2014

Projection Equilibrium: Definition and Applications to Social Investment and Persuasion (longer older version with private projection and auctions)

Kristof Madarasz, *London School of Economics and Political Science*

Projection Equilibrium: Definition and Applications to Social Investment and Persuasion

Kristóf Madarász (LSE) [1]

First Circulated Version, May, 2013. This Version, January 2015 (minor revision 10/15).

[1]First online version July 2014. I would like to thank audiences at Arizona, Bonn, Columbia, Harvard, Princeton, Stockholm, UC Los Angeles, UC San Diego, Yale, ITAM, Wash U, Southampton, Royal Holloway, Berlin Behavioral Seminar 2011, Central European University, European Behavioral Economics Meeting Berlin 2013, ESSET Gerzensee 2013, SITE 2015, and Pedro Bordalo, Peter Bossaerts, Colin Camerer, Jeff Ely, Ignacio Esponda, Erik Eyster, Marina Halac, Paul Heidhues, Philippe Jehiel, Navin Kartik, George Loewenstein, Wolfgang Pesendorfer, Marek Pycia, Drazen Prelec, Matthew Rabin, Luis Rayo, Larry Samuelson, Adam Szeidl, Balazs Szentes, Andrei Shleifer, Tomasz Strzalecki and Jörgen Weibull for comments, as well as the hospitality of the Department of Economics at Harvard. All errors are mine.

Abstract

People underappreciate the extent to which their information is private. I incorporate such information projections into the solution of Bayesian games. In the context of social investments, people underestimates the uncertainty others face about their preferences and too often conclude that others have antagonistic preferences. Even if everyone prefers mutual investment, no one invests, and though behaving identically, each comes to believe that none else values mutual investment. In the context of communication, the model predicts credulity: persuasion by an advisor with a known incentive to exaggerate the truth, nevertheless, induces uniformly inflated expected posteriors. Credulity results when receivers have sufficient financial literacy and the conflict is limited, and an increase in the former, as well as, a decrease in the latter, can systematically lower receiver welfare. I extend the model to incorporate also ignorance projection and apply it to common-value trade. I show that the predictions match the data better than BNE or cursed equilibrium.

Keywords: Perspective Taking, Social Conflict, Organizational Apathy, Credulity in Persuasion, Financial Literacy, Under-Bluffing in Trade.

1 Introduction

> *"The only true voyage of discovery,..., would be not to visit strange lands but to possess other eyes, to behold the universe through the eyes of another..."* Marcel Proust, the Captive (1923)

Strategic responses to informational differences are key to most economic activity. The usual assumption here is that people form unbiased views about the extent to which the perspectives of others differ from their own, and appropriately adjust their behavior to such differences. In contrast, a large body of evidence shows that the typical person too often acts as if others had access to the same information she did. Such information projections – empathy gaps in perspective taking – will impact the outcomes of social interactions in many domains commonly analyzed via Bayesian games.

Direct evidence for information projection in beliefs dates back to the classic work of Piaget, for example, Piaget and Inhelder (1948), pointing to an egocentric bias in people's 'theory of mind,' that is, their insufficient tendency to attribute different beliefs to others than what they themselves hold. In a well-known study, Wimmer and Perner (1983) demonstrate that young children too often act as if lesser-informed others shared their superior information. Birch and Bloom (2007) showed that the same kind of mistake is present among Yale undergraduates in slightly more complex tasks.[1]

Such robust and widely documented phenomena, as the curse of knowledge, (in double auctions, Camerer et al., 1989; in communication, Newton, 1990, Epley et al. 2004); the hindsight bias (Fischhoff, 1975), the outcome bias (Baron and Hershey, 1988), or the illusion of transparency (Gilovich et al., 1998, 2000) are all consistent with the idea that the typical person acts as if she exaggerated the probability with which others knew her private information. Madarász (2012) offers a more extensive review of the evidence and introduces the idea of information projection into monotone learning problems. In a strategic context, Samuelson and Bazerman (1985) study behavior in common-value bilateral trade and find evidence consistent with the idea that sellers and buyers act as if they ignored the informational asymmetry that existed between them.

The goal of this paper is to incorporate informational projections into the solution of Bayesian games. A key issue when considering biased forecasts about others in strategic settings is that here higher-order perceptions may also matter. Accounting for such considerations, the paper develops a parsimonious, but fully specified formulation to study the implications of this phenomenon. To illustrate such consequences, I consider applications to problems of social investment, persuasion, and common-value trade.

[1] These studies are often referred to 'false belief tasks.' The phenomenon is also discussed in neuroscience and psychiatry under the term *mentalization*. For example, Allen, Fonagy, and Baterman (2008) argue that perspective-taking is the fundamental common feature among the many versions of adult psychoteraphy.

Model Section 2 presents the model. I consider Bayesian games with partitional information where people receive different information about the state. A person who projects information misperceives her opponent's strategy set in that she has an exaggerated belief that if she can condition her behavior on the knowledge of an event, so can her opponent. The extent of this false belief is characterized by the parameter $\rho \in [0, 1)$.

When considering such biased forecasts about others in strategic settings, one needs to specify not only how each player thinks about the information of her opponent, but also how she thinks her opponent would behave based on that information. This, in turn, depends on a player's view of her opponent's view of herself. Before turning to the main model, I first describe a notion of *private* information projection equilibrium, whereby people do not anticipate the biases of others. After presenting the model private projection I turn to the main model where projection is public.

To model projection in a parsimonious manner, I distinguish between the real and the projected versions of each player. The real version of a player conditions his strategy on his true information. The fictional projected (super) version, real only in his opponent's imagination, instead conditions his strategy on his and this opponent's joint information. A player who projects to degree ρ believes that her opponent is such a projected super version, as opposed to the real version, with probability ρ.

Two properties characterize the model. First, in equilibrium, projection is *all-encompassing*: a real player assigns probability ρ to her opponent being the projected version who knows everything she does, including the fact that she is her real version. If the true game is poker where, in reality, each player only knows the value of his or her own hand, a biased Judith believes that with probability ρ Paul is the projected super version who knows both his and her hand, and also that Judith does not know Paul's hand. Projected Paul is then believed to best-respond to Judith's real strategy given such information. Second, in equilibrium, a player's belief about how her opponent might behave is consistent with how this opponent actually behaves. Each player assigns probability $1 - \rho$ to her opponent being the real version, and thus, to her opponent's true strategy. Despite players being biased, nothing happens in equilibrium that would be inconsistent with what players think might happen.

These two properties imply that the predicted behavior is such that people act as if they partially anticipated the biases of others. Due to consistency, each player acts as if she anticipated that her opponent was biased and projected onto her. Due to all-encompassing projection, exactly proportional to the extent that she herself projects, a player underestimates the true extent to which her opponent has biased views about her strategy. All resulting differences in higher-order perceptions are, however, solely a function of the degree of projection ρ, allowing the model to provide a tight characterization. After presenting the model, I establish existence and present some basic properties.

Social Investment In Section 3, I apply the model to the problem of social investment. Partnerships in trade, Williamson (1979), friendships, cooperation in large organizations, or the formation of political associations require people to pool resources and invest into joint assets.

Investing with someone who has matching goals and would reciprocate investment (a positive type) is a source of gain. Investing with someone who is opportunistic or has opposing goals and would not reciprocate investment is a source of loss. Such investments are risky because people face uncertainty regarding the preferences others' preferences. A key determinant in this setting is trust: the belief that one's partner is the former instead of the latter type.

To illustrate, consider a simple dating example. Two people are sitting at a bar. Each is privately informed about his or her own preference for a match. Each can independently decide to make a move (invest) or not (not invest). If neither makes a move, each gets the outside options. If both invest, a match is formed. If only one of them makes a move, the other accepts if interested, and rejects otherwise. In case of a rejection, the proposer incurs a loss such as pain incurred when being rejected. Investment is risky because neither player knows whether the other is interested or not.

A strategically analogous problem arises when a buyer and a seller need to decide to invest into a relationship-specific asset, not knowing whether the other party is reciprocal or opportunistic. Similarly, it arises when a member of an organization decides whether to voice dissent with the leadership (norm) or stay silent and act loyal in front of an other member not knowing whether this other member disapproves of the leadership (status-quo), or is loyal and would want to report or punish dissent.

By projecting information, a person fails to appreciate the extent to which the other party faces the same kind of uncertainty about her preferences and goals as she does about his preferences. In the dating context, an interested Judith too often thinks that Paul should know that she is interested, and exaggerates how often Paul should make a move if he is interested. Since Judith still does not know Paul's preferences, she now finds it relatively more important to protect herself from the loss in case Paul were to reject her. At the same time, an interested Paul, in a symmetric situation, reasons similarly. Therefore, neither invests, but both conclude that the other is not interested. Even if they both prefer mutual investment into the relationship, they behave identically, but each concludes that the other party is an opportunistic type.

In particular, two mistakes affect social attitudes and the formation of trust in equilibrium. First, conditional on *any* outcome in the game, a positive type underestimates her opponent's preference for mutual investment. Second, players develop false antagonism on average: a positive type on average concludes that her opponent is more negative than she thought, and a negative type concludes on average that her opponent is more positive than she thought. The model, thus, predicts an *ex ante* expected false negative correlation between the direction of a player's own preferences and her inference about the direction of the preferences of others. Those who privately oppose the norm will always come to strictly exaggerate the public support for the norm. Those who support the status-quo will always be (at least weakly) too suspicious that the silence of others just masks their preference for dissent. Projection, thus, leads to the exaggeration of social conflict.

As a corollary, in a setting with repeated encounters between people, I specify environments

in which continued interaction leads to fully efficient investment under Bayesian assumptions, it leads to no investment, given any positive degree of repeated information projection. Even if all players are positive, none invests, but they all and always come to believe with near certainty that the everyone else is a negative type. Even if all players want to deviate from the status quo, they all come to believe that all others support the status quo and become apathetic about organizational change. Such false uniqueness is the consequence of the above differential attribution of identical behavior to oneself and to others. I discuss comparative static implications in the context of various applications and relate the predictions to an empirical phenomenon described in psychology under the rubric of 'pluralistic ignorance', for example, Prentice (2007).

I conclude by generalizing the setting and show that above underestimation and false antagonism remain true both if initial investments are substitutes or complements. Hence, to the extent that a player's valuation of a social asset increases in her perception of how much her opponent values mutual investment, the model predicts too little trust in partnerships and a general undervaluation of social assets.

Persuasion In Section 4, I apply the model to a simple problem of strategic communication. Bayesian communication under rational expectations has two general properties: by providing information it improves the expected welfare of receivers, and it is purely informative in that the receivers, on average, are never fooled. In an environment with a commonly known conflict of interest and a commonly known distribution of the quality of the good sold , the model, nevertheless, implies a systematic violation of both of these properties.

A sophisticated advisor sends a cheap talk message to a receiver about a statement being true or false. A doctor advises a patient or a broker advises an investor about the suitability of a drug or a financial product. The advisor's preferences are misaligned towards claiming that the statement is true. Receivers have private information about the cost at which they can verify the advisor's recommendation. Differences in such costs might reflect private information about one's (financial) expertise or background knowledge determining the cost of accessing additional sources to evaluate the sender's advice. If the receiver decides to verify, she learns whether the sender lied or not, and the sender suffers a loss (of business) in case he did.

A biased receiver projects her private information and exaggerates the probability with which the sender knows how costly it would be for her to verify the sender's message. In turn, the receiver exaggerates the extent to which the sender's incentive to lie is tailored to her privately known verification cost as opposed to the publicly known distribution thereof. In equilibrium, the lower the receiver's cost of verification, the more confident she is that a positive recommendation by the sender is truthful.

While in equilibrium receiver types with the highest level of expertise, that is, the lowest cost of verification, always check and learn the truth, types with sufficient but not too much expertise will always be overconfident and overinvest in the asset. Types with little or no expertise at the same time will be (weakly) in disbelief and may underinvest in the asset. I show, however, that as long as the conflict between the parties is sufficiently high, or the asset is

sufficiently complex to evaluate, the model, nevertheless, predicts *uniform credulity*: all receiver types are too optimistic when receiving a positive recommendation. For any positive degree of projection, persuasion now predictably inflates the receiver's ex ante expected posterior and leads to overinvestment.

Understanding the mechanism through which persuasion leads to credulity is potentially key for evaluating such advocated policies as capping the conflict between a financial advisor and an investor or improving the financial education of investors.

Key to the above mechanism is that it is the joint presence of a limited conflict and sufficiently low verification costs (sufficiently high financial literacy of receivers) which is necessary for projection to lead to uniformly credulous illusions. While full literacy and no conflict always maximizes welfare, the comparative statics are non-monotonic. The comparative static properties of the model imply that both improving financial education and decreasing the conflict from sufficiently low initial levels will systematically lower ex ante expected receiver welfare.

I conclude with endogenizing the conflict and partially also the complexity of the asset by invoking the seller of the asset. I show that any partial cap on the conflict may have very limited effectiveness. In fact, the seller-optimal way to induce uniform credulity will minimize the conflict and ensure that the asset is not trivial but too difficult to evaluate either. Unless the conflict is too large, or financial education is sufficiently low, given any degree of projection, the receiver can have a strictly positive willingness to pay for advice which only reduces her expected welfare.

Projection equilibrium In Section 5 I combine information projection – the underappreciation of the positive side of the information gap – with a notion of *ignorance projection* – the underappreciation of the negative side of the information gap. Here, a player projects both what she knows and what she does not know, exaggerating the probability that her opponent conditions his strategy on exactly the same set of events as she does. I derive implications of this combined model of projection equilibrium for the classic problem of common-value trade, Akerlof (1970). Consistent with the evidence – for example, Samuelson and Bazerman (1985) and Holt and Sherman (1994) – when a privately informed seller has the bargaining power, the model predicts non-altruistic truth-telling and underbidding relative to the buyer's acceptance behavior. In contrast, when the uninformed buyer has the bargaining power, the model predicts overbidding in a classic situation of the winner's curse, and underbidding in a classic situation of the loser's curse. I compare the predictions of the model with the evidence and show that a very small degree of projection robustly provides a better fit with the data than BNE or cursed equilibrium.

