2016

Projection Equilibrium: Definition and Applications to Social Investment, Communication, and Trade (revised)

Kristof Madarasz

Projection Equilibrium: Definition and Applications to Social Investment, Communication, and Trade

Kristóf Madarász[1]

First version: May 2013. This version: December 2016.

[1]First online version July 2014. I would like to thank seminar audiences at Arizona, Columbia, Harvard, Princeton, Yale, UC San Diego, UC Los Angeles, Utah, Wash U, Bonn, CEU, Essex, UCL, LSE, Royal Holloway, Stockholm, Southampton, ITAM, Berlin Behavioral Seminar 2011, European Behavioral Economics Meeting Berlin 2013, ESSET Gerzensee 2013, SITE 2015, Roland Bénabou, Peter Bossaerts, Colin Camerer, Jeff Ely, Marina Halac, Philippe Jehiel, Navin Kartik, Wolfgang Pesendorfer, Drazen Prelec, Matthew Rabin, Luis Rayo, Larry Samuelson, Joel Sobel, Balazs Szentes, Adam Szeidl, Tomasz Strzalecki, Jörgen Weibull for comments. All errors are mine. Contact: k.p.madarasz@lse.ac.uk; LSE, Houghton Street, London, UK.

Abstract

I develop a simple model of egocentric thinking for Bayesian games. In risky cooperation settings, people under-invest in relationships and too often infer antagonistic preferences. When no one supports a rule, no one speaks up, but everyone concludes that everyone else supports it. Arbitrarily unpopular rules gradually appear as arbitrarily popular even if dissent is almost free. In strategic communication, advice is deceptive. The easier the receiver could check the sender's lie, the more credulous he becomes. Financial education boosts credulity and lowers receiver welfare. In common-value trade, sellers underbluff, buyers are cursed, and the model closely matches existing data.

Keywords: Theory of Mind, Perspective Taking, Free Speech, False Antagonism, Paranoia, Credulity, Financial Literacy, Trade.

1 Introduction

Perspective taking is a key aspect of strategic behavior. While the usual assumption is that people fully appreciate differences in perspectives that arise due to differences in information, the evidence shows that the typical person too often acts as if others shared her perspective and knew what she did. Since such egocentric thinking — empathy gaps in informational perspective taking — is likely to shape behavior in settings commonly analyzed via Bayesian games, this paper incorporates this phenomenon into strategic settings and explores its relevance both theoretically and empirically.

Evidence for informational projections dates back to Piaget and Inhelder (1948) studying children's 'theory-of-mind', that is, their (limited) ability to attribute different beliefs to others than what they themselves hold. In a classic study, Wimmer and Perner (1983) demonstrate that young children too often act as if lesser-informed others shared their informationally superior perspectives. Birch and Bloom (2007) show that the exact same mistake is present among Yale undergraduates in slightly more complex tasks.[1] Such robust phenomena as the curse of knowledge (Camerer et al., 1989), the hindsight bias (Fischhoff, 1975; Biais and Weber, 2008), the outcome bias (Baron and Hershey, 1988), and the illusion of transparency (Gilovich et al., 1998) all support the idea that people project their private information onto others. Madarász (2012) reviews some of the evidence and develops the notion of information projection for pure inference problems where people exaggerate the probability that if they have observed a signal so did others.

Incorporating egocentric thinking into strategic models, however, raises an issue entirely absent from inference problems. In a game, Judith's beliefs about Paul's information entail her beliefs about Paul's beliefs about her information. In poker, a player may wrongly believe that others know her hand, but what matters is also whether she thinks others realize that she thinks this way and whether she thinks that others wrongly think that she may know their hands. Specifying self-referential perceptions is, thus, essential. A general model of projection must then tie together such higher-order perceptions and, given a clear empirical motivation, do so in a relatively parsimonious fashion.

While information projection implies biased views about the information of others, it leaves open the question of just how biased a player's view of her opponent's view of herself is. Each fully anticipating the biased views of others is clearly incompatible with the very idea of projection. Not anticipating others' biased views at all, however, may mean that people's expectations are fully misspecified vis-a-vis how others actually behave. While in Appendix A I present such a naive case where projection is private, by

[1] As Epley et al. (2004, pp. 327) also point out "Piaget recognized, adults come to view the world less egocentrically than children, although they do not outgrow their childhood tendencies altogether." The experimental paradigm employed by Wimmer and Perner (1983) is often referred to as the 'false-belief task,' see e.g., Baron Cohen et al. (1985) linking below average performance on this task to autism and Frith and Corcoran (1996) to schizoid paranoia.

tying together the extent to which people project onto others and the extent to which they fail to anticipate others' projections onto them, the main model offered by this paper achieves a tight balance between these two extremes. Here, projection satisfies an intuitive property of being all-encompassing and the extent to which Judith projects onto Paul is proportional to the extent to which she underestimates Paul's projection onto her. After presenting the model, I turn to three distinct applications.

Social Investments. In many settings, such as partnerships in trade or the formation of social and political associations**,** the return on investing with someone depends on one's partner's privately known preference for reciprocating investment. To fix ideas, consider a dating example. Judith and Paul are sitting at a bar. Each is privately informed about his or her interest in a match, and each can decide whether to make a move (invest). If both do, a match is formed. If neither does, each gets the outside option. If only one does, the other accepts if interested. Otherwise he rejects and the proposer incurs a loss.

An analogous situation arises when two members of an organization meet and each needs to decide whether to speak up against a prevailing norm or rule (invest) — such as a homophobia, a corporate practice, or Stalin's rule of the Party — or stay silent and act loyal towards the status quo, each being privately informed whether she opposes the rule, hence, would benefit from forming a bond with someone who also opposed the rule, or is loyal to the rule, hence, would not want to dissent, but instead would seek to punish explicit dissent, e.g., by reporting it to a central authority.

In such settings while Judith's willingness to make a move (dissent) increases in her confidence that Paul would welcome her move, it decreases in her confidence of how confident Paul is that she would welcome a move by him. If Judith is interested in Paul, then by projecting her private information onto Paul, she exaggerates Paul's incentive to make a move in case he is also interested. As a consequence, she finds it relatively more important to stay out and avoid a potential rejection by him. In a symmetric situation Paul reasons similarly. An increase in projection then increases each interested party's expectation that the other party will invest which, in turn, decreases actual investments.

Two systematic inferential mistakes accompany such underinvestment. First, if Judith is interested in Paul, she comes to underestimate Paul's interest in her given any contingency arising in equilibrium. Second, on average, each player comes to exaggerate the probability with which the preferences of others point in the opposite direction as her own. If Judith is interested in Paul, she too often concludes that he is not interested in her. If she is not interested in him, she too often concludes the reverse. Similarly, if Judith opposes Stalin, she attributes Paul's silence too much to his genuine loyalty and on average exaggerates the probability that Paul supports Stalin. If instead she supports Stalin, she attributes Paul's silence too much to his fear, and on average exaggerates the probability that Paul secretly opposes Stalin.

In the context of repeated encounters these effects jointly imply that even if the

potential loss from speaking up goes to zero over time, provided it does so sufficiently gradually, none ever speaks up, but all those who oppose the rule conclude that they are alone with their preferences. It is then exactly when none supports the rule that everyone concludes that everyone else supports it. Furthermore, the majority of the group always concludes on average that the majority supports the status-quo rule independent of the margin by which this is true or false in reality.

Based on this mechanism, I then consider an organization whose leadership wants to ensure that its members act loyal towards a given rule. I show that even if this rule is arbitrarily unpopular and intensely disliked, that is, for most members of the organization the benefit from deviating or dissenting from this rule in front of each other is very high, the leadership can still always secure the illusion whereby this rule appears arbitrarily popular to all those who oppose it. Furthermore, it can do so even if the extent to which it can suppress dissent, that is, even if the expected punishment of a dissenter ensuing dissent in front of a loyalist, is arbitrarily small, e.g., even if at any point in time speech is almost free.

If the expected punishment for dissenting in front of a loyalist is small, then introducing a sufficiently unpopular rule directly still causes most opponents of this rule to speak up and reveal the rule's unpopularity despite projection. At the same time, by moving towards such a rule sufficiently gradually over time, the organization can always achieve the above objective.

Starting with a rule that is at most mildly disliked by anyone, no one finds it worthwhile to dissent. Given projection, however, all those who oppose this rule underestimate others' opposition to it. Exploiting such underestimation, the organization can then replace the initial rule with a more radical one, one that is commonly known to be somewhat more intensely disliked by all those who already disliked the initial rule, and still ensure that no one dissents. Proceeding along this line sufficiently gradually, while people's dislike of the subsequent rules increases, their perception of others' opposition to these rules decreases. Hence, still no one dissents. Unless speech is absolutely free, this slippery slope eventually guarantees that all prefer to publicly endorse a rule that privately almost all may immensely hate. I conclude by linking the model's predictions to various pieces of evidence, e.g., a phenomenon termed pluralistic ignorance, Prentice (2007).

Advice. In Section 4, I apply the model to strategic communication. Under general Bayesian assumptions, e.g., Crawford and Sobel (1982), by virtue of unbiased inference, two general properties characterize communication: receivers must benefit from advice and are never fooled by it *on average.* At the same time, the evidence suggests that receivers are persuaded too easily and may systematically lose from access to advice, e.g., Bergstresster et al. (2009). In an environment with a commonly known distribution of the payoff relevant state egocentric thinking indeed implies a systematic violation of both of these properties.

A sophisticated sender tells a client whether a statement is true or false. The sender

has an incentive to claim that the statement is true, but before making a decision, the client can check, at a cost privately known to the client, whether the sender lied. If the client checks, he learns the truth, and if the sender lied, the sender suffers a loss. A projecting client interprets the sender's message in a too egocentric manner. If his realized cost of checking is relatively high, he exaggerates the sender's incentive to lie, and is dissuaded by advice. If his realized cost is relatively low, he underestimates her incentive to lie, and is persuaded too easily. Nevertheless, both if the conflict between the parties is not too high (but not too low either) or if the ex ante distribution of checking costs is not too high (but checking is also not too trivial), the model predicts uniform credulity: each client type is successfully deceived in that each type's ex ante expected posterior confidence in the statement being true is inflated above the prior.

Key, however, is not simply the emergence of credulity and disbelief in advice, but given an endogenous mechanism, the comparative static predictions on how the underlying environment affects the extent of such mistakes and the client's welfare. In particular, in the wake of the Great Recession many have argued that financial education was the essential tool to improve clients' financial outcomes and how much they benefit from financial advice, e.g., Lusardi (2013), Dodd-Frank Act (2010). Yet the evidence on such a channel being effective is mixed at best, e.g., Hastings et al. (2013). Evidence further shows that those who are successfully persuaded to buy into fraudulent investment schemes have higher financial literacy on average than non-victims, NASD Fraud Report (2006).

Contrary to such common wisdom, but consistent with these observations, the comparative static predictions of the model imply that while financial education always helps an unbiased client, in the biased case it is precisely valuable financial literacy which allows advice to be deceptive. The lower is the ex ante distribution of checking costs, the greater is the scope for advice to be deceptive. In a large class of environments the easier it is for the client to fact check the sender on average, the higher is the client's financial literacy, the greater is the client's equilibrium credulity and the lower is his welfare. Relatively easy to check lies are the most deceptive.

When endogenizing the conflict between the sender and the client, by invoking the seller of the asset, I also show that if selling is sufficiently profitable, advice is always deceptive for any positive degree of the conflict. As the mistake becomes full, not only is the client's welfare always lower with advice than without, but the easier it is on average for the client to fact check the sender, the lower is the client's welfare and the higher is the seller's profit.

Projection equilibrium. In Section 5 I generalize the model to incorporate also the combined presence of information projection — the underappreciation of the positive side of the information gap — and *ignorance projection* — the underappreciation of the negative side of this gap. I derive implications to the classic problem of common-value trade, Akerlof (1970). Consistent with the experimental evidence – e.g., Samuelson and Bazerman (1985), Holt and Sherman (1994) – when a privately informed seller

has the bargaining power, the sender bluffs too little, and when the uninformed buyer has the bargaining power, the buyer ignores selection and is cursed. I compare the model's prediction with the data and the predictions of BNE and cursed equilibrium, Eyster and Rabin (2005).

This paper incorporates a fundamental egocentric mistake into strategic settings. Despite its simplicity and tight specification, the model provides novel insights and offers a more unified explanation of a variety of empirical observations from distinct domains. Section 6 concludes discussing direct evidence for the model's assumptions, Danz, Madarász, and Wang (2015), and further applications.

2 Setup

This section develops the model. For ease of exposition, I 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 an associated strictly positive prior $\pi \in \Delta\Omega$. 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 $\Gamma = \{\Omega, \pi, P_i, A_i, u_i\}$.

To introduce information projection, consider the following correspondence:

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

describing the coarsest common refinement of the two players' partitions. If an event, $E \subseteq \Omega$, 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. The formulation will imply that a person who projects her information has an exaggerated belief that whenever she can condition her strategy on an event, so can her opponent.[2]

To incorporate information projection in a parsimonious manner, I distinguish between the regular and the projected versions of each player i. The *regular* 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 i's and j's joint information, that is, she chooses

[2] Since a state encodes all payoff relevant information, in a common-value environment with private signals a player may then also project information about her opponent's expected preferences. If common values are positively related, she may then exaggerate the closeness of the expected preferences, if they are negatively related, she may instead exaggerate their distance.

a strategy from the set:

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

All real players are regular. Below, the operator $\circ$ denotes the mixture of two lotteries, BR 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 opponents' strategies.

