test).

Task-based estimates. Finally, to explore whether our model has predictive power within our data, we examine how well the ρ estimates, inferred from *all* tasks ($l = 1, ...10$), predict the wedge between the agents' own first- and second-order beliefs in any given task. Formally, we first take each agent's ρ_{A_j}, *estimated* over all tasks, and calculate that agent's *predicted* second-order belief given Eq.(4) for each task l. We then average this prediction across all agents for this given task l.

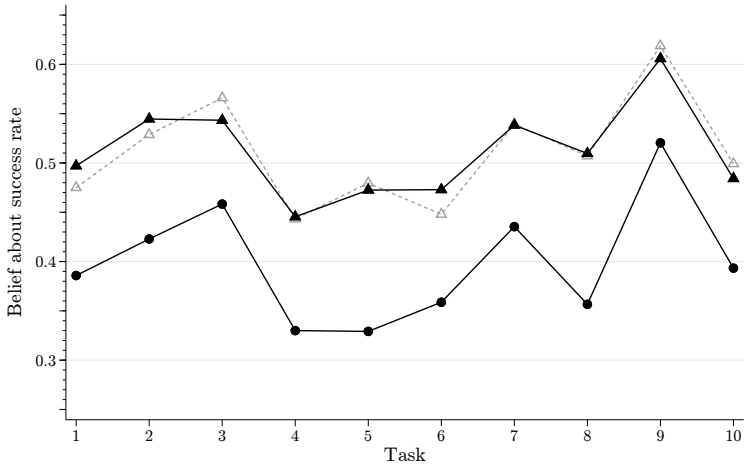

Figure 3: **Task-level predictions.** The solid line with circle [triangle] markers shows the agents' average *elicited* first-order [second-order] belief for each task. The dashed line shows the agents' average *predicted* second-order belief for each task.

Figure 3 shows the agents' average *elicited* versus *predicted* second-order beliefs per task (as well as the elicited average first-order belief). The correlation between the predicted and actual second-order beliefs across tasks amounts to 0.88. The

model has a remarkable ability to predict the key elicited beliefs across the sampled tasks as well.[19]

In sum, the data not simply reveals that people anticipate but underestimate the basic projection of others, but provides remarkable support for the specific logic and precise predictions of our model in a design where the predictions comprise a very small subset of the set of possible observations.[20]

4 Costly Dissent and False Antagonism

Adherence to norms, obedience to authority, or the formation of friendships often depend on people's inferences about the preferences of others. When expressing preferences truthfully is a dominant strategy, such inferences shall quickly convergence to the truth. In settings where expressing preferences in a certain direction carries a potential downside, such as being punished, the data suggests that people instead often systematically misperceive the preferences of others, e.g., Katz et al. (1931), Miller and McFarland (1987), Shelton and Richeson (2005), Bursztyn et al. (2020). In particular, the literature documented the emergence of pluralistic ignorance: "the phenomenon that occurs when people erroneously infer that they feel differently from their peers, even though they are behaving similarly." (Prentice, 2007).

We now apply the model to a class of entry games where behavior and inference about others critically depend on what a player believes others believe about her privately-known preference. We describe how projective thinking leads to the emergence of these perceptions and provide key comparative statics on how such misperceptions about the genuine popularity of a norm interact with the cost of dissent (lack of free speech) or the perceived hostility of others with the cost of a wrong move.

4.1 Setup

Consider a class of entry games. Upon each player i privately learning her own valuation $\theta_i \in \mathbb{R}$, where each θ_i is an i.i.d. draw from an uniform density over $[\underline{\theta}, \overline{\theta}]$, with $\underline{\theta} < 0 < \overline{\theta}$, each decides whether to enter (dissent) or to stay out (act loyal). If both players enter, each receives his or her own valuation. If both stay out, each

[19] The share of unexplained variance in the empirical estimates is only 26.8%.

[20] For a measure and discussion of predictive success comparing these sets see, e.g., Selten (1991), Beatty and Crawford (2011), or Fudenberg et al. (2024).

receives the outside or status quo option normalized to zero. The rest of the game is described as follows:

	In	Out
In	θ_i , θ_{-i}	$g(\theta_i, \theta_{-i})$, $f(\theta_{-i})$
Out	$f(\theta_i)$, $g(\theta_{-i}, \theta_i)$	0 , 0

The key distinction is between negative and positive values. A negative-value player, $\theta_i < 0$, strictly prefers to stay out.[21] A positive-value player strictly prefers to enter (stay out) if her opponent is positive (negative). Specifically, if $\theta_i \geq 0$,

- Sorting: $f(0) = 0$ and $f' < 1$,
- Risk: $g(\theta_i, \theta_{-i}) \geq 0$ if and only if $\theta_{-i} \geq 0$.

For ease of intuition, we first consider a leading example where we further assume that for a positive-value player one-sided entry leads to a constant loss if her opponent is negative, and is an almost perfect substitute of mutual entry if her opponent is positive. Formally, given $\theta_i > 0$, if $\theta_{-i} < 0$, then $g(\theta_i, \theta_{-i}) = -c$, and if $\theta_{-i} > 0$, then $f(\theta_i) = \gamma\theta_i$ and $g(\theta_i, \theta_{-i}) = \gamma\theta_i$, with $0 < \gamma < 1$, and $\gamma \to 1$ corresponding to the case of perfect substitutes. We return to the more general game and show that our predictions hold there as well (Proposition 4). We describe two simple interpretations.

♢ **At the Bar.** Judith and Paul can each make a move. If both do, a match is formed. If only one makes a move, and the other values a match positively, a match is formed with a slight delay discounting payoffs by γ; but if the other values a match negatively, the proposer incurs a loss of c such as the emotional pain or cost associated with a rejection.

♠ **Costly Dissent.** A person either opposes (positive value) or agrees with (negative value) a prevailing norm such as Stalin's leadership of the party. When two people meet, each can dissent or act loyal. If Judith agrees with the norm, she acts loyal. If she opposes it, she gains when dissenting in front of Paul if he also opposes the norm, as they may form a coalition or experience a sense of liberation, but loses if Paul supports the norm because he may punish or report explicit dissent

[21] Formally, if $\theta_i < 0$, then $f(\theta_i) > \theta_i$ and $g(\theta_i, \theta_{-i}) < 0$.

and c corresponds to the punishment for reported dissent. For further interpretations in terms of investment in partnerships, see Madarasz (2016). [22]

In these contexts with two-sided private information, each piece pertains to a person's own personal attitude. We then consider the case where people project their information, but not their ignorance, that is, Paul understands that Judith knows her own attitude towards the norm or him. Hence we employ information projection equilibrium from Section 2.

4.2 Equilibrium

Below, π_0 denotes i's prior on θ_{-i}, π_1^ρ her posterior on the same, and π_1 the true posterior distribution given the players' real behavior in the game.

Proposition 2 *Equilibrium is always in cutoff strategies. Consider the ρ-IPE correspondence as a function of γ. In the unique limit point of this correspondence, as $\gamma \to 1$, player i enters iff $\theta_i \geq \theta^{*,\rho}$ where:*

$$\theta^{*,\rho} = \sqrt{\frac{c\,|\underline{\theta}|}{1-\rho}}.$$

Furthermore, in this limit, if $\rho > 0$,

I. $\pi_1^\rho[\theta_{-i}|\theta_i, a] <_{fosd} \pi_1[\theta_{-i} \mid \theta_i, a]$ *given any* $a \in A$ *and* $\theta_i > 0$;[23]

II. $E[\pi_1^\rho|\theta_i] <_{fosd} \pi_0$ *if* $\theta_i > 0$, *and* $E[\pi_1^\rho|\theta_i] \geq_{fosd} \pi_0$ *if* $\theta_i \leq 0$.

Shyness. A positive Judith's willingness to enter increases in her confidence that Paul is positive, but decreases in her confidence that Paul knows that she is positive. By projecting, she exaggerates a positive Paul's incentive to enter and finds it relatively more important to stay out and avoid a loss. In a symmetric situation, a positive Paul reasons similarly. Projection increases each positive player's expectation that the other party shall enter and, in turn, increases the cutoffs for entry (silence). A positive player then perceives equilibrium to be 'more separating' than it actually is. At the same time, a negative player, who always underestimates the probability with which her opponent enters, perceives equilibrium to be 'more

[22] Note that while the setup describes bilateral interactions, it applies to such interactions taking place pair-wise between members of a group. In friendship, preferences may depend on the pairing. In dissent, each player's preference is constant across pairings.

[23] If $a = \{a_i = in; a_j = out\}$, this relation is weak. In all other cases it is strict.

pooling' than it actually is. Below we describe the non-Bayesian consequences of these facts that imply a systematic violation of the martingale property.

Conditional False Antagonism I. A positive player underestimates her opponent's valuation conditional on *any* outcome. If she opposes the norm, seeing Paul dissent, she is too convinced he dissented knowing that she would not report him, and seeing him stay silent, she is too convinced that he is genuinely loyal.

Average False Antagonism II. Updating is also antagonistic *on average*. Each player comes to exaggerate that her opponent has the opposite preference than she does. If Judith opposes the norm, she exaggerates how often Paul should dissent and over-infers genuine loyalty from his silence. If she supports the norm, she exaggerates how often he should stay silent, and underinfers loyalty from his silence. A form of paranoia arises whereby each person exaggerates the probability that the preferences of others point in the opposite direction as her own.

A further non-Bayesian comparative static follows. Let $E[\overline{\pi}_1^{\rho}]$ be i's ex-ante expected posterior probability estimate that j is a positive-value player, and $\overline{\pi}_0$ denote the corresponding prior probability. Analogously, let $E[\overline{\pi}_1^{\rho,+}]$ ($E[\overline{\pi}_1^{\rho,-}]$) denote the same, conditional on i being positive (negative), e.g., Stalin's opponents' (supporters') average view of Stalin's unpopularity.

Proposition 3 *If $\rho > 0$, then $E[\overline{\pi}_1^{\rho}]$, $E[\overline{\pi}_1^{\rho,+}]$, and $E[\overline{\pi}_1^{\rho,-}]$ are all decreasing in c.*

In the unbiased case, the martingale property implies that each player's ex-ante expected posterior is the same as her prior, i.e., $E[\overline{\pi}_1^{0}] = \overline{\pi}_0$ irrespective of c and θ_i. Under projective thinking, the less free speech is, the more both those who support and those who oppose Stalin conclude that he is genuinely popular. Similarly, the more costly a wrong move is, the more people conclude that others do not want friendship with them – irrespective of whether they themselves want friendship. An increase in c increases (decreases) the wedge between a positive (negative) player's perception of the probability that her opponent shall enter and the true probability thereof and her inference from his silence.

4.3 Dynamics

To further describe implications of the above comparative static, consider now a dynamic repetition of the game with changing cost of dissent, specifically a weakly

decreasing sequence $\underline{c} = \{c_t\}_{t=1}^{T}$ with $c_T > 0$.[24] For simplicity, we focus on myopic repetitions: in each round t, players care only about the payoff of that round, but recall the history of past interactions. At the end of each round, each player updates her belief both about the opponent's preference and his information based on the realized action profile (and possibly also on her own payoff) – and such updating per se is commonly known. Here, the natural psychological assumption is that players project to some extent at the beginning of each new encounter: at the beginning of each round t, player i believes that with some probability ρ her opponent becomes his projected version and learns her valuation if he has not yet - and such projection is common knowledge. Let $\Pr^{\rho}_{\underline{c}}$ denote the true probability that, conditional on both players being positive in a pair, at least one enters by the end of the sequence. We use the same notation as before now indexed by t.

Corollary 2 *Fix any $\underline{c}$ and $\rho > 0$. $\Pr^{\rho}_{\underline{c}}$ is decreasing in ρ and $\underline{c}$. Furthermore, for each t, $E[\overline{\pi}_t^{\rho,+}] < \overline{\pi}_0 \leq E[\overline{\pi}_t^{\rho,-}]$ and there exist $0 < \alpha^{+}_{t,\underline{c}} < \alpha^{-}_{t,\underline{c}} < 1$ such that*

a. if $c_{t+1}/c_t > \alpha^{+}_{t,\underline{c}}$, then $E[\overline{\pi}^{\rho,+}_{t+1} - \overline{\pi}^{\rho,+}_{t}] < 0$; else, $E[\overline{\pi}^{\rho,+}_{t+1} - \overline{\pi}^{\rho,+}_{t}] > 0$;

b. if $c_{t+1}/c_t > \alpha^{-}_{t,\underline{c}}$, then $E[\overline{\pi}^{\rho,-}_{t+1} - \overline{\pi}^{\rho,-}_{t}] = 0$; else, $E[\overline{\pi}^{\rho,-}_{t+1} - \overline{\pi}^{\rho,-}_{t}] > 0$.