The beliefs people use to predict and understand the behavior of others is central to economics. Motivated directly by the evidence, the paper offers a simple and fully specified model incorporating informational projections into strategic behavior. Operating in a fully canonical framework, the model provides a common and general way of formulating this wedge between the true and perceived informational differences. At the same time, its parsimony allows one to tightly link the theoretical and empirical consequences of limited informational perspective-

taking across many domains.

2 Model

This section develops the main 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 π. Player i's information about ω is given by a standard information partition $P_i : \Omega \to 2^{\Omega}$; her finite action set is A_i; and her bounded payoff is $u_i(a, \omega) : A \times \Omega \to \mathbb{R}$, where $a \in A = \times_i A_i$ is an action profile. In short, the game is summarized by the tuple $\Gamma = \{\Omega, \pi, P_i, A_i, u_i\}$.

To introduce information projection, I express the joint information of the two players i and j. Specifically, consider the following correspondence:

$$P^+(\omega) = \{\, \widehat{\omega} \in \Omega \;\mid\; \widehat{\omega} \in P_i(\omega) \cap P_j(\omega) \,\} \text{ for all } \omega \in \Omega. \tag{1}$$

The above correspondence is also partitional and describes the coarsest common refinement of the two players' partitions – that is, the information distributed between these two players. Note that this partition is the unique one to capture the players' joint information. 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 that is known in a given state under P^+ is known given the pooled information of the two players.[2] This joint information, thus, corresponds to the natural object to capture the idea of information projection; it will imply that a person who projects her private information has an exaggerated belief that whenever she can condition her strategy on an event, so can her opponent.[3]

To incorporate such information projection in a parsimonious manner, I distinguish between the regular and the projected versions of each player i. The real *regular* version of i conditions his strategy on his true information. Formally, he 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* (super) version of player i – who is real only in the imagination of player j – conditions his strategy on the joint information of i and j. Formally, he chooses a strategy from the set

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

[2] Note that the unique knowledge operator $K : 2^{\Omega} \to 2^{\Omega}$ associated with partition $P_i(\omega)$, is given by $K_i(E) = \{\omega \mid P(\omega) \subseteq E\}$ describing the set of states ω where event E is known. The knowledge operator corresponding to the joint information P^+ is uniquely defined by $K^+(E) = \{\omega \mid \cap_i P_i(\omega) \subseteq E\}$.

[3] Note that, since all payoff-relevant facts are encoded in ω, to the extent that player i has information about her own taste, or the taste of her opponent, she projects her private *information* about preferences as well. For example, in positive common-value environments this may imply exaggerating the correlation between one's own valuation and the valuation of one's opponent. See Section 5 for further discussion.

In reality, all players are regular. Fictional, projected versions enter only into people's beliefs about each other.

Information projection by a player j corresponds to a mistaken belief that, with probability $\rho \in [0,1)$, player i is a projected, as opposed to a regular, version. For ease of notation below, I first assume the degree of projection to be common across players, but then immediately extend the definition to heterogeneous projection.

Notation Below, the operator $\circ$ denotes the mixture of two probability-weighted lotteries. The operator BR denotes the standard best-response operator; its subscript always refers to the set of strategies over which the there indexed player maximizes her expected utility; its argument refers to this player's belief about her opponents' strategies to which she wishes to best respond.

2.1 Private Information Projection

As a brief digression, before presenting the main model in Section 2.2 where projection is public, as described in the introduction, I briefly consider a private version of projection. This solution violates the two key properties of the main model: limited consistency and the all-encompassing features of projection. Nevertheless, discussing private projection briefly may help translate some psychological intuitions related to projection. The material, here, however can be skipped, and the reader may wish to jump to Section 2.2.

To present the definition, I need to distinguish between strategies that are played in equilibrium, and strategies that describe players' beliefs about how their opponents behave. The strategy profile that describes people's view of their opponent's behavior is given by a mixture of two strategy profiles. In particular, each player i will believe that with probability $1-\rho$ her opponent picks a strategy $\sigma^0_{-i} \in S_{-i}$ and with probability ρ a strategy σ^+_{-i} form S^+_{-i}. This lottery will be denoted by $(1-\rho)\sigma^0_{-i} \circ \rho\sigma^+_{-i}$.

Definition 1 *A strategy profile $\sigma^\rho \in S_i \times S_j$ is a private ρ information projection equilibrium (PIPE) of Γ if there exists strategy profiles $\sigma^0 \in S_i \times S_j$ and $\sigma^+ \in S^+_i \times S^+_j$ such that for all i,*

1.
$$\sigma^\rho_i \in BR_{S_i}((1-\rho)\sigma^0_{-i} \circ \rho\sigma^+_{-i})$$

2.
$$\sigma^0_{-i} \in BR_{S_{-i}}(\sigma^0_i) \text{ and } \sigma^+_{-i} \in BR_{S^+_{-i}}(\sigma^0_i)$$

■ The definition corresponds to a parametric extension of BNE. Above, the strategy profile σ^0 always describes a BNE of Γ. This profile describes people's initial shared view of the behavior in the game. If $\rho = 0$, then the definition reduces to that of BNE of Γ. If $\rho > 0$, the model deviates from that of BNE. Specifically, given σ^0, describing how players initially expect each other to behave, a biased player i mistakenly assigns probability ρ to her opponent best responding to her strategy, σ^0_i, by conditioning his action on their joint information in

the game, that is choosing $\sigma^+_{-i} \in BR_{S^+_{-i}}(\sigma^0_i)$. She assigns the remaining probability to her opponent acting as before, σ^0_{-i}. Player i's private IPE strategy σ^ρ_i is then a best-response to such wrong beliefs.

■ In a private information projection, people do not anticipate the biases of others. Given an initially common BNE of the game, each player believes that her opponent expects her to play according to her strategy in that equilibrium. A biased player i then comes to believe that her private information has been unexpectedly leaked to her opponent with probability ρ. At the same time, she maintains the belief that her opponent never thinks that she attaches positive probability to such leakage. Judith, thus, thinks that with probability ρ Paul best responds to her original equilibrium strategy using their joint information and that with probability $1-\rho$ Paul acts as he was initially supposed to. Judith's best response to this perception constitutes her private ρ information projection equilibrium strategy.

■ A private IPE consists of a minimal deviation from a BNE of the game in the following sense: it is *only* a player's first-order belief about her opponent's strategy that is changed relative to an underlying BNE of Γ. All higher-order beliefs about strategies remain the same. In particular, player i thinks that player j plays σ^0_j for sure and thinks that player j thinks that player i plays σ^0_i for sure and so on. This means that the belief that opponent picks a strategy from S^+_{-i} as opposed to S_{-i} enters only into first-order beliefs about strategies.

■ Note that in this definition, players best respond to a misspecified theory of their opponent's behavior – one that does not contain in its support the truth. This aspect of a private information projection equilibrium links this model to level-k models of strategic behavior. In both cases a person's theory of her opponent's theory of how she behaves need not cover how she actually behaves. At the same time, these wrong theories are derived from a common heuristic about play in the game. An important difference here is that the above expectations are anchored to a BNE of the underlying true environment. Furthermore, a private information projection equilibrium operates through a misperception of the opponents' strategy space, while level-k approaches, e,g., Crawford and Irriberi (2008), leave such perceptions intact.

■ Finally, identifying the set of $\rho-$PIPE is straightforward given a BNE of the game in the sense that it involves calculating only individual best-responses. This feature makes it relatively easy to apply it to settings where the set of BNE is well understood.

2.1.1 Example 1: Zero-sum Games

To illustrate the model, consider a hide-and-seek game. Each player picks one of two locations: A or B. If the defender is strong, $\omega = 0$, she wins iff the players pick the same location. If she is weak, $\omega = \omega_w > 0$, then even if they both pick A, the defender wins only with probability $1-\omega_w$. When the defender is weak location A is her Achilles heel. Formally,

attacker/defender	A	B
a	$\omega, 1-\omega$	$1, 0$
b	$1, 0$	$0, 1$

(2)

D-Day (Calais-paradox) To illustrate, consider one of history's most noted zero-sum games: the landing of the Allies in France on June 6th of 1944, (D-day). Here the Allies had the choice to land at Calais (A) or Normandy (B). The Axes had to decide to concentrate troops at one of these two locations. There was good reason to believe that Calais would be the easier terrain for an attack. German forces occupying both locations also had some private information whether this was true or not. The historic success of D-day is often attributed to the Axes' expectation that an attack would take place at Calais causing them to try to defend that location.

Suppose the state is the defender's private information and the ex-ante each state is equally likely. The table below summarizes the defender's strategy in the unbiased and the fully biased case. Since the attacker has no private information he mixes symmetrically, in both settings.

defender	weak	strong	EU_D^ρ
$\rho = 0$	B	A	$\frac{1}{2}$
$\rho = 1$	A	B	$\frac{1}{2} - \frac{\omega_w}{4}$

Under the unique BNE, the defender *hides* optimally behind her private information: she defends A when strong, and B when weak. Hence, she never defends her Achilles heel, and wins half of the time irrespective of ω_w. In contrast, a fully biased defender always plays her Achilles heel whenever she has one. Thinking that her information has leaked to the attacker, but that he does not realize that she recognizes this, when she is strong, she expects him to attack at location A; when she is weak, she expects him to attack at location B. Her best response is then to defend A when weak, and to defend B when strong.

The next observation implies that even as it becomes ex-ante virtually certain that the defender is weak, $p \to 1$, while the BNE converges to the defender mixing symmetrically, any ρ-$PIPE$ converges to defending her Achilles heel for sure. Specifically,

Claim 1 *Note that for any p, $\sigma_2^\rho(A \,|weak) = 1$, iff $\rho > 0$.*

Finally, let me present a reversal of the key informational comparative static result under BNE. For any given prior π over ω, I compare the defender's ex-ante equilibrium winning probability in two cases: (i) the defender is privately informed about ω as above, (ii) she has no private information, thus, only knows the prior π . The second game is one with symmetric information, hence information projection does not alter predictions. I now state a 'chocking' effect describing the reversal of the informational comparative static: while in the Bayesian case, private information always has positive value for the defender, in the fully biased case, it always has negative value to her.

Claim 2 (Negative Value of Private Information) *For all π, if $\rho = 0$, the defender wins more often in (i) than in (ii), if $\rho = 1$, the reverse is true.4*

2.1.2 Example 2: IPV Auctions

As a second-example, consider a symmetric independent private-value auction. Suppose each player's valuation is distributed according to some π over a finite set of valuations, $v_1 < v_2 < .. < v_N$. The classic Bayesian result in this setting is r*evenue equivalence*: the seller's expected equilibrium revenue is independent of whether a first- or a second-price auction is adopted, for example, Riley (1989). The result below shows that revenue equivalence is systematically violated for any positive degree of information projection. Consistent with much of the existing evidence - for a survey see Kagel (1995) - there is always a $\rho - PIPE$ where players over-bid in the first-price auction relative to the second-price auction. Furthermore, this increase in revenue is discontinuous as one moves from no-bias to any positive bias.

Claim 3 *If $\rho = 0$, revenue equivalence holds. If $\rho > 0$, there always exist a PIPE such that the first-price auction generates discretely higher revenue than the second-price auction.*

Note first that a second-price auction has an ex post equilibrium. This implies that the Bayesian predictions are unchanged. Consider now the first-price auction. The key feature of the BNE is that players appropriately shield their bids below their valuations. By projecting information a player comes to believe that if her opponent has a lower valuation than hers, he now has an incentive to bid higher. In contrast, if he has a higher valuation, he now has an incentive to bid lower. Both of these (fictional) effects imply that a biased player has an incentive to increase her bid. If valuations are discrete, then each type bids on an interval that has positive measure. Hence by projecting information the increase in revenue is discrete when moving from the case of $\rho = 0$, to the case where $\rho > 0$.[4]

2.2 Definition

I turn to the main model. Below, σ^ρ describes the predicted strategy profile of the players – the strategy profile of the real, regular versions. Since, in reality, people can condition their true strategies only on the information that they truly have, this strategy profile is an element of the true strategy space. It is supported by a profile σ^+ describing the imagined behavior of the projected versions.

Definition 2 *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 all i,*

1.

$$\sigma_i^\rho \in BR_{S_i}\left\{(1-\rho)\sigma_{-i}^\rho \circ \rho\sigma_{-i}^+\right\} \tag{3}$$

2.

$$\sigma_{-i}^+ \in BR_{S_{-i}^+}\{\sigma_i^\rho\} \tag{4}$$

[4] Note that because this is a private value environment, cursed equilibrium here makes the same predictions as BNE.

If $\rho = 0$, each player has correct forecasts about the behavior of her opponent, and the predictions of IPE collapse to that of the BNE for Γ. If $\rho > 0$, each real player mistakenly assigns probability $(1 - \rho)$ to the actual strategy of her opponent and probability ρ to the strategy of her projected opponent. Here, each player has potentially mistaken forecasts and assigns positive probability to her opponent's playing a strategy which is also conditioned on her own private information and is a best response to her true strategy. I now describe two defining features of the model.

All-encompassing Projection First, projection is all-encompassing: the real player i believes that her projected opponent knows that she is regular for sure. This is reflected in Eq. (4). In other words, a biased player believes that her projected opponent knows also what information she has. If the true game is poker, in which each player truly sees only his/her own card, Judith believes that with probability ρ Paul knows both the value of her card and the fact that she does not know the value of his card. Since a player always knows what she herself knows, this feature implies, consistent with the psychological logic of information projection, that projection is not based on an arbitrary distinction between the content of a person's private information and her information about what she knows, but rather applies to both of these equally.

Consistency Second, each player's expectation about her opponent's play is consistent, in a limited way, with how her opponent actually plays. This is reflected in Eq.(3). Each regular player assigns probability $1-\rho$ to her opponent's behaving in the way that this opponent always behaves. Thus, in equilibrium, nothing happens that explicitly contradicts a player's theory of how her opponent may behave. This remains true even if players observe joint payoffs. What happens in the game is part of what each player thinks can happen in the game. The deviation from the standard model of forming appropriate beliefs is simply that, despite potential evidence to the contrary, she expects something to happen, based on her egocentric mistake, that may never happen or may happen with a different probability than expected.