2.1 Definition

The main approach where projection is partially public, adopted throughout this paper, is motivated by an alternative specification where projection is *private*. Under private projection players fully fail to anticipate each other's projections. Private projection describes a more radical departure from equilibrium logic and violates the two key properties of the main approach. Crucially, since, as discussed in more detail in the conclusion, Danz, Madarász and Wang (2015) provide direct evidence for the model's prediction that people not only project, but act as if they partially anticipated each other's projections, and since the main model is able to incorporate projection in a way that stays close to the standard equilibrium logic, to provide a more focused presentation, I describe the private alternative only in Appendix A.

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

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

and

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

■ If $\rho = 0$, players form correct expectations about each other's strategies and the predictions of the model collapse to those of BNE in any given game Γ. If $\rho > 0$, each player i mistakenly assigns probability ρ to her opponent best responding to her *true* strategy conditioning it on his and her joint information in the game. She assigns the remaining probability $1-\rho$ to her opponent playing the strategy he truly plays. Let me describe the two key properties of the model.

■ First, projection is *all-encompassing*. Judith acts as if she believed that the projected version of Paul knew her perception of the distribution of information in the game. In poker, the projected version of Paul is believed to know that Judith does not know his hand, that Judith thinks that he might know her hand, etc. Since a player always knows what she herself knows, this assumption is consistent with the very idea of information projection.

■ Second, each player's expectation is *consistent*, in a limited way, with feedback. Nothing happens in the game that would explicitly contradict a player's theory of how

her opponent may behave. Instead each player expects something to happen that may never happen or happen with a different probability than expected.

■ These properties jointly imply that in equilibrium each player acts as if she partially anticipated, but *underestimated* her opponent's projection. Judith estimates Paul's average projection onto her to be $(1-\rho)\rho$, thus, underestimating by ρ^2 its true extent. The same structure of underestimation holds along the belief hierarchy but its extent decreases and eventually vanishes.[3]

■ The definition immediately extends to heterogenous projection. Here, a different ρ_i replaces ρ in Eq.(1) for each i. If $\rho_i = 0$, player i is sophisticated and fully anticipates her opponent's projection. Otherwise she estimates it to be $(1-\rho_i)\rho_j$ underestimating it proportional to the degree of her own projection ρ_i.

2.2 Example 1: Zero-sum Games

To illustrate, consider a hide-and-seek game, such as serving in tennis or a military operation, with two locations, A and B, and two ex ante equally likely states, $\omega \in \{\omega_s, \omega_w\}$, such that a strong defender, $\omega_s = 0$, wins iff the players pick the same location, while a weak defender, $\omega_w = w$, wins only with probability $1-w$ even if players both pick location A. When the defender is weak A is her Achilles heel. Only the defender knows whether she is strong or weak.

Overplaying the Achilles Heel. The next table describes the game, as well as, the defender's strategy and expected payoff in the unbiased and biased cases:

attacker/defender	A	B	if	weak	strong	$\text{EU}_{\text{D}}^{\rho}$
a	$\omega, 1-\omega$	$1,0$	$\rho = 0$	B	A	$\frac{1}{2}$
b	$1,0$	$0,1.$	$\rho > \frac{w}{2-w}$	$\frac{1}{2-w}\text{A} \circ \frac{1-w}{2-w}\text{B}$	$\frac{1}{2}\text{A} \circ \frac{1}{2}\text{B}$	$\frac{1}{2} - \frac{w}{4(2-w)}$.

In the unbiased case, the defender optimally hides behind her private information; she never protects her Achilles heel (and prefers strictly not to). Hence, her payoff is independent of w. A biased defender instead too often thinks that the attacker has figured out her strength and that he knows that she thinks this way. She then overprotects her Achilles heel: she protects A strictly more often when she is weak than when she is strong. While a biased attacker also projects, since he expects the projected defender to optimally hide behind her private information, he mixes symmetrically both in the unbiased and in the biased cases.[4]

Choking. The above points to a choking effect. If players were symmetrically uninformed about ω, the defender's chance of winning would be $\frac{1}{2} - \frac{w}{4(2-\frac{1}{2}w)}$ independent of ρ. Hence, in the unbiased case the defender *gains*, and does so increasingly in w,

[3] Formally, one can construct the (real) players' iterative higher-order expectations about say player j being a projected version. The k-th element of this iteration is $\sum_{s=1}^{k}(-1)^{s-1}\rho^s$. The sub-sequence of odd elements, referring to expectations of real i, is decreasing in k. The sub-sequence of even elements, referring to expectations of real j, is increasing in k. Both converge to $\rho/(1+\rho)$, and the adjacent discrepancy, which is always ρ^k, vanishes as k increases.

[4] The projected attacker plays $\{\frac{1}{2}a \circ \frac{1}{2}b\}$ if ω_{s}, and $\{ha \circ (1-h)b\}$ if ω_{w} where $h = \frac{1}{2-w} + \frac{(1-\rho)w}{2\rho(2-w)}$.

in the biased case she *loses*, and does so increasingly in w, from access to private information.

2.3 Discussion

In contrast to the private approach to projection described in the Appendix, the model of ρ-IPE has a straightforward heterogenous prior equilibrium interpretation whereby each player commits a specific egocentric mistake and misperceives the true distribution of information in the game; believes that with probability ρ her opponent becomes his all-encompassing projected version, but she herself may never become such a version.

The next proposition ensures existence and asserts that the model delivers differential predictions from the unbiased case only if players are asymmetrically informed. Furthermore, it shows that while a BNE that is also an ex post equilibrium — an equilibrium from which no player has an incentive to deviate even after observing the state ex post — often does not exist, when it does, it is immune to information projection.

Proposition 1 *1. For any* Γ *and* ρ*, a* ρ*-IPE exists. 2. If* $P_i = P_j$*, then the set of* ρ*-IPE is independent of* ρ*. 3. If* σ^0 *is an ex post equilibrium of* Γ*, then it is a* ρ*-IPE of* Γ *for all* ρ*.*

Related Literature. The model is related to alternative approaches where players form wrong theories of each other. In particular, Jehiel (2005) and Jehiel and Koessler (2008) study analogy-based expectations equilibria, ABEE, while Eyster and Rabin (2005) study cursed equilibrium, CE. The mistake postulated by the current model differs both in terms of its order and in terms of its direction.

Under both ABEE and CE each player has correct beliefs about the beliefs of her opponent and a mistakenly theory only of the link between her opponent's actions and his beliefs (the mistake is 'zeroth' order). In addition, a cursed Judith thinks that her opponent's strategy is *coarser* than it actually is. In poker she correctly thinks that Paul knows his hand and only his hand, but mistakenly thinks that Paul does not condition his strategy on his own hand. In contrast, under IPE each player has mistaken beliefs also about the beliefs of her opponent (the mistake is 'first' order). Furthermore, an information projecting Judith thinks that Paul's strategy is *finer* than it actually is; she thinks that Paul also conditions his strategy on her hand.

Crucially, ABEE and CE are closed by the common identifying assumption that, while players may have wrong expectations about the action distribution of others state-by-state, such expectations are always correct on average.[5] In contrast, under IPE such

[5] Formally, in any game Γ, ABEE postulates that i's expectation of j's strategy, $\widehat{\sigma}_j$, in any given state ω, must satisfy:

$$\widehat{\sigma}_j(\omega) = \frac{\sum_{\omega' \in Q_i(\omega)} \pi(\omega')\sigma_j(\omega')}{\sum_{\omega' \in Q_i(\omega)} \pi(\omega')},$$

where $\sigma_j(\omega')$ is j's *true* strategy in state ω' and Q_i is some *partition* of Ω. Hence, it must be true that i's expectations of j's action distribution and j's action distribution *on average* are the same. Formally, $E_\pi[\widehat{\sigma}_j(\omega)] = E_\pi[\sigma_j(\omega)]$ must hold, irrespective of the details of the analogy partition Q_i, only by virtue

expectations might well be wrong also on average. The key qualitative predictions in this paper are all based on such wrong average action expectations, e.g., misperceiving a pooling equilibrium to be a separating one.

Example 1 illustrates the above point. As long as the defender has correct beliefs about the attacker's information, hence realizes that his strategy is constant across states, then no matter what her specific belief about the attacker's strategy may be, it is a strictly dominated strategy for her to overprotect her Achilles heel. Hence, in all alternative models mentioned above, she must play A strictly less often when she is weak than when she is strong. The model then also differs from an application of the level-k logic to Bayesian games, Crawford and Iriberri (2008) since in such models, irrespective of the specification of level-0 behavior, players again have correct beliefs about each other's information, hence, a level-k defender still must also underplay her Achilles heel.

When extending the model in Section 5 to the mistake of ignorance projection, leading to the simultaneous presence of overly fine and overly coarse expectations, I return to a comparison of the current egocentric approach and that of coarse thinking and cursed equilibrium.

3 Social Investment

> "None of the emperor's clothes had ever before received such praise." — Hans Christian Andersen, The Emperor's New Clothes (1837).

At the heart of many social interactions is a dilemma where each person's return on investing depends on her partner's privately known preference for investment. Upon each player i receiving a private signal describing her valuation of mutual investment, $\theta_i \in \mathbb{R}$, each decides whether to invest or to stay out. If both invest, each receives his or her own valuation. If both stay out, each receives the outside (status quo) option. The game is described as follows:

$$\begin{array}{cccc} & & \text{In} & \text{Out} \\ \text{In} & & \theta_i \ , \ \theta_{-i} & g(\theta_i, \theta_{-i}) \ , \ f(\theta_{-i}) \\ \text{Out} & & f(\theta_i) \ , \ g(\theta_{-i}, \theta_i) & 0 \ , \ 0, \end{array} \tag{3}$$

where each θ_i is an i.i.d. draw from a uniform density π_0 over a finite interval $[\underline{\theta}, \overline{\theta}]$ with $\underline{\theta} < 0 < \overline{\theta}$. Let $l = \overline{\theta} / |\underline{\theta}|$.

The key distinction is between negative and positive types. A player with a negative signal prefers not to invest herself. A player with a positive signal prefers to invest if her opponent also invests, and this preference is increasing in her signal.

Sorting: $f(0) = 0$; if $\theta_i < 0$, then $f(\theta_i) > \theta_i$; if $\theta_i > 0$, then $f'(\theta_i) < 1$.

of Q being a partition. Fully cursed equilibrium is equivalent to an ABEE where $Q_i = P_i$ for all i. The same property also always holds for partially cursed equilibrium, see, Eyster and Rabin (2005).

Second, only a positive type benefits from investing alone and does so iff her opponent is also positive.

Investment Risk: if $\min\{\theta_i, \theta_{-i}\} \geq 0$, then $g(\theta_i, \theta_{-i}) \geq 0$; else $g(\theta_i, \theta_{-i}) < 0$.

The normal form of Eq.(3) also describes a set of sequential games. Fix the payoffs as terminal ones, but assume that if only i invests, $-i$ can take a second unmodeled action. For example, suppose that a positive player always reciprocates her opponent's investment and a negative one always continues to stay out. Payoffs can now depend only on the action profile and each player's own private type. Below, I utilize such a sequential interpretation.

3.1 Main Specification

The following specification allows me to present the main insights:

$\min\{\theta_i, \theta_{-i}\} \geq 0$	in	out	$\min\{\theta_i, \theta_{-i}\} < 0$	in	out
in	θ_i, θ_{-i}	$\gamma\theta_i, \gamma\theta_{-i}$	in	θ_i, θ_{-i}	$-c,\ f(\theta_{-i})$
out	$\gamma\theta_i, \gamma\theta_{-i}$	$0, 0$	out	$f(\theta_i), -c$	$0, 0$

where $\gamma \to 1$ from below, and $c > 0$.[6] I describe three interpretations.

♢ **At the Bar.** Judith and Paul can each decide whether to make a move. If both do, a match is formed. If only one does, the other accepts if he values a match positively, a positive type, and again a match is formed with a slight delay discounting payoffs by γ. Otherwise he rejects and the proposer incurs a loss of c such as the pain or the cost of a futile move.

♣ **Trust in Trade.** Trading partners can invest in a relationship-specific asset. While each type may benefit from her partner's investment, this holds as long as $f(\theta_i) \geq 0$, negative types are opportunistic and prefer neither to invest themselves nor reciprocate investment. Hence, a positive type only wants to invest if she sufficiently trusts her opponent to be also positive. Here, c can correspond to the loss from being held up, the extent to which ex ante contracts are incomplete.

♠ **Costly Dissent.** A member of an organization either agrees with (negative type) or disagrees with (positive type) a prevailing rule. When two members meet, each can decide whether to speak up against and deviate from this rule (invest), such as a corporate rule, affirmative action, homophobia, or Stalin's leadership of the Party, or stay silent and act loyal. If a member agrees with the rule, he acts loyal. If he disagrees with the rule, he gains when dissenting in front of someone who also disagrees with it. They may form a coalition or experience a sense of liberation. Dissenting in front of a loyalist, however, leads to a loss of c. A loyalist may punish or report explicit dissent and the dissenter may be fired, ostracized, or persecuted. Here, c can then correspond

[6] The fact that $\gamma < 1$ ensures that the sorting assumption is satisfied. I then consider the limit of the equilibrium set as $\gamma \to 1$. As Proposition 5 shows, all qualitative results hold away from the limit as well.

to the intensity with which dissent is reported to, and then monitored and punished by a central authority, e.g., the extent to which speech in the organization is not free.