Entry is decreasing in projective thinking dynamically as well. Consider now the comparative statics with respect to $\underline{c}$. Opponents of the norm exaggerate how often others should dissent, but, as a dynamic consequence, come to underestimate the fraction of others who oppose the norm. In turn, if the drop in c_t is smaller than a threshold, there is still too little dissent relative to their expectations, and their antagonistic inference grows on average. If instead the drop is larger, they are too surprised by how common dissent is and their false sense of antagonism shrinks on average, but is never fully eliminated. Since a loyalist always finds dissent in front of her too surprising, her antagonistic inference grows (at least weakly) over time. In turn, if the drop in c_t from one round to the next is small, average inference is *negative* over time; if it is large, however, average inference instead becomes *positive* over time. This implies the next result.

Corollary 3 *Given any $\rho > 0$, suppose that $c_t \geq \frac{\overline{\theta}^2}{|\underline{\theta}|}(1-\rho)^t$ for all t. No one ever enters, but $E[\overline{\pi}_t^{\rho,+}]$ and $E[\overline{\pi}_t^{\rho}]$ are strictly decreasing in t. Furthermore,*

[24] Assuming that $\underline{c}$ is weakly decreasing is without loss of generality as any sequence where $c_{t+1} > c_t$ for some t, will be strategically equivalent to an identical sequence with $c_{t+1} = c_t$.

1. *each positive player develops false uniqueness:* $\lim_{t\to\infty} E[\overline{\pi}_t^{\rho,+}] = 0$;

2. *the majority concludes that the majority is negative:* $\lim_{t\to\infty} E[\overline{\pi}_t^{\rho}] \leq \frac{1}{4}$.

In the above environments, equilibrium remains pooling exactly because positive types perceive it to be separating. Since, as long as none enters, negative types maintain unbiased estimates, it is then exactly when none supports the norm that everyone concludes that everyone else supports it (that nobody she wants to be friends with wants to become friends with her). Furthermore, as ensured by the fact that $E[\overline{\pi}_t^{\rho}] \leq \frac{1}{4}$, the (silent) majority of the group on average also always concludes that the majority supports the norm independent of the margin by which this is true or false in reality.

4.4 Relation to the Evidence

The predictions are consistent with the phenomenon of pluralistic ignorance whereby people adopt opposing attribution of identical behavior to self and the other. We now discuss a more detailed link between the evidence and our prediction.

False Antagonism and Intergroup Interactions. Consider a group of people with i.i.d. preferences towards friendship with each other randomly divided into *two* sub-groups. Suppose that in-group members have a greater chance of knowing each other's preferences, but face uncertainty about out-group members' preferences. Proposition 2 (Proposition 4 for the general case) implies that, on average, people will mistakenly conclude that out-group members they want to be friends with are distinctly less likely to want to be friends with them than in-group members who they want to be friends with. Consistent with this prediction, Shelton and Richeson (2005) find that students at Princeton and U Mass desired having more interracial friendships, but significantly underestimated out-group members' interest in interracial friendship. Consistently, they attributed their own lack of initiation to the fear of being rejected, but that of the out-group members to their genuine lack of interest.

Corollary 2 also predicts that while the initial interactions shall increase such differential false antagonism vis-a-vis outgroup members, whether further interactions *increase or decrease* this depends on how fast c drops. If the drop is large, precisely because of this initial accumulation of false antagonism, intergroup interactions will have predictable positive effects on such perceptions. Otherwise they

will exacerbate them. This may also help explain a puzzling discrepancy between the negative findings of the 'intergroup interaction' studies, documenting increased intergroup avoidance following single or shorter interactions with strangers in more rigid settings, and the positive findings of the 'intergroup contact' studies, which typically involve longer interactions in closer relationships between people who may already know each other, documenting the opposite, e.g., MacInnis and Page-Gould (2015) for a summary.[25]

False Antagonism and Norms. In an illustrative study, measuring attitudes towards affirmative action, Van Boven (2000) finds that only a minority of Cornell undergraduates anonymously surveyed supported affirmative action, while the most common attitude was to oppose it (27% versus 46%). At the same time, students on average believed that support for affirmative action was the more common attitude among their peers (47% versus 30%). In line with the false antagonism predicted, those who *opposed* affirmative action had a significantly *higher* estimate of the support for affirmative action among their peers (58%) than did those who supported affirmative action (40%). Thus, the evidence supports the *joint* prediction of false antagonism and the reversal between the majority's perception of the majority's preference when the latter opposes the status quo norm (Proposition 2; Proposition 4 for the general case).

In the context of political attitudes in Hong Kong, Cantoni et al. (2016) present similar findings. Consistent with Proposition 2 (Proposition 4 for the general case) and the above comparative static with respect to uncertainty about the preferences of others, they find that people's perceptions of others' attitude towards authoritarianism are significantly *negatively correlated* with their own attitudes. However, also consistent with our model, their own attitudes and the perception of their close friends' attitudes, with whom they presumably interact under effective free speech ($c = 0$), are *uncorrelated*.

Disciplinary Organizations and Shy Revolutions. The logic of our result (Corollary 3) also implies that if the leadership of an organization wants to ensure loyalty to a rule by punishing dissent, given projective thinking, it can do so with less and less formal enforcement as self-censorship outlives actual censorship as members of the organization become increasingly apathetic (in the unbiased case, the intensity of disciplinary sanctions has to remain constant over time). The same logic also describes how, even when punitive resources are more scarce, c_t is low for any t, in

[25]For a review of the evidence on intergroup interactions, see, e.g., Pettigrow and Tropp (2006).

the biased case, the introduction of less and less popular (more extreme) versions of a norm or rule over time can be successful, provided the norm becomes more extreme sufficiently gradually.[26] At the same time, suppose that at some time t there is a sufficiently large relative drop in c_t, e.g., speech becomes free or a secret ballot on upholding the status quo norm or rule is held. Dissent then comes as a great surprise to all those who dissent and their perception of the popular support for the norm or rule, growing until then, erodes. While there are many historical examples, e.g., Kuran (1995) on the widespread surprise among opponents of the regime in East Germany in 1989, such a surprise was also on display more recently during the UK's 2016 referendum on leaving the EU and in the election of Donald Trump in 2016.[27]

4.5 Entry Games

Let's return to the general class of games introduced. To characterize the implications, a distinction between complement and substitute entries is needed. Entries are substitutes (complements) if, *conditional* on both players having positive valuations, Judith's gain from making a move is higher (lower) when Paul does not make a move compared to the case where he does.

Definition 2 *Entries are substitutes (complements) if* $\theta_i - f(\theta_i) < (>)\ g(\theta_i, \theta_{-i})$ *whenever* $\min\{\theta_i, \theta_{-i}\} \geq 0$.

In the applications discussed, when both players are positive, the gain from one-sided entry versus the status quo is at least as large as the gain from simultaneous versus one-sided entry. Here entry choices are substitutes. There are of course settings, such as when time pressure is significant, where the relationship may be

[26] Benabou (2013) describes a complementary logic, via a mechanism of mutually assured delusion induced by wishful thinking, on why a dictator may not need to exert constant censorship to implement extreme policies because citizens adjust their beliefs to rationalize the status quo. The paper shows that this mechanism may also lead to a form of collective fatalism.

[27] The odds offered by Betfair always overwhelmingly favored a Remain victory, and in a sample of 12,369 voters, LordAschroftPolls found that "seven voters in ten expected a victory for Remain, including a majority of those who voted to leave.". Remains implied chance of victory never dropped below 63%, averaged somewhat above 70%, and was still above 80% on the day of the referendum (suggesting that the price is not mostly due to speculation). See, respectively, https://betdata.io/historical-odds/uk-eu-referendum-2016 for the data on Betfair and http://lordashcroftpolls.com/2016/06/how-the-united-kingdom-voted-and-why/#more-14746 for the poll. Similarly, in the months leading up to the 2016 US Presidential election, including the day before the election, prediction market prices always indicated that Hillary Clinton was the heavy favorite to win. See, https://www.predictit.org/markets/detail/1234/Who-will-win-the-2016-US-presidential-election.

the reverse and there entry choices are complements. To present the results, suppose that f and g are continuously differentiable and g is non-decreasing in each element of the type vector. The loss from a wrong move (cost of dissent) now corresponds to a decrease in a positive player's payoff from one-sided entry when her opponent is negative. Accordingly, consider a change in $g(\theta_i, \theta_{-i})$ which leaves the value of this function unaffected if $\theta_i, \theta_{-i} > 0$, but decreases the value of this function if $\theta_i > 0 > \theta_{-i}$. We call *any* such change a decrease in g^-.

Proposition 4 *Equilibrium is in cutoff strategies.*

1. *If entries are substitutes, there is a unique symmetric equilibrium. It is increasing in ρ. Furthermore, $E[\overline{\pi}_1^\rho]$, $E[\overline{\pi}_1^{\rho,+}]$, $E[\overline{\pi}_1^{\rho,-}]$, are all decreasing in g^- iff $\rho > 0$.*

2. *If entries are complements and $g_2 = 0$, all equilibria are symmetric, the lowest is decreasing in ρ, the second-lowest, if it exists, is increasing in ρ.*

3. *If $\rho > 0$, in all symmetric equilibria,*

   - $\pi_1^\rho[\theta_{-i}|\theta_i, a] <_{fosd} \pi_1[\theta_{-i} \mid \theta_i, a]$ *for any given $a \in A$ and $\theta_i > 0$,*[28]
   - $E[\pi_1^\rho|\theta_i] <_{fosd} \pi_0$ *if $\theta_i > 0$, and $E[\pi_1^\rho|\theta_i] \geq_{fosd} \pi_0$ if $\theta_i \leq 0$.*

If entries are substitutes, there is a unique symmetric equilibrium and the same comparative static with respect to ρ and with respect to any increase in the cost of a wrong move holds again, as in Propositions 3. If entries are complements, there may be multiple symmetric equilibria with the lowest decreasing, while the second-lowest, if it exists, increasing in ρ. Nevertheless, *all* symmetric equilibria, irrespective of entries being substitutes or complements, exhibit both conditional (Part I) and average (Part II) false antagonism, as in Proposition 2. Hence, the dynamic consequences of the model also extend.

5 Trade

As the last application, we consider the classic problem of trade with asymmetric information, e.g., Akerlof (1970), Samuelson (1984), Myerson (1985). This problem is at the heart of many economic applications and is commonly analyzed assuming

[28] If $a = \{a_i = in, a_{-i} = out\}$ this relation is again weak.

rational response to asymmetric information. The empirical literature, however, documented key departures from these rational predictions, e.g., Samuelson and Bazerman (1984, 1985), Ball et al. (1991), Holt and Sherman (1994), Kagel and Levin (2002), Fudenberg and Peysakhovich (2016). Specifically, better-informed traders appear to under-utilize their informational advantage while less-informed traders appear to under-appreciate their informational disadvantage and fall prey, for example, to the classic winner's curse. Below, we describe how projection equilibrium can provide an unified explanation of these findings. We also compare our predictions to that of cursed equilibrium (Eyster and Rabin, 2005), ABEE with the coarse partition, which is often motivated as an explanation of the less-informed party's behavior in this problem.

Setup. The seller values an asset at some non-negative real q, the buyer at $w(q) = mq + x$, where q is drawn according to a commonly known continuous and strictly positive density f over $[a, b]$. Its realization is observed only by the seller. The ex-ante value of the asset to the buyer is $\overline{q}$ and to the seller is $\overline{w}$ and trade is ex-ante beneficial, $\overline{w} > \overline{q}$. Given unbiased beliefs, interim efficient and individually rational trade is impossible if $\overline{w} < b$ (Myerson, 1985). Since in this setting with one-sided private information such information typically relates to a common value component of an external asset, or the interactions are less personal, we consider the implications of projection equilibrium.

5.1 Informed-offer Game

Suppose that the informed party has the bargaining power and makes a take-it-or-leave it offer $p_s(q)$ which the uninformed party can accept or reject.[29] To focus on the main insight, we assume for the analysis of this game that trade is strictly beneficial and this benefit is non-decreasing in q, that is, $m \geq 1$ and $w(a) > a$.

In the unbiased case, $\rho = 0$, pure separation cannot arise in equilibrium. The seller would never name the lower of any two prices if both were accepted (with the same probability) by the buyer. Under general conditions, the seller-optimal mechanism, also the social surplus maximizing one, corresponds to the pooling equilibrium of this game, where he names $p_s(q) = p^*$ if $q \leq p^*$, and makes no serious offer (offer that is accepted with positive probability), otherwise (Samuelson, 1984).

[29]The game here is a simple sequential-move game with observable moves by the players. It is then straightforward to impose the standard restriction of perfectness for the uninformed party's off-equilibrium path beliefs maintaining projective thinking, which we do.

Trade occurs at a single price, and for all lower q the seller bluffs, that is, sells at a price which exceeds the buyer's ex-post valuation, $w(q)$. Projective thinking, $\rho > 0$, changes the conclusions above.