Partial Anticipation A logical implication of the above two properties is that the predicted behavior is consistent with an interpretation whereby players partially anticipate the biases of others. Each player plays as if she anticipated that her opponent projects onto her, but proportional to the extent that she herself projects, she always underestimates the opponent's projection. Given all-encompassing projection, Judith believes that, with probability ρ, Paul has correct beliefs about her strategy. Given consistency, Judith believes that with probability $1 - \rho$ Paul wrongly believes that she knows the value of his card with probability ρ. In sum, Judith expects Paul to believe that Judith knows the value of his card with probability $\rho - \rho^2$ on average. Instead, Paul believes that Judith knows the value of his card with probability ρ.[5],

[5] In a similar fashion, one can construct the real players' iterative higher-order beliefs about, say, player j being the projected version as follows: the first-order belief is the probability that real i assigns to j being the projected version; the second-order belief is the probability that real j assigns to the *expected* probability that player i assigns to player j being super, the third-order belief is the probability

Heterogeneity The definition extends immediately to differentially biased players. Heterogenous projection is described by a vector ρ with a potentially different ρ_i replacing ρ for each i in Eq. (3). If $\rho_i = 0$, then player i is unbiased. Given the consistency property, an unbiased player is fully sophisticated and has correct forecasts about her opponent's strategy given her information. Under heterogeneous projection, Judith – player j – acts as if she expected Paul – player i – to assign probability $(1 - \rho_j)\rho_i$ to Judith being the projected super version on average. Her underestimation of the extent to which Paul thinks she is super is, thus, proportional to the degree of her own mistake ρ_j. If $\rho_j = 0$, Judith is unbiased, which then implies that she is sophisticated and fully anticipates Paul's misperception.

Evidence A companion paper, Danz, Madarász, and Wang (2014), directly tests the partial anticipation aspect of projection in an agency setting.[6] Consistent with earlier evidence, they find that people mistakenly project onto others. They also find that at the same time, people anticipate that others will projection onto them. Finally, they find that while people anticipate the projection of others, they underestimate its extent. By considering heterogenous projection, their design allows one to measure both jointly and separately the extent to which people project and the extent to which people underappreciate the projection of others. Consistent with the above logic, they find that the degree to which subjects project and the degree to which they underestimate the projection of others are remarkably similar.

2.3 Discussion

Let me turn to some basic properties of the model. The first claim establishes existence.

Proposition 1 *For any Γ and ρ, a $\rho - IPE$ exists.*

The next corollary points out that the model delivers differential predictions only to the extent that players are differentially informed.

Corollary 1 *If $P_i(\omega) = P_j(\omega)$ for all ω, the set of $\rho-IPE$ for Γ is independent of ρ.*

While in games with symmetric information, projection has no bite, as long as at least one player has private information it can affect predictions. Furthermore, even in games with one-sided private information, that is, when P_i is a strict refinement of P_j, the degree to which the lesser-informed player projects information already matters. This is true by virtue of the

that real i assigns to the *expected* probability that player j assigns to the expected probability that player i assigns to player j being super. The k^{th} element of this sequence is given by $\sum_{s=1}^{k}(-1)^{s+1}\rho^s$. In this sequence, (i) the sub-sequence of odd elements is decreasing in k, (ii) the sub-sequence of even elements is increasing in k. In words, real j assessment is increasing and real i's is decreasing in k. Furthermore, (iii) each odd element is larger than the subsequent even element, but both converge to $\rho/(1+\rho)$. Hence, the same pattern of underestimation holds as above, but the discrepancy, which is always ρ^k, vanishes as k increases.

[6] Technically, given the nature of the task in their design, Danz et al. (2014) only allows them to estimate the model of projection equilibrium as introduced in Section 5. The structure of higher-order perceptions, and hence the qualitative nature of partial anticipation there is the same as here.

all-encompassing nature of projection in the model: the extent to which the lesser-informed player projects information governs her anticipation of the projection of her opponent which then affects play. If ρ_j is the degree to which player j projects, then if $\rho_j = 1$, the lesser informed player acts as if she believed that the better informed player did not project.

The next claim shows that a BNE of Γ which is also an ex post equilibrium – an equilibrium where no player has an individual incentive to deviate even after observing the state – is also an information projection equilibrium for any ρ. In contrast to a $\rho-$IPE, an ex post equilibrium often does not exist, but when it does, it is projection-proof.

Proposition 2 *If a BNE is an ex post equilibrium, then it is also a $\rho-$IPE for all ρ.*

A converse of the above claim is not true.[7] Even if all BNE of a game are ex-post equilibria, IPE can extend the set of predictions and lead, for example, to illusory coordination.[8]

Related Literature This paper relates to other approaches that study players who exhibit an explicitly wrong theories of the behavior of others. In particular, Jehiel (2005) and Jehiel and Koessler (2008) study a framework of analogy-based expectations equilibria. Eyster and Rabin (2005) study the notion of cursedness. The identifying assumption in all of these approaches is that while each player has potentially coarse theory of her opponent, her expectations about the strategy of her opponent is correct *on average*. Instead, the key assumption in this model is that a player has wrong expectations about her opponent's strategy, on average. Each player forms an egocentric view of her opponent's beliefs and strategy which systematically deviates from the strategy on average where exactly such a deviation is governed by the parameter ρ. The model thus systematically violates the identifying assumption of these approaches. Since information projection applies through the misperception of the opponent's strategy set, it also clearly differs from the application of level-k approaches in Bayesian games – for example, Crawford and Irriberi (2008) – because these maintain the assumption that people have a correct understanding of informational differences.

Note that the logic of information projection differs markedly from the logic of cursedness. A cursed player perceives informational differences correctly, but underappreciates the extent to which her opponent conditions his choice on his own private information. Instead, information projection points to an exaggeration of the extent to which a player thinks that her opponent conditions his choices not on his but also on her information. She forms wrong beliefs about

[7] In the case of PIPE, this converse is true: if all BNE are ex post equilibria, then the set of ρ PIPE is independent of ρ.

[8] To illustrate, consider the following game with a symmetric prior and the state being the column player's private information:

$$\begin{array}{ccccc} \omega_1 & R & L & \omega_2 & R & L \\ T & 1,1 & 0,0 & , \quad T & 1,1 & 0,0 \\ B & -3,3 & -3,3 & B & 0,0 & 2,2 \end{array}$$

The unique BNE, also an ex post equilibrium, is given by the profile $\{T; R(\omega_1), R(\omega_2)\}$. In contrast, if $\rho > 1/3$, there is a $\rho-$IPE given by $\{T; R(\omega_1), L(\omega_2)\}$ because now a projecting column player can expect the row player to play B in state ω_2.

the information of others, but forms coherent beliefs about how others would behave given her misperception.

Finally, the interim beliefs described in this model are also consistent with a heterogenous prior interpretation where initially each player assigns zero probability to herself becoming a projected version, but probability ρ to her opponent becoming the relevant projected version. In this interpretation, the model describes to a tight directional divergence in perceptions as a function of the true data generating process.

3 Social investment

Efficient outcomes between trading partners, as in the case of the classic hold-up problem (Williamson 1979), friendships, the formation of a political or social associations, typically require partners to pool resources and make investments into a relationship specific or joint social asset. Here, the return on one's investment depends on the goals and preferences of one's partner. Investing with someone who has matching goals and would reciprocate investment is a source of gain. Investing with someone who is opportunistic and would prefer not to reciprocate such investment, is a source of loss. Hence, whenever people face uncertainty regarding the goals and preferences of others, a key component of such interactions is trust: the belief that one's opponent is the former as opposed to the latter type.

When contracts are incomplete or badly enforced, such trust is a key determinant of bilateral exchange and it plays an important role in cooperation and efficient exchange in large organizations, for example, La Porta et al. (1997). As Arrow (1972) argued, "virtually every commercial transaction has within itself an element of trust, certainly any transaction conducted over a period of time. It can be plausibly argued that much of the economic backwardness in the world can be explained by the lack of mutual confidence." Confidence that others have matching rather than opposing goals is often a first-order determinant of social and political outcomes, or the presence of intergroup conflict.

3.1 Setup

Consider a general social investment problem. Upon each player i privately observing her type θ_i, players decide whether to invest (enter) or not (stay out) in the relationship. If both invest, each receives a net payoff equal to her type. If both stay out, each receives the outside option. The game is described as follows:

$$\begin{array}{lcc} & \text{In} & \text{Out} \\ \text{In} & \theta_1 \ , \ \theta_2 & g(\theta_1, \theta_2) \ , \ f(\theta_2) \\ \text{Out} & f(\theta_1) \ , \ g(\theta_2, \theta_1) & 0 \ , \ 0, \end{array} \tag{5}$$

where each θ_i is i.i.d. given a uniform density on some $[\theta_{\min}, \theta_{\max}]$ with $\theta_{\min} < 0 < \theta_{\max}$. For ease, I adopt the notation that $\theta_{\min} = -n$ and $\theta_{\max} = x$.

Let me describe the assumptions governing the rest of the payoffs. The key strategic distinction is between positive and negative types: positive types are reciprocal and prefer mutual entry to their opponent entering alone. Negative types, in contrast, are opportunistic and have the reverse preference. In particular, the following sorting condition holds:

1. Sorting Let $f(0) = 0$, and $f_1 < 1$.[9]

The second assumption is that unless both players are positive, one-sided investment leads to a loss to the investing party investing relative to the outside option.

2. Investment Risk If $\min\{\theta_i, \theta_{-i}\} < 0$, then $g(\theta_i, \theta_{-i}) < 0$.

Finally, I also impose some monotonicity assumptions on g, in particular, $g(0, \theta_{-i}) = 0$ and $g_1 \geq 0$ if $\theta_i > 0$, and $g_2 \geq 0$.

Two further remarks are in order.

a. Above only positive types prefer mutual investment to the outside option. This assumption can be relaxed. Suppose a player – independent of her own type and her action – receives a benefit b whenever her opponent enters.[10] The analysis below then holds for any $b \geq 0$. For example, if $b > n$, mutual investment Pareto dominates the outside option given *any* type profile. Here, if both players were known to be negative types, the game would be a Prisoner's Dilemma with a dominant strategy outcome of mutual Out and a social optimum of mutual In. For some of the applications, I make use of that fact that $b > 0$.

b. Specifications of the above normal form can be equivalently described as a sequential game where a player's payoff depends only on her own type and the action profile. Specifically, assume that players first play the above game. Payoffs are the same as before except in the case of one-sided investment, that is, when only one of the players enters. If, say, only i entered, now $-i$ can decide to reciprocate investment. Player $-i$'s payoff is still given by $f(\theta_{-i})$. If a positive $-i$ always reciprocates investment, and a negative one never does, then, the assumptions on $g(\theta_i, \theta_{-i})$ can now be satisfied by virtue of player $-i$'s second action as opposed to his type. The specification described in the main example below allows for this sequential interpretation, hence, I invoke it in the applications.

3.2 Main Example

Consider the specification below which will allow me to highlight the main results and intuitions.

[9] For all of the analysis, for the case where $\theta_i < 0$, it is sufficient to assume that $f(\theta_i) > \theta_i$.

[10] Formally, if $-i$ chooses In, i's payoff is $\theta_i + b$ if i also chooses In, and $f(\theta_i) + b$ if i chooses Out.

$\min\{\theta_i, \theta_{-i}\} \geq 0$	In	Out	$\min\{\theta_i, \theta_{-i}\} < 0$	In	Out
In	θ_1, θ_2	$\gamma\theta_1, \gamma\theta_2$	In	θ_1, θ_2	$-c,\ f(\theta_2)$
Out	$\gamma\theta_1, \gamma\theta_2$	$0,\ 0$	Out	$f(\theta_1), -c$	$0,\ 0$

(6)

where $\gamma \to 1$, and $c > 0$.[11] The following examples describe some applications.

♡ **At the Bar**. ($b = 0$). Judith and Paul are sitting at a bar. Each decides to make a move or stay out. If both make a move, a match is formed. If both stay out, they get the outside option. If only one player, say Judith, makes a move, Paul, can accept or reject it. If Paul is interested, a positive type, he accepts, and again a match is formed with a slight delay discounting payoffs by γ. If Paul is not interested, a negative type, he rejects and Judith now incurs a cost of c associated with the shame or embarrassment of being rejected, or simply with the cost of investing in a futile move.

♠ **Trust in Trade.** ($b > n$) Trading partners, such as a buyer and a seller, need to invest in a relationship-specific asset to maximize benefits from trade (Williamson 1979). While each would benefit from mutual investment, relative to the outside option, a one-sided investment which is not reciprocated leads to a loss of c for the investing party. At the same time, an opportunistic party benefits more if only the other party invests. Hence, an opportunistic type would never himself invest or reciprocate investment. This leads to the classic *hold-up problem*: if a party believes that her partner is opportunistic (negative), she has no incentive to invest either. In contrast, if one's opponent is reciprocal (positive), one-sided investment is always reciprocated, so a positive player who believes that her opponent is also positive would want to invest.[12] Here, c the loss from being held up may increase in the extent to which ex ante promises about future investment choices are not enforceable ex post. A similar situation may arise in negotiations between a creditor and a borrower, the former deciding whether or not to roll over sovereign debt, the latter deciding whether or not to adopt economic reforms.

♣ **Costly Dissent.** A member of an organization either disagrees with (a positive type), or agrees with (a negative type) an existing norm, the status quo or a prevailing business practice. When two members meet, each can voice dissent and deviate from the norm (In), or stay silent and act loyal (Out). If a member agrees with the norm, he acts loyal. If he disagrees, he gains if he expresses dissent it in front of someone who also disagrees with the norm. They might form a coalition or merely experience a sense of liberation. When dissenting in front of a loyalist, however, the dissenter experiences a loss of c. The loyalist might punish or report the dissenter,

[11] The fact that $\gamma < 1$ ensures that the sorting assumption is satisfied.