Note that while the setup describes bilateral interactions, it applies to such interactions taking place pair-wise between members of a group. In friendship, i.i.d. types may depend on the pairing. In dissent, each player's i.i.d. type is still constant across pairings. In both, people's views about the preferences of others is oft the result of a series of such bilateral interactions.

3.2 Equilibrium

I turn to the predictions. Below, π_1^ρ refers to player i's posterior of her opponent's valuation upon observing the realized action profile and her payoff; E to the ex ante expectation operator given the *true* distribution of signals and actions in equilibrium; π_1 to the true posterior distribution in equilibrium; and $>_{fosd}$ to the partial order of first-order stochastic dominance.

Proposition 2 *For any given ρ, ρ-IPE is unique. Player i enters iff $\theta_i \geq \theta^{*,\rho}$ where:*

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

Furthermore, if $\rho > 0$, for each i,

I. $\pi_1^\rho[\theta_{-i} \mid \theta_i, a] <_{fosd} \pi_1[\theta_{-i} \mid \theta_i, a]$ *given any* $a \in A$ *and* $\theta_i > 0$;[7]

IIa. $E[\pi_1^\rho \mid \theta_i] <_{fosd} \pi_0$ *for any given* $\theta_i > 0$;

IIb. $E[\pi_1^\rho \mid \theta_i] \geq_{fosd} \pi_0$ *for any given* $\theta_i < 0$.

Under-Investment. Judith's willingness to move is increasing in her confidence that Paul would welcome her move, but is decreasing in how confident she thinks Paul is that she would welcome a potential move by him. If Judith is interested in Paul, by projecting this information onto Paul, she exaggerates Paul's incentive to make a move. Since her gain from reciprocating Paul's move is almost as high as making a move simultaneously with him, she finds it relatively more important to stay out and avoid a potential rejection. In a symmetric situation, Paul reasons similarly. An increase in projection increases each positive player's expectation that the other party will invest and, in turn, decreases actual investments.

I. Conditional Antagonism. Each positive player underestimates her opponent's type conditional on any outcome arising in equilibrium. If Judith is interested in Paul, when seeing Paul make a move, she is too convinced he did so only because he knew she would welcome it; when seeing him stay out, she is too convinced that he is not interested.

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

II a & b. Average Antagonism. On average each player wrongly infers antagonistic preferences. If Judith opposes Stalin, she exaggerates how often Paul should speak up in front of her. She overinfers sincere loyalty from Paul's silence, and, on average, becomes too convinced that he supports Stalin. If Judith supports Stalin, she instead underinfers loyalty from Paul's silence and, on average, becomes too convinced that he opposes Stalin. 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 non-Bayesian comparative static follows. Let $E[\overline{\pi}_1^{\rho}]$ be a player's ex ante expected probability estimate of her opponent's being a positive type. Let $\overline{\pi}_0$ denote the corresponding prior estimate.

Corollary 1 *For any $\rho > 0$, $E[\overline{\pi}_1^{\rho}]$ is strictly decreasing in c. If $c > \overline{c}(\rho)$, then $E[\overline{\pi}_1^{\rho}] < \overline{\pi}_0$.*

An increase in c decreases investments and increases positive types' over-inference, but decreases negative types' under-inference from others' lack of investment. Hence, it leads to more negative updating on average. I return to this in Section 3.5.

3.3 Dynamics

Suppose now that the above opportunity to interact repeats itself, but the cost of a wrong move may change over time. Specifically, consider a dynamic repetition of the game, but allow for a weakly decreasing sequence $\underline{c} = \{c_t\}_{t=1}^{T}$ with $c_T > 0$.[8] For simplicity, I focus on myopic repetitions: in each round t players care only about the payoff of that round, but recall the history of past interactions. 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 probability ρ her opponent becomes his projected version and learns her valuation. This (mistaken) perception of the real players' is still common knowledge.

The prediction for each round is unique. Let $\Pr_{\underline{c}}^{\rho}(m)$ denote the true probability that, conditional on both players being positive in a pair, at least one invests by the end of the sequence. This measure also describes the extent to which equilibrium, in reality, reveals the sign of each player's signal; if it is 0, equilibrium is pooling; if it is 1, it fully reveals the sign of each type's signal.

Corollary 2 $\Pr_{\underline{c}}^{\rho}(m)$ *is decreasing in ρ. Furthermore,* $\Pr_{\underline{c}}^{0}(m) = 1 - c_T/l\overline{\theta}$.

Investment is decreasing in the bias dynamically as well. Furthermore, in the unbiased case, as c_T goes to zero, equilibrium is always revealing; all efficient matches are formed and players always learn whether others support or oppose the rule. In contrast, in the biased case, false antagonism reinforces underinvestment which may prevent any

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

investment even in this limit. To describe the logic, let $E[\overline{\pi}_t^{\rho,+}]$ denote player i's ex ante expected round-t probability estimate that her partner is positive conditional i being positive, e.g., Stalin's opponents' expected average opinion of Stalin's unpopularity; let $E[\overline{\pi}_t^{\rho,-}]$ denote the analogous estimate conditional on i being negative, e.g., Stalin's supporters' expected opinion of Stalin's unpopularity; let $E[\overline{\pi}_t^{\rho}]$ be the analogous unconditional estimate.[9]

Corollary 3 *Fix any $\underline{c}$ and $\rho > 0$. For each t, $E[\overline{\pi}_t^{\rho,+}] < \overline{\pi}_0 \leq E[\overline{\pi}_t^{\rho,-}]$ and there exist $0 < \alpha_t^+ < \alpha_t^- < 1$ such that*

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

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

In a BNE the martingale property ensures that each player's ex ante expected estimate of the preferences of others remains constant over time. In the biased case, these estimates are predictably distorted in a 'paranoid' manner and average antagonism holds in each round. Consider positive players. Due to projection from each round to the next, opponents of Stalin exaggerate how often those who also oppose Stalin should speak up in front of them. As a dynamic consequence, however, they underestimate the fraction of others who oppose Stalin. In turn, if the drop in c_t is smaller than a threshold, there is too little new dissent in round $t+1$ relative to their expectations, and their paranoia grows on average. If instead it is greater than this threshold, the opponents of the rule are too surprised by how common dissent or a positive response to dissent is and their paranoia shrinks on average, but is never fully eliminated. Consider now negative players. Since a supporter of Stalin always finds dissent in front of her (weakly) too surprising, her expected paranoia grows (weakly) over time. This implies the next result.

Corollary 4 *Given any $\rho > 0$, suppose that $c_t \geq \bar{l\theta}(1-\rho)^t$ for all t. No one ever invests, but $E[\overline{\pi}_t^{\rho,+}]$ and $E[\overline{\pi}_t^{\rho}]$ are strictly decreasing in t. Furthermore,*

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, positive players adopt differential attribution of identical behavior to self and other and there is no investment even as c_T vanishes. In the context of friendship, no one makes a move, but each concludes that none she wants to be friends with wants friendship with her. In the context of dissent, no one dissents, but all those who oppose the rule conclude that they are alone with their preferences. Since, as long as none dissents, negative types maintain unbiased estimates, it is then exactly when none supports the rule that everyone concludes that everyone else supports it.

[9] Formally, $E[\overline{\pi}_t^{\rho,+}] = E[\overline{\pi}_t^{\rho} \mid \theta_i \geq 0]$ and $E[\overline{\pi}_t^{\rho,-}] = E[\overline{\pi}_t^{\rho} \mid \theta_i < 0]$.

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 rule independent of the margin by which this is true or false in reality.

3.4 Implications to Dissent

> "Nothing appears more surprizing than the easiness with which the many are governed by the few; and the implicit submission, with which men resign their own sentiments and passions to those of their rulers. When we enquire by what means this wonder is effected, we shall find, that, as Force is always on the side of the governed, the governors have nothing to support them but opinion [...] and this maxim extends to the most despotic and most military governments, as well as to the most free [...] all the farther power he [the tyrant] possesses must be founded either on our own opinion, or on the presumed opinion of others." — David Hume, Of The First Principles of Government (1741).

> "It is proper to take alarm at the first experiment on our liberties." — James Madison, Memorial and Remonstrance against Religious Assessments (1785).

Norm Falsification. The persistence of a law or a norm of behavior in a group, such as a particular form of discrimination, doctrine, custom, often depends not on the genuine popularity of the corresponding rule, but on people's perception of this rule's popularity. Such perceptions about the preferences of others pin down people's expectations of the return on being disobedient or on pursuing change. Indeed as Hume (1741) suggest, the perceived threat of one's peers voluntary enforcement of a rule is often even necessary for formal sanctions to have sufficient binding force. Corollary 3 implies that as long as dissent is not absolutely free and people are not fully unbiased, if $c, \rho > 0$, true and perceived public opinion systematically diverge; those who oppose the status quo rule always underestimate others' opposition to it. Such a misperception may have important consequence in many domains.[10] Let me describe two direct consequences of Corollary 4.

Disciplinary Organization. Suppose that the leadership of an organization wants to ensure that its members act loyal towards a given rule desired by the leadership. Suppose that the leadership can set c_t in each round t. The organization can devote resources to encourage loyalists to report dissent and then to monitor such reports and potentially punish reported dissenters.[11] A higher intensity of such disciplinary

[10] This misperception may matter in models of political transition, e.g., Acemoglu and Robinson (2001), but also in models of representative democracy where the set of policies voters can choose from is determined endogenously based on people's perception of the preferences of others, e.g., Besley and Coate (1997), thus, there, it may discourage popular alternatives appearing on the ballot.

[11] The secret employment of citizens to report the potential dissent of others is a practice not only of the totalitarian states of the former Soviet block or Nazi Germany, but is (was) present in many other organizations.

sanctions, a higher c, however, requires a greater sacrifice of organizational resources devoted to monitoring dissent. The next result then describes the optimal way for the organization to ensure loyalty while minimizing the resources devoted to suppressing dissent over time.

Proposition 3 *Suppose that an organization wants to minimize* $\sum_{t=1}^{T} c_t$ *subject to* $\Pr_{\underline{c}}^{\rho}(m) = 0$. *The optimal* $\underline{c}$ *is unique and is given by* $c_t^*(\rho) = l\overline{\theta}(1-\rho)^t$. *Hence,*
if $\rho = 0$, $\lim_{T\to\infty} \sum_{t=1}^{T} c_t^*(\rho)/T = l\overline{\theta}$;
if $\rho > 0$, $\lim_{T\to\infty} \sum_{t=1}^{T} c_t^*(\rho)/T = 0$.

The optimal sequence ensuring full loyalty is unique. In the unbiased case, the intensity of disciplinary sanctions remains constant over time and its long-term average level must go to infinity as the expected share of those who oppose the rule becomes full. In the biased case, loyalty is, however, maintained gradually less by a formal threat and more by those opposing Stalin (or homophobia) becoming more and more convinced that it is pointless to criticize Stalin since everyone else thinks he is great. Self-censorship outlives actual censorship; opponents of the rule become apathetic. Optimal sanctions go to zero and an arbitrarily unpopular rule, once formally sanctioned, persists as an (essentially) self-sustaining norm.

Slippery Slope under Almost Free Speech. Crucially**,** Proposition 3 relies on the organization's ability to impose a potentially very high c at least initially. At the same time, for most organizations, even the most militant ones, adopting such intense sanctions is typically prohibitively costly. Indeed suppose now that the intensity with which the organization can monitor and punish dissent at any point in time is limited, i.e., that the following constraint must hold:

$$c_t \leq \varepsilon \text{ for all } t$$

Given such a constraint, Proposition 2 implies that if a rule introduced directly in round 1 was sufficiently likely to be unpopular and intensely disliked (high $l\overline{\theta}$ relative to ε), then most opponents of this rule would speak up and reveal the rule's unpopularity despite projection. Nevertheless, by approaching this desired rule sufficiently gradually over time— for example, escalating a mild form of discrimination or vilification of a minority group to full-fledged genocide — the organization can always create the illusion whereby the desired rule eventually appears arbitrarily popular to all those who oppose it no matter how small ε is, e.g., even if speech is almost free.

To describe the logic, let the desired rule r^* be one where each player i's dislike of this rule is still an i.i.d. draw from a commonly known uniform density over some $[\underline{\theta}, \overline{\theta}]$ just as before. Consider a sequence of rules $\underline{r} = \{r_t\}_{t=1}^{T}$ approaching this rule over time, that is, suppose that at the beginning of each round t, rule r_t replaces rule r_{t-1} with $r_T = r^*$. Specifically, take the same repetition of the game as before, but assume that each player i's type evolves according to a commonly known sign-preserving process:

$$\theta_{i,t+1} = \alpha_t \theta_{i,t} \text{ if } \theta_{i,t} \geq 0 \text{ and } \theta_{i,t+1} < 0 \text{ if } \theta_{i,t} < 0 \text{ for each } t. \tag{4}$$

where $\alpha_t \in \mathbb{R}^+$. Each rule r_t represents the same stage game as before, but in round t, $\theta_{i,t}$, that is, player i's dislike of rule r_t, replaces θ_i for each i. Again, let $\Pr^{\rho}_{r^*}(m)$ be the true extent to which equilibrium reveals whether a player supports or opposes r^* by the end of the sequence.

Proposition 4 *Consider any* $\varepsilon > 0$.

If $\rho = 0$, *then* $\Pr^{0}_{r^*}(m) \geq 1 - \varepsilon/l\overline{\theta}$.