Proposition 5 *1. For any $\rho > 0$, there exists a projection equilibrium where $p_s^\rho(q) = w(q)$ and the buyer accepts all prices $p \leq \overline{p} = min\{\frac{w(a)-\rho a}{1-\rho}, \overline{w}\}$ for sure and any higher price p with some probability $z^\rho(p)$. 2. Let Π_s^ρ be the seller's maximal ex-ante expected payoff in such an equilibrium for any given ρ. Then Π_s^ρ smoothly increases in ρ and $lim_{\rho \to 1} \Pi_s^\rho = \overline{w}$.*

In the above fully revealing projection equilibrium, the informed-party holds the uninformed party to his ex-post valuation, thus he never bluffs, and the uninformed party accepts all lower prices for sure, thus pure separation holds. Two more properties characterize the above prediction. First, for all lower q the seller *underbids* relative to his payoff-maximizing bid given the buyer's true acceptance behavior. Second, projective thinking allows for more efficient and individually rational trade than the unbiased Bayesian upper bound. As long as there is sufficient projecting, the seller's maximal ex-ante expected payoff, and social surplus, exceed their unbiased upper bounds. As the bias becomes full, trade can always become efficient with the projecting seller extracting the full surplus.

By projecting, the informed party underestimates the return on bluffing. The buyer partially anticipates this and, since projection is all-encompassing, she believes that the projected seller both knows that she is uninformed and is himself uninformed thus unable to condition his offer on q. The bound $\overline{p}$ – along with the decreasing acceptance probability of higher price offers – then ensures that neither the real nor the projected seller wants to bluff.[30]

Evidence. Evidence for the informed-offer game matching our assumptions comes from Samuelson and Bazerman (1984, 1985). They consider the additive lemons problem where $m = 1$, $x = 30$, and $q \sim U[0, 100]$. In the unbiased case, the seller-optimal mechanism is the pooling equilibrium with $p^* = 60$ and an ex-ante expected seller payoff of 60 (a social gain from trade of 10). Proposition 5 implies a separating equilibrium with $p_s^\rho(q) = q + 30$ and $\overline{p} = min\{\frac{30}{1-\rho}, 80\}$ and potentially greater gains from trade. The empirical findings are consistent with this prediction

[30] If only the buyer projected, then instead of pure separation, 'excess' pooling could follow where the buyer is willing to accept a higher pooling price than in the unbiased case. If only the seller projected, the same equilibrium structure would hold as above but one could relax $\overline{p} \leq \overline{w}$.

and are inconsistent with BNE.[31] The most common offer in the data, for any given q, is $p_s(q) = w(q) = q + 30$ (54% of offers). In addition, 89% of offers involve no bluffing, i.e., $p_s(q)$ is in $[q, w(q)]$. There is very little evidence on pooling on 60 (5.5% of all offers). Consistent with our prediction, but inconsistent with BNE, for all $q \leq 40$, sellers significantly *underbid* relative to what their empirical payoff-maximizing strategy would be given the buyers' actual behavior ($p < 0.01$; $p = 0.04$ for $q = 40$). At the same time, we can reject such underbidding for higher values of q. Similarly, the seller's expected payoff of \$66 is significantly higher than the upper bound on the seller's equilibrium payoff under BNE, \$60 ($p < 0.01$), and the gains from trade and efficiency are substantially improved relative to the unbiased case. The buyer's expected payoff is \$2.8, close to the prediction of zero.

5.2 Uninformed-offer game

Suppose now that it is the uninformed party who has the bargaining power and makes a take-it-or-leave-it offer p_b (the buyer optimal mechanism in the unbiased case). Suppose that the buyer's unbiased optimal bid, the BNE prediction p_b^0, is unique. Let p_b^ρ denote the bid predicted by projection equilibrium.

Proposition 6 *1. If $p_b^0 < \overline{q}$, then either $p_b^\rho = p_b^0$ or $p_b^\rho \geq \overline{q}$ and there is a winner's curse, $p_b^{\rho \to 1} = \overline{q}$. 2. If $p_b^0 > \overline{q}$, then $p_b^\rho \in [\overline{q}, p_b^0]$ and there is a loser's curse, $p_b^{\rho \to 1} = \overline{q}$. 3. If $p_b^0 = \overline{q}$, then $p_b^\rho = p_b^0$ for all ρ.*

By projecting, the uninformed party underestimates the probability that the seller is informed and that trade is subject to selection. When the optimal bid would be below (above) the seller's ex-ante cost, the buyer instead falls prey to the winner's (loser's) curse. The analysis here is simplified in that both the real and the projected versions of the seller have a strictly dominant strategy. The buyer always achieves a (weakly) lower payoff here than in the unbiased case. In fact, she will often lose from trade and thus is better off if the informed party has the bargaining power. As $\rho \to 1$, our model always predicts trade, but the buyer always loses from such trade whenever $E[w(q)|q \leq \overline{q}] < \overline{q}$.

Evidence. Samuelson and Bazerman (1984) study the same additive lemons problem as discussed above with $q \sim U[0, 100]$ and $w(q) = q + 30$ also in the

[31] For the dis-aggregated data see Figures 3.2a and 3.2b for the seller-offer game and Figure 1.2 for the buyer-offer game in Samuelson and Bazerman (1984). The data is presented in histograms only so we use the lower bound of each range.

uninformed-offer case. A number of classic studies also consider the multiplicative lemons problem $w(q) = 1.5q$ where $q \sim U[a, b]$ and the uninformed party has the bargaining power. Samuleson and Bazerman (1984) consider $U[0, 100]$ while Holt and Sherman (1994) consider three specifications: (i) $a = 1, b = 3$, (here, $p_b^0 = \overline{q}$, no curse); (ii) $a = 1.5, b = 6$, (here, $p_b^0 < \overline{q}$, winner's curse); (iii) and $a = 0.5, b = 1$, (here, $p_b^0 > \overline{q}$, loser's curse). Below we describe the data for these specifications along with the unique predictions of (i) BNE; (ii) full projection equilibrium, $\rho \to 1$, PE; and (iii) fully cursed equilibrium, CE (Eyster and Rabin, 2005) – ABEE (Jehiel and Koessler, 2008) with the coarse analogy partition. For projection equilibrium, we also list the threshold value ρ^* above which the projection equilibrium is constant in ρ at PE. Partially cursed equilibrium here spans the interval between BNE and CE (which is always closer to the data than any partially cursed equilibrium). We then test the equivalence between the data and these models (p-values reported in parentheses in the table).[32]

Specification, $U[a, b]$	BNE	**Data**	PE	ρ^*	CE
1. additive, $[0, 100]$	30 (< 0.01)	**55**	50 (< 0.01)	$1/16$	40 (< 0.01)
2. multiplicative, $[0, 100]$	0 (< 0.01)	**49**	50 (0.32)	0.2	37.5 (< 0.01)
3. multiplicative, $[1, 3]$	2 (0.33)	**2.3**	2 (0.33)	0	2 (0.33)
4. multiplicative, $[1.5, 6]$	3 (< 0.01)	**3.77**	3.75 (0.79)	0.02	3.56 (0.02)
5. multiplicative, $[.5, 1]$	1 (< 0.01)	**0.75**	0.75 (0.54)	0.2	0.81 (< 0.01)

Table 3: **Model tests** on the data of Samuelson and Bazerman (1984) & Holt and Sherman (1994) (p-values in parentheses)

The data is again consistent with the PE predictions in all conditions except one (specification 1). However, even in specification 1, the modal offer (23% of all offers) is the exact same as the PE prediction of 50.[33] Furthermore, 71% of bids were in $[50, 80]$, 48% in $[60, 80]$, consistent with projective thinking and some surplus

[32] We are very grateful to Charles Holt for sharing their data. We retrieved data from 48 of the 50 subjects in Holt and Sherman (1994), which yielded virtually the same results whenever we replicated their analysis. All tests are based on average values per subject; all results are qualitatively the same when signed-rank tests are applied instead of t-tests.

[33] Fudenberg and Peysakhovich (2016) also study an isomorphic problem to specification 1, with $x = 3$ and $q \sim U[0, 10]$, and exactly as predicted by PE, find an average bid of 5.1 with 95% confidence interval of $[4.88, 5.41]$. They also study the same additive lemons problem with $x = 6$. Here, projection equilibrium spans $[5, 6]$ and they find the 95% confidence interval of bids in $[5.72, 6.27]$, while here CE is 5.5.

sharing motive, but inconsistent with any surplus sharing motive under rational expectations and players respecting dominance (where any such offer leads to a strictly negative payoff for the buyer as long as the informed sellers accepts offers that provide a positive surplus for them). Essentially no subject bid above 80 (1.5%).[34] PE predicts a negative buyer payoff only in Specification 2 and indeed that is the only specification where the buyer's empirical payoff is significantly below 0 ($p < 0.01$).[35] Finally, the threshold value on ρ in all conditions is at most 0.2. Note that in our experiment described in Section 3, 70% subjects have a ρ greater than 0.2. In specification 3, all models make the same prediction which is not significantly different from the data. The data, however, rejects both BNE and CE (thus also partial cursedness) in *all* other specifications.

Cursedness and Projection We now compare the implications of projective thinking to that of cursedness this setting. In the informed-offer game, since the buyer has no private information, a cursed seller always has correct beliefs about the buyer's strategy state-by-state, and cursedness has no direct impact on the informed party. A cursed seller has the same best response as an unbiased one, making pure separation and underbidding, as predicted under ρ-PE, infeasible. The logic of cursedness points in the direction of pooling and bluffing, but the empirical evidence clearly contradicts this.[36]

In the uninformed-offer game, both cursedness and projective thinking imply that the buyer underestimates selection, but their logic differs sharply. Below we show that whether their predictions point in the same or opposite directions depends on whether trade is subject to adverse, $m > 0$, or advantageous selection, $m < 0$. Under adverse selection, there is a positive relationship between the seller's cost and the buyer's valuation; under advantageous selection, there is a negative relationship between these two. For example, in the context of insurance, those with higher risks may care less and have a lower rather than higher willingness to pay, e.g.,

[34]The presence of such a surplus sharing motive is present in all specfications of Samuelson and Bazerman (1984). In specification 2, 75% of bids were in $[50, 75]$ again consistent with projection equilibrium and some surplus sharing motive, but inconsistent with such a motive under rational expectations. It is also present in the buyer-offer condition described above.

[35]Ball et al. (1991) also study the above specification 2, but allow for 20 rounds of learning. They find that even after such learning the mean reported bid was 52.61 again consistent with the PE prediction of 50. Furthermore, they find that there is no substantive change in the average bid over the learning period.

[36]On the possibility of extending the idea of cursedness to more general extensive-form games, see, e.g., Cohen and Li (2023).

Einav and Finkelstein (2011). Since in these settings partial cursedness is again between the unbiased and the fully cursed predictions, and the analogous fact holds for projection equilibrium, we again describe CE and PE. To focus on the main insight, for simplicity, we assume that f is uniform.

Corollary 4 *The following comparisons hold.*

- ***No Selection***. *If $m = 0$, then $CE = BNE$; while $PE > BNE$ if $\overline{w} < b$ and $PE < BNE$ if $\overline{w} > b$.*
- ***Adverse Selection***. *If $m > 0$, CE and PE deviate from BNE in the same direction. Furthermore, $|CE - BNE| \leq |PE - BNE|$.*
- ***Advantageous Selection***. *If $m < 0$, CE and PE deviate from BNE in the opposite directions.*

A cursed buyer always has correct beliefs about the seller's average strategy and thus about the marginal cost of increasing her offer. She only wrongly thinks that conditional on acceptance, her benefit will be $\overline{w}$ irrespective of her offer. Instead, a projecting buyer underestimates (overestimates) the probability of acceptance following a bid below (above) $\overline{q}$.

In case of private values, no selection ($m = 0$), CE is then always equivalent to BNE. Instead, a projecting buyer's offer deviates from her optimal bid in the direction of the seller's average cost $\overline{q}$. All else equal, the projecting buyer then overbids when her valuation is low and underbids when her valuation is high. Note that this case is isomorphic to the classic monopoly problem where the monopolist (isomorphic to the buyer above) with cost x faces standard demand uncertainty about q (isomorphic to the seller above) and posts a price. Our model predicts that the monopolist will price in a way that depends too little on her costs, over-pricing the good when her cost is low and under-pricing it when her cost is high, as she targets customers' mean valuation ignoring the true elasticity of demand.

In the presence of selection, a cursed buyer overestimates the marginal benefit of raising her offer at the optimal bid if $w(p_b^0) < \overline{w}$, and underestimates it if the reverse holds. Under adverse selection, the sign of $w(p_b^0) - \overline{w}$ is the same as that of $p_b^0 - \overline{q}$ and projective thinking and cursedness predict the same directional deviation from the optimal bid, but, as is consistent with the evidence in Table 1, the deviation under projection is always larger. Under advantageous selection, the sign of $w(p_b^0) - \overline{w}$

is the opposite of $p_b^0 - \overline{q}$ and the two models predict opposite deviations from the optimal bid.