[12] As mentioned in footnote 7, the analysis is unchanged if $f(\theta_i) > 0$ for all $\theta_i < 0$. Here, $f(\theta_i) > 0$ now holds for all θ_i. It follows then that before deciding to invest, each type has an incentive to convince her opponent that she is a positive reciprocal type. This is true because now each type benefits from her opponent investing initially. Hence, pre-play communication cannot reduce the uncertainty about preferences.

causing the dissenter to be ostracized, fired, or persecuted.

3.3 Equilibrium

The next proposition characterizes the predictions. Below, $E^{\rho}_{\sigma^\rho}$ is the expectation operator describing a ρ-biased player's expectation in equilibrium σ^ρ. Similarly, $E^{0}_{\sigma^\rho}$ refers to the true expectation operator given the true distribution of behavior in equilibrium. Finally, E_0 refers to the prior expectation of a variable.

Proposition 3 *For any ρ, there is a unique $\rho - IPE$. Each player i enters iff $\theta_i > \theta^{*,\rho}$ where*

$$\theta^{*,\rho} = \sqrt{\frac{nc}{1-\rho}}.$$

Furthermore,

I. for any $\rho > 0$, $E^{\rho}_{\sigma^\rho}[\theta_{-i} \mid \theta_i, a] < E^{0}_{\sigma^\rho}[\theta_{-i} \mid \theta_i, a]$, given any $a \in A$ and $\theta_i > 0$;

IIa. for any $\rho > 0$, $E^{0}_{\sigma^\rho}[E^{\rho}_{\sigma^\rho}[\theta_{-i} \mid \theta_i, a]] < E_0[\theta_{-i} \mid \theta_i]$ if $\theta_i > 0$;

IIb. for any $\rho > 0$, $E^{0}_{\sigma^\rho}[E^{\rho}_{\sigma^\rho}[\theta_{-i} \mid \theta_i, a]] \geq E_0[\theta_{-i} \mid \theta_i]$ if $\theta_i < 0$.

Equilibrium is unique and is given by cutoff strategies. Projecting information causes each player to underestimate the uncertainty her opponent faces about her privately known motives or preferences. This then implies the following results on actions and the assessment of the attitudes of others.

Under-Entry To describe the prediction on actions, return to the bar example. An interested Judith now exaggerates the extent to which Paul knows that she is interested. Since the projected Paul always enters if interested, she exaggerates the probability with which Paul will make a move. Since Judith still does not know whether or not Paul is interested, and because conditional on Paul being interested, her payoff from reciprocating a move by Paul is almost as high as making a move at the same time as he does, it now becomes relatively more important for her to stay out and reduce the risk of embarrassment in case Paul were to reject her move. By symmetry, the same holds for Paul. An increase in projection decreases each player's willingness to invest and, thus, overall investment. Eventually, if the bias is sufficiently high, no type invests. At the same time, each positive type increasingly expects investment by the opponent if interested.

Underestimation A positive type (trustworthy reciprocal trading partner, an interested Judith) always becomes too pessimistic about Paul's type no matter what happens in equilibrium. When seeing Paul enter, Judith too often thinks that Paul invested only because he knew that she would at least reciprocate such an investment. Hence, she underestimates the extent to which initial entry is good news about Paul's interest. When seeing Paul stay out, she is too convinced that Paul is opportunistic, as opposed to fearing being held up. Since a

biased positive player believes that her opponent uses a lower average cutoff than he actually does, she underestimates him in *all* action contingencies arising in equilibrium.[13]

False Antagonism The model predicts biased assessments not only conditional, but also on average, that is, in the ex ante expected sense. In equilibrium, beliefs no longer follow a martingale, instead each player always becomes more convinced that others have opposing preferences.

First, each positive type underestimates her opponent on average. A positive Judith exaggerates the probability that Paul will adjust his action to her preferences and invest if he is positive. Thus, she over-infers from seeing Paul stay out, and under-infers from Paul investing. Since the decision to invest is positively correlated with Paul's type, Judith comes to underestimate Paul on average. Second, each negative type exaggerates the probability that her opponent will adjust his action to her preferences, and stay out even if he is positive. Thus, she over-infers from Paul's investing initially, and under-infers from Paul staying out. Hence, a negative type overestimates her opponent's type on average.[14]

A player who opposes the norm, too often concludes that others are loyalists, while a loyalist too often suspects others to be potential dissenters, on average. Judith becomes more convinced that Paul is interested in her exactly when she is not interested and in him, and she becomes more convinced that Paul is not interested in her exactly when she is interested in him. In short, information projection introduces an ex ante predictable false *negative* relation between one's own type and the perceived type of one's opponent: on average, a player always mistakenly concludes that others are less similar to her than she originally thought given the prior.

3.4 Dynamics

The willingness to enter into a relationship decreasing in loss from being held up c both in the biased and in the unbiased case. It is, then, natural to consider settings in which the opportunity repeats itself with a decreasing c. Such a decrease could correspond to: (i) a wrong move being less costly in an informal than in a formal environment; (ii) an improvement in the enforceability of ex ante promises; or (iii) weakening disciplinary actions following reported dissent.

Consider a dynamic repetition of the exact same game over time t, except assume a changing value of c. Specifically, consider a strictly decreasing sequence $\underline{c} = \{c_t\}_{t=1}^{T}$. For simplicity, I focus on myopic repetition: in each period t, players care only about the payoff of that period, but are able to recall the history of past interactions. In this context, the natural psychological assumption is that players project to some extent at the beginning of each new encounter independent of the history. That is, at the beginning of each t, each player believes that with some probability $\rho > 0$ her information privately leaks to the opponent at the beginning of that

[13] The above results remains true even if a player fully observes her own payoff ex post. In the contingency where Judith chooses in and Paul chooses out, if Judith also observes her own payoff her underestimation is now weak. In all other cases, however, even after observing her own payoffs ex post, her underestimation is strict for any $c, \rho > 0$.

[14] These results again remain true even if a player observes her own payoff ex post.

period t.[15]

Suppose that in each round, players play according to the unique $\rho-$IPE in that round.[16] Let then $\Pr^{\rho}(M \mid \underline{c})$ be a measure of efficiency describing the true ex ante probability that, conditional on both players being positive, at least one party invests by the end of the sequence $\underline{c}$, that is, a match is formed. Finally, let $q^{\rho}_{\underline{c}}$ be the true ex-ante expected posterior probability that a positive player assigns to her opponent being positive by the end of sequence $\underline{c}$, and let q_0 be the analogous prior probability.

Corollary 2 *Suppose $\rho = 0$. For any $\underline{c}$, $\Pr^{0}(M \mid \underline{c},) = \max\{1 - c_T n/x^2, 0\}$ and $q^{0}_{\underline{c}} = q_0$*

In the Bayesian case, matching is efficient: as the loss from being held up goes to zero, all positive types match with certainty. Furthermore, beliefs are always unbiased, and as c_T vanishes, players always correctly learn whether their opponents have similar or opposite attitudes. The matching probability is also history-independent and depends only on the last element of any sequence $\underline{c}$. The next corollary, based on Proposition 3, shows that given any positive degree of repeated projection, the reverse can hold. Here, as c_T vanishes, no matches are formed, but with ex ante probability of one, each positive type comes to wrongly conclude that her opponent is almost surely a negative type.

Corollary 3 (False Uniqueness) *For any $\rho, \tau > 0$, there exists a strictly decreasing $\underline{c}$ with $c_T = \tau$ such that $\Pr^{\rho}(M \mid \underline{c}) = 0$ and $q^{\rho}_{\underline{c}} \leq \tau$. If this is true for $\underline{c}$, it is also true for any $\underline{c}'$ such that $c'_t \geq c_t$ for all t.*

The above result specifies environments where even if c_t vanishes, no type ever invests, but each positive type always concludes that her opponent is a negative type. Key to this corollary is the logic that information projection leads to differential attribution of identical behavior to self and others. While each positive type attributes her own lack of entry to her fear of her opponent being opportunistic, she attributes the identical behavior of her opponent to the opponent actual being opportunistic.

The statement of the corollary focuses on the beliefs and actions of positive types. At the same time, in the above limit, negative types maintain correct views about their opponents. This is true because the projected version of a negative type's opponent now behaves the same way as her real opponent does. They both always stay out. Hence, a negative type correctly makes no inferences from her opponent's behavior. Finally note that, here, beliefs of all types are fully self-confirming even when observing full feed-back on payoffs.

3.5 ♣ Discussion

Let me turn to a discussion of the above results in the context of some applications. Although the interactions above described bilateral situations, they can be equally applied to such bilateral interactions taking place pair-wise between all members of a community.

[15] The results hold a fortiori if $\underline{c}$ is not strictly decreasing.

[16] Since equilibrium is in cutoff strategies, and is independent of the value of x, by Proposition 3, the uniqueness continues to apply.

■ **Norm Falsification** In the context of dissent, the above results imply that those who oppose the norm misattribute the silence of others to their genuine loyalty. When speech is free, $c = 0$, people learn the truth about the attitudes of others; when such 'speech' is not free, Proposition 3 predicts a *systematic* wedge between the privately held support for a norm and the perceived public support for this norm: those who privately oppose the norm will predictably come to exaggerate the public support for it. Under the conditions of Corollary 3, even if almost all opposes the norm – such as business practice, homophobia, political correctness, leadership, such as Stalin's choices within the Politburo of the Communist Party – they all, nevertheless, come to predictably believe that everyone else supports the norm. The non-dissenting, for example, silence or publicly cheering, majority will conclude with probability one that they belong to a minority opposing the norm. The aggregate private opinion and the perception of public opinion will predictably diverge.[17]

■ **Disciplinary Organizations** The predictions may matter for understanding organizations that sanction dissent. Such organizations will not only maintain obedience, but also create a false sense of loyalty of the other members. The proof of Corollary 3 implies that the sanction c can be removed over time without risking dissent provided that such decrease is sufficiently gradual. For example, the organization can save on the cost of running a disciplinary organization which is likely to increase in c. Given the false pessimism of positive types, the deterrent due to the size of c can be gradually replaced by the increasing pessimism of those who would like to change the status quo. Self-censorship will outlive effective censorship due to the misattribution of the silence of others to their loyalty. The above logic then implies a form of organizational apathy which may prevail even when voice is only slightly risky relative to acting loyal.[18]

■ **Shy Revolutions** The results point to a non-monotone comparative static with respect to $\underline{c}$. First, consider a sequence $\underline{c}$ which satisfies Corollary 3. By the end of round T, all positive types are almost surely convinced that everyone else is negative. Suppose, now, that in round $T + 1$, there is a further drop and $c_{T+1} = 0$. Now, all efficient matches are formed discontinuously. Such unexpected mass investment, for example, mass dissent, comes as a surprise to all those who opposed the norm.

More generally, consider what happens as c_t drops to c_{t+1}. There are two effects. First, due to projection from one round to the next, positive types exaggerate the probability of entry by others. They are too surprised by how little entry there is relative to their expectations, and become too pessimistic. Second, as a dynamic consequence of projection, positive types

[17] The logic differs markedly from that of herding in social learning. First, in sequential social learning, there is no direct strategic interaction, while here co-ordination risk is the key. Second, the identifying assumption of rational social learning is that, in the relevant ex ante expected sense, players must develop unbiased beliefs about the state. Instead, here, people surely develop wrong beliefs in the relevant ex ante expected sense. Finally, a herd is formed when people believe that others act on the *same* preferences as they do, e.g., that they all prefer the better restaurant, and not on the *opposite* preferences, e.g., the belief that nobody Judith wants to be kissed by would actually want to kiss her.

[18] Hirschman (1970) discusses the critical role of voice versus loyalty for organizational change.

underestimate the fraction of others who are also positive. Thus, given a small drop from c_t to c_{t+1}, the negative wedge between the private support for the norm and the private perception of the public support for it is still reinforced. If the drop from c_t to c_{t+1} is sufficiently large, however, the second effect dominates. Now all types who prefer to deviate from the status quo are too surprised by seeing the large fraction of people who invest (dissent), which then reduces, or even completely eliminates, this wedge.

■ **Mistrust** In the context of trade, Proposition 3 implies the presence of a psychological hold-up problem. Exactly when trust is key, $c > 0$, trustworthy trading partners come to exaggerate the likelihood that others are opportunistic. They misattribute the lack of investment of their partners to be the result of opportunistic preferences as opposed to the fear of being held up. Even if the potential loss from being held up decreases, for example because institutions improve, people remain reluctant to invest due to their mistaken pessimism. When such beliefs are passed on over time, Corollary 3 implies that a gradual improvement of institutions may have significantly less impact on efficient exchange, maintaining low levels of output, than more dramatic or early reforms.

■ **Mistaken Segregation** False antagonism in Proposition 3 is the consequence of informational differences and limited perspective taking. This allows for a further comparative static. Let there be two groups. Suppose that the distribution of preferences is independent of group membership, the only asymmetry being that each person can read the attitudes of in-group members, but faces uncertainty regarding the attitudes of out-group members, perhaps due to cultural differences. Proposition 3 now implies that a person will come to conclude that in-group members are more likely to have matching objectives or attitudes, for example, to want to be friends when one wants to be friends, than out-group members. Uncertainty about others leads to pessimism about the extent to which others have compatible goals. At the same time, due to all-encompassing projection, people exaggerate the extent to which out-group members have correct views about their attitudes.

■ **Evidence** The above predictions of information projection equilibrium are consistent with a generally discussed empirical phenomenon described under the rubric *pluralistic ignorance*, a term coined by Katz and Allport (1931) in their study of fraternities, and 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 an illustrative study, Prentice and Miller (1993) showed that undergraduates at Princeton rated the average comfort levels of others, including their friends', with the prevailing drinking norm on campus, as significantly higher than their own and, hence, than that of reality. O'Gorman (1975) documented a similar wedge in Whites' preference towards racial segregation and their perception of other Whites' preference for it.[19]

Miller and Mcfarland (1987) provide consistent evidence from the lab. Students had to complete a comprehension task of a very difficult text. In the unconstrained treatment, stu-

[19] Based on a national survey form 1968 the ratio of the perceived fraction of whites who preferred segregation, as estimated by whites, versus the true fraction was 2.6.

dents, seated in small groups, had an option to publicly leave the room, and seek clarification outside. In the constrained treatment, no such option was present. Although none actually left the room in the unconstrained treatment, students rating of their own relative ability was significantly lower in the unconstrained treatment compared to that in the constrained one. This is consistent with students, the majority whom would have greatly benefitted from clarification, attributing their own lack of asking questions to their fear of embarrassment, but the identical behavior of others to their superior comprehension.