If $\rho > 0$, *there always exists a* $\underline{r}$ *such that* $\Pr^{\rho}_{r^*}(m) = 0$ *and* $E[\overline{\pi}^{+,\rho}_T] \leq \varepsilon$.

In the unbiased case players always form unbiased estimates about the preferences of others. In turn, as the ratio of ε over $\overline{\theta}l$ goes to zero, equilibrium smoothly converges to one that fully reveals the sign of each player's type just as before. Almost free speech is close to free speech.

In the biased case, there instead always exists a finite sequence leading to r^* such that no one ever dissents. Starting with a rule r_1 that is at most mildly disliked by anyone relative to ε, no one finds it sufficiently worthwhile to speak up. This lack of initial dissent, however, causes all those who oppose this rule to underestimate others' opposition to it. When replacing this initial rule with one that is somewhat more intensely disliked by all those who disliked r_1 such underestimation ensures that still no one speaks up even if ε remains the same. Proceeding along this line sufficiently gradually over time, while people's dislike of the subsequent rules increases, their perception of others' opposition to these rules decreases. Unless speech is absolutely free, $\varepsilon = 0$, this slippery slope always ensures that eventually all prefer to appear loyal towards an arbitrarily unpopular and intensely disliked rule because even if privately all despise it, they perceive others support as sincere.[12]

Shy Revolutions. Both of the above propositions are based on the gradual accumulation of false antagonism by positive types. Suppose now that at some T there is a sufficiently large relative drop in the potential loss from dissent, e.g., a secret ballot on upholding the status quo rule is held, or an intensely disliked new rule is introduced too abruptly. Corollary 3 implies that dissent comes as a great surprise to all those who dissent and their perception of the popular support for the rule, growing until then, erodes potentially discontinuously.[13] Similarly, an outside observer who did not think people projected would also be surprised.

[12] The fear that any violation of the freedom of speech will lead to a decay into tyranny appeared in the arguments of the founding fathers of the US, e.g., the admonition of Benjamin Franklin, that "freedom of speech is a principal pillar of a free government; when this support is taken away, the constitution of a free society is dissolved, and tyranny is erected on its ruins." B. Franklin, Freedom of Speech and the Press (1737).

[13] An element of surprise is characteristic of many historic instances of regime change. For example, Kuran (1995) reports that a year after the Fall of the Berlin Wall, despite the benefit of hindsight, 76% of those surveyed in former East Germany, claimed that even shortly before the Fall they thought that such a change was inconceivable. This was also true amongst churchgoers who were disproportionately active in the demonstrations.

3.4.1 Implications for Intergroup Contact

Consider a random division of people with i.i.d. preferences towards friendship with each other into two sub-groups. Suppose that in-group members can read each other's preferences, but face uncertainty about out-group members' preferences. Proposition 2 implies that on average each person will mistakenly conclude that out-group members are distinctly less likely than an in-group member to want to be friends with her when she wants friendship. Corollary 3 implies that if further intergroup contact occur in high-risk environments (high c), it raises, if it occurs in low-risk environments (low c), it predictably decreases such hostile misattributions.

3.5 Evidence

The predictions are consistent with an empirical findings discussed in psychology under the rubric of *pluralistic ignorance* (Allport, 1924) defined as "the phenomenon that occurs when people erroneously infer that they feel differently from their peers, even though they are behaving similarly", Prentice (2007). In the context of norms for example, Prentice and Miller (1993) found that undergraduates at Princeton rated their peers' average comfort level with the prevailing heavy drinking norm on campus as significantly higher than their own, hence, than that of reality.

Affirmative Action. Van Boven (2000) documents a similar effect, but provides more detailed data. Studying Cornell undergraduates's attitudes towards affirmative action, using anonymous surveys, he finds that only a minority of students supported affirmative action (27%) while the most common attitude was to oppose it. At the same time, consistent with Corollary 4, students on average believed that support for rather opposition towards affirmative action was the more common attitude amongst their peers. Crucially, in-line-with the false antagonism predicted by Corollary 3, those who *opposed* affirmative action had a significantly higher estimate of the support for affirmative action amongst their peers than did those who supported affirmative action. The data, thus, supports the *joint* prediction of false antagonism and the reversal between the majority's perception of the majority's preference and the majority's true preference when the latter opposes the status quo rule.[14]

Evidence from the lab is also consistent with the predictions. In an illustrative lab study, Miller and McFarland (1987) presented students with a difficult text. In the unconstrained treatment, students, seated in small groups, could publicly leave the room and seek clarification in case "they had any serious difficulty understanding the material" (high c). In the constrained one, no such option was present. While no student left the room, consistent with Corollary 1, students on average rated their own knowledge of the topic as significantly lower than that of the average group member in

[14] For related evidence on the majority's misperception of the majority's preference when the latter opposes the status quo, see, e.g., O'Gorman (1975) in the context of Whites' support for racial segregation in 1968, or Shamir and Shamir (1997) in the context of Israelis attitude towards territorial concessions.

the unconstrained, but not in the constrained treatment.

In the context of friendship, 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. Consistent with the model, 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.[15]

Brexit. Finally the predictions may help understand recent evidence from the UK's 2016 referendum on leaving the EU. The odds offered by Betfair, the world's largest internet betting exchange, always overwhelmingly favored the victory of Remain. Its implied chance of victory never dropped below 63%, averaged somewhat above 70%, and was above 80% on the day of the referendum. Similarly, 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."[16] Such a misprediction is consistent with the predicted false antagonism whereby some may not have felt comfortable speaking up against the prevailing norm of inclusion, but misattributed others' identical reluctance too much to their support of this norm.

While clearly any single piece of evidence can be interpreted in multiple ways, the totality of the evidence provides support for the predictions.

3.6 Investment Games

Let me return to the more general case. To characterize the implications, a distinction between complement and substitute (initial) investments is needed.

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

Investments are substitutes if, conditional on both players having positive valuations, Judith gains more from making a move (initially) if Paul does not make a move (initially) than when he does. Investments are complements if the reverse holds. Since the benefit of simultaneous investment relative to reciprocated investment is typically very small, such as in the case of dating or in the case of dissent, social investments are predominantly substitutes. For a more complete analysis, however, I analyze both cases below. To present the results, suppose that f and g are continuously differentiable and a positive type's return from one-sided (initial) investment is weakly increasing in the type profile, i.e., $g_1(\theta_i, \theta_{-i}) \geq 0$ and $g_2(\theta_i, \theta_{-i}) \geq 0$ if $\theta_i > 0$.

[15] In the psychology literature, pluralistic ignorance is sometimes contrasted with a so-called *false consensus* effect, defined as: "false consensus is revealed when people making a particular choice consider this choice more common than do people making the opposite choice." Kunda (1999, p.397). Although the model could then be described as one that predicts both endogenously depending on the true informational environment, when invoking such evidence in support of the model serious caution is needed. This correlation in *conditional*, as opposed to *average*, predictions documented by such false consensus is often perfectly warranted. Hence, such evidence is genuinely hard to interpret.

[16] See, respectively, http://politicalodds.bet/eu-referendum? for the date on Betfair and http://lordashcroftpolls.com/2016/06/how-the-united-kingdom-voted-and-why/#more-14746 for the poll.

Proposition 5 *All ρ-IPE are given by cutoff strategies.*

1. *If investments are substitutes, there is a unique symmetric equilibrium and it is increasing in ρ.*

2. *If investments 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,*

   I. $\pi_1^\rho[\theta_{-i} \mid \theta_i, a] <_{fosd} \pi_1[\theta_{-i} \mid \theta_i, a]$ *given any* $a \in A$ *and* $\theta_i > 0$;[17]

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

If investments are substitutes, a positive player's willingness to invest decreases in the perceived probability that her opponent invests. Given the logic outlined before, this leads to under-investment in the unique symmetric equilibrium. If investments are complements, the same willingness increases in the perceived probability that one's opponent invests. Here, all equilibria are symmetric, but there may be multiple such equilibria. The lowest decreases in ρ, while the second-lowest, if it exists, increases in ρ. Crucially, *all* symmetric equilibria again exhibit both conditional and average false antagonism, hence, their dynamic implications continue to hold.

4 Advice

I now turn to the second application and consider strategic advice. A general feature of Bayesian communication, given rational prior beliefs, e.g., Crawford and Sobel (1982), is that receivers benefit from advice and are never fooled by it *on average*. While the incentive to strategically distort recommendations is commonly understood in many contexts, growing evidence suggests that even in such contexts receivers are persuaded too easily and lose rather than gain from access to advice. For example, in the context of financial advice, Bergstresser et al. (2009) show that investors buy broker-recommended funds that deliver lower risk-adjusted returns than directly-sold funds even before subtracting the fee charged for such recommendations.

Crucially, while a form of receiver naiveté has been considered in the literature before, e.g., Kartik et al. (2007), the model not simply predicts credulity, as well as its opposite, disbelief, endogenously, but, by linking it to underlying portable mistake, provides comparative static predictions on how the strategic environment affect their extent and shape the receiver's welfare. Such comparative statics are both positively and normatively key.

In particular, in the wake of the Great Recession, many have argued that greater *financial literacy* was essential to improve client's financial outcomes and how much

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

they benefit more from financial advice, e.g., Lusardi (2013).[18] Although financial education has for long been present at various levels of the educational curriculum, the 2010 Dodd-Frank Act further mandates the establishment of "the Office of Financial Education, which shall develop a strategy to improve the financial literacy of consumers." [19]

In contrast, as Hastings et al. (2013) point out in their general review of the existing evidence, "of the few studies that exploit randomization or natural experiments, there is at best mixed evidence that financial education improves financial outcomes."[20] Furthermore, the National Association of Securities Dealers (Investor Fraud Study, 2006), using data from law enforcement, directly compared the characteristics of those who were successfully persuaded to buy into fraudulent investment schemes to a randomly selected pool of non-victims. They concluded that "a major hypothesis going into the survey was that investment fraud victims do not know as much about investing concepts as non-victims and would therefore score lower on financial literacy questions. In fact, the study found the exact opposite: investment fraud victims scored higher than non-victims on [all] eight financial literacy questions."[21]

4.1 Setup

Timing. Upon meeting the receiver (client, patient, decision maker), the sender (financial advisor, doctor, lobbyist) privately learns whether a statement is true, $\{\theta = 1\}$, or false, $\{\theta = 0\}$. She then sends a message about this to the receiver. Upon obtaining the recommendation, the receiver decides whether to fact check this recommendation at some cost c. If he checks, he learns whether or not the statement is true. Finally, the receiver takes an action y. For simplicity, I assume the prior on θ to be symmetric.[22]

Verification. The receiver's cost of checking c is drawn from a commonly known distribution, cdf F, with a strictly positive density over $[0, \infty)$. Its realization is the receiver's private information. A first-order stochastic decrease in F (a decrease in F, henceforth), corresponds to a lower distribution of checking costs and is interpreted as a greater financial literacy of the receiver, or a greater simplicity of the statement to be evaluated vis-a-vis the receiver's background knowledge.

Investment. The receiver's action $y \in [0, 1]$ is aimed at maximizing his expected utility. This may correspond to the fraction of resources allocated to buying or selling an asset, or promoting or blocking a policy. To keep the analysis fully transparent, I assume that the optimal action simply equals the receiver's posterior confidence that the statement is true. This is captured by the standard assumption that the receiver's

[18] See also, e.g., Bernanke (2011).

[19] See Dodd-Frank Wall Street Reform and Consumer Protection Act. H.R. 4173. Title X - Bureau of Consumer Financial Protection 2010, Section 1013.

[20] Hastings et al. (2013, p. 361). See also Willis (2011).

[21] See, p. 5, NASD (2006) Final Report.

[22] None of the qualitative results depend on the prior being symmetric.

utility from investment, potentially only observed with noise ex post, is:

$$u_r(y, \theta) = -(y - \theta)^2.$$

Conflict of Interest. The sender gets a bonus $B > 0$ whenever she issues a positive recommendation. At the same time, if the receiver checks and finds out that the sender lied, the sender incurs a loss (of business, reputation, or a regulatory fine) $S > 0$. I normalize $S = 1$ and interpret B in proportional terms. To make the analysis non-trivial, I then also assume that $B < 1$. All of the above is common knowledge.

Welfare. When discussing receiver welfare I take the standard ex ante expected perspective: welfare equals the receiver's true expected utility from investment minus the potential verification cost incurred in equilibrium. I refer to the receiver's expected utility when acting only on his prior as the receiver's welfare without advice.

4.2 Bayesian Case

Consider the unbiased case. The sender tells the truth if the statement is true and lies with probability p^0 if it is false. The receiver checks a positive recommendation iff her cost is lower than some threshold c^0. Below, $E_\theta[y_c^{*,0}]$ denotes the true ex ante expected equilibrium investment of type c. The receiver's prior confidence is $\overline{y}$.

Claim 1 *If $\rho = 0$, the receiver checks iff $c \leq c^0(B, F)$. The sender lies with probability $p^0(B, F) > 0$. An increase in B and in F increases both $c^0(B, F)$ and $p^0(B, F)$. Communication is neutral, $E_\theta[y_c^{*,0}] = \overline{y}$ for all c.*

Neutrality. While both a lower conflict and higher financial literacy lead to more information transmission, communication is always neutral: the receiver's ex ante expected posterior confidence in the proposition being true is the same as his prior. This is a general consequence of the martingale property of correct Bayesian updating in a BNE; persuasion is purely informative and never shifts the receiver's ex ante expected posterior.