Projecting Valuations. Finally, the data is also inconsistent with the hypothesis that players mistakenly think that others have the same valuations (e.g., the seller thinking that the buyer's utility from the good is the same as his cost of producing it), as opposed to, or in conjunction with, the same information as they do. In the data, informed sellers bid the buyers' *higher* conditional valuations, and uninformed buyers bid the sellers' *lower* unconditional costs. They both act as if they fully exploited the correct and binding individual rationality constraints of others given differences in valuations, but ignored differences in information. More generally, note that our model allows for general state dependent preferences, $u_i(\omega, a)$, and indeed predicts an initial mis-estimation of others' preferences in any setting with interdependent values (Milgrom and Weber, 1982; which may at first sight appear as if they projected valuations, that is, they underestimated differences in interim valuations). It predicts that people will mis-estimate others' interim preferences as a function of their private information. In the above context, whether people under- or over-estimate others' interim valuation as a function of their own depends on whether ex-post valuations are positively or negatively related. If $m > 0$, the informed party will exaggerate the other party's valuation when his own valuation is high and underestimate it when his own valuation is low, while if $m < 0$, he will do the reverse. When $m = 0$, he will have the correct prediction. More general implications can be derived in settings with interdependent values.[37]

6 Conclusion

Motivated by robust evidence, this paper proposes a parsimonious but general model of limited perspective taking for games. We provide a direct experimental test of the implied tight partial distortion in higher-order beliefs in a general design and find clear support for it. The model also helps account for empirical puzzles in a variety of strategic problems. Future research can expand these applications, describe general theoretical consequences, or assess the model's relevance in various domains from voting and information aggregation, e.g., Feddersen and Pesendorfer

[37]Note also that projecting valuations fully implies no trade, as trade is a result of differences in valuations. Projecting valuations partially should again reduce trade relative to trade under correct beliefs. In contrast, a robust observation of the data is *excess* trade whenever $p_b^0 < \overline{q}$, e.g., specifications 1-4 in Table 1.

(1996) to strategic communication, e.g., Sobel (2020).[38] We conclude by mentioning some settings where the model may be directly helpful.

False Consensus in Macroeconomic Expectations. In many macroeconomic contexts, what matters is not simply people's forecast of a common variable, such as inflation, but their expectation of others' forecast of this common value and related higher-order beliefs. In leading models, people's expectations is a combination of a public and a private signal. Our model predicts that people will then exhibit a *false consensus effect* in such forecast problems and overreact to their private signals. In a survey of firm managers, Coibion et al. (2021) elicited both first- and second-order beliefs about inflation expectations. One of their key findings is that "managers disagree with each other about the level of inflation but do not realize how much they disagree." Our model further predicts that a person will partially anticipate that others will exhibit such a false consensus regarding those forecasts and such partial anticipation may then affect pricing and the transmission of shocks.

Relative Overconfidence and Conflict. The strategic implications of relative overconfidence also depend on the structure of higher-order beliefs. Both in congested entry (Camerer and Lovallo, 1999), as well as in bargaining (Yildiz, 2003), key predictions depend on the shape of such higher-order beliefs, e.g., whether others know and agree with one's overconfidence, etc. Our model provides guidance for such settings. It suggests that people will anticipate but *underestimate* the extent of the relative overconfidence of others.

Social Learning and Coordination. Projective thinking can also affect social learning. Suppose two people who are trying to learn the common state of the world (e.g. the urn picked) each receive i.i.d. signals about the state (e.g. colored balls from the urn picked) and can communicate with each other. If they engage in projective thinking, they may underinfer from the other belief updating in the same direction and not realize that the other person received additional signals (Conlon et al., 2022). If there is uncertainty both about what signals the other person received and how the other person updates beliefs based on the signals (e.g. Bayesian or non-Bayesian), projective thinking could also lead to erroneous inference about each other's belief updating rules.

[38] For an application of the current model to deception and credulity in strategic communication, see Madarasz (2023).

References

[1] Akerlof, George. (1970). "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics,* 84: 488–500.

[2] Ball, Sheryl, Max Bazerman, and John Carroll. (1991). "An Evaluation of Learning in the Bilateral Winner's Curse." *Organizational Behavior and Human Decision Processes*, 48: 1–22.

[3] Bénabou, Roland (2013). "Groupthink: Collective Delusions in Organizations and Markets." *Review of Economic Studies*, 80(2): 429-462.

[4] Blanco, Mariana, Dirk Engelmann, Alexander K. Koch, and Hans-Theo Norman. (2010). "Belief Elicitation in Experiments: Is There a Hedging Problem?"*Experimental Economics*, 13(4), 412-438.

[5] Beatty, Timothy K. M., and Ian A. Crawford. (2011). "How Demanding Is the Revealed Preference Approach to Demand? "*American Economic Review*, 101 (6): 2782–95.

[6] Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott. (2020). "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia."*American Economic Review*, 110(10): 2997-3029.

[7] Camerer, Colin, and Dan Lovallo. (1999). "Overconfidence and Excess Entry: An Experimental Approach. "*American Economic Review*, 89(1): 306-318,

[8] Camerer, Colin, George Loewenstein, and Martin Weber. (1989). "The Curse of Knowledge in Economic Settings: An Experimental Analysis."*Journal of Political Economy*, 97(5): 1234–1254.

[9] Cantoni, Davide, David Y. Yang, Noam Yuchtman, and Jane Zhang. (2016). "The Fundamental Determinants of Anti-Authoritarianism. "mimeo, LMU Munich. http://www.davidecantoni.net/pdfs/hk_democrats_20161116.pdf

[10] Coase, Ronald. (1960). "The Problem of Social Cost "*Journal of Law and Economics*, 3: 1-44.

[11] Cohen, Shani, and Shengwu Li (2023). "Sequential Curse Equilibrium " *Working Paper.*

[12] Coibion, Oliver, Yuriy Gorodnichenko, Saten Kumar, and Jane Ryngaert. (2021). "Do You Know that I Know that You Know. . . ? Higher-Order Beliefs in Survey Data." *Quarterly Journal of Economics*, 136(3): 1387–1446.

[13] Conlon, John, Malavika Mani, Gautam Rao, Matthew Ridley, and Frank Schilbach. (2022). "Not Learning from Others." *Working Paper.*

[14] Cooter, Robert. (1991). "Economic Theories of Legal Liability." *Journal of Economic Perspectives*, 5 (3): 11–30.

[15] Crawford, Vincent, and Nagore Iriberri (2007). "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions? "*Econometrica,* 75(6): 1721–1770.

[16] Danz, David. (2020). "Never Underestimate Your Opponent: Hindsight Bias Causes Overplacement and Overentry into Competition." *Games and Economic Behavior*, 124: 588–603.

[17] Danz, David, Lise Vesterlund, and Alistair J. Wilson. (2022). "Belief elicitation and behavioral incentive compatibility."*American Economic Review*, 112(9): 2851–2883.

[18] Danz, David, Kristóf Madarász, and Stephanie Wang. (2018). "The Biases of Others: Projection Equilibrium in an Agency Setting. " CEPR Discussion Paper No. DP12867.

[19] Davison, Anthony, and David Hinkley (1997). "Bootstrap Methods and Their Application". Cambridge University Press.

[20] Epley, Nicolas, Keysar Boaz, Leaf Van Boven, and Thomas Gilovich.(2004). "Perspective Taking as Egocentric Anchoring and Adjustment." *Journal of Personality and Social Psychology*, 87(3): 327-339.

[21] Einav, Liran, and Amy Finkelstein. (2011). "Selection in Insurance Markets: Theory and Empirics in Pictures. "*Journal of Economic Perspectives*, 25 (1): 115-38.

[22] Esponda, Ignacio. (2008). "Behavioral Equilibrium in Economies with Adverse Selection." *American Economic Review,* 98(4): 1269–91.

[23] Eyster, Erik, and Matthew Rabin. (2005). "Cursed Equilibrium." *Econometrica*, 73(5): 1623–1672.

[24] Feddersen, Timothy J., and Wolfgang Pesendorfer (1995). "The Swing Voter's Curse "*American Economic Review*, 86(3), 408-424.

[25] Fischhoff, Baruch. (1975). "Hindsight / foresight: The Effect of Outcome Knowledge On Judgment Under Uncertainty." *Journal of Experimental Psychology: Human Perception and Performance*, 1, 288–299.

[26] Fudenberg, Drew, and Alex Peysakhovich. (2016). "Recency, Records and Recaps: Learning and Non-Equilibrium Behavior in a Simple Decision Problem."*ACM Transactions on Economics and Computation (TEAC)*, 4 (4), 1-18.

[27] Fudenberg, Drew, Wayne Gao, and Annie Liang. (2024). "How Flexible is that Functional Form? Quantifying the Restrictiveness of Theories."*Review of Economics and Statistics*.

[28] Harley, Erin. (2007). "Hindsight Bias in Legal Decision Making."*Social Cognition*, 25 (1): 48-63.

[29] Holt, Charles, and Roger Sherman. (1994). "The Loser's Curse." *American Economic Review*, 84(3): 642–652.

[30] Hume, David. (1741). "Of the First Principles of Government." Essay V in: *Essays, Moral and Political.* A. Kincaid, Edinburgh.

[31] Jehiel, Philippe. (2005). "Analogy-Based Expectations Equilibrium." *Journal of Economic Theory,* 123: 81–104.

[32] Jehiel, Philippe, and Frederick Koessler. (2008). "Revisiting Games of Incomplete Information with Analogy-Based Expectations." *Games and Economic Behavior*, 62: 533–557.

[33] Kagel, John H., and Dan Levin. (2002). *Common Value Auctions and the Winner's Curse.* Princeton University Press

[34] Kaplow, Louis, and Steven Shavell. (2002). "Economic Analysis of Law " *Handbook of Public Economics.* Auerbach and Feldstein, eds. vol. 3, 1661-1784.

[35] Katz, Daniel, Allport, Floyd, and Jenness, M. B. (1931). Students' attitudes; a report of the Syracuse University reaction study. Craftsman Press.

[36] Kuran, Timur. (1995). *Public Lies and Private Truth.* Harvard University Press.

[37] Loewenstein, George, Don Moore, and Roberto Weber. (2006). "Misperceiving the Value of Information in Predicting the Performance of Others."*Experimental Economics*, 9(3), 281-95.

[38] MacInnis, Cara, and Elizabeth Page-Gould (2015). "How Can Intergroup Interaction Be Bad If Intergroup Contact Is Good? Exploring and Reconciling an Apparent Paradox in the Science of Intergroup Relations. "*Perspectives on Psychological Science*, 10(3), 307–327.

[39] Madarász, Kristóf. (2012). "Information Projection: Model and Applications."*Review of Economic Studies*, 79: 961–985.

[40] Madarász, Kristóf. (2016). "Projection Equilibrium: Definition and Applications to Social Investment and Persuasion."mimeo LSE, CEPR D.P. 10636.

[41] Madarász, Kristóf. (2023). "Limited Perspective Taking in Strategic Communication."mimeo, LSE.

[42] McKelvey, Richard D., and Thomas R. Palfrey. (1995). "Quantal Response Equilibrium for Normal Form Games "*Games and Economic Behavior*, 10(1), 6-38.

[43] Milgrom, Paul and Robert Weber. (1982). "A Theory of Auctions and Competitive Bidding."*Econometrica*, 50(5): 1089–122.

[44] Miller, Dale, and Cathy McFarland. (1987). "Pluralistic Ignorance: When Similarity is Interpreted as Dissimilarity." *Journal of Personality and Social Psychology*, 53(2), 298–305.

[45] Myerson, Roger. (1985). "Analysis of Two Bargaining Problems with Incomplete Information. "in Game Theoretic Models of Bargaining, ed. by A. Roth. Cambridge, U.K.: Cambridge University Press

[46] O'Gorman, Hubert. (1975). "Pluralistic Ignorance and White Estimates of White Support for Racial Segregation." *Public Opinion Quarterly,* 39 (3): 313–30.

[47] Pettigrew T. F., Tropp L. R. (2006). A meta-analytic test of intergroup contact theory. Journal of Personality and Social Psychology, 90, 751–783

[48] Piaget, Jean, and Bärbel Inhelder. (1948). *The Child's Conception of Space.* Translated (1956). London: Routledge and Kegan Paul.

[49] Posner, Richard. (1998). "Rational Choice, Behavioral Economics, and the Law." Stanford Law Review 50(5): 1551–75.

[50] Prentice, Deborah, and Dale Miller. (1993). "Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm." *Journal of Personality and Social Psychology*, 64: 243–256.

[51] Prentice, Deborah. (2007). "Pluralistic Ignorance." *Encyclopedia of Social Psychology,* eds. Roy Baumeister and Kathleen Vohs, pp. 674-675, Sage Publications, Inc.