In the context of friendship formation, Shelton and Richeson (2005) find that students at Princeton and U. Mass desired having more interracial friendships, but attributed the lack of their own initiative to the fear of rejection and the lack of initiative by members of the other racial group to their lack of interest while also significantly underestimated the outgroup people's interest in interracial friendship relative to the truth. In the context of political change, Kuran (1995) presents anecdotal evidence consistent with the idea that dissenters' overestimation of the popularity of the status quo is a common force preventing social change.[20]

3.6 Investment Games

Let me return to the more general case introduced at the beginning of this section. To characterize the implications of the model, a distinction between complement and substitute initial investments (entry) is needed.

Definition 3 *Initial investments are substitutes (complements) if* $\theta_i - f(\theta_i) < (>)\ g(\theta_i, \theta_{-i})$ *whenever* $\min\{\theta_i, \theta_{-i}\} > 0$.

Initial investments are *substitutes* if, conditional on both players being positive, it is more important to invest initially if one's opponent does not invest initially than if he does. Note that if both players are positive types, then initial investment is always at least reciprocated. Hence, in the former case, no investment causes a player to forgo all gains relative to the outside option. In the latter case, no investment causes a player to forgo the benefit of investing simultaneously as opposed to just reciprocating investment. Initial investments are substitutes if the former loss is larger than the latter. Initial investments are *complements* if the reverse is true.

In the main example, $\gamma \in [0, 1)$ governs the degree of substitutability. If $\gamma \to 1$, investments are (almost) perfect substitutes. In fact, for all $\gamma > 0.5$, initial investments remain substitutes, and *all* qualitative statements of Proposition 3 continue to hold including uniqueness. If $\gamma < 0.5$, investments are complements. To illustrate the complement case here, consider a different bar example setting $\gamma = 0$. Suppose that Judith and Paul each need to decide whether or not to go to the bar. Only if both go to the bar do they have the opportunity to really enjoy each other's company. If only one goes, the other learns about this. If this other is not interested all is the same as before. If they are both interested, there is still no shame cost, but by the time

[20] Examples include the unexpected popular support for the Solidarity Movement in Poland in the elections of 1981; or the fact that a year after the Fall of the Berlin Wall in 1989, over 70% of those surveyed said they were totally surprised by such a change.

the other party learns about this and rushes to the bar, the night is almost over, leaving both essentially with their outside options.

For most economic applications, the real gain from investing in a relationship specific asset is with respect to the outside option, hence, investment initial investments are likely to be substitutes. For a more complete analysis, however, I analyze both cases below.

Proposition 4 *For any $\rho > 0$, all equilibria 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, and the lowest one is decreasing in ρ.*

3. *Both if investments are substitutes or complements,*

3a in all equilibria $E^{\rho}_{\sigma^\rho}[\theta_{-i} \mid \theta_i, a] < E^{0}_{\sigma^\rho}[\theta_{-i} \mid \theta_i, a]$ for any $a \in A$ and $\theta_i > 0$;

3b in all equilibria $E^{0}_{\sigma^\rho}[E^{\rho}_{\sigma^\rho}[\theta_{-i} \mid \theta_i]] < E_0[\theta_{-i} \mid \theta_i]$ for any $\theta_i > 0$;

3c in all equilibria, $E^{0}_{\sigma^\rho}[E^{\rho}_{\sigma^\rho}[\theta_{-i} \mid \theta_i]] \geq E_0[\theta_{-i} \mid \theta_i]$ for any $\theta_i < 0$.

By projecting information a player underestimates the uncertainty her opponent faces, thus, underappreciates the extent to which the behavior of others is impacted by such uncertainty. If initial investments are substitutes, a player's willingness to enter decreases in the perceived probability that her opponent enters. A positive type now exaggerates the probability with which her opponent shall invest initially if he is also a positive type. Thus, her own willingness to enter is decreased. This leads to under-entry in the unique symmetric equilibrium. If initial investments are complements, a player's willingness to enter increases in the probability that her opponent enters. In this case, multiple symmetric equilibria may exist. The perceived return on initial investment is now potentially exaggerated since a biased positive type exaggerates the probability with which her opponent invests initially. In fact, the lowest equilibrium, the one with the highest probability of entry, now decreases in ρ. This leads to over-entry relative to the lowest BNE. The second-lowest equilibrium cutoff, if it exists, however, increases in the degree of projection, leading to under-entry relative to the second-lowest BNE.

Crucially, the qualitative predictions on dynamic formation of social attitudes, as described in Proposition 3, hold in both cases: any equilibrium exhibits underestimation by positive types in any contingency. Similarly, any equilibrium exhibits false antagonism by all types, on average, as well. A positive type's assessments are always too negative both conditionally and on average.[21] A negative type's assessment is too positive on average. Both of these result from underestimating the investment risks others face.

[21] Again in the contingency where only one player enters underestimation may be weak.

Undervaluation of Social Assets The fact that in all equilibria a reciprocal type comes to underestimate her opponent's valuation in all contingencies implies a general failure of trust: those who would invest in the joint asset always become too skeptical about how much their partners would want to invest in it. Even if a match is formed, each player must now underestimate how much her opponent actually values mutual investment, as opposed to opportunistically free-riding on the opponent's investment. Such beliefs are often critical for a person's willingness to protect the relationship or further invest in it. To the extent that one's valuation of the joint asset increases in one's belief of how strongly one's partner's prefers mutual investment, projection leads to the undervaluation of social assets.[22]

4 Persuasion

I now turn to the second application of the model and consider a simple setting of strategic communication. Persuasion and expert advice is central to many domains. While an incentive to distort advice may exist, a puzzle remains as to why, people may fail to sufficiently discount strategically distorted recommendations. For example, Malmendier and Shanthikumar (2007, 2014) provide evidence that small investors take positively strategically distorted positive recommendations too much at face value and make biased investments. Della Vigna and Kaplan (2007) offer evidence in the context of political persuasion through media bias and voting decisions. In the context of costly financial advice, Bergstresser et al. (2009) provide evidence that investors are willing to pay for advice which results only in lower risk-adjusted returns, Mullianathan et al. (2012) conduct an audit field study and show that advisors in some cases provide recommendations that make investors worse off.

A general feature of Bayesian communication is that receivers are never fooled on *average*. When the martingale property of correct Bayesian beliefs holds in a BNE, communication is informative but never by itself shifts the ex ante expected posteriors. In this section, I consider a classic sender-receiver game without commitment. A sophisticated sender provides advice to a receiver (investor) whether a proposition is true or false. The receiver can verify the sender's recommendation at a privately known cost c. Although the sender has a commonly known incentive to claim that the proposition is true, the model nevertheless predicts uniform credulity. If the conflict is sufficiently large, or it is sufficiently difficult to verify the sender's message, persuasion always leads to uniformly exaggerated posteriors and overinvestment on

[22] The empirical literature also documents a so-called *'false consensus'* effect. The evidence on such a false consensus effect, however, is very mixed. For a survey see, e.g., Dawes and Mulford (1996). Furthermore, it is not linked to a more disciplined approach clarifying *what* such false consensus may be about.

Note, that the result on false uniqueness above is about one's own preferences. Importantly, it is the result of a mechanical false consensus effect about others's views about one's own preferences caused by information projection. Furthermore, in this strategic context, projection also leads to a false consensus effect about actions in the following sense: a positive type who enters exaggerates the probability with which her opponent will enter; a negative type who stays out exaggerates the probability with which her opponent stays out. Hence, here, false consensus in prediction is the very condition for false antagonism in inference.

average.[23]

Understanding the mechanism through which persuasion may inflate expectations and lead to credulity is potentially key. For example, in the UK, regulation since 2013 aims at capping the direct commission that financial advisors may receive from the producers of the asset which can be interpreted as a cap on the conflict between the sender and the receiver.[24] Similarly, many have argued that an increase in financial education can ameliorate biased financial decisions. The model implies that it is the very presence of only limited conflict and sufficient financial literacy by receivers which creates the right misperception and drives receivers to make credulous decisions. The key comparative static predictions show that increasing financial education, or lowering the conflict between the sender and the receiver can systematically *fuel* such credulous beliefs and lower receiver welfare. By endogenizing the conflict and the complexity of the asset, when invoking its producer, I also describe that *any* partial cap on the conflict may have very limited effectiveness and still allow for receivers to want to pay for welfare-reducing advice.

4.1 Setup

Timing. The sender privately learns whether a proposition is true, $\{\theta = 1\}$, or false, $\{\theta = 0\}$. She then provides advice via cheap talk. A lobbyist recommends to a politician whether or not to support a given policy; a financial expert recommends to an investor whether the investor should buy or sell a security; and a doctor recommends to a given patient whether or not this patient should take a certain drug. Only the sender knows the truth value of the proposition. Upon receiving advice, the receiver can verify the sender's statement at some cost c privately known to the receiver. If she verifies the message, she learns the value of θ. If she does not, she knows only the recommendation. Finally, the receiver takes an action y. For simplicity, I assume the prior on θ to be symmetric.[25]

Verification. The receiver's cost of verification c is drawn according to a strictly positive density $f(c)$ over $[0, \infty)$. Its realization is the receiver's private information. Differences in costs may reflect private information about the receiver's financial literacy, or his cost of accessing additional sources that help him assess the sender's advice. A higher distribution of costs, that is, a first-order stochastic increase of the cdf F (an increase in F, henceforth), corresponds to a lower distribution of receiver expertise or lower financial education. Equivalently, it can be interpreted as greater complexity of the asset to be evaluated.

Investment. The receiver takes an action $y \in [0, 1]$ to maximize her expected utility. This action could correspond to the fraction of resources allocated into buying or selling a given portfolio, the amount of resources invested in promoting or blocking a policy. To keep

[23] Exogenously invoked naive listeners who ignore conflict of interest have been considered by e.g., Kartik, Ottaviani, and Squintani (2007). Instead, here, credulity arises *endogenously* as a function of the exogenous parameters describing the conflict between the sender and the receiver and the complexity of the problem.

[24] See, for example, https://www.handbook.fca.org.uk/instrument/2011/2011_54.pdf .

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

the analysis fully transparent, I assume that the optimal action equals the receiver's posterior confidence that the proposition is true, $\{\theta = 1\}$. This is captured by the standard assumption that the receiver's payoff is determined by the loss function:

$$u_r(y, \theta) = -(y - \theta)^2. \tag{7}$$

Conflict of Interest. The conflict between the sender and the receiver is such that the sender gets a bonus $B > 0$ – potentially from the seller of the asset to be introduced later – anytime she issues a positive recommendation, independent of θ. At the same time, if the receiver decides to check and finds out that the sender has lied to him, the sender, for example, the doctor or the lobbyist, incurs a loss (of business or reputation) of size S. Without loss of generality, I normalize $S = 1$. Hence, B is always interpreted in proportional terms relative to S. Furthermore, to make the analysis non-trivial, I assume that $B < 1$.

Welfare. When discussing receiver welfare (henceforth welfare), I take the standard ex ante expected perspective. The receiver's welfare is given by the expected loss minus the potential verification cost incurred taking expectations, given the *true* distribution of actions in equilibrium.

4.2 Bayesian Case

Consider the unbiased case. There is a unique equilibrium: the sender tells the truth if $\theta = 1$ and lies with probability p^0, if $\theta = 0$. The receiver checks a positive recommendation iff her cost is below a certain threshold c^0 and never checks a negative recommendation. Below, $E_\theta[y_c^{*,0}]$ denotes the true ex ante expected equilibrium investment (confidence) of receiver type c. I denote the prior confidence by $\overline{y}$.

Proposition 5 *Let $\rho = 0$. In the unique equilibrium, the receiver checks iff $c \leq c^0(F, B)$, and the sender lies with probability $p^0(F, B) > 0$. An increase in F or B increases $c^0(F, B)$ and $p^0(F, B)$. Communication is neutral, $E_\theta[y_c^{*,0}] = \overline{y}$ for all c.*

Neutrality. In equilibrium, each type either checks a positive recommendation or discounts it proportional to the true probability of a lie. To maintain balanced incentives, a greater conflict, or lower average receiver expertise (more complexity) induces more lying and more checking. A key feature of the BNE is that persuasion is neutral: the ex ante expected confidence of each type is the same as his prior. This is a direct and general consequence of the martingale property of Bayesian equilibrium beliefs. Although advice is valuable to the receiver, Bayesian communication is purely informative and never shifts *average* posterior beliefs.

4.3 Persuasion under Projection

Consider, now, a biased receiver ($\rho_R = \rho$) and an unbiased, thus sophisticated, sender ($\rho_S = 0$). A biased receiver, thus, exaggerates the extent to which the sender knows his private information, that is, his cost of verification c. The unbiased sender fully anticipates this. Persuasion is no longer neutral. Instead, it leads to two kinds of mistakes: *credulity*, whereby

a positive recommendation is taken too much at face value, and *disbelief*, whereby a positive recommendation is interpreted with too much skepticism by the receiver. In the former case, persuasion successfully inflates confidence, on average – belief updating forms a strict sub-martingale. In the latter case, persuasion effectively decreases confidence, on average – belief updating forms a strict super-martingale.