4.3 Advice under Projection

Consider, now, a biased client ($\rho_R = \rho$) and, for simplicity, an unbiased, thus, sophisticated sender ($\rho_S = 0$). Persuasion is no longer neutral. Instead, it leads to two opposing mistakes: *credulity*, whereby the receiver believes a positive recommendation too much and overinvests on average, and *disbelief*, whereby the receiver believes a positive recommendation too little and underinvests on average.

Proposition 6 *For any $\rho > 0$, ρ-IPE is unique. There exist $0 < c_1^\rho < c_2^\rho \leq c_3^\rho$ such that*

for $c < c_1^\rho$ persuasion is neutral, $E_\theta[y_c^{*,\rho}] = \overline{y}$;
for $c \in [c_1^\rho, c_2^\rho)$ credulity holds, $E_\theta[y_c^{*,\rho}] > \overline{y}$;

for $c \in (c_2^\rho, c_3^\rho)$ *strict disbelief holds,* $E_\theta[y_c^{*,\rho}] < \overline{y}$*;*
for $c \geq c_3^\rho$ *(weak or strict) disbelief holds,* $E_\theta[y_c^{*,\rho}] \leq \overline{y}$*;*
c_1^ρ *is decreasing and* c_3^ρ *is increasing in* ρ*.*

By projecting his private information onto the sender, the receiver exaggerates the extent to which the sender tailors the truthfulness of her message to his actual cost c and partially neglects the fact that the sender is constrained to lie to all types to the same extent, $p^\rho(B, F)$. The projected sender is then perceived to lie according to a monotone function $p^+(B, c)$ which is strictly increasing in c on some positive interval $[c_1^\rho, c_3^\rho]$. In turn, the lower is the receiver's actual cost of checking, the more he trusts a positive recommendation.

The receiver's checking strategy matches these perceptions. Types for whom checking is sufficiently cheap still always check. Middle types $[c_1^\rho, c_2^\rho)$ check too little, relative to their true benefit of checking in equilibrium, but are too credulous when hearing a positive recommendation and overinvest on average. Finally, types above c_2^ρ check (weakly) too much, relative to their true value of checking, but when they don't, they are (weakly) too skeptical when hearing a positive recommendation, thus, underinvest on average.

4.4 Uniform Credulity

While in the biased case some types are always credulous, the scope for disbelief is limited. I refer to the case in which all receiver types are at least weakly credulous and a strictly positive measure of them are strictly credulous, as *uniform credulity*. The next proposition shows that such uniform credulity holds whenever the conflict is not too low and also whenever the sender's statement is not too trivial to check on average.

Proposition 7 *Suppose that* $\rho > 0$*,*

1. *if* $B \geq \overline{B}(\rho, F)$*, uniform credulity holds;* $\overline{B}(\rho, F) < 1$ *and it is decreasing in* ρ *with* $\lim_{\rho \to 1} \overline{B}(\rho, F) = 0$*;*
2. *if* $F \geq_{fosd} \overline{F}(\rho, B)$*, uniform credulity holds and if it holds at* F *given* ρ*, it also holds at* F *given any* $\rho' > \rho$*.*[23]

To see the logic, note that an increase in B or in F always reduces real information transmission. Hence, eventually, the sender always lies when the statement is false and disbelief no longer applies. At the same time, all receiver types for whom checking is not a dominated choice still believe a positive recommendation too much. Since such types check too little, uniform credulity follows. Advice is unambiguously deceptive and leads to strict overinvestment. Finally, an increase in ρ increases the set of environments where such uniformly credulity holds.

[23] Since fosd is only a partial order, multiple such $\overline{F}(\rho, B)$ exist. Below, $\overline{F}(\rho, B)$ refers to *any* such distribution.

Proposition 7 establishes the threshold on the size of the conflict and an upper-bound on the amount of financial literacy under which uniform credulity holds. Recall, however, that in our setting the conflict is also always limited, i.e., $B < 1$, and checking is never a dominated choice for all types, i.e., F has full support. If either of these conditions is violated, then, irrespective of ρ, the sender has a commonly known dominant strategy to lie. It follows that neither credulity nor disbelief applies. Hence, it is the joint presence of a positive but limited conflict and a sufficiently easy, but not too easy to evaluate statement, valuable but non-full literacy (perhaps an MBA degree), which leads to uniform credulity.

4.5 Welfare

Let me turn to the comparative static predictions. In the unbiased case the receiver always strictly gains from advice, it is never toxic, and both a lower conflict and higher financial literacy improves the receiver's welfare. The next result describes sufficient conditions for these key comparative statics to be reversed.

Proposition 8 *For $\rho = 0$, an increase in B or in F decreases welfare. For any $\rho > 0$,*

1. *if $B \geq \overline{B}(\rho, F)$, an increase in B decreases expected investment and strictly increases welfare;*

2. *if $F \geq_{fosd} \overline{F}(\rho, B)$ and $B < \frac{1}{2}$, an increase in F, which does not change $F(\frac{1-\rho}{(2-\rho)^2})$, decreases expected investment and strictly increases welfare.*

Comparative Static with B. An increase in the conflict — a decrease in the sender's potential loss from lying S — decreases information transmission and increases checking. Both constitute a negative welfare force. In the biased case, however, there is a third countervailing effect. A higher conflict tampers credulity; it induces more checking and decreases credulous types' overinference from a positive recommendation, and reduces investment. If uniform credulity holds, an increase in B does not change real information transmission. Hence, it unambiguously raises welfare.

Uniform credulity is only sufficient, but not necessary for the above result. Since a higher conflict, always increases checking, it decreases the scope for biased investments both for credulous types and those in disbelief. This positive effect can outweigh the two negative ones and often lead to higher welfare even if uniform credulity does not hold.[24]

Comparative Static with F. A lower distribution of checking costs, higher literacy, always increases information transmission, decreases checking, and also the cost of checking. In the biased case there is again a countervailing force. All else equal, the

[24] To illustrate, consider the following setting. Let $B = 0.24$. Assume that $f(c) = 2$ if $c \leq \frac{1}{4}$, and let $f(c)$ be an arbitrary density otherwise. If $\rho = 0$, then $p^0 = 0.16$ and the receiver's welfare is -0.07. If $\rho = 0.9$, then $p^\rho = 0.94$ and the receiver's welfare is -0.23. If instead $B' \geq 0.36$, the receiver's welfare under $\rho = 0.9$ is always greater than -0.23; e.g., if $B' \geq 1$, it is -0.18.

easier it is for the receiver to check, the more he believes the sender. If the conflict is not too large, $B < \frac{1}{2}$, there is always sufficiently little checking in equilibrium such that, all else equal, credulous types lose more from becoming more credulous and investing more than how much they gain from a lower cost of checking. As long as uniform credulity holds, and the measure of types for whom checking is always sufficiently cheap is unchanged, the highest such type is bounded from above by $(1-\rho)/(2-\rho)^2$, higher literacy, or a simpler statement to be evaluated, increases credulity and investment and decreases welfare.

Uniform credulity is again only sufficient but not necessary. A lower distribution of checking costs always increases the scope for credulity, and potentially for disbelief, and may often decrease welfare even if uniform credulity does not hold.[25]

4.6 Endogenous Conflict

Above I specified conditions under which both a greater ease with which the receiver can fact check the sender and a lower conflict induces more deception and lower welfare. To sharpen the former result, I now endogenize the conflict by invoking the seller of the asset.

Suppose that before the resolution of any uncertainty the seller of the asset pledges to pay the sender a bonus B whenever the sender issues a positive recommendation.[26] The seller wants to maximize her profit given by the receiver's expected investment into the asset y, times some positive markup $\gamma > 0$, minus the expected transfer to the sender:

$$R(\rho, B, F) - T(\rho, B, F) = \gamma E_{c,\theta}[y_c^{*,\rho}] - t^* B,$$

where t^* is the probability with which the sender makes a positive recommendation in equilibrium. The chosen value of B is again common knowledge.

What is the optimal bonus that a sophisticated seller, one who correctly understands the sender's and the receiver's true behavior in equilibrium, would want to set? Let $B^*(\rho, F, \gamma)$ denote the seller-optimal conflict.

Proposition 9 *If $\rho = 0$, then $B^*(0, F, \gamma) = 0$. If $\rho > 0$, then $0 < B^*(\rho, F, \gamma) \leq 1$ and $E_{c,\theta}[y_c^{*,\rho}] > \overline{y}$ iff $\gamma \geq \overline{\gamma}(\rho, F)$ where $\lim_{\rho \to 1} \overline{\gamma}(\rho, F) = 0$.*

In the unbiased case, the seller-optimal conflict is zero. Since communication is neutral, a positive bonus does not shift the receiver's expected investment; it only introduces noise in communication and does so at a pure cost to the seller. In the biased case, instead, by choosing a conflict that is sufficiently large, but not too large, advice always inflates the receiver' expected confidence in the asset, thus, persuasion

[25] To illustrate, consider the same setting as in the previous footnote, except assume that $\widetilde{f}(c) = 2$ if $c \leq \frac{10}{121}$, $\widetilde{f}(c) = 1$ if $c \in [\frac{10}{121}, \frac{1}{4}]$, and $\widetilde{f}(c)$ can again be any density otherwise. It is easy to see that the corresponding $\widetilde{F}$ can fosd F. The receiver's welfare under any such $\widetilde{F}$ is -0.22.

[26] Considering unconditional payments leaves the analysis unchanged.

leads to excess demand and strict overinvestment. If the markup is not too low, the seller then always chooses such a conflict and ensures that advice is deceptive.

When Financial Education always Backfires. The extent to which the sender can deceive the receiver depends on how easy it is for the receiver to check on average. Let me then conclude with a final comparative static. As the bias becomes full, the Bayesian result is always fully reversed. The value of advice for the receiver is always negative. Furthermore, the receiver's welfare is strictly decreasing in his financial literacy.

Corollary 5 *Given any $\gamma > 0$,*

if $\rho = 0$, the receiver's welfare is strictly higher with advice than without advice;

if $\rho \to 1$, the receiver's welfare is strictly lower with advice than without advice, and any decrease in F decreases the receiver's welfare and increases the seller's profit.

As egocentric thinking becomes full, each receiver type fully neglects the behavior of other receiver types. Since there are always types for whom checking is too costly, but this is not internalized by any type who can afford to check, there is never enough checking for sender to want to be truthful when the statement is false, and uniform credulity holds given any positive conflict. In turn, the seller-optimal conflict and equilibrium checking vanishes. At the same time, all types for whom checking is not a dominated choice, still believe a positive recommendation too much. Advice unambiguously lowers welfare. Since the lower is c the more the receiver believes the sender to be truthful, *any* increase in the receiver's financial literacy lowers his welfare, but increases the seller's profit. Paradoxically, the simpler it is for the receiver to check the sender, the more he suffers from wrongly believing her lies.

All the above results rely on it not being *too* difficult to fact check the sender's lies, e.g., the receiver having valuable financial literacy, as ensured by the fact that F has full support. If instead F was concentrated only on types for whom checking was an ex ante dominated choice, then irrespective of projection, and given any $B > 0$, it would be common knowledge that the sender had no incentive to be truthful. In turn, communication would again be neutral, and the seller-optimal conflict would again be zero, hence, the receiver's welfare would be maximal.[27]It follows that given any $\rho > 0$, under the seller-optimal conflict, valuable financial literacy can only hurt the receiver and boost the seller's profit.

5 Projection Equilibrium

So far I focused on the notion of information projection in strategic contexts. The logical counterpart of information projection is a phenomenon I term *ignorance projection:* the exaggerated belief that if one does not know an event, others do not know it either. In many domains people will likely project their information without projecting

[27] Formally, for any such F, $B^*(\rho, F, \gamma) = 0$ for any γ and ρ.

their ignorance. At the same time, the formalism developed in this paper allows me to immediately extend the model to incorporate the *joint* presence of information and ignorance projection as well.

If Judith projects both her information and her ignorance, she exaggerates, the probability with which Paul can condition his strategy on the exact same set of events that she can. Formally, the projected version of player i — still real only in the imagination of player j — now chooses a strategy from the set:

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

In each state ω, the projected version of i knows the same set of events that j does.

Definition 3 *A strategy profile $\sigma^\rho \in S_i \times S_j$ is a ρ projection equilibrium (PE) of Γ if there exists $\sigma^\pm \in S_i^j \times S_j^i$ such that for each i,*

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

and

$$\sigma_j^i \in BR_{S_j^i}\{\sigma_i^\rho\}. \tag{6}$$

In the above definition, projection is still partially public, the same way as before, and the definition again satisfies the same two properties as before. In poker, the difference from ρ-IPE is simply that Judith now acts as if she thought that the projected version of Paul knew her hand, but only her hand, that is, he knew exactly what she did.

Nested Model. The models of IPE and PE can be nested within a single one. Specifically, suppose that each real player j assigns probability ρ^+ to i choosing his strategy from the set S_i^+, probability $\rho^\pm$ to i choosing it from S_i^j, and probability $1-\rho^+-\rho^\pm > 0$ to i being regular and, in equilibrium, playing the strategy i actually plays. Maintaining all-encompassing projection as before, if $\rho^\pm = 0$, the nested model collapses to IPE; if $\rho^+ = 0$, it collapses to PE.

Note that ignorance projection corresponds to overly coarse expectations about one's opponent's play. Hence, it has a closer link to the idea of coarse thinking captured by ABEE and cursed equilibrium. At the same time the order of these mistakes still differs as before. Projection is still a mistake about the beliefs of others (a 'first-order' mistake), as opposed to only about the link between the beliefs and the actions of others (a 'zeroth-order' mistake). As a consequence, it is now perfectly compatible for Judith to have overly fine expectations about Paul's strategy vis-a-vis *her* private information, but overly coarse expectations about his strategy vis-a-vis *his* private information. The application I turn to illustrates this.