[52] Pronin, Emily, Daniel Lin, and Lee Ross. (2002). "The Bias Blind Spot: Perceptions of Bias in Self Versus Others. " *Personality and Social Psychology Bulletin*, 28(3), 369-381.

[53] Pronin, Emily. (2008). "How We See Ourselves and How We See Others." *Science*, 320: 1177-1180.

[54] Rachlinski, Jeffrey. (1998) "A Positive Psychological Theory of Judging in Hindsight, " *University of Chicago Law Review*, 65(2): Article 4.

[55] Rensink, Ronald A., J. Kevin O'Regan, and James. J. Clark. (1997). "To See or Not to See: The Need for Attention to Perceive Changes in Scenes." *Psychological Science*, 8(5), 368-373.

[56] Samuelson, William. (1984). "Bargaining under Asymmetric Information." *Econometrica*, 995-1006.

[57] Samuelson, William F., and Max H. Bazerman (1984). "The Winner's Curse in Bilateral Negotiations."*Working Paper.*

[58] Samuelson, William F., and Max H. Bazerman (1985). "Negotiation under The Winner's Curse. "*Research in Experimental Economics*, vol. 3, Vernon L. Smith, ed., Greenwich, CT: JAI Press.

[59] Selten, Reinhard. (1991). "Properties of a Measure of Predictive Success." *Mathematical Social Sciences*, 21:153-167.

[60] Shelton, Nicole, and Jennifer Richeson. (2005). "Intergroup Contact and Pluralistic Ignorance." *Journal of Personality and Social Psychology*, 88(1): 91–107.

[61] Simons, Daniel J., and Daniel T. Levin. (1997). "Change Blindness." *Trends in Cognitive Sciences*, 1(7), 261-267.

[62] Sobel, Joel. (2020). "Lying and Deception in Games." *Journal of Political Economy*, 128(3), 907-947.

[63] Sunstein, Cass R., Christine Jolls, and Richard H. Thaler. (1998) "A Behavioral Approach to Law and Economics. " 50 *Stanford Law Review*, 1471-1550.

[64] Van Boven, Leaf. (2000). "Political Correctness and Pluralistic Ignorance: The Case of Affirmative Action." *Political Psychology,* 21(2): 267-276.

[65] Yildiz, Muhamet (2003). Bargaining without a Common Prior: An Immediate Agreement Theorem. *Econometrica*, 71, 793-811.

Appendix A. Extension -Online

Consider an N-player game of the form described before. We first extend the model of ρ-IPE. Let $S = \prod_{i \in N} S_i$ be the true finite strategy space. Each player i now imagines a projected version for each of her opponents. Since the information of players i and j differ, the projected version of k, as imagined by i, differs from the projected version of k, as imagined by j. Formally, let the strategy set of the projected version of player k, as imagined by player i, be:

$$S_k^{+i} = \{\sigma_k(\omega) \mid \sigma_k(\omega) : \Omega \to \Delta A_k \text{ measurable with respect to } P_k(\omega) \cap P_i(\omega)\}.$$

We denote the generic element of the set S_k^{+i} by σ_k^{+i}. Let $S^{+i} = \prod_{k \neq i \in N} S_k^{+i}$ denote the strategy space of the $N-1$ projected opponents of player i. We denote the generic element of this set S^{+i} by σ^{+i}. Finally, we denote the restriction of σ^{+i}, containing all parts of this profile except for σ_k^{+i}, for some $k \neq i$, by σ_{-k}^{+i}.

In the definition below projection occurs as a binary event whereby each player i believes that either all of her opponent's are regular or all are projected. Furthermore, the projected version of k, as imagined by player i, believes that player j is also the projected version of player j as imagined by player i.

Definition 3 *A strategy profile $\sigma^\rho \in S$ is a ρ-IPE of Γ if for each $i \in N$ there exist $\sigma^{+i} \in S^{+i}$ such that*

$$\sigma_i^\rho \in BR_{S_i}\{(1-\rho)\sigma_{-i}^\rho \circ \rho\sigma^{+i}\}, \tag{5}$$

and for each $k \neq i$

$$\sigma_k^{+i} \in BR_{S_k^{+i}}\{\sigma_i^\rho, \sigma_{-k}^{+i}\}. \tag{6}$$

The definition continues to satisfy the same all-encompassing and consistency properties as before. The extension of ρ-PE is analogous. It is obtained by replacing each S_k^{+i} with S_k^i, as defined in Section 2, and then S^{+i} with $S^i = \prod_{k \neq i \in N} S_k^i$ in the above definition.

Appendix B. Proofs – Online

We first formally present the predictions of projection equilibrium for Section 3. 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 ω with respect to the *true* distribution of actions and signals in the game. Since reference agents (those only solving the basic task) have no relevant strategic interactions and are ex ante equivalent, we can introduce a representative reference agent and denote them by $\overline{A}$.

All players perform 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 thus given by $A_P = D \times [0,1]$. The action set of the strategically-active agent involves two estimates tasks and is thus given by $A_A = D \times [0,1] \times [0,1]$. Since for no player i does the payoff from choosing x_i directly interact with the payoff from the estimation, we denote this payoff by $f(x_i, \omega)$ and normalize it to be one if the solution is a success and zero otherwise. Let $f^*(x_i, \omega)$ be the equilibrium value of this payoff. With a slight abuse of notation, we can then represent the average success rate of the representative agent by $E[f^*(\omega, x) \mid P_{\overline{A}}(\omega)] = \phi$.

Claim 1. The ex-ante expected estimate of ϕ by a ρ_P-biased principal, using the definition of projection equilibrium, equals to:

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

thus, given the law of iterated expectations, this equals $E[b_P^I] = \rho_P(d + \phi) + (1 - \rho_P)\phi = \phi + \rho_P d$, as stated in Claim 1.

Claim 2 The ex-ante expected first-order belief of ϕ by the agent is analogous to the above and is then given by:

$$E[\rho_A E[f^*(\omega, x) \mid P_A(\omega)] + (1 - \rho_A) E[E[f^*(\omega, x) \mid P_{\overline{A}}(\omega)] \mid P_A(\omega)]],$$

which then becomes $E[b_A^I] = \rho_A \phi + (1 - \rho_A)\phi = \phi$, establishing the first part of Claim 2. Now consider the agent's ex-ante expected second-order belief, her estimate of the principal's estimate of ϕ. Using the definition of projection equilibrium, this equals:

$$\rho_A E[E[f^*(\omega, x) \mid P_A(\omega)]] + (1 - \rho_A) E[b_P^I].$$

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 representative agents in a fully correlated fashion. The second part comes from the linearity of the expectations operator and the law of

iterative expectations. Rearranging the above expression, the above then equals $E[b_A^{II}] = \rho_A\phi + (1-\rho_A)(\phi + \rho_P d) = \phi + (1-\rho_A)\rho_P d$, as stated in Claim 2.

Proof 1 (Corollary 1) *1. is immediate. 2. Note that if P_i refines P_j, then $BR_{S_j^i}$ is equivalent to $BR_{S_j^+}$ and $BR_{S_i^+} = BR_{S_i}$. 3. Suppose that σ^0 is a BNE that is also an ex-post equilibrium of Γ, that is, for each i, ω and $a_i \in A_i$:*

$$u_i(\sigma_i^0(\omega), \sigma_{-i}^0(\omega), \omega) \geq u_i(a_i, \sigma_{-i}^0(\omega), \omega).$$

Let for each i and $k \neq i$, $\sigma_k^{+i} = \sigma_k^0$. This is feasible since $S_k \subseteq S_k^{+i}$ for all k and i. Furthermore, $\sigma_k^0 \in BR_{S_k^{+i}}(\sigma_{-k}^0)$ for each k and i. Hence, σ^0 is a ρ-IPE of Γ for any ρ.

Example 1 *To provide an example of a BNE that is an ex post equilibrium but is not a PE, consider the following game in Table 4 where the state is the private information of the row player and there is a symmetric prior. It follows that $(B; b, a)$ is the unique BNE which is also an ex-post equilibrium (and IPE for any ρ). Instead, the unique PE for, ρ sufficiently large, is $(A; b, a)$.*

ω_1	A	B	ω_2	A	B
a	$2,1$	$1,0$	a	$4,0$	$2,1/2$
b	$6,1$	$4,2$	b	$3,5$	$1,0$

Table 4: Example

Proof 2 (Proposition 1) *Suppose that $\rho > 0$, let $\underline{z}_\rho = (z-\rho)/(1-\rho)$. Note that when the de jure threshold is z, the agent abstains when $s \in [\underline{z}_\rho, z)$ iff $b < s(1-\rho)$ also holds because he thinks that here only the real but not the projected version of the principal holds him liable.*

1. If $b \geq z(1-\rho)$, given de jure threshold z, there does not exist $s \in [\underline{z}_\rho, z)$ such that the agent abstains, and the agent abstains iff $s > \max\{z, b\}$. To show that $z^* = z$ is unique, consider de jure threshold $z' < z$. There then exist $s \in [z', z)$ and $b \in [z(1-\rho), s]$ such that the agent abstains. Consider now threshold $z' > z$. There, given $\rho > 0$, then exists $s \in (z, z']$ and $b \leq z$ such that the agent engages.

2. If $b < z(1-\rho)$, then $max\{z, b\} = z$, and, given de jure threshold z, there exists $s \in [\underline{z}_\rho, z)$ such that the agent abstains. When the de jure threshold is raised

to z^r, the agent abstains iff $s \geq max\{z, b\}$ since when $s \geq z$, then $b < s(1-\rho)$ by assumption and if $s < z$ the agent rightly believes that he will not be liable. To show that $z^* = z^r$ is unique, note that for any threshold $z' < z^r$, there exists $s \in [(z'-\rho)/(1-\rho), z)$ such that the agent abstains. Similarly, for any threshold $z' > z^r$ there exists $s \in [z, (z'-\rho)/(1-\rho))$, such that the agent engages.

Proof 3 (Proof of Proposition 2) *To simplify the notation, let $x = \overline{\theta}$, $n = -\underline{\theta}$, and $r = (\overline{\theta} - \underline{\theta})^{-1}$ in all proofs below. The projected version of any player i has a dominant strategy: she enters iff $\min\{\theta_i, \theta_{-i}\} \geq 0$. Proposition 4 further shows that all equilibria are in cut-off strategies for the real versions as well.*

Let θ_i^ρ denote (real) i's equilibrium cutoff. Given this cutoff type's indifference between 'in' and 'out', θ_i^ρ must satisfy:

$$\rho(x(\theta_i^\rho - \gamma\theta_i^\rho) - nc) + (1-\rho)((x - \theta_{-i}^\rho)(\theta_i^\rho - \gamma\theta_i^\rho) + \theta_{-i}^\rho(\gamma\theta_i^\rho) - nc)) = 0. \qquad (7)$$

Rearranging terms one obtains that: $\theta_i^\rho = nc(x(1-\gamma) + \theta_{-i}^\rho(1-\rho)(2\gamma-1))^{-1}$. Substituting in the symmetric equation for θ_{-i}^ρ, then taking $\gamma \to 1$, the unique solution is $\theta_i^\rho = \sqrt{cn/(1-\rho)}$. When no interior solution exists, we assume, wlog, that $\theta_i^\rho = x$.

I. If $\theta_i > 0$, then player i's expectation of the average cutoff used by $-i$ is always lower than $-i$'s true cutoff given any $\rho > 0$. If $a \neq \{a_i = in, a_{-i} = out\}$, observing payoffs provides no additional information, hence, here $\pi_1^\rho[\theta_{-i} \mid \theta_i, a] <_{fosd} \pi_1[\theta_{-i} \mid \theta_i, a]$. If $a = \{a_i = in, a_{-i} = out\}$, observing payoffs leads i to form unbiased posterior beliefs because i now always learns the sign of $-i$'s valuation, and if this sign is positive, i also learns that $-i$ could not have been the projected version. Hence, here, $\pi_1^\rho[\theta_{-i} \mid \theta_i, a] = \pi_1[\theta_{-i} \mid \theta_i, a]$.