Proposition 6 *For any $\rho > 0$, equilibrium is unique. There exist $0 < c_1^\rho < c_2^\rho \leq c_3^\rho$, such that*
(i) for $c < c_1^\rho$, persuasion is still neutral and $E_\theta[y_c^{*,\rho}] = \overline{y}$;
(ii) for $c \in [c_1^\rho, c_2^\rho)$, credulity holds and $E_\theta[y_c^{*,\rho}] > \overline{y}$;
(iii) for $c \in (c_2^\rho, c_3^\rho)$, disbelief holds and $E_\theta[y_c^{*,\rho}] < \overline{y}$;
(iv) for $c \geq c_3^\rho$, (weak) disbelief holds and $E_\theta[y_c^{*,\rho}] \leq \overline{y}$.

To provide intuition, note that a biased receiver too often thinks that the advisor sees through him and knows how costly it would be for him to verify her recommendation. Such a receiver exaggerates the extent to which the sender tailors the truthfulness of her message to his privately known type, as opposed to the commonly known distribution thereof. Since the real sender knows only the distribution of c, she is constrained to lie to each receiver type to the same extent. By exaggerating the extent to which the sender's message is conditioned on his privately known type, the receiver's perception of the extent to which the sender lies to him is decreasing in his cost c in equilibrium. The easier it would be to verify the message, the more he thinks that the sender must be truthful.[26]

In equilibrium, the real sender lies with probability p^ρ. The projected sender is perceived to lie according to a monotone function $p^+(c)$ strictly increasing from zero to one on a positive interval, $[c_1^\rho, c_3^\rho]$. It follows that there always exists a type $c_2^\rho \in [c_1^\rho, c_3^\rho]$ such that for this type the projected version's lying frequency matches the real sender's. This type develops correct beliefs on average despite projection. At the same time, all types below c_2^ρ believe a positive recommendation too much. All types above c_2^ρ are too skeptical. The verification strategy of the receiver matches these perceptions. Types below c_1^ρ always check; types in $[c_1^\rho, c_3^\rho]$ check probabilistically, consistent with their increasing belief about the extent to which the projected sender lies to them; types above c_3^ρ never check, consistent with their belief that the projected sender always lies to them. As a result, receivers with the highest expertise (lowest costs) always check and make correct decisions. Middle cost types, those with sufficient but not full literacy, overinvest on average. For such types, persuasion predictably boosts expected confidence. Finally, persuasion decreases the expected confidence of types with little or no literacy. Such types underinvest on average.

[26] **True Leakage** The results on belief distortions do not depend on the perception of 'leakage' per se, but on *exaggeration* of such a perception due to projection. To see this, suppose that there was a true commonly known probability $\alpha \in [0, 1]$ with which the sender privately learned the receiver's type before making a recommendation. Although the equilibrium, here, would have a similar structure to that in Proposition 6, persuasion would, nevertheless, still be neutral for all types.

4.4 Uniform Credulity

An implication of the above logic is that while credulity is always present, disbelief is a limited phenomenon. Below, I refer to the case in which all receiver types are at least weakly credulous and a strictly positive measure is strictly credulous as *uniform credulity*. Here, persuasion unambiguously increases ex ante expected confidence, causing all receiver types to (at least weakly) overinvest. The next result claims that given any degree of information projection, uniform credulity follows, provided that the conflict is not too low or verification is not too cheap in expectation.

Proposition 7 *For any $\rho > 0$,*

1. *If $B \geq \overline{B}(\rho, F)$, uniform credulity holds. Furthermore, $\overline{B}(\rho, F) < 1$ and is decreasing in ρ with $\lim_{\rho \to 1} \overline{B}(\rho, F) = 0$.*

2. *There exists $\overline{F}(\rho, B)$ such that for any $F \geq \overline{F}(\rho, B)$, uniform credulity holds. Furthermore, if it holds given ρ, it also does for any $\rho' > \rho$.*[27]

To provide intuition, note that as the conflict increases, the amount of information transmitted decreases. To maintain balanced incentives, now, a greater measure of types need to check. Disbelief, however, is limited by the extent to which there is actual information transmission. If the conflict is sufficiently large, the receiver's checking behavior will not counterbalance the sender's incentive to lie, and disbelief no longer applies. At the same time, as long as $B < 1$, projection implies that all receiver types who check underestimate their own incentives to do so, and strictly overinvest on average. In short, although the sender always lies, all receiver types for whom it is ever rationalizable to check underestimate the probability with which she does so. As ρ increases, the threshold value of the conflict above which uniform credulity holds decreases.

Since an increase in F (in the sense of fosd) makes it harder to provide incentives for the sender to transmit information, the same logic applies if the asset is sufficiently complex to evaluate.

4.5 Welfare

I now turn to welfare. In the Bayesian case, a decrease in the conflict, or an increase in receivers financial literacy, increases the amount of information transmitted by the sender and decreases the verification cost incurred by the receiver; hence, it increases receiver welfare. In contrast, in the presence of information projection these comparative statics will systematically reverse. The next result establishes sufficient conditions where such reversal is always true.

Proposition 8 *If $\rho = 0$, a decrease in B or a decrease F always increases welfare. For any $\rho > 0$,*

[27] Since first-order stochastic dominance is only a partial order, multiple such $\overline{F}(\rho, B)$ exist. Below, $\overline{F}(\rho, B)$ refers to any such distribution.

1. *if* $B \geq \overline{B}(\rho, F)$*, a decrease in* B *strictly decreases welfare;*

2. *if* $F \geq \overline{F}(\rho, B)$ *and* $B < \frac{1}{2}$*, a decrease in* F *which does not change* $F(\frac{1-\rho}{(2-\rho)^2})$ *strictly decreases welfare.*

For any $\rho \geq 0$, welfare is still maximal if the conflict is zero or c is zero. Given any $\rho > 0$, however, comparative static results are non-monotonic. In fact, it is the combination of limited conflict and sufficient financial literacy which creates the most scope for credulity and overly optimistic investments. To see this, note that credulous types always check too little relative to their true interests because (a) they underestimate the value of checking, and (b) overinvest in the absence of checking. A decrease in such credulous types' perception of the sender's incentive to lie then results in bolder investments and constitutes a negative welfare force.

Comparative Static with B. In the unbiased case, a decrease in the conflict raises welfare because (i) it increases information transmission, and (ii) induces less costly checking. Given information projection, less checking (iii) also leads to more-biased investments. Whenever uniform credulity holds, all receiver types (at least weakly) underestimate the sender's true probability of lying and check too little. Now, channel (i) is unaffected, but a decrease in B leads to less checking and more biased investments, which unambiguously reduces welfare.

Even if uniform credulity does not hold, the welfare of *some* types must increase in B as long as $\rho > 0$. Since types right above c_1^ρ make the most distinctly overoptimistic investment choices, they enjoy a discontinuously lower expected utility than types just below c_1^ρ since such types always check. An increase in B increases c_1^ρ, hence, it improves the welfare of these originally most credulous types discretely. At the same time, it does not change the welfare of types below the original c_1^ρ. The overall welfare effect here, however, depends on further assumptions.

Comparative static with F. In the unbiased case, greater financial literacy (lower complexity) increases welfare because (i) it mechanically decreases verification costs, and (ii) increases information transmission. Under information projection, however, (iii) it also creates more scope for credulity since, all else equal, a receiver's excess confidence is decreasing in c. If uniform credulity holds, an increase in F lowers the perceived, but not the real amount of information transmitted. If the conflict is not too great, there is always sufficiently little checking such that the benefit of more checking following an increase in F is always higher than the loss due to the higher cost of checking. Holding the measure of types who always check under uniform credulity constant, higher financial literacy fuels credulous expectations and must reduce welfare. The same can hold even if uniform credulity is not satisfied.

4.6 Endogenous Conflict and Complexity

An implication of Proposition 8 is that the receiver's welfare with advice can be lower than his welfare without advice, where the latter is defined as the ex ante expected utility of the receiver when simply acting on the prior on θ without any further information, that is, simply taking action $\overline{y}$. Furthermore, this is true in a setting where the conflict, as well as the true

distribution of θ, are common knowledge.[28] Let me now briefly illustrate these points through endogenizing the value of the conflict B and partially also the complexity of the problem F by invoking the seller of the asset.

So far, the conflict and complexity were exogenous. Suppose, now, that, before the resolution of any uncertainty, the seller – the manufacturer of the drug or the asset – pledges to pay the sender – the doctor or the advisor – some bonus $B \in \mathbb{R}^+$ whenever the sender makes a positive recommendation. As before, the chosen value of B is common knowledge. Suppose that the seller's expected profit is simply the ex ante expected investment – the aggregate demand for buying the drug or the asset – times some markup γ minus the transfer to the sender:

$$R(\rho, B) - B = \gamma E_{c,\theta}[y_c^{*,\rho}] - B. \tag{8}$$

What is the optimal B that an unbiased, hence sophisticated seller, who understands the receiver's true behavior, would want to offer?

In the unbiased case, the seller-optimal conflict is always zero. Since persuasion is neutral, providing a bonus is a pure cost for the seller. In the biased case, a limited increase in B will induce endogenous credulity and increase aggregate confidence and demand. Let $B^*(\rho, F, \gamma)$ denote the seller-optimal bonus. The next corollary shows that, as long as the markup on the asset is not too low, the seller always wants to choose an intermediate value for the size of the conflict.

Corollary 4 *If $\rho = 0$, then $B^*(0, F, \gamma) = 0$. If $\rho > 0$, then $B^*(\rho, F, \gamma) = 0$ if $\gamma \leq \overline{\gamma}(\rho, F)$ and $0 < B^*(\rho, F, \gamma) < 1$ if $\gamma > \overline{\gamma}(\rho, F)$.*

Finally, consider the case whether the seller can also affect the distribution F. As an extreme assumption, suppose that the seller can pick not only B but also any $F \in \Delta\mathbb{R}^+$satisfying full support. While a full analysis on the seller-optimal joint design of B and F is beyond the scope of the analysis, let me conclude with a partial one. Suppose the seller wants to implement uniform credulity. The seller can then maximize expected profit by minimizing the size of the bonus B and picking an F concentrated on sufficiently low, but not too low, cost types.

Corollary 5 *Let $\rho > 0$. The seller-optimal way of inducing uniform credulity (i) minimizes B subject to $B > 0$ and (ii) concentrates F at $(1-\rho)/(2-\rho)^2$. Here, the receiver's welfare is always lower than without advice.*

The above corollary shows that any partial cap on the conflict still allows for exploitative persuasion. As long as uniform credulity holds, the cutoffs c_1^ρ and c_3^ρ are independent of F and B. Note then two facts. First, given any B, investment in the asset is decreasing in c as long as

[28] Since it is common knowledge, mandatory disclosure of the conflict B here is ineffective, and credulity arises endogenously despite this fact. In a setting with exogenously invoked mechanical naive receivers who are assumed to simply take recommendations at 'face value', Ottaviani and Indherst (2012) advocate mandatory disclosure as a way to eliminate credulity.

$c > c_1^\rho$. Second, the frequency of verification is increasing in the conflict B, thus overinvestment is decreasing in B. Hence, the seller's revenue is maximized when B goes to zero while F is concentrated on c_1^ρ subject to the sender still gaining more from misreporting than from truth-telling. Minimizing B and concentrating F on c_1^ρ, such that $B > F(c_1^\rho)/(F(c_1^\rho) + 1 - F(c_3^\rho))$ still holds, creates uniform credulity most conducive for average overinvestment. This way, the seller takes full advantage of credulity induced by information projection and does so at the lowest possible cost. One can then compare the seller's and the receiver's expected welfare in three cases: (i) without advice, (ii) with advice in the seller-optimal design in the unbiased case, and (iii) in the limit of the above constrained seller-optimal setting with advice.

	Receiver's welfare	Seller's expected profit
no advice	$-\frac{1}{4}$	$\frac{\gamma}{2}$
advice, $\rho = 0$	0	$\frac{\gamma}{2}$
advice, $\rho > 0$	$-\frac{1-\rho+0.5\rho^2}{(2-\rho)^2}$	$\frac{\gamma}{1+(1-\rho)}$

In the unbiased case, advice benefits the receiver. In the biased case, advice hurts the receiver and boosts the seller's profit given any positive ρ.

Paying for Advice Consider now the ex ante interpretation of the model whereby the biased receiver wrongly believes that the realization of c will privately leak also to the sender with probability ρ. For any $\rho > 0$, such a receiver would still be willing to pay a positive amount for advice ex ante. The perceived welfare consequence of advice is still strictly positive for any feasible F. Hence, the receiver is willing to pay a non-trivial amount for advice which then only reduces his welfare. Despite a commonly known compensation structure of the sender, a commonly known distribution of the quality of the asset, given information projection, investment following financial advice leads to worse outcomes than the outcomes the investor could achieve without advice, that is, simply acting on his prior, *even* when ignoring the amount paid for such advice.[29]

5 Projection Equilibrium

So far, I focused on information projection. Its logical counterpart is *ignorance projection:* the mistaken belief that if one does not known an event, others do not know it either. Taken together, information and ignorance projection imply total projection, that is, an exaggerated belief that others know the same events she does. While in many domains people may project their information without projecting their ignorance, and the direct evidence supports the presence of information projection, the formalism of Section 2 allows one to incorporate the *joint* presence of information and ignorance projection. I turn to the resulting solution of projection equilibrium.

[29] A recent literature considers the benefit for the sender of committing to a a communucation rule ex ante in a Bayesian setting, e.g., Rayo and Segal (2010). In this model, given projection, it is exactly the *lack* of such commitment which allows the seller to take advantage of the receiver's endogenously arising credulity.

If player j projects both her information and her ignorance, she exaggerates the probability with which player i can condition his strategy on the same set of events as she can. Formally, the projected version of player i – who is real in the imagination of player j – now chooses a strategy from the set:

$$S_i^j = \{\sigma_i(\omega) \mid \sigma_i(\omega) : \Omega \to \Delta A_i \text{ measurable with respect to } P_j(\omega)\}. \tag{9}$$

In each state, this fictional projected version of i knows the events that real j knows and only those events.