5.1 Adverse Selection

As the last application, consider the classic problem of common-value trade of the kind introduced by Akerlof (1970). The seller owns an object and values it at q, while the buyer values it at $w(q)$. Quality q is drawn uniformly from $[q_0, q_0 + r]$ with mean $\overline{q}$, and its realization is observed only by the seller. This fundamental problem has been studied empirically by the study of Samuelson and Bazerman (1985) and a small literature following it. In particular, the classic study of Samuelson and Bazerman (SB, henceforth) considers both the case where the informed seller and where the uninformed buyer has the bargaining power. I link the predictions to the data, arguing that it suggests both overly fine and overly coarse expectations along the egocentric logic of the model, and also to alternative explanations.

5.2 Additive Lemons Problem

Seller-Offer: **Under-Bluffing.** Following SB, suppose first that $w(q) = q + x$. Consider the case where the seller has the bargaining power and makes a take-it-or-leave-it (TIOLI) price offer $p_s(q)$ which the buyer can accept or reject. In a perfect BNE the seller cannot sell different qualities at different prices for sure (or with equal positive probabilities). Since the seller would never have an incentive to name the lower of any two such prices, hence, such a pricing would not be incentive compatible. This well-known fact limits trade. In contrast, the next proposition shows that projection makes the seller too reluctant to bluff which relaxes the relevant IC constraint and expands trade.[28]

Proposition 10 *For any $\rho \geq 0$, there exists a ρ projection equilibrium where $p_s^\rho(q) = q + x$ and the buyer accepts any price $p \leq \overline{p} = \min\{\frac{x}{1-\rho}, \overline{q} + x\}$ for sure and any price $p > \overline{p}$ with probability $e^{-(p-\overline{p})/x}$.*

Two properties characterize the above prediction. First, the seller engages in non-altruistic truth-telling: her price fully reveals the object's quality, but leaves no rent to the buyer. Second, she bluffs too little. She appears selfish, but collects less rent than she could, given the buyer's actual behavior.

To see the logic, note that by projecting her information, the seller exaggerates the probability with which the buyer can detect an overpriced offer. Such an overly fine expectation about the buyer's strategy limits the seller's incentive to bluff. The buyer partially anticipates the seller's projection. At the same time, by virtue of *all-encompassing* projection, the he thinks that the projected version of the seller knows that he does not know q, but by projecting his ignorance, he also thinks that she cannot base deviations on the realizations of q. The bound on $\overline{p}$, then ensures that neither the real nor the projected version of the seller has an incentive to deviate from truth-telling.

[28] The standard restriction of perfectness for off-equilibrium path beliefs, given projection, is always satisfied.

Evidence. The findings of SB, in their setting, $x = 30$, and $q \sim U[0, 100]$, are in-line with the above predictions. Half of the sellers bid exactly $p_s(q) = q+x$. In addition, sellers significantly underbid relative to what their empirical payoff maximizing strategy would have been given the buyers' acceptance behavior. Nearly three-quarters of the buyers accepted all prices below $\overline{q} + x = 80$.

Finally, note that if $\rho = 0$, the seller's maximal profit is achieved when the seller only sells at $p_s = 60$, e.g., Samuelson (1984). If ρ is not too low, it is easy to see that the above projection equilibrium always generates higher seller profit and higher overall surplus.[29]

Buyer-Offer: **Winner's Curse.** Suppose now that the buyer has the bargaining power and makes a TIOLI price offer p_b that the seller can accept or reject. The analysis is now simplified because both the real and the projected versions of the seller have dominant strategies. The real seller accepts a price iff it is greater than q, the projected seller accepts a price iff it is greater than $\overline{q}$. In the above specification studied by SB, this implies the following claim:

Claim 2 *In the unique ρ projection equilibrium, the buyer's bid is given by $p_b^\rho = 30$ if $\rho \leq 1/16$ and by $p_b^\rho = 50$ if $\rho > 1/16$.*[30]

In the unique projection equilibrium, if $\rho < 1/16$, the buyer names the same prices as under the unique BNE, otherwise, the buyer names a price that equals the seller's unconditional valuation $\overline{q}$. Since the buyer projects her ignorance, and believes that with probability ρ the seller would accept any price above $\overline{q}$, independent of q, he underappreciates selection and falls prey to the classic winner's curse.

Evidence. The empirical findings of SB again match the predictions. In their data the most common empirical bid was 50. Less than 17% of bids were below 40, nesting the BNE prediction, and a non-trivial fraction of bids were in [60,80]. Under correct expectations, bidding above 60 leads to strictly negative expected payoff for the buyer. In contrast, bidding below 80 can still lead to positive perceived payoff under projection. Fudenberg and Peysakhovich (2013) also study empirically the isomorphic problem with $x = 3$ and $q \sim U[0, 10]$. Exactly as predicted by projection equilibrium, the average empirical bid in their sample was 5.1.

Cursedness versus Projection. Cursed equilibrium, $CE(\chi)$, parametrized by the degree of cursedness χ, is often motivated by addressing uninformed players behavior in adverse selection problems, and, to the best of my knowledge, of the equilibrium alternatives to BNE, cursed equilibrium provides the tightest match of this data.[31] Let me then compare the predictions of projection equilibrium to that of CE.

[29] If $\rho \geq \frac{5}{8}$, then $\overline{p} = \overline{q} + 30 = 80$, and the seller's expected payoff in the ρ-PE of Proposition 10 is 72.3. Under the seller-optimal BNE, the seller's expected payoff is 68.

[30] The prediction of ρ-PE is identical to that of the nested model with $\rho^{\pm} = \rho$, allowing for any $\rho^{+} < 1 - \rho$. The same holds for the multiplicative lemons problems.

[31] Fully cursed equilibrium is equivalent to the ABEE with the private information analogy partition.

Consider first the seller-offer game. In a cursed equilibrium the sender knows that the buyer does not know q and has correct expectations about the buyer's distribution of actions on average. Since, however, the buyer is uninformed, thus, plays the same strategy in each state, a cursed seller must still have correct expectations about the buyer's strategy. In turn, no under-bluffing, of the sort empirically observed can exist.

In the buyer-offer game the predictions of $CE(\chi)$ on the buyer's price p_b span the interval of $[30, 40]$, as a function of the degree of cursedness χ, with 40 being the fully cursed prediction. Hence, cursed equilibrium accounts for 17% of behavior in the data of SB. Similarly, if $x = 3$ and $q \sim U[0, 10]$, then $CE(\chi) \in [3, 4]$.[32] Finally, here, the predictions of projection equilibrium, over all possible values of ρ, are also concentrated on only two price points lending power to its predictions.

5.3 Multiplicative Lemons Problem

Evidence. Holt and Sherman (1994) test a multiplicative specification of the lemons problem where the buyer's valuation is given by $w(q) = 1.5q$. They focus only on the buyer-offer game. Table 1 below, adopted directly from Eyster and Rabin (2005), describes the BNE prediction, b^0, the fully cursed prediction, $b^{\chi=1}$, with partially cursed equilibrium spanning the interval between these two. The unique ρ projection equilibrium price, b^ρ is again concentrated on only two points. It equals the BNE prediction if $\rho \leq \rho^*$; and equals $\overline{q}$ otherwise. The average empirical bid is denoted by $\overline{b}$.

	q_0	r	b^0	$b^{\chi=1}$	$b^{\rho>\rho^*}$	$\overline{b}$	ρ^*
No curse	1	2	2	2	2	2	0
Winner's curse:	1.5	4.5	3	3.56	3.75	3.78	0.02
Loser's curse:	0.5	0.5	1	0.81	0.75	0.74	0.07

In the winner's curse condition, the prediction of the model matches the data almost exactly for *any* ρ greater than 2%. In the loser's curse condition, it matches it for *any* ρ greater than 7%. The reason that such a small degree of projection is sufficient to robustly generate such substantial deviations from the BNE is that the buyer's gain from trade conditional on selection is much lower than the buyer's gain from trade in the absence of selection.

Finally, Ball et al. (1991) also study the case where $w(q) = 1.5q$ and allow for 20 rounds of learning. Here, in reality, any positive bid by the buyer leads to a strictly negative expected earning for him as long as the seller respects dominance. The analogue of Table 1 is:

q_0	r	b^0	$b^{\chi=1}$	$b^{\rho>\rho^*}$	$\overline{b}$	ρ^*
0	1	0	0.375	0.5	0.55	0.2

[32] Esponda (2008) also proposes a boundedly rational equilibrium model for adverse selection settings. In the setting studied by SB, his model predicts that $p_b \in [20, 30]$.

Both cursedness and an ignorance projection describe similar deviations in these contexts but their logics differ. A cursed buyer may act as if he thought that the seller knowingly played a dominated strategy. Instead a projecting buyer always acts as if he believed that the seller acted rationally given her (lack) of information. To illustrate a qualitative consequence of this, suppose now that $w(q) = mq$ with $0 < m < 1$ that is, the buyer always values the object strictly less than the seller does. Here, both BNE and projection equilibrium still predict that the buyer must bid zero. Instead a fully cursed buyer can still name a strictly positive price.

Projecting Valuations. Finally, the data is inconsistent with the hypothesis that players mistakenly think that others have the same valuations, 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 valuations. They both acted as if they exploited the correct and fully binding individual rationality constraints given differences in valuations, but ignored informational differences.

6 Conclusion

Motivated directly by robust psychological evidence, this paper introduces a potential key form of egocentric thinking into strategic settings. I demonstrate the model's relevance by exploring its implications to three distinct settings, and show that the model helps provide a more unified explanation of empirical phenomenon even in purely canonical settings such as common-value trade.

A key aspect of the proposed model is that projection is partially public and players act as if they partially anticipated, but underestimated each other's biases. Danz, Madarász, and Wang (2015) directly tests the model, including its partial anticipation aspect, in the context of a visual task. They find not only that people significantly project their information onto others, and that they significantly anticipate others' projection onto them, but, crucially, also that they significantly underestimate the extent to which others project onto them. Furthermore, the empirical extent to which people project onto others, as predicted to be ρ, and the empirical extent to which they underestimate others' projection onto them, as predicted to be ρ^2, are remarkably consistent with the logic proposed in this paper.

Future research can explore the generality of this finding, and directly assess the model's empirical relevance in a great variety of domains. It would also be interesting to explore the extent to which the model may offer a unified understanding of a variety of empirical findings in social psychology, clarify their empirical content, and help more easily integrate these parsimoniously into economics.

Limited informational perspective taking may matter in many settings not covered in this paper. I believe that projection equilibrium can provide many insights in such contexts. Future research can generalize and refine the model's predictions for a host of

strategic signalling problems, as well as, poker. A context where information projection may be particularly important is mechanism design. Optimal mechanisms are often concerned with the appropriate provision of economic rents in exchange for agents' private information. One may then consider the performance of various mechanisms in settings where agents' demand for such rents are distorted in a manner specified by the model. Along these lines, Madarász (2014b) adopts a dynamic extension of the model and explores its implications to classic problems of dynamic bargaining.

7 Appendix A — Extensions (For Online Publication)

7.1 Private Projection

In the main approach adopted throughout this paper projection is partially public. Let me briefly describe an alternative specification where instead projection is fully private. I again present it for two-players and generalize it for N-player games below.

Definition 4 *Let $\sigma^0 \in S_i \times S_j$ be a BNE of Γ. A strategy profile $\sigma^\rho \in S_i \times S_j$ is a ρ-private information projection equilibrium (PIPE) of Γ if there exists $\sigma^+ \in S_i^+ \times S_j^+$ such that for each i and j,*

$$\sigma_i^\rho \in BR_{S_i}\{(1-\rho)\sigma_j^0 \circ \rho\sigma_j^+\}, \tag{7}$$

and

$$\sigma_j^+ \in BR_{S_j^+}\{\sigma_i^0\}. \tag{8}$$

■ In a ρ-PIPE players' higher-order expectations about each other's play are all anchored to a common BNE profile σ^0 of Γ. Each player i believes that her opponent expects her to play her part of this profile σ_i^0, and also expects that her opponent plays his part of this profile, that is σ_j^0, with probability $1-\rho$. At the same time, each i comes to mistakenly believe that with probability ρ her information is unexpectedly shared with her opponent who then best responds to σ_i^0 given his and her joint information in the game. Player i's strategy σ_i^ρ is a best response to such mistaken beliefs.[33]

■ In a ρ-PIPE it is only each player's first-order belief about her opponent's information (and play) that is perturbed relative to the unbiased case. Projection is fully private. In poker, Judith believes that, with probability ρ, Paul knows her hand but believes that she still thinks that he never knows her hand. Judith projects the content of her private information, but not her thoughts about what information she thinks Paul has. Analogously, Judith still expects Paul to believe that she does not know her hand. Hence, she also fails to anticipate any projection by Paul.

■ Under private projection, given that it corresponds to a minimal departure where only players first-order expectations about the information of others is perturbed, each

[33] For the model's predictions to collapse to those of BNE for $\rho = 0$, one also needs to impose the restriction that $\sigma^{\rho=0} = \sigma^0$ to avoid the possibility of miscoordination.

player may act as if she thought that her opponent had *fully* misspecified expectations about her strategy, that is, expectations that did not contain in their support her actual strategy. Each player may, in turn, form fully misspecified expectations about her opponent's strategy. The fact that projection is private naturally implies that this alternative violates both the all-encompassing nature of projection, and the limited consistency property.