II a&b. Suppose first that players only observe the realized action profile. Let $\Pr(in)_{\theta_i}^\rho$ denote real θ_i's perception of the probability with which $-i$ enters. Let $\Pr(in)$ be the corresponding true probability. Since the martingale property of beliefs must hold with respect to the perceived probability, it follows that for any θ_i and θ_{-i}:

$$\pi_0(\theta_{-i}) = \Pr(in)_{\theta_i}^\rho \pi_1^\rho[\theta_{-i} \mid \theta_i, a_i^{\theta_i}, a_{-i} = in] + (1 - \Pr(in)_{\theta_i}^\rho)\pi_1^\rho[\theta_{-i} \mid \theta_i, a_i^{\theta_i}, a_{-i} = out],$$

where $a_i^{\theta_i}$ is θ_i's equilibrium action. Let's define

$$\Delta_{\theta_i}^\rho(\theta_{-i}) \equiv \pi_1^\rho[\theta_{-i} \mid \theta_i, a_i^{\theta_i}, a_{-i} = in] - \pi_1^\rho[\theta_{-i} \mid \theta_i, a_i^{\theta_i}, a_{-i} = out].$$

Note that $\int_{-n}^{x} \Delta_{\theta_i}^\rho(\theta_{-i}) d\theta_{-i} = 0$ and $\Delta_{\theta_i}^\rho(\theta_{-i})$ is increasing in θ_{-i}. The wedge between

the prior and the ex ante expected posterior of type θ_i is then given by this function multiplied by a scalar:

$$\pi_0(\theta_{-i}) - E[\pi_1^\rho(\theta_{-i}) \mid \theta_i] = (\Pr(in)_{\theta_i}^\rho - \Pr(in))\Delta_{\theta_i}^\rho(\theta_{-i}). \tag{8}$$

If $\theta_i > 0$, then $\Pr(in)_{\theta_i}^\rho > \Pr(in)$, hence, $E[\pi_1^\rho \mid \theta_i] <_{fosd} \pi_0$. If $\theta_i < 0$, then $\Pr(in)_{\theta_i}^\rho \leq \Pr(in)$, where equality holds only if $\Pr(in) = 0$, hence, $E[\pi_1^\rho \mid \theta_i] \geq_{fosd} \pi_0$.

Suppose now that players also observe their realized payoffs. If $a \neq \{a_i = in, a_{-i} = out\}$, the analysis is unchanged since i makes no additional inferences. If $a = \{a_i = in, a_{-i} = out\}$, player i forms unbiased beliefs as outlined above. The probability of such an action profile arising in equilibrium, however, conditional on any realization of θ_i is bounded away from 1. Hence, the result follows

Proof 4 (Proof of Proposition 3) *Follows from the proof of Proposition 4.*

Proof 5 (Proof of Corollary 2) *Following entry by either of the players, the continuation game has dominant strategies. Suppose now that there is no entry till the end of round t. At the beginning of round $t+1$, a positive θ_i's belief about θ_{-i} is given by a density that equals some constant v_t^ρ on $[0, x_t^\rho]$ and some constant y_t^ρ on $[-n, 0]$, where x_t^ρ is the symmetric cutoff of round t, conditional on no entry till $t-1$. Since this piece-wise constant density is strategically equivalent to a uniform density on $[-n', x_t^\rho]$ given some $n' > 0$, the uniqueness of the limiting ρ-IPE for each t follows immediately from Proposition 2. If $\rho = 0$, then $v_t^0 = y_t^0$ and, using Eq.(7), $x_t^0 = \sqrt{nc_t}$. If $\rho > 0$, then:*

$$y_t^\rho / v_t^\rho = y_{t-1}^\rho / (1-\rho) v_{t-1}^\rho. \tag{9}$$

Re-writing Eq.(7), re-weighting terms with the corresponding densities and solving for the unique fix point, one obtains that:

$$x_{t+1}^\rho = \min\{\sqrt{\frac{c_{t+1} n}{1-\rho} \frac{y_t^\rho}{v_t^\rho}}, x_t^\rho\}. \tag{10}$$

Thus,the cutoff of round $t+1$, conditional on no investment till t, x_{t+1}^ρ is increasing in ρ. Hence, $\Pr_{\underline{c}}^\rho(m)$ is decreasing in ρ Following entry in any round t, players' estimates of their opponents remain constant. Suppose now that there is no entry

till $t \geq 0$ where we simply denote x by x_0.[39] *Let $\rho > 0$.*

1. Notice that $E[\overline{\pi}^{\rho,+}_{t+1} \mid$ no entry till $t]$ is given by:

$$1-\frac{x^\rho_{t+1}+n}{x^\rho_t+n}[\frac{x^\rho_t-x^\rho_{t+1}}{x^\rho_t}\int_{-n}^{0}\frac{1}{n+x^\rho_{t+1}}d\theta_{-i}+\frac{x^\rho_{t+1}}{x^\rho_t}\int_{-n}^{0}\frac{y_{t,\rho}}{y^\rho_t n+(1-\rho)v^\rho_t x^\rho_{t+1}}d\theta_{-i}],$$

since if only i invests, then from observing her own payoff, she develops an unbiased estimate of θ_{-i}. Differentiating the above with respect to x^ρ_{t+1}, one gets:

$$-\frac{x^\rho_{t+1}n(x^\rho_{t+1}v^\rho_t(1-\rho)+2ny^\rho_t)}{x^\rho_t(n+x^\rho_t)}\frac{y^\rho_t-(1-\rho)v^\rho_t}{\left(y^\rho_t n+(1-\rho)v^\rho_t x^\rho_{t+1}\right)^2}<0,$$

where the inequality follows from the fact that, given Eq.(9), $v^\rho_t \leq y^\rho_t$. Hence, since x_{t+1} is increasing in c_{t+1}, it follows that $E[\overline{\pi}^{\rho,+}_{t+1} \mid$ no entry till $t]$ is decreasing in c_{t+1}. If $x^\rho_{t+1}=0$, then $E[\overline{\pi}^{\rho,+}_{t+1}]=\overline{\pi}_0$ since, here, equilibrium fully reveals the direction of each player's preference. Hence, since $c_t > 0$, it follows that $E[\overline{\pi}^{\rho,+}_{t+1}] < \overline{\pi}_0$ for all $t \geq 0$. Furthermore, if $x^\rho_{t+1}=x^\rho_t$, then $E[\overline{\pi}^{\rho,+}_{t+1}] < E[\overline{\pi}^{\rho,+}_t]$, and if $x^\rho_{t+1}=0$, then $E[\overline{\pi}^{\rho,+}_{t+1}] > E[\overline{\pi}^{\rho,+}_t]$. Hence, by continuity, there is a unique $\alpha^{\rho,+}_{t,\underline{c}} \in (0,1)$ such that if $c_{t+1}=\alpha^{\rho,+}_{t,\underline{c}}c_t$, then $E[\overline{\pi}^{\rho,+}_{t+1}]=E[\overline{\pi}^{\rho,+}_t]$.

2. Notice that $E[\overline{\pi}^{\rho,-}_{t+1} \mid$ no entry till $t]$ is given by:

$$1-\frac{x^\rho_{t+1}+n}{x^\rho_t+n}\int_{-n}^{0}[n+x^\rho_{t+1}+\textstyle\sum_{s=1}^{t+1}(x^\rho_{s-1}-x^\rho_s)(1-(1-\rho)^s)]^{-1}d\theta_{-i},$$

because any negative player becomes increasingly more convinced that her opponent has learned that she is negative, thus, stays out irrespective of his valuation. Differentiating the above with respect to x^ρ_{t+1}, one gets that:

$$-\frac{n}{n+x^\rho_t}\frac{(n+x^\rho_t)(1-(1-\rho)^{t+1})+\sum_{s=1}^{t}(x^\rho_{s-1}-x^\rho_s)(1-(1-\rho)^s)}{\left[n+x^\rho_{t+1}+\sum_{s=1}^{t+1}(x^\rho_{s-1}-x^\rho_s)(1-(1-\rho)^s)\right]^2}<0,$$

since x^ρ_s is weakly decreasing in s. Hence, $E[\overline{\pi}^{\rho,-}_{t+1} \mid$no investment till $t]$ is decreasing in c_{t+1}. It follows that $E[\overline{\pi}^{\rho,-}_{t+1}] \geq \overline{\pi}_0$ for all $t \geq 0$, where the inequality is strict iff $x^\rho_{t+1} < x$. Furthermore, if $x^\rho_{t+1}=x^\rho_t$, or equivalently, if $c_{t+1}=\alpha^{\rho,-}_{t,\underline{c}}c_t$, then $E[\overline{\pi}^{\rho,-}_{t+1}]$ $=E[\overline{\pi}^{\rho,-}_t]$; if $x^\rho_{t+1} < x^\rho_t$, then $E[\overline{\pi}^{\rho,-}_{t+1}] > E[\overline{\pi}^{\rho,-}_t]$. Finally, since if $x^\rho_{t+1}=x^\rho_t$, then $E[\overline{\pi}^{\rho,+}_{t+1}] < E[\overline{\pi}^{\rho,+}_t]$, it follows that $\alpha^{\rho,-}_{t,\underline{c}} > \alpha^{\rho,+}_{t,\underline{c}}$

[39] Since $c_t > 0$ for all t, the ex ante probability of a player not investing till the end of round t in equilibrium despite having a positive valuation is bounded away from zero.

Proof 6 (Proof of Corollary 3) *Iterating Eq.(10) and Eq.(9) from* $t = 1$ *on, it follows that if* $c_t \geq \frac{x^2}{n}(1-\rho)^t$ *for all* t*, then there is no entry in any* t*. Furthermore, it follows that* $\lim_{t\to\infty} E[\overline{\pi}_t^{\rho,+}] = 0$*. Since along this sequence* $x_t^\rho = x_{t+1}^\rho$ *it also follows that* $\lim_{t\to\infty} E[\overline{\pi}_t^\rho] = \lim_{t\to\infty}[\frac{n}{n+x}E[\overline{\pi}_t^{\rho,+}] + \frac{x}{n+x}E[\overline{\pi}_t^{\rho,-}]] = \frac{n}{n+x}\frac{x}{n+x} \leq \frac{1}{4}$ *where the last inequality follows from the fact that* $0 \leq (x-n)^2 = x^2 - 2xn + n^2$

Proof 7 (Proof of Proposition 4) *1. The projected version of* i *enters iff* $\min\{\theta_i, \theta_{-i}\} \geq 0$ *. Given any fixed strategy* σ_{-i}*, let* z_{-i} *be the true unconditional probability with which real* $-i$ *enters. For real* i *with a given valuation* $\theta_i > 0$*, the perceived expected utility difference between 'in' versus 'out' is:*

$$\rho(rx(\theta_i - f(\theta_i)) + \int_{-n}^{0} rg(\theta_i, \theta_{-i})d\theta_{-i}) + (1-\rho)(z_{-i}(\theta_i - f(\theta_i)) + (1-z_{-i})E[g(\theta_i, \theta_{-i}) \mid \sigma_{-i}(\theta_{-i}) = out]). \quad (11)$$

Differentiating the above with respect to θ_i*, one gets a strictly positive number since* $f' < 1$ *and* $g_1(\theta_i, \theta_{-i}) \geq 0$*, for any* $\theta_i > 0$*. Hence, equilibrium must be in cutoff strategies.*

Consider now the best-response function of real i*,* $\beta^\rho(\theta_{-i}) : [0, x] \to [0, x]$*. By the implicit function theorem, since Eq.(11) is continuously differentiable in* $\theta_{-i} > 0$*, the slope of* $\beta^\rho(\theta_{-i})$*, evaluated at some point* $(\widehat{\theta}_i, \widehat{\theta}_{-i})$*, is:*

$$\frac{(1-\rho)r(\widehat{\theta}_i - f(\widehat{\theta}_i) - g(\widehat{\theta}_i, \widehat{\theta}_{-i}) - \int_{-n}^{\widehat{\theta}_{-i}} g_2(\widehat{\theta}_i, \theta_{-i})d\theta_{-i})}{\rho r(x(1-f'(\widehat{\theta}_i)) + \int_{-n}^{0} g_1(\widehat{\theta}_i, \theta_{-i})d\theta_{-i}) + (1-\rho)(z_{-i}(1-f'(\widehat{\theta}_i)) + \int_{-n}^{\widehat{\theta}_{-i}} rg_1(\widehat{\theta}_i, \theta_{-i})d\theta_{-i})}$$

The denominator is strictly positive. The numerator is strictly negative if investments are substitutes, and strictly positive if investments are complements and $g_2 = 0$*.*

2. By the intermediate value theorem a symmetric equilibrium must exist because $h(\theta_{-i}) \equiv \beta^\rho(\theta_{-i}) - \theta_{-i}$ *is continuous with* $h(0) \geq 0$ *and* $h(x) \leq 0$*, and the players' best-response functions are mirror images of each other given the 45-degree line. If investments are substitutes,* $\beta^\rho(\theta_{-i})$ *is strictly decreasing and there is a unique symmetric equilibrium. If investments are complement,* $\beta^\rho(\theta_{-i})$ *is strictly increasing and all equilibria must be symmetric since* $\theta_i = \beta^\rho(\theta_{-i}) > \beta^\rho(\theta_i) = \theta_{-i}$ *cannot hold if* $\beta^\rho(\theta_{-i})$ *is increasing.*

3. Consider the comparative static with respect to ρ. Suppose that $(\theta_i^\rho, \theta_{-i}^\rho)$ constitutes a symmetric ρ-IPE. Since $g(\theta_i, \theta_{-i}) < 0$ if $\min\{\theta_i, \theta_{-i}\} < 0$, and $f(0) = 0$, it must be that $\theta_i^\rho, \theta_{-i}^\rho > 0$. Rewriting Eq.(11), one gets that an internal equilibrium cutoff must satisfy:

$$\overbrace{\rho[\int_0^{\theta_{-i}^\rho} r(\theta_i^\rho - f(\theta_i^\rho) - g(\theta_i^\rho, \theta_{-i}))d\theta_{-i}]}^{V} + \int_{\theta_{-i}^\rho}^{x} r(\theta_i^\rho - f(\theta_i^\rho))d\theta_{-i} + \int_{-n}^{\theta_{-i}^\rho} rg(\theta_i^\rho, \theta_{-i})d\theta_{-i}] = 0. \tag{12}$$

If investments are substitutes, Term V is strictly negative. Holding $(\theta_i^\rho, \theta_{-i}^\rho)$ fixed, the LHS of Eq.(12) is strictly decreasing in ρ. Hence, the unique symmetric equilibrium cutoff must increase in ρ.