Definition 4 *A strategy profile $\sigma^\rho \in S_i \times S_j$ is a ρ-projection equilibrium of Γ if there exists $\sigma^\pm = \{\sigma_i^j, \sigma_j^i\} \in \{S_i^j \times S_j^i\}$ such that for all i,*

1.

$$\sigma_i^\rho \in BR_{S_i}\{(1-\rho)\sigma_{-i}^\rho \circ \rho\sigma_{-i}^i\} \tag{10}$$

2. and

$$\sigma_{-i}^i \in BR_{S_{-i}^i}\{\sigma_i^\rho\}. \tag{11}$$

The definition satisfies the two key properties as before: projection is all-encompassing, and the limited consistency property holds. First, the projected version of Paul, as imagined by Judith, knows that Judith is regular. Second, each real player assigns probability $1-\rho$ to her opponent's real strategy. The difference from IPE is that Judith now wrongly thinks that with probability ρ Paul knows the *same* set of events as she does. In the case of poker, the projected version of Paul knows Judith's hand, but not his own hand. Thus, he faces the same kind of uncertainty about this as real Judith. The structure of higher-order perceptions is the same as before, that is, people again display partial anticipation of the biases of others. The definition again immediately extends to heterogeneous projection. Finally, equivalent versions of Proposition 1 and Corollary 1 continue to hold.

Nested Model Crucially, the two models can be nested within a single one. Specifically, suppose that each real player j assigns probability ρ^+ to player i choosing his strategy from the set S^+; she assigns probability $\rho^\pm$ to player i choosing his strategy from the set S_i^j; and assigns probability $1-\rho^+-\rho^\pm > 0$ to i being regular. Suppose, as before, that projection is all-encompassing: the real player j believes that both of the above projected versions of player i know that j is real for sure. If $\rho^\pm = 0$, this joint model collapses to that of information projection equilibrium. If $\rho^+ = 0$, the joint model collapses to that of projection equilibrium.

5.1 Trade

As the last application, I consider the predictions of projection equilibrium to the classic problem of common-value trade with asymmetric information, Akerlof (1970). The informed party, the seller or the target company, values the object of quality q at q. The uninformed party, the buyer or the acquiring the company, values it at $w(q)$. If $w(q) > q$, it is common knowledge

that there are benefits from trade. Quality q is drawn from a density π, and its realization is observed only by the seller.

Strategic behavior in this fundamental setting has been explored experimentally by Samuelson and Bazerman (1985) and a literature following it. The remainder of this section derives the predictions of projection equilibrium to this problem and compares its empirical fit with that of BNE and cursed equilibrium often motivated by addressing biased decision making in such settings with adverse selection.

5.1.1 Additive Lemons Problem

Samuelson and Bazerman (1985, S&B henceforth) study an additive specification where $w(q) = q + x$ with $x > 0$ and π is uniform on some $[a, b]$ with mean $\overline{q}$. They study both the case where the seller has the bargaining power and the case where the buyer does.[30]

Seller-Offer When the seller has the bargaining power, the seller makes a take-it-or-leave-it (TIOLI) price offer $p_s^\rho(q)$ which the buyer can accept or reject. A key feature of any BNE of this seller-offer game is that the seller cannot sell different qualities at different prices for sure, or even with the same probability. Such pricing is not incentive compatible. An unbiased seller who fully appreciates the informational asymmetry would never name the lower of any two of such prices. The seller's incentive to bluff then limits her ability to sell and greatly reduces the efficiency of trade. By projecting information, a biased seller mistakenly thinks that the buyer maybe able to detect a bluff. This increases the scope for truth-telling and efficient trade. Specifically, the following result holds.

Proposition 9 *For any $\rho \geq 0$, there exists a ρ projection equilibrium where $p_s^\rho(q) = q + x$, and the buyer accepts any price below $\overline{p} = \min\{\frac{x}{1-\rho}, \overline{q} + x\}$ for sure and any higher price p with probability $e^{-(p-\overline{p})/x}$.*

Two properties characterize the above prediction. First, the seller engages in non-altruistic truth-telling: the seller's price fully reveals the quality, but leaves no rent for the buyer. Second, the seller under-bids relative to buyers' actual acceptance behavior. Since all prices below $\overline{p}$ are accepted by the buyer with certainty, the seller leaves money on the table. In sum, all qualities below a certain threshold are sold for sure, high quality items are sold with decreasing probabilities, and all benefits realized through trade go to the seller.

The above ρ PE is supported by the strategies of the fictional projected player versions. The projected buyer – whom the seller believes to knows q – has a dominant strategy to accept a price p if and only if $p \leq q + x$. The projected seller – whom the buyer believes not know q, but who, given all-encompassing projection, is believed to know that the buyer does know q – bids $\overline{q} + x$. The bound on $\overline{p}$ is then determined by whether the IC constraint due to the deviation of the real seller, or that of the projected seller binds.[31]

[30] Since, here, some offers need not be on the equilibrium path, I assume that the standard restriction of perfectness holds.

[31] Note that this result relies on the joint presence of information and ignorance projection, because the projected seller cannot base deviations on the realization of q.

Data The two properties of the model's prediction match the evidence closely. S&B study the case where $a = 0, b = 100$, and $x = 30$. They find that the the most common bidding strategy of the sellers is $p_s(q) = q + 30$. Furthermore, sellers significantly underbid relative to what their payoff maximizing strategy would be, given the buyers' actual acceptance behavior. In particular, the acceptance probability in the data is fairly flat for any price below 80, but declines more quickly after that.[32] Finally, if $\rho = 0$, the seller's maximal revenue is attained in the equilibrium in which the seller sells only objects with qualities lower than 60 at a single price of $p = 60$. It is easy to see that if ρ is sufficiently high, the above equilibrium generates higher revenue and greater social efficiency than this Bayesian optimal one.

Buyer-Offer Let me turn to the buyer-offer game where it is the uninformed buyer who makes a TIOLI price offer p_b which the seller can accept or reject. The analysis is simplified since both the real and the projected seller have dominant strategies. The real seller accepts p_b iff it is greater than q; the projected seller, who does not know q, accepts it iff it is greater than $\overline{q}$. This is true because given all-encompassing projection, the projected seller is believed to know that the buyer does not know q. Hence, the buyer's perceived expected utility when bidding p_b is simply

$$EU_b^\rho = \begin{array}{ll} (1-\rho)\Pr(q < p_b)(E_\pi[w(q) \mid q < p_b] - p_b) + \rho 0 & \text{if } p_b \leq \overline{q} \\ (1-\rho)\Pr(q < p_b)(E_\pi[w(q) \mid q < p_b] - p_b) + \rho(E_\pi[w(q)] - p_b) & \text{if } p_b > \overline{q}. \end{array} \tag{12}$$

Given the specification of S&B, this implies the following claim.

Claim 4 *In the unique ρ projection equilibrium, the buyer's bid is given by*

$$p_b^\rho = 30 \text{ if } \rho \leq 1/16, \text{ and } p_b^\rho = 50 \text{ if } \rho > 1/16.$$

If ρ is small, the unique prediction of PE is identical to the unique prediction of BNE. If it is greater than 0.062, the buyer bids the seller's unconditional valuation. A projecting buyer now underappreciates negative selection. As a result, the buyer overbids, buys more often but realizes a smaller expected payoff than in the unbiased case. In other words, he falls prey to the classic 'winner's curse.'[33]

Data The data matches the predictions closely. S&B find that the most common is in fact 50. Furthermore, less than 17 percent of bids are in $[30, 35]$. A non-trivial fraction of bids are above 60. Under correct expectations, bidding above 60 leads to strictly negative expected earnings for the buyer. In contrast, bidding below 80 still leads to positive perceived earnings under projection for any $\rho > 1/16$.[34]

[32] Here, $\overline{p} = 80$ for any $\rho > \frac{5}{8}$.

[33] The predictions of the nested model are isomorphic with the predictions of projection equilibrium by setting $\rho^\pm = \rho$ and allowing ρ^+ to be any number smaller than $1 - \rho^\pm$.

[34] One of the treatments of Fudenberg and Peysakhovich (2013) also studies an additive lemons problem with $a = 0$, $b = 10$, and $x = 3$. The average bid is again 5.1.

In the buyer-offer game, cursed equilibrium, $CE(\chi)$, also predicts plausible deviations from the BNE. The predictions of cursed equilibrium span the interval $[30, 40]$ as a function of the degree of cursedness χ; 40 being the fully cursed prediction. Projection equilibrium, thus, robustly matches the data better than BNE or cursed equilibrium given any $\rho > 1/16$.

5.1.2 Multiplicative Lemons Problem

Holt and Sherman (1994) test a multiplicative specification where $w(q) = 1.5q$ and π is uniform on $[q_0, q_0 + r]$. They focus only on the buyer-offer game. Table below characterizes the predictions of the unique projection equilibrium in the three conditions studied experimentally and calibrated by Eyster and Rabin (2005).

	$[r]$	$[q_0]$	$[m]$	$b(\chi=0)$	$b(\chi=1)$	$b(\rho>\rho^*)$	ρ^*	$\bar{b}$
No Curse	2	1	1.5	2	2	2	0	2
Winner's Curse	4.5	1.5	1.5	3	3.5	3.75	0.02	3.78
Loser's Curse	0.5	0.5	1.5	1	0.81	0.75	0.07	0.74

Table 1: Holt and Sherman (1994), Eyster and Rabin (2005).

In Table 1, the average empirical bid is $\bar{b}$; the unique prediction of BNE corresponds to $b(\chi = 0)$, the unique fully cursed prediction to $b(\chi = 1)$, with CE(χ) spanning the interval between these two. The unique ρ-projection equilibrium is identical to BNE if $\rho < \rho^*$; and equals $b(\rho > \rho^*) = \overline{q}$ if$\rho > \rho^*$. Projection equilibrium, thus, matches the data almost perfectly and robustly. In the winner's curse condition, this is true for *any* $\rho > 0.02$, and in the loser's curse condition, for *any* $\rho > 0.07$. The reason that such a small degree of projection leads to such substantial deviation in bidding behavior is that the buyer's gain from trade conditional on selection is much smaller than the gain from trade without selection. In the winner's curse condition, slightly under-estimating selection leads to substantial overbidding. In the loser's curse condition, it leads to substantial underbidding.[35],[36]

Cursedness versus Ignorance Projection In the buyer-offer game, both cursedness and ignorance projection imply empirically plausible deviations from BNE. Their predictions and logic differ. A cursed buyer has correct expectations about the seller's information, but

[35] Ball, Bazerman, and Carroll (1991) study a close variant of this multiplicative specification and also allow for multiple rounds of learning. Here, the relevant threshold is $\rho > \rho^* = 0.12$

r	q_0	m	$b(\chi=0)$	$b(\chi=1)$	$b(\rho>\rho^*)$	$\bar{b}$
1	0	1.5	0	0.375	0.5	0.55

[36] The predictions of the nested model are again isomorphic with the predictions of projection equilibrium by setting $\rho^{\pm} = \rho$ and allowing ρ^{+} to be any number smaller than $1 - \rho^{\pm}$.

mistakenly thinks that with probability χ the seller's acceptance is independent of q, for example, the buyer believes that the seller might accept a price lower than the seller's privately known valuation. A buyer who projects her ignorance mistakenly thinks that with probability ρ the seller does not know q, hence, accepts any price greater than $\overline{q}$. A projecting buyer has mistaken beliefs about the seller's information, but has coherent beliefs about how such a seller might act given that information. These differences imply that the two will have qualitatively different predictions in many settings. For example, as ER (2005) note, that $CE(\chi)$ predicts a strictly positive bid even if $m < 1$, that is, even if the buyer always values the object strictly less than the seller. Projection equilibrium, instead, here, would always predict a bid of 0.

Projecting Valuation Finally, the data is inconsistent with the hypothesis that players mistakenly think that others have the same valuations, as opposed to the same information, as they do. As presented, informed sellers bid the buyers' higher conditional valuations. Uninformed buyers bid the sellers' lower unconditional valuations. They both act as if they exploited the right binding individual rationality constraints, ignoring informational differences.

6 Conclusion

A wealth of direct evidence shows that people fail to fully appreciate informational differences and too often thinks that others can condition their choices on their private information. This paper incorporates this general mistake into the solution of Bayesian games. Incorporating informational projections into the analysis of strategic problems may shed novel light on a number of economic outcomes in contexts not covered in this paper. For example, it is likely to affect bargaining outcomes, behavior in contests, information aggregation in committees and juries, or trading in markets with asymmetric information. Similarly, informational projections will affect people's demand and supply of information – as in the cases of search and signalling. Future research can extend the findings presented and consider implications to a variety of other problems.

A context where the wedge between true and perceived informational differences may be particularly important is mechanism design. When designing optimal incentives, a key concern is the optimal provision of information rents. The presence of projection will affect agents' demand for information rents and, by modifying key incentive compatibility constraints, may alter the shape of optimal trading mechanism. As in Section 5, this will affect the scope for truth-telling and efficiency and may alter such classic results for bilateral trade as the Bayesian upper-bound identified by Myerson and Satterthwaite (1983).

In this vein, Madarász (2014b) extends the current model to sequential bargaining with observable moves, and shows that the presence of even minimal projection can significantly alter the seller-optimal way to sell an object, Myerson (1981). The model provides a strong rationale for haggling over commitment to posted prices or price-schedules. The model predicts a full reversal of the classic Coasian property of bargaining. The existing evidence rejects the Bayesian comparative static results but, instead, is consistent with the model. Further dynamic extensions of the model to social learning or to consumers' perception of the value of

their privacy and firms' dynamic contracting responses to projection-based misperceptions may be particularly fruitful.

The portability of the model allows one to assess the empirical implications of this phenomena in many strategic settings. For example, in the context of a simple agency setting, Danz, Madarász and Wang (2014) find strong support for the model by directly eliciting beliefs. The model may offer a more unified explanation of a variety of seemingly unrelated or contradictory empirical findings in social psychology and provide a clear testable ex ante hypothesis, including comparative static results, as to when they may or may not occur.