To illustrate, let me return to Example 1. The defender's strategy and expected payoff are now described by the following table:

defender	weak	strong	EU_D^ρ
BNE	B	A	$1/2$
PIPE $\rho > \frac{w}{2-w}$	A	B	$1/2 - w/4$

A biased defender thinks that with probability ρ the attacker has figured out her strength but not that she has figured this out about him and still expects her to follow her BNE strategy. If she is weak, she then expects the projected attacker to attack at B; if she is strong, she expects him to attack A. Hence, if her bias is not too small, she plays the exact reverse of her BNE strategy; she always protects her Achilles heel (and strictly prefers to). Since the attacker does not anticipate the defender's mistake, he may continue to mix symmetrically. In turn, the same choking effect holds as before *a fortiori*.

7.2 N-Player Extensions

Consider an N-player game of the form described before. I first extend the model of ρ-IPE. Let $S = \prod_{i \in N} S_i$ be the true strategy space given a finite set of actions for each player. Each player i now imagines a projected version for each of her opponents. Since the information of players i and j may 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:

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

I 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. I denote the generic element of this set S^{+i} by σ^{+i}. Finally, I 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 versions. Furthermore, the projected version of k, as imagined by player i, believes that j is also the projected version of player j as imagined by player i.

Definition 5 *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*

$$\sigma_i^\rho \in BR_{S_i}\{(1-\rho)\sigma_{-i}^\rho \circ \rho\sigma^{+i}\}, \tag{9}$$

and for each $k \neq i$

$$\sigma_k^{+i} \in BR_{S_k^{+i}}\{\sigma_i^\rho, \sigma_{-k}^{+i}\}. \tag{10}$$

The definition continues to satisfy the same all-encompassing and consistency properties as before. Existence follows from the same argument 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 5, and then S^{+i} with $S^i = \prod_{k\neq i\in N} S_k^i$ in the above definition.

■ The extension of ρ-PIPE is also analogous to that of ρ-IPE. Given a BNE profile $\sigma^0$of Γ, it is obtained by replacing $\{(1-\rho)\sigma_{-i}^\rho \circ \rho\sigma^{+i}\}$ with $\{(1-\rho)\sigma_{-i}^0 \circ \rho\sigma^{+i}\}$ in Eq.(9), and replacing $\{\sigma_i^\rho, \sigma_{-k}^{+i}\}$ with $\{\sigma_{-k}^0\}$ in Eq. (10).

8 Appendix B – Proofs

Proof of Proposition 1. 1. Best-response correspondences are upper-hemicontinuous and convex and the existence of the models follow from Kakutani's fixed-point theorem. 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 .

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 5 further shows that all equilibria are in cut-off strategies for the real versions as well.

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

$$r(\rho(x(\theta_i^\rho - \gamma\theta_i^\rho) - nc) + (1-\rho)((x - \theta_{-i}^\rho)(\theta_i^\rho - \gamma\theta_i^\rho) + \theta_{-i}^\rho(\gamma\theta_i^\rho) - nc)) = 0. \tag{11}$$

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 interior solution is $\theta_i^\rho = \sqrt{cn/(1-\rho)}$. When no interior solution exists, I 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 $\Pr(in)^\rho_{\theta_i}$ denote real θ_i's perception of the probability with which $-i$ invests. Let $\Pr(in)$ be the corresponding true probability. Since the martingale property of beliefs must hold in a ρ-IPE with respect to the *perceived* probability, it follows that for any θ_i:

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

where $a_i^{\theta_i}$ is θ_i's equilibrium action. For any given ρ and θ_i, let's define the function $\Delta^\rho_{\theta_i}(\theta_{-i})$: $[-n, x] \to \mathbb{R}$ as:

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

Note that $\int_{-n}^{x}\Delta^\rho_{\theta_i}(\theta_{-i})d\theta_{-i} = 0$ and $\Delta^\rho_{\theta_i}(\theta_{-i})$ 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:

$$\pi_0 - E[\pi_1^\rho \mid \theta_i] = (\Pr(in)^\rho_{\theta_i} - \Pr(in))\Delta^\rho_{\theta_i}(\theta_{-i}). \tag{12}$$

If $\theta_i > 0$, then $\Pr(in)^\rho_{\theta_i} > \Pr(in)$, hence, $E[\pi_1^\rho \mid \theta_i] <_{fosd} \pi_0$. If $\theta_i < 0$, then $\Pr(in)^\rho_{\theta_i} \leq \Pr(in)$, 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. Given the symmetry of the prediction and the fact that valuations are drawn independently across players, 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 of Corollary 1. Follows directly from the proof of Corollary 3 .

Proof of Corollary 2. Following investment by either of the players, each player has a dominant continuation strategy: invest iff $\min\{\theta_i, \theta_{-i}\} \geq 0$. Suppose now that there is no investment 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 investment 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 ρ-IPE for each t follows immediately from Proposition 2. If $\rho = 0$, then $v_t^0 = y_t^0$ for each t and, using Eq.(11), $x_t^0 = \sqrt{nc_t}$. If $\rho > 0$, then:

$$y_t^\rho / v_t^\rho = y_{t-1}^\rho / (1-\rho)v_{t-1}^\rho. \tag{13}$$

Re-writing Eq.(11), re-weighting terms with the corresponding densities and solving for the unique fix point, one obtains that:

$$x^{\rho}_{t+1}=\min\{\sqrt{\frac{c_{t+1}n}{1-\rho}\frac{y^{\rho}_t}{v^{\rho}_t}},x^{\rho}_t\}. \tag{14}$$

Thus,the cutoff of round $t+1$, conditional on no investment till t, x^{ρ}_{t+1} is increasing in ρ. Hence, $\Pr^{\rho}_{\underline{c}}(m)$ is decreasing in ρ .

Proof of Corollary 3. Following investment in any round t, players' estimates of their opponents remain constant. Suppose now that there is no investment till $t \geq 0$ where I simply denote x by x_0.[34] Let $\rho > 0$.

1. Notice that $E[\overline{\pi}^{\rho,+}_{t+1} \mid \text{no investment till } t] = E[\overline{\pi}^{\rho}_t \mid \theta_i \geq 0, \text{ no investment till } t]$ 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 expression with respect to x^{ρ}_{t+1}, one gets that:

$$-\frac{x^{\rho}_{t+1}n(x^{\rho}_{t+1}v^{\rho}_t(1-\rho)+2ny^{\rho}_t)}{x^{\rho}_t(n+x^{\rho}_t)}\frac{y^{\rho}_t-(1-\rho)v^{\rho}_t}{(y^{\rho}_t n+(1-\rho)v^{\rho}_t x^{\rho}_{t+1})^2}<0,$$

where the inequality follows from the fact that, given Eq.(13), $v^{\rho}_t \leq y^{\rho}_t$. Hence, since x_{t+1} is increasing in c_{t+1}, it follows that $E[\overline{\pi}^{\rho,+}_{t+1} \mid \text{no investment till } t]$ is decreasing in c_{t+1}. If $x^{\rho}_{t+1}=0$, then $E[\overline{\pi}^{\rho}_{t+1} \mid \theta_i \geq 0, \text{ no investment till } t] = \overline{\pi}_0$ since, here, equilibrium fully reveals the direction if each player's preference. Hence, $E[\overline{\pi}^{\rho,+}_{t+1}] < \overline{\pi}_0$ 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 $E[\overline{\pi}^{\rho,-}_{t+1} \mid \text{no investment till } t] = E[\overline{\pi}^{\rho}_t \mid \theta_i < 0, \text{ no investment till } t]$ 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:

$$-\frac{n}{n+x^{\rho}_t}\frac{(n+x^{\rho}_t)(1-(1-\rho)^{t+1})+\sum_{s=1}^{t}(x^{\rho}_{s-1}-x^{\rho}_s)(1-(1-\rho)^s)}{[n+x^{\rho}_{t+1}+\sum_{s=1}^{t+1}(x^{\rho}_{s-1}-x^{\rho}_s)(1-(1-\rho)^s)]^2}<0,$$

[34] 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.

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

Proof of Corollary 4. Iterating Eq.(14) and Eq.(13) from $t = 1$ on, it follows that if $c_t \geq xl(1-\rho)^t$ for all t, then there is no investment in any t. Furthermore, from Eq.(13) it follows that $\lim_{t\to\infty} E[\overline{\pi}_t^{\rho,+}] = 0$. Since along this sequence $x_t^\rho = x_{t+1}^\rho$ for each t, 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 of Proposition 3. $\Pr_{\underline{c}}^\rho(m) = 0$ implies $x_t^\rho \geq x$ for all t. If $\rho = 0$, then in optimum this constraint binds for each t. If $\rho > 0$, by backward induction, the constraint must bind in round T. Suppose it first binds in round $t+1$. One can then decrease c_t by some Δ such that it binds already in round t and, given Eq.(14), increase c_{t+1} by at most $(1-\rho)\Delta$ for the constraint to still bind in $t+1$, and leave all else unaffected. Since $(1-\rho)\Delta < \Delta$, a non-binding constraint in round t cannot be optimal. Hence, $c_{t+1}^*(\rho) = c_t^*(\rho)(1-\rho)$ must hold for any t, and $\lim_{T\to\infty} \sum_{t=1}^T c_t^*(\rho)/T = 0$ iff $\rho > 0$.

Proof of Proposition 4. If $\theta_{i,T}$ is an i.i.d. draw from a uniform density on $[-n, x]$, then, given the transition process of Eq.(4), $\theta_{i,1}$ can always be represented as an i.i.d. draw from a piece-wise constant density $y_1 = (n+x)^{-1}$ on $[-n, 0]$ and v_1 on some $[0, x_1]$ where $v_1 = \frac{x}{x_1(n+x)}$.

Consider $\rho = 0$. From the proof of Proposition 2 it follows that the cutoff of round 1 is given by $\sqrt{c_1 n y_1 / v_1} = \sqrt{c_1 n x_1 / x}$. Conditional on no investment in round 1, given Eq.(4) and Eq.(14), the cutoff of round 2 is given by $\min\{\sqrt{c_1 n x_1/x}, \sqrt{c_2 n \alpha_1 x_1/x}\}$. By induction on t, the (potentially non-internal) cutoff of round T, conditional on no investment till then, is $\min\{\sqrt{c_1 n x_1/x}, \sqrt{c_2 n \alpha_1 x_1/x}, ..., \sqrt{c_T n \underline{\alpha} x_1/x}\}$ where $\underline{\alpha} = \prod_{t=1}^T \alpha_t$. Since $r_T = r^*$, thus, $x_1 \underline{\alpha} = x$, it follows that $\sqrt{c_T n \underline{\alpha} x_1/x} = \sqrt{c_T n}$. Hence, any player i with valuation $\theta_{i,T}$ greater than $\sqrt{c_T n}$ invests at least once along sequence $\underline{r}$. Since $c_t \leq \varepsilon$ for all t, it then follows that $\Pr_{r^*}^0(m) \geq 1 - \varepsilon n/x^2 = 1 - \varepsilon/xl$.

Consider now $\rho > 0$. Let $c_t = \varepsilon$ for all t. Let x_1 be such that $x_1 \leq \sqrt{\varepsilon n y_1/(1-\rho)v_1}$. Since $v_1 = \frac{x}{x_1(n+x)}$, this is equivalent to $x_1 \leq \varepsilon n/x(1-\rho)$. From Proposition 2 it follows that there is no investment at $t = 1$. Given Eq.(13), the unique (possible non-internal) equilibrium cutoff of $t = 2$ is $\sqrt{\varepsilon n \alpha_1 x_1/x(1-\rho)^2}$. Simple re-arrangements imply that if $\alpha_1 \leq (1-\rho)^{-1}$, then this cutoff is greater than $\alpha_1 x_1$ and there is no investment in $t = 2$ either. Set $\alpha_t = (1-\rho)^{-1}$ for all $t \leq \bar{t}$ and set x_1 to be the largest number such that $x_1 = x(1-\rho)^{\bar{t}-1}$ for some integer $\bar{t}$ while $x_1 \leq \varepsilon/l(1-\rho)$ is still satisfied. Finally, set $\alpha_t = 1$ for all $t > \bar{t}$. The logic then follows from above .

Proof of Proposition 5. 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 any real player with a given valuation $\theta_i > 0$, the perceived expected utility difference between 'in' versus 'out' is:

$$\rho(rx(\theta_i - f(\theta_i)) + \int_{-n}^{0} rg(\theta_i, \theta_{-i})d\theta_{-i}) + (1-\rho)(z_{-i}(\theta_i - f(\theta_i)) + (1-z_{-i})E[g(\theta_i, \theta_{-i}) \mid \sigma_{-i}(\theta_{-i}) = out]). \quad (15)$$

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.(15) 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.(15), one gets that an internal equilibrium cutoff must satisfy:

$$\overbrace{\rho[\int_0^{\theta_{-i}^\rho} r(\theta_i^\rho - f(\theta_i^\rho) - g(\theta_i^\rho, \theta_{-i}))d\theta_{-i}]}^{V} + \int_{\theta_{-i}^\rho}^{x} r(\theta_i^\rho - f(\theta_i^\rho))d\theta_{-i} + \int_{-n}^{\theta_{-i}^\rho} rg(\theta_i^\rho, \theta_{-i})d\theta_{-i}] = 0. \quad (16)$$

If investments are substitutes, Term V is strictly negative. Holding $(\theta_i^\rho, \theta_{-i}^\rho)$ fixed, the LHS of Eq.(16) is strictly decreasing in ρ. Hence, the unique symmetric equilibrium cutoff must increase in ρ.