To show the comparative static with respect to g^-, note first that Eq.(12) implies that a decrease in g^- increases the equilibrium cutoff and decreases $Pr(in)$. Note that $Pr(in)_{\theta_i}^\rho - Pr(in) = \rho(1 - Pr(in))$ if $\theta_i > 0$ while $Pr(in)_{\theta_i}^\rho - Pr(in) = -\rho Pr(in)$ if $\theta_i < 0$. Hence this value increases as g^- decreases. For any θ_i,

$$\overline{\pi}_0 - E[\overline{\pi}_{\theta_i}^\rho] = [Pr(in)_{\theta_i}^\rho - Pr(in)][E[\overline{\pi}_{\theta_i}^\rho | in] - E[\overline{\pi}_{\theta_i}^\rho | out]]. \tag{13}$$

Since both parts of the product on the RHS increase when g^- decreases, this then implies the comparative static since $\overline{\pi}_0$ is independent of g^-.

If investments are complements, Term V is strictly positive. Holding $(\theta_i^\rho, \theta_{-i}^\rho)$ fixed, the LHS of Eq.(12) is strictly increasing in ρ. Since $\theta_{-i}^+(\theta_i) = 0$ for any $\theta_i > 0$, and $\beta^\rho(0)$ is independent of ρ, an increase in ρ shifts $\beta^\rho(\theta_{-i})$ down. Since $\beta^\rho(0) > 0$ must hold, the lowest equilibrium cutoff, the first intersection of $\beta^\rho(\theta_{-i})$ with the 45-degree line, is decreasing in ρ. The second intersection, if it exists, is increasing in ρ since $\beta^\rho(\theta_{-i})$ is strictly increasing in θ_{-i}.

4. Since $\theta_i^\rho > 0$ must hold for each i, conditional and average false antagonism both follow from the proof of Proposition 2

Proof 8 (Proof of Proposition 5) *Suppose that the real seller names $p_s(q) = w(q)$ and the projected seller names $\overline{w}$. Let $q^* = w^{-1}(\overline{p})$. The projected buyer knows q hence is only indifferent between accepting or rejecting $p(q) = w(q)$, otherwise has a dominant strategy. Consider the real seller's incentive to deviate when his type is below q^* to a price weakly below $\overline{p}$. This is never beneficial as long as $w(a) \geq (1 - \rho)\overline{p} + \rho a$.*

Consider now $z(v)$, expressed in terms of the direct reporting of value v given $p_s(v) = w(v)$, such that $max_v z(v)w(v) + (1 - z(v))q$ is maximized at $v = q$ for each $q \geq q^*$. Solving this, as long as $m \neq 1$, one gets that $lnz(v) = -\int_{q^*}^{v} \frac{m}{(m-1)s+x} ds$ for any $v \geq q^*$ — where this $z(v)$ is the result of the probabilistic mixture of the randomization by the real and the projected buyer. Straightforward calculations show that this solution also satisfies the second-order condition. If $m = 1$, then $z(v) = e^{-(v-q^*)/x}$ for $v \geq q^*$. Note there are also no incentives for deviations below q^* by construction for types $q \geq q^*$. For types $q < q^*$, deviating upwards to a type $q' \geq q^*$, it follows that $w(q^*) \geq z(q')w(q') + (1 - z(q'))q^* > z(q')w(q') + (1 - z(q'))q$, hence, since $(1 - \rho)w(q^*) + \rho q \leq w(q)$ when $q < q^*$, such a deviation is not perceived to be profitable either.

Finally, the maximal seller revenue in this class of equilibria is given when the composition of $z(v)$, given as the $(\rho, 1 - \rho)$ probabilistic mixture of the acceptance probabilities of the projected and the real buyer, is such that the projected buyer provides the minimal necessary acceptance probability given that the real buyer provides the maximal possible one. As ρ increases the latter can then increase and tend to 1 as $\rho \to 1$.

Proof 9 (Proof of Proposition 6) *1. If $p_b^0 < \overline{q}$, by revealed preference, the buyer's perceived payoff-maximizing offer, conditional on it being strictly below $\overline{q}$, is still p_b^0. The benefit of increasing it weakly above $\overline{q}$ increases with ρ, and $p_b^{\rho \to 1} \to \overline{q}$ since $\overline{w} > \overline{q}$. 2. If $p_b^0 > \overline{q}$, then again, by revealed preference, the buyer never wants to raise the price above p_b^0, and again, $p_b^{\rho \to 1} \to \overline{q}$.*

Proof 10 (Proof of Corollary 4) *1. If $m = 0$, then $BNE = CE = min\{(a + x)/2, b\}$ and $PE = (a + b)/2$ since $\overline{q} < \overline{w}$ by assumption.*

2. If $m < 0$, then $x > 0$ and $w(a) > a$ must hold given $\overline{q} < \overline{w}$ thus all models predict a bid weakly above a. Here, $BNE = min\{((a + x)/(2 - m)), b\}$, and $CE = min\{(a + \overline{w})/2), b\}$. Suppose first that both BNE and CE are internal. Then $CE - BNE = (m/(2(2-m)))(b-\overline{w})$ while $PE-BNE = (2/(2(2-m)))(b-\overline{w})$. If instead $CE = b$, then $a + \overline{w} \geq 2b$, but then $a + x > (a + b)(1 - \frac{m}{2})$, thus $BNE > \overline{q} = PE$. Finally, if $BNE = b$, then $a + x \geq 2b - mb$. Suppose now that here $CE < b$, but then $a + x < 2b - m(\frac{a+b}{2})$ also needs to hold which is impossible since $m < 0$, hence $CE = b$ then too.

3. If $m > 0$, consider first the case where $m \geq 2$. Here, either $BNE = b$ when $\overline{w} \geq b$ or $BNE < a$ when $\overline{w} < b$ and the statement follows in both cases (here we assume that if $BNE < a$, then it is held constant across comparisons). Consider now the case that $m < 2$. Suppose now that $w(a) > a$, then all models predict a bid weakly above a. Assume first that BNE is internal. This implies that CE is internal too. Then $CE - BNE = (m/(2(2-m)))(b-\overline{w})$ while $PE - BNE = (2/(2(2-m)))(b-\overline{w})$ and the statement follows. Consider now the case where $BNE = b$. If $CE = b$, then the statement follows. If $CE < b$, then $a + \overline{w} < 2b$, but if $BNE = b$, then $\overline{w} \geq b$ must hold by interim individual rationality and the statement follows again. Finally, if $w(a) \leq a$, then, either $BNE \leq a$, or $BNE = min\{((a+x)/(2-m)), b\}$. In the former case $\overline{w} < b$ must hold thus $|BNE - CE| = (a+\overline{w})/2 - BNE < |BNE - PE| = (a+b)/2 - BNE$ must hold too. In the latter case, the same argument applies as above.

For Online Publication

Appendix C Alternative Models and Mechanisms for Experimental Evidence

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 unbiased BNE. A cursed principal should never exaggerate the agent's performance, on average, and a cursed agent should never anticipate any systematic misprediction by the principal, on average. Instead, we find both patterns and do so explicitly by eliciting beliefs directly. The same applies to the model of Esponda (2008). Note, that QRE (McKelvey and Palfrey, 1995) 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.[40]

Risk Aversion. We find no evidence that risk aversion matters for the subjects' choices. (See Tables E.1 and D.1 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 underestimate

[40]Note also that, in contrast to the defining feature of the model of biased but coherent social beliefs, under a cursed equilibrium players' behavior need not be consistent with a coherent belief hierarchy.

the success rate of others 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 beliefs. 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.

Appendix D Supplementary analysis

D.1 Stated beliefs of the principals

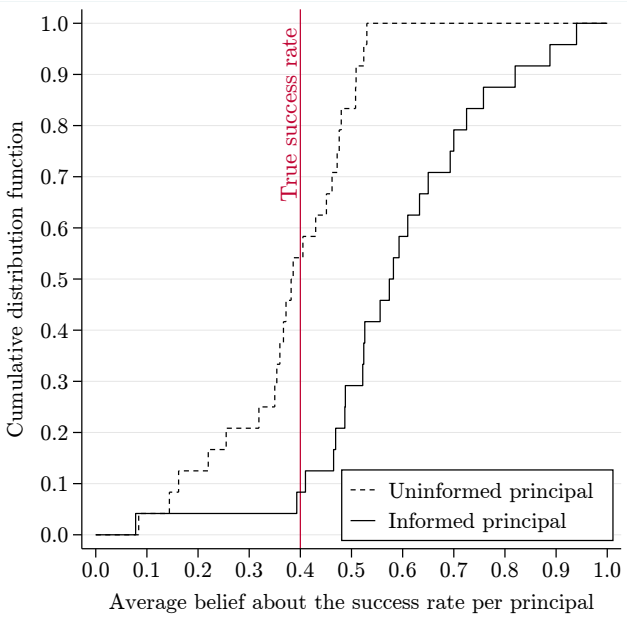

Figure D.1: Distribution of average first-order belief per principal in the informed and the uninformed treatment. A Kolmogorov-Smirnov test of the distributions of average individual beliefs between treatments yields $p < 0.001$.

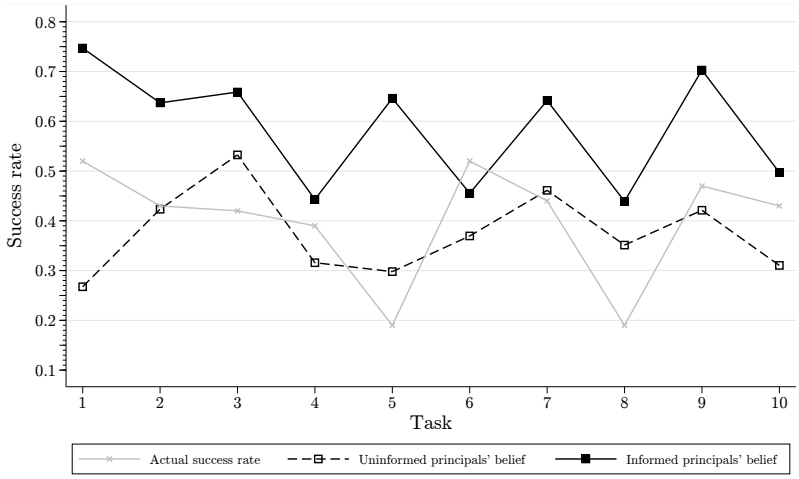

Figure D.2: Average belief of informed and uninformed principals and actual success rate per task.

D.2 Stated beliefs of the agents

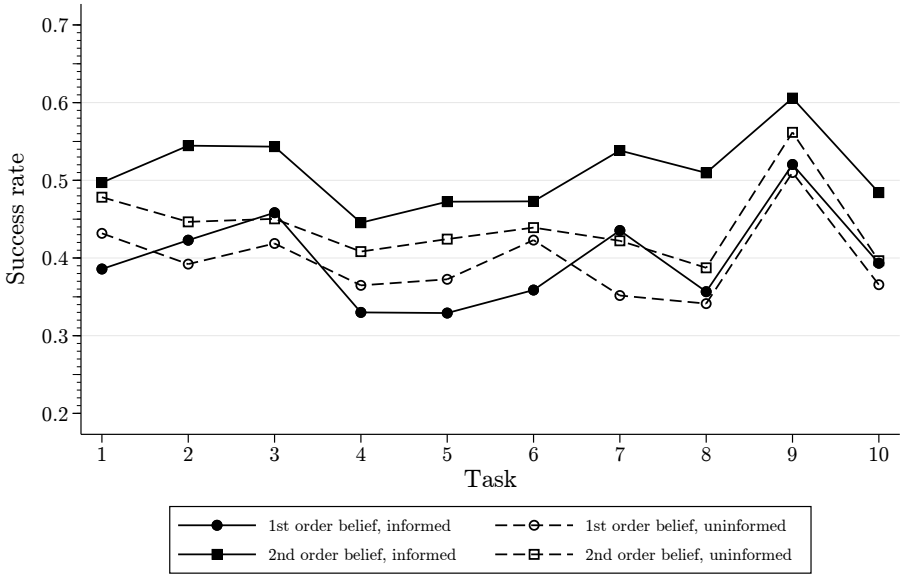

Figure D.3: Agents' first-order belief (guess of success rates) and second-order belief (guess of principals' belief) by task and treatment.