7 Appendix

7.1 Appendix A: Multi-Player Extension

Lastly, consider an N-player game Γ. Below, I define the extension of information projection equilibrium.[37] The extension to projection equilibrium is perfectly analogous. Now each player i has a collection of projected opponents, one for each opponent. Furthermore, since the information of players i and j differ, the projected version of player k, as imagined by player i, differs from the projected version of k, as imagined by player j. The projected version of Sam – as imagined by Judith – knows Sam's hand and Judith's hand. The projected version of Sam – as imagined by Paul – knows Sam's hand and Paul's hand.

As introduced in Section 2, let the strategy set of the projected version of player i – as imagined by player j – be

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

This set again consists of the strategies player i could choose from if he could condition his behavior on the joint information of players i and j. I denote the generic element of this set by σ_i^{i+j}. Let

$$S^{+j} = \prod_{i \neq j} S_i^{i+j}$$

be the strategy set of the $N-1$ fictional projected opponents of player j. I denote the generic element of this set by σ^{+j}. Lastly, I denote the restriction of a profile σ^{+j}, containing all of its elements, except for some σ_k^{+j}, by σ_{-k}^{+j}. Finally, let $S = \prod_{i=1}^{N} S_i$ be the set of strategies of the real players.

In the definition below, player j believes that all projected versions of her opponents occur in a perfectly correlated manner; she believes that with probability ρ all her opponents are projected versions and with probability $1-\rho$ they are all regular versions. Projection is again all-encompassing: each projected opponent of player j knows that j is real and the same limited consistency property holds as before. Finally, I assume, for simplicity, that, consistent with player j's belief that projection occurs in a perfectly correlated manner, player j believes

[37] In the case of private projection such extension is straightforward since deviations from a BNE σ^0 are uncoordinated.

that all her projected opponents believe that they are facing her other projected opponents. If the true game is poker, a biased Judith thinks that with probability ρ both Sam and Paul know her hand. Furthermore, to maintain transparency, Judith believes that each of her projected opponents believes that all other players are the corresponding projected versions as well; projected Paul, who knows her hand believes that Sam knows her hand as well. Let me then turn to the definition.

Definition 5 *A strategy profile $\sigma^\rho \in S$ is a ρ information projection equilibrium of Γ if for each i there exist a $\sigma^{+i} = \{\sigma_j^{j+i}\}_{j \neq i} \in S^{+i}$ where*

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

and for each $j \neq i$

$$\sigma_j^{j+i} \in BR_{S_j^{j+i}}\{\sigma_i^\rho, \sigma_{-j}^{+i}\}.$$

The extension of projection equilibrium is analogous. It is obtained by replacing each S_i^{i+j} by S_i^j, as defined before, and S^{+j} with $S^j = \prod_{i \neq j} S_i^j$. Equivalent versions of Proposition 1 and Corollary 1 continue to hold.

7.2 Appendix B

Proof of IPV. As shown by Riley (1989), the BNE of the first-price auction is given by an efficient mixed-strategy equilibrium where each payoff type mixes over an interval of positive measure such that different payoff types mix over intervals that are non-overlapping. Consider now a σ^+ with the following properties. If the fictional super player $-i$ has a lower valuation than his opponent, $\theta_{-i} < \theta_i$, then he will bid higher than the regular player $-i$ with the same valuation θ_{-i}, if the lowest value of the support over which type θ_i mixes is lower than θ_{-i}. If the fictional super player has a weakly higher payoff type than his opponent, $\theta_{-i} \geq \theta_i$, then he will under-bid, and bid the highest value of the support over which θ_i mixes under $\sigma_i^0(\theta_i)$. Consider now the biased player's best response.

$$b^*(\theta_i) \in \arg\max E_\pi[\rho \Pr(win \mid b, \sigma_{-i}^+(\theta_i)) + (1-\rho)\Pr(win \mid b, \sigma_{-i}^0)](\theta_i - b)$$

Note first that since the auction was efficient under σ^0, the equilibrium probability of winning was zero conditional on the opponent having a higher valuation. In contrast under σ^+ it is positive if the bidder bids above the relevant part of σ^0. At the same time, bidding lower than under σ^0, the probability of winning is lower than in the case where $\rho = 0$. It is thus easy to see that bidding below the lowest value of the support over which this payoff type was mixing under the BNE cannot be an equilibrium. Finally, note that if one's opponent has the same valuation as she, an event that happens with positive probability given the finite support, then given the indifference condition under BNE and the deviation of such an informed type, it is

now strictly beneficial to bid above the original bid. The discontinuity in the revenue result arises from the fact that for any $\rho > 0$, a biased player will not bid below the highest point of the interval on which she was supposed to mix under the BNE combined with the fact that all such intervals have positive measure. .

Proof of Zero-Sum Games. The derivation of the private $\rho - IPE$ follows directly from algebra. Specifically, it is given by:

$p < \frac{1}{2}$	weak	strong	EU_D
$\rho = 0$	B	$\frac{1}{2-2p}$A$\circ\frac{1-2p}{2-2p}$B	$\frac{1}{2}$
$\rho = 1$	A	B	$\frac{1-\omega_w p}{2}$
	$\frac{1}{2}a \circ \frac{1}{2}b$	$\frac{1}{2}a \circ \frac{1}{2}b$	

and

$p > \frac{1}{2}$	weak	strong	EU_D
$\rho = 0$	$\frac{2p-1}{p(2-\omega_w)}A\circ\frac{1-\omega_w p}{p(2-\omega_w)}$B	A	$\frac{1-\omega_w p}{2-\omega_w}$
$\rho = 1$	A	B	$\frac{1-\omega_w}{2-\omega_w}$
	$\frac{1}{1+\omega_w}a\circ\frac{\omega_w}{1+\omega_w}b$	$\frac{1}{1+\omega_w}a\circ\frac{\omega_w}{1+\omega_w}b$	

To see the revenue result, note that in case the defender does not have private information, her expected utility (winning probability) is $\frac{1-p\omega_w}{2-p\omega_w}$. To show the result, note that $\frac{1-p\omega_w}{2-p\omega_w} \leq \frac{1}{2}, \frac{1-p\omega_w}{2-\omega_w}$. At the same time, $\frac{1-\omega_w p}{2-\omega_w} > \frac{1-\omega_w}{2-\omega_w}, \frac{1-\omega_w p}{2}$ where the latter follows from the fact that $1 > \omega_w p$. .

Proof of Proposition 1. Note that since the best-response correspondences of the perceived game are upper hemicontinous and convex the existence of IPE (PE) follows from Kakutani's theorem .

Proof of Corollary 1. If $P_i = P_j$, then $P^+ = P_i = P_j$. For any $\sigma^0 \in BNE(\Gamma)$, consider σ^+ where $\sigma_i^+(\omega) = \sigma_i^0(\omega)$ for each ω and i. This strategy supports $\sigma^0(\omega)$ as a ρ IPE for any given ρ since $S_i^+ = S_i$ for each i. By the same token, any σ^ρ which is a $\rho - IPE$ of Γ must also be a BNE of Γ. The logic immediately extends to projection equilibrium since, here, now $S_i^j = S_i$ for each i and j .

Proof of Proposition 2. Suppose σ^0 is a BNE and is also an ex-post equilibrium. Then, for each i and any $\sigma_i' \in S_i$

$$u_i(\sigma_i^0(\omega), \sigma_{-i}^0(\omega), \omega) \geq u_i(\sigma_i'(\omega), \sigma_{-i}^0(\omega), \omega) \text{ for all } \omega \in \Omega$$

Consider now $\sigma_i^+(\omega) = \sigma_i^0(\omega)$ for each ω and i. It follows that $\sigma_i^+ \in BR_{S_i^+}(\sigma_{-i}^0)$ since $S_i \subseteq S_i^+$. Hence, it follows that σ^0 is a $\rho - IPE$ for any ρ .

Proof of Proposition 3. The proof of Proposition 4 shows that equilibrium is in cut-off strategies both for the real and for the projected players. If θ^{ρ}_{-i} is player $-i$'s cutoff, then player i is indifferent between In and Out at θ^{ρ}_{i} satisfying

$$\rho(x(\theta^{\rho}_{i}-\gamma\theta^{\rho}_{i})-nc)+(1-\rho)((x-\theta^{\rho}_{-i})(\theta^{\rho}_{i}-\gamma\theta^{\rho}_{i})+\theta^{\rho}_{-i}(\gamma\theta^{\rho}_{i})-nc)=0 \tag{13}$$

Solving for θ^{ρ}_{i}, one obtains that

$$\theta^{\rho}_{i}=\frac{cn}{x(1-\gamma)+\theta^{\rho}_{-i}(1-\rho)(2\gamma-1)}. \tag{14}$$

Substituting in the symmetric equation for θ^{ρ}_{-i}, then taking $\gamma\to 1$, the unique interior solution is $\theta^{\rho}_{i}=\sqrt{nc/(1-\rho)}$.[38]

1. If $\theta_i>0$, then $\theta^{\rho}_{-i}>\theta^{+}_{-i}(\theta_i)=0$ for all $c,\rho>0$. This implies strict underestimation given any $a\in A$ since the projected opponent uses a strictly lower cutoff than the real opponent. Underestimation continues to hold even if player i observes her own payoff ex post, except when $(a_i,a_{-i})=(in,out)$. If $(a_i,a_{-i})=(in,out)$, underestimation is weak after player i observing her own payoff since whenever i learns that $-i$ is positive, i knows that $-i$ must have been the real version, thus develops correct beliefs. In all other cases, observing her own payoff contains no additional information.

2. Let $\Pr(in)^{\rho}_{\theta_i}$ be the perceived probability that type θ_i assigns to player $-i$'s entering in equilibrium. Let $\Pr(in)$ be the true probability of such an event. For each θ_i, the martingale property of beliefs holds with respect to this *perceived* probability in equilibrium. Hence, by the law of total probability,

$$E_0[\theta_{-i}|\,\theta_i]=\Pr(in)^{\rho}_{\theta_i}E^{\rho}_{\sigma^{\rho}}[\theta_{-i}|\,\theta_i,a_i^{\theta_i},a_{-i}=in]+(1-\Pr(in)^{\rho}_{\theta_i})E^{\rho}_{\sigma^{\rho}}[\theta_{-i}|\,\theta_i,a_i^{\theta_i},a_{-i}=out] \tag{15}$$

where $a_i^{\theta_i}$ is the action taken by θ_i in equilibrium. Let

$$\Delta^{\rho}(\theta_i)\equiv E^{\rho}_{\sigma^{\rho}}[\theta_{-i}|\,\theta_i,a_i^{\theta_i},a_{-i}=in]]-E^{\rho}_{\sigma^{\rho}}[\theta_{-i}|\,\theta_i,a_i^{\theta_i},a_{-i}=out]$$

denote the difference between the conditional mean estimate of type θ_i when observing player $-i$ choose In and the conditional mean estimate versus when observing player $-i$ choose Out. Note, that $\Delta^{\rho}(\theta_i)>0$ must hold for any θ_i and any $\rho\geq 0$. Consider now the difference between the prior mean estimate and the true ex ante expected posteriori mean estimate of type i. This is given by,

$$E_0[\theta_{-i}|\,\theta_i]-E^{0}_{\sigma^{\rho}}[E^{\rho}_{\sigma^{\rho}}[\theta_{-i}|\,\theta_i]]=\Delta^{\rho}(\theta_i)(\Pr(in)-\Pr(in)^{\rho}_{\theta_i}). \tag{16}$$

If $\rho=0$, the RHS of Eq.(16) must be zero for any θ_i. Suppose $\rho>0$. If $\theta_i>0$, then $\theta^{\rho}_{-i}>\theta^{+}_{-i}(\theta_i)$; hence, $\Pr(in)^{\rho}_{\theta_i}>\Pr(in)$.This implies that the RHS of Eq.(16) is strictly negative. If $\theta_i<0$, then $\theta^{\rho}_{-i}\leq\theta^{+}_{-i}(\theta_i)$; hence $\Pr(in)^{\rho}_{\theta_i}<\Pr(in)$. This implies that the RHS of Eq.(16) is positive .

[38] I adopt the convention that when no interior solution exists, then $\theta^{\rho}_{i}=x$.

Proof of Corollary 2. Suppose there is no entry by either players until round $t-1$. Since equilibrium is unique and is in cutoff strategies for any x, it follows that in round t, player i's belief about $-i$'s type must be given by some uniform density $[x_{t-1}, -n]$. It, thus, follows from Proposition 3, that if $\theta^0_{i,t}$ is the cutoff used by player i in round t conditional on no entry until round $t-1$, then $\theta^0_{i,t} = \sqrt{nc_t}$. Hence, $\Pr^0(M \mid \underline{c}) = \max\{1 - nc_T/x^2, 0\}$ since, by round T all types greater than$\sqrt{nc_T}$ must have entered .

Proof of Corollary 3. Let $q^\rho_{t,-i}$ be the probability that player i assigns to the event that $\{\theta_{-i} |\, \theta_{-i} > 0\}$, conditional on no entry until round t. Let $z^\rho_{t,-i}$ be the probability that the real player $-i$ chooses In in round t, again, conditional on no entry until round t. Straightforward algebra shows that player i's indifference cutoff in such a round t is

$$\theta^\rho_{i,t} = \frac{(1 - q^\rho_{t,-i})c_t}{(1-\rho)\,(q^\rho_{t,-i} - z^\rho_{t,-i})}, \tag{17}$$

which is decreasing in $q^\rho_{t,-i}$ and increasing in c_t.

Suppose that $c_t > (1-\rho)x(q^\rho_{t,-i})/(1 - q^\rho_{t,-i})$ for all $t < M(\rho)$. Then by symmetry and uniqueness, it follows from Eq.(17) that $z^\rho_{t,-i} = 0$ for all $t < M(\rho)$. Furthermore, as long as $\rho > 0$,

$$q^\rho_{t,-i} = \frac{q^\rho_{t-1,-i}(1-\rho)}{(1-\rho) + (1 - q^\rho_{t-1,-i})\rho} < q^\rho_{t-1,-i} \text{ for all } t < M(\rho) \text{ and } i, \tag{18}$$

hence, c_t can be a strictly decreasing sequence. Set $c_T = \tau > 0$. Since the belief sequence given by Eq.(18) converges to 0 as $M(\rho)$ goes