If investments are complements, Term V is strictly positive. Holding $(\theta_i^\rho, \theta_{-i}^\rho)$ fixed, the LHS of Eq.(16) 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 of Claim 1. Since the benefit of checking is constant in c, the receiver adopts a cutoff strategy. The indifferent type is $c^0 = p^0/(1+p^0)^2$, hence, $c^0 = \min\{F^{-1}(B), c_{\max} \equiv 1/4\}$.

Proof of Proposition 6. Suppose that $\rho > 0$. First, $p^+(c)$ must be weakly increasing. If for some $c'' > c'$, $p^+(c'') < p^+(c')$ was true, type c'' would have a strictly lower incentive to check than type c'.This would imply that $p^+(c'') > p^+(c')$; a contradiction. Second, $p^+(c)$ must also be continuous. A jump in $p^+(c)$ at some $\widehat{c}$ would imply that a type just above $\widehat{c}$ checked strictly more than $\widehat{c}$, but checking cannot strictly increase in c because then $p^+(c)$ could not be weakly increasing. Hence, there exist c_1^ρ and c_3^ρ such that the receiver always checks iff $c < c_1^\rho$, and never checks iff $c > c_3^\rho$. Furthermore, $p^+(c_1^\rho) = 0$, $p^+(c_3^\rho) = 1$ and $p^+(c)$ must strictly increase on $[c_1^\rho, c_3^\rho]$ with types in $[c_1^\rho, c_3^\rho]$ checking with probability B.

Finally, let me show that c_3^ρ is increasing and c_1^ρ is decreasing in ρ. Suppose instead that for some $\rho' > \rho$ it was true that $c_3^{\rho'} < c_3^\rho$. Then $p^{\rho'} < p^\rho$ since c_3^ρ must increase both in p^ρ and in ρ separately because $p^\rho \leq p^+(c_3^\rho)$. It then also follows that $c_1^{\rho'} < c_1^\rho$ because c_1^ρ is increasing in p^ρ and decreasing in ρ separately because $p^\rho > p^+(c_1^\rho)$. But if both $c_1^{\rho'} < c_1^\rho$ and $c_3^{\rho'} < c_3^\rho$, then, $F(c_1^{\rho'})(B-1) + (1-F(c_3^{\rho'}))B > F(c_1^\rho)(B-1) + (1-F(c_3^\rho))B$, hence, the sender has a strictly higher incentive to lie under ρ' which implies that $p^{\rho'} > p^\rho$; a contradiction. Suppose now that for some $\rho' > \rho$ it was true that $c_1^{\rho'} > c_1^\rho$. It follows that $p^{\rho'} > p^\rho$ which then implies that $c_3^{\rho'} > c_3^\rho$. But if both $c_1^{\rho'} > c_1^\rho$ and $c_3^{\rho'} > c_3^\rho$, then, for the same reason as above, $p^{\rho'} < p^\rho$; a contradiction .

Proof of Proposition 7. The sender's incentive condition, for any interior $p^\rho \in (0,1)$, is given by:

$$B = F(c_1^\rho)/(1 - F(c_3^\rho) + F(c_1^\rho)). \tag{17}$$

The LHS of Eq.(17) is increasing in B. Hence p^ρ, thus, also c_3^ρ and c_1^ρ, must increase in B. Since $c_3^\rho \leq c_{\max}$, if B is sufficiently high, Eq.(17) can no longer hold. Instead, $c_2^\rho = c_3^\rho$ binds, and $p^\rho = 1$. The stated threshold $\overline{B}(\rho, F)$ follows and also that $\overline{B}(\rho, F) < 1$ because $F(c_{\max}) < 1$. Finally, by Proposition 7, c_3^ρ increases and c_1^ρ decreases in ρ, hence, $\overline{B}(\rho, F)$ is decreasing in ρ.

Let's rewrite Eq.(17) as:

$$B = F(c_3^\rho)B + F(c_1^\rho)(1 - B). \tag{18}$$

All else constant, an increase in F decreases the RHS of Eq.(18). Hence, c_3^ρ and p^ρ must increase in F. Since such an increase is again bounded, the stated existence of $\overline{F}(\rho, B)$ follows and, given Proposition 7, if $B > F(c_3^\rho)B + F(c_1^\rho)(1 - B)$, the same holds given any $\rho' > \rho$, a valid $\overline{F}(\rho, B)$ for ρ is also valid for all $\rho' > \rho$.

Proof of Proposition 8. The case of $\rho = 0$ is immediate. If uniform credulity holds, $c_1^\rho = (1-\rho)/(2-\rho)^2$ and $c_3^\rho = c_{\max}$. Let $y_c^\rho(+)$ denote type c's investment conditional on receiving a positive recommendation and not checking. If $c \in (c_1^\rho, c_3^\rho)$, then $y_c^\rho(+) = (1+\overline{p}^\rho(c))^{-1}$ where $\overline{p}^\rho(c)$ is:

$$\overline{p}^\rho(c) = \left(1-2c-\sqrt{1-4c}\right)/2c. \tag{19}$$

Let $E[u^\rho|c]$ be type c's welfare. For a $c \in [c_1^\rho, c_3^\rho]$, $E[u^\rho|c]$ is:

$$B(-c) - 0.5(1-B)((1-y_c^\rho(+))^2 + y_c^\rho(+)^2) = (2B-1)(c_{\max}-c) - c_{\max}\,, \tag{20}$$

where I used Eq. (19) and some re-arrangements. The above expression is increasing in B. Consider now $c \notin [c_1^\rho, c_3^\rho]$. Here, $E[u^\rho|c]$ is independent of B. Hence, an increase in B decreases investment and increases welfare.

Note if $B < \frac{1}{2}$, Eq.(20) implies that $E[u^\rho|c]$ is strictly increasing in c on $[c_1^\rho, c_3^\rho]$. Furthermore, for all $c > c_{\max}$, $E[u^\rho|c]$ is constant in c since those types never check and $y_c^\rho(+) = \overline{y}$. Hence, an increase in F which leaves $F((1-\rho)/(2-\rho)^2)$ unaffected shifts probability weight to types with lower investment and higher welfare .

Proof of Proposition 9. Since uniform credulity holds if $B \geq \overline{B}(\rho, F)$, there exists $\widehat{\gamma}(\rho, F)$ such that $\widehat{\gamma}(\rho, F)[R(\rho, \overline{B}(\rho, F), F) - \overline{y}] = \overline{B}(\rho, F)$. Hence, $B^*(\rho, F, \gamma) > 0$ anytime that γ is larger than some $\overline{\gamma}(\rho, F)$. Furthermore, if $B^*(\rho, F, \gamma) > 0$, then $E_{c,\theta}[y_c^{\rho,*}] > \overline{y}$ must hold. Since $\lim_{\rho \to 1} \overline{B}(\rho, F) = 0$, then $\lim_{\rho \to 1} \overline{\gamma}(\rho, F) = 0$.

Proof of Corollary 5. The case of $\rho = 0$ is immediate. As $\rho \to 1$, $c_1^\rho \to 0$ and $\overline{B}(\rho, F) \to 0$, thus, uniform credulity always holds. Hence, $B^*(\rho, F, \gamma) < \frac{1}{2}$ also holds and $E[u^\rho|c]$ is now strictly increasing for all $c \leq c_{\max}$. Since $y_c^\rho(+)$ is now globally decreasing in c, a decrease in F decreases receiver welfare and increases the seller's expected profit .

Proof of Proposition 10. The real and the projected buyer both accept any price on the equilibrium path and reject any price greater than $q_0 + r + x$. The projected seller names a price of $\overline{q} + x$. Consider deviations by the real seller. If $q + x < \overline{p}$, then (i) deviating upward to some price $p < \overline{p}$ leads to a payoff bounded from above by $(1-\rho)\overline{p} + \rho q \leq q + x$; (ii) deviating upward to some price $p > \overline{p}$ generates an expected payoff of:

$$(1-\rho)(pe^{-(p-\overline{p})/x} + q(1 - e^{-(p-\overline{p})/x})) + \rho q < q + x,$$

since here $pe^{-(p-\overline{p})/x} + qe^{-(p-\overline{p})/x} \leq q + x$. If $q + x \geq \overline{p}$, then naming a price of $p = q + x$ maximizes $pe^{-(p-\overline{p})/x} + q(1 - e^{-(p-\overline{p})/x})$. Consider now deviations by the projected seller. Deviating to a price $p > \overline{q} + x$ is not profitable since for the projected seller $\overline{q} + x$ maximizes $pe^{-(p-\overline{p})/x} + \overline{q}(1 - e^{-(p-\overline{p})/x})$. Deviating to a price $p < \overline{p}$ leads to a loss both if $\overline{p} = \overline{q} + x$ and if $\overline{p} = x/(1-\rho)$ because in the latter case $p = \overline{q} + x$ is again optimal for the projected seller .

References

[1] Acemoglu, Daron, and James Robinson. (2001). "A Theory of Political Transitions." *American Economic Review,* 91: 938-963.

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

[3] Allport, Floyd H. (1924). *Social Psychology.* Boston: Houghton Mifflin.

[4] 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.

[5] Baron, Jonathan, and John Hershey. (1988). "Outcome Bias in Decision Evaluation." *Journal of Personality and Social Psychology*, 54(4): 569–579.

[6] Baron-Cohen, Simon, Leslie Alan, and Uta Frith. (1985). "Does the Autistic Child Have a "Theory of Mind"?" *Cognition*, 21: 37-46.

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

[8] Bernanke, Ben. (2011). *Statement by Chairman Bernanke on Financial Literacy.* www.federalreserve.gov/newsevents/testimony/bernanke20110420a.htm

[9] Bergstresser, Daniel, John Chalmers, and Peter Tufano. (2009). "Assessing the Costs and Benefits of Brokers in the Mutual Fund Industry." *Review of Financial Studies*, 22(10): 4129-4156.

[10] Besley, Timothy, and Stephen Coate. (1997). "An Economic Model of Representative Democracy." *Quarterly Journal of Economics*, 112, 85–114.

[11] Biais, Bruno and Martin Weber. (2010). "Hindsight Bias, Risk Perception and Investment Performance." *Management Science*, 55(6), 1018-1029.

[12] Birch, Susan, and Paul Bloom. (2007). "The Curse of Knowledge in Reasoning About False Beliefs." *Psychological Science*, 18(5): 382–386.

[13] 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.

[14] 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.

[15] Crawford, Vincent, and Joel Sobel. (1982). "Strategic Information Transmission." *Econometrica*, 50: 1431-1451.

[16] Danz, David, Kristóf Madarász, and Stephanie Wang. (2015). "The Biases of Others: Anticipating Informational Projection in an Agency Setting." mimeo LSE and U of Pittsburgh.

[17] 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.

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

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

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

[21] Frith, Christopher, and Rhiannon Corcoran. (1996). "Exploring 'Theory of Mind' in People with Schizophrenia." *Psychological Medicine*, 26: 521-530.

[22] Fudenberg, Drew, and Alex Peysakhovich. (2013). "Recency, Records and Recaps: Learning and Non-Equilibrium Behavior in a Simple Decision Problem." mimeo Harvard.

[23] Kartik, Navin, Marco Ottaviani, and Francesco Squintani. (2007). "Credulity, lies, and costly talk." *Journal of Economic Theory*, 134: 93–116.

[24] Kunda, Ziva. (1999). *Social Cognition.* MIT Press.

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

[26] Gilovich, Thomas, Victoria Medvec, and Kenneth Savitsky. (1998). "The Illusion of Transparency: Biased Assessments of Others' Ability to Read One's Emotional States." *Journal of Personality and Social Psychology*, 75(2): 332–46.

[27] Hastings, Justine, Brigitte Madrian, and William Skimmyhorn. (2013). "Financial Literacy, Financial Education, and Economic Outcomes." *Annual Review of Economics*, 5: 347–73.

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

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

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

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

[32] Lusardi, Annamaria. (2013). *Testimony Before the U.S. Senate Committee on Health, Education, Labor and Pension.* April 24, 2013. http://www.help.senate.gov/imo/media/doc/Lusardi.pdf.

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

[34] Madarász, Kristóf. (2014a). "Projection Equilibrium: Definition and Applications to Social Investment and Persuasion." mimeo LSE, CEPR D.P. 10636.

[35] Madarász, Kristóf. (2014b). "Bargaining under the Illusion of Transparency." CEPR D.P. 10327.

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

[37] NASD Investor Education Foundation. (2006). *Investor Fraud Study Final Report.* https://www.sec.gov/news/press/extra/seniors/nasdfraudstudy051206.pdf .

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

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

[40] 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.

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

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

[43] Samuelson, William, and Max Bazerman. (1985). "The Winner's Curse in Bilateral Negotiations." Working Paper MIT Sloan; also in *Research in Experimental Economics*, vol. 3, Vernon L. Smith, ed., Greenwich, CT: JAI Press.

[44] Shamir, Jacob, and Michal Shamir. (1997). "Pluralistic Ignorance Across Issues and Over Time." *Public Opinion Quarterly*, 61 (2): 227-260.

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

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

[47] Willis, Lauren. (2011). "The Financial Education Fallacy." *American Economic Review*, 101: 429-434.

[48] Wimmer, Heinz and Joseph Perner. (1983). "Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception." *Cognition*, 13(1): 103–128.