Table D.1: Mean individual differences in second-order belief (guess of principal's belief) and first-order belief $b^A_{1,i}$ (guess of success rate) by treatment and further controls.

Dependent variable	$(b^{II}_{A_i} - b^{I}_{A_i}) = L^{-1}\sum_t (b^{II}_{A_i l} - b^{I}_{A_i l})$				
(OLS)	(1)	(2)	(3)	(4)	(5)
Treatment	0.068***	0.067***	0.089***	0.073***	0.090***
(1-informed)	(0.019)	(0.019)	(0.024)	(0.020)	(0.024)
Gender		0.013	0.047		0.045
(1-female)		(0.020)	(0.029)		(0.030)
Treatment×Gender			−0.062		−0.056
			(0.040)		(0.041)
Coef. risk aversion				−0.006	−0.004
(DOSE)				(0.006)	(0.006)
Constant	0.044***	0.040**	0.030*	0.048***	0.034*
	(0.014)	(0.015)	(0.016)	(0.014)	(0.017)
N	47	47	47	47	47
R^2	0.220	0.228	0.270	0.236	0.278
F	12.720	6.490	5.289	6.787	4.035

Note: Values in parentheses represent standard errors. Asterisk represent p-values: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

D.3 Test of Equivalence

This section provides a structural test of the hypothesis of equivalence of $\rho_A = \rho_P$. projection. We use a random-coefficient model to capture general heterogeneity in projection bias in the population. In the unrestricted model, the distributions of projectivity can be different for principals and agents. The parameters of the unrestricted model are $\Theta_{UR} = \{\rho_P, \rho_A, \lambda_\rho, \lambda_b\}$, where ρ_P and ρ_A denote the mean degree of projection bias in the principal and agent population, respectively, and λ_ρ and λ_b are role-unspecific precision parameters governing the variance of individual projection and noise in response, as we detail below. We also estimate the model under *equivalent projection*—i.e., with restricted parameters $\rho_P = \rho_A$, i.e., the parameters of the model are $\Theta_R = \{\rho, \lambda_\rho, \lambda_b\}$. A comparison of the restricted and the unrestricted specification provides a test of the proposed equivalence.

Since our variables of interest (and their first moments when modeled as random

coefficients) have bounded support on $[0,1] \subset \mathbb{R}$, our econometric model makes repeated use of the beta distribution. For a straightforward interpretation of estimated parameters, we will use Ferrari and Cribari-Neto's (2004) parameterization of the beta distribution $x \sim \text{Beta}(\mu, \lambda)$ with density

$$f(x;\mu,\lambda) = \frac{\Gamma(\lambda)}{\Gamma(\lambda\mu)\Gamma(\lambda(1-\mu))} x^{\lambda\mu-1}(1-x)^{\lambda(1-\mu)-1}, \tag{14}$$

where the first parameter μ is the expected value of x, and the second parameter λ is a precision parameter that is inversely related to the variance of x, $\text{var}(x) = \mu(1-\mu)/(1+\lambda)$.[41] That is, conditional on the mean μ, higher values of λ translate into a lower variance.

We now specify two basic structural assumptions of the econometric model. 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. We can therefore derive only the ex ante expected biased estimate. However, we also 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 l are:

$$\begin{aligned} b^I_{P_i l} &\sim \text{Beta}(\mu_{P_i l}, \lambda_b), \\ b^{II}_{A_j l} &\sim \text{Beta}(\mu_{A_j l}, \lambda_b), \\ &\text{where} \\ \mu_{P_i l} &= \rho_{P_i} + (1-\rho_{P_i})\phi_l, && \text{(see Claim 1)} \\ \mu_{A_j l} &= \phi_l + (1-\rho_{A_j})\rho_P(1-\phi_l). && \text{(see Claim 2)} \end{aligned} \tag{SA1}$$

Our second structural assumption serves to capture individual heterogeneity in

[41] The standard $Beta(\alpha, \beta)$ parameterization is obtained by setting $\mu = \alpha/(\alpha+\beta)$ and $\lambda = \alpha+\beta$ in (14).

the degree of projection bias and accounts for repeated observations on the individual level. To this end, we use a random-coefficient model in which individual degrees projection in the principal and the agent populations follow a beta distributions with

$$\begin{aligned} \rho_{P_i} &\sim \text{Beta}(\rho_P, \lambda_\rho), \\ \rho_{A_j} &\sim \text{Beta}(\rho_A, \lambda_\rho). \end{aligned} \tag{SA2}$$

We impose a restriction 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.[42]

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

$$L_{k_i}(\rho_{k_i}, \lambda_b) = \textstyle\prod_l f_b\left(b_{k_i l}; \mu_{k_i l}(\rho_{k_i}), \lambda_b\right).$$

Hence, the unconditional probability amounts to

$$L_{k_i}(\rho_k, \lambda_\rho, \lambda_b) = \int \left[\textstyle\prod_l f_b\left(b_{k_i l}; \mu_{k_i l}(\rho_{k_i}), \lambda_b\right)\right] f_\rho\left(\rho_{k_i}; \rho_k, \lambda_\rho\right) d\rho_{k_i}. \tag{15}$$

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

$$l(\rho_P, \rho_A, \lambda_\rho, \lambda_b) = \textstyle\sum_k \sum_i \log L_{k_i}(\rho_k, \lambda_\rho, \lambda_b). \tag{16}$$

We estimate the parameters in (16) by maximum simulated likelihood (Train, 2009; Wooldridge, 2010).[43] Table D.2 shows the estimation results for the unrestricted model ($\rho_P \neq \rho_A$) in the left column and the restricted model with ($\rho_P = \rho_A$) in the

[42] We tested this assumption ex post by comparing the estimated beta distribution with the empirical distribution of individual estimates of projection bias in Figure 2. We found no significant differences between the estimated beta and the empirical distribution of individual estimates, neither for the principals ($p = 0.996$ from K-S test) nor for the agents (0.149), nor when pooling participant roles ($p = 0.444$).

[43] The estimation is conducted with GAUSS, Aptech Systems. 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).

Table D.2: Maximum likelihood estimates of projection bias ρ based on Claim 1 and 2.

	Unrestricted model with heterogeneous ρ ($\rho_P \neq \rho_A$)		Restricted model with homogeneous ρ ($\rho_P = \rho_A$)	
Parameter	Estimate	Conf. interval	Estimate	Conf. interval
ρ_P	0.340^{***}	$[0.257, 0.423]$	0.337^{***}	$[0.262, 0.413]$
ρ_A	0.354^{***}	$[0.134, 0.574]$		
λ_ρ	2.741^{***}	$[0.962, 4.520]$	2.717^{***}	$[0.971, 4.463]$
λ_b	5.348^{***}	$[4.673, 6.023]$	5.347^{***}	$[4.672, 6.022]$
N	480		480	
$\ln L$	95.601		95.588	

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 $p = 0.8389$.

right column.[44,45]

Focusing on the unrestricted model first, we make three observations. First, the principals' average degree of projection is estimated to be $\widehat{\rho}_P = 0.340$, with a confidence interval of $[0.257, 0.423]$. This estimate indicates the relevance of projection bias: the unbiased BNE, which is the special case of $\rho_P = 0$, is clearly rejected. Second, the agents' average degree of projection bias is estimated to be 0.354, with a confidence interval of $[0.134, 0.574]$. The $\widehat{\rho}_A = 0.354$ 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, under-anticipate the principals' level of projection.

Crucially, the estimated parameters of the degree of projection are not significantly different between the principals and the agents ($p = 0.869$). Furthermore, the log likelihood of the unrestricted model is very close to that of the restricted model (bottom row of Table D.2), and standard model selection criteria (e.g., Bayesian

[44]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 (16) appears to assume a global (and unique) maximum at the estimated parameters.

[45]The results are very similar when including tasks for which only the principals' beliefs where elicited (a second set of ten tasks where the agents made investment decisions instead of stating their beliefs, see Appendix E).

information criterion) favor the parsimonious model of homogeneous projection over the unrestricted model. In short, the data are consistent with the structure of biased beliefs implied by projection equilibrium, that is, a joint account—and a common source—of the basic mistake and the mistake in the anticipation of this basic mistake in others.

Appendix E Investment decisions of the agents

We collected further choice data in addition to the belief data that has been in the focus of our analysis. Following the ten tasks with belief elicitation, for a second set of ten change-detection tasks, the agent decided between two investment options, A and B. Option A provided a sure payoff of EUR 4. Option B was a lottery where the agent received EUR 10 if the principal's estimate was not more than ten percentage points higher the true success rate; otherwise, the lottery paid EUR 0. This binary choice is, implicitly, 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 subjective expected return is decreasing in the wedge between the agent's second- and first-order belief. We will refer to this choice as the agents' investment decision. The agents were paid based on a randomly chosen round out of twenty. If it was a round involving belief elicitation, they were paid in the manner described in Section 3.1. If it was a round involving the investment decision, they were paid based on the scheme described above.

We also ran separate sessions with investment choices only, 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 (i.e., instead of belief elicitation, the agents made the investment choice also for the first ten tasks of the session). Figure E.1 shows the distribution of individual investment rates in the informed and the uninformed treatment. Agents with informed principals invested at a significantly lower rate (39.2%)than agents with uninformed principals (67.3%, $p < 0.001$).[46]

The agents in the informed treatment, relative to agents in the uninformed treat-

[46] We pool the sessions with belief elicitation and those without. There is no significant difference in investment rates between sessions with and without belief elicitation (t-tests yield $p = 0.756$ and $p = 0.699$ for the informed and uninformed treatment, respectively) and the treatment difference in the investment rate is significant also when focusing on the sessions w/o belief elicitation (investment only; $p = 0.008$) or when focusing on tasks that were used for belief elicitation in other sessions (first part of sessions with investment only; $p = 0.013$). There are no significant time trends in the investment rates.

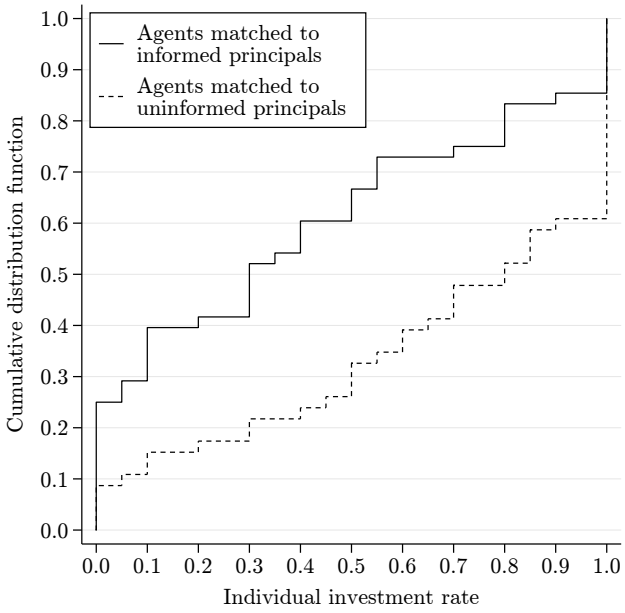

Figure E.1: Distribution of individual investment rates in the informed and the uninformed treatment.

ment, shy away from choosing an option whose payoff decreases—in the sense of first-order stochastic dominance—in the principal's belief. This finding 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 their principal had access to the solution, while the agent in the uninformed treatment was told that their 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 beliefs.[47]

[47]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.

Table E.1: Regressions of individual investment rates on treatment, gender, and risk attitude.

Dependent variable	Individual investment rate				
(OLS)	(1)	(2)	(3)	(4)	(5)
Treatment	-0.281^{***}	-0.279^{***}	-0.255^{**}	-0.259^{***}	-0.254^{**}
(1-informed)	(0.075)	(0.075)	(0.102)	(0.077)	(0.102)
Gender		-0.059	-0.032		-0.048
(1-female)		(0.075)	(0.108)		(0.109)
Treatment×Gender			-0.053		-0.009
			(0.151)		(0.157)
Coef. risk aversion				-0.026	-0.024
(DOSE)				(0.022)	(0.023)
Constant	0.673^{***}	0.698^{***}	0.687^{***}	0.695^{***}	0.715^{***}
	(0.053)	(0.063)	(0.071)	(0.057)	(0.076)
N	94	94	94	94	94
R^2	0.134	0.140	0.141	0.147	0.151
F	14.230	7.390	4.920	7.813	3.960

Note: Values in parentheses represent standard errors. Asterisks represent p-values: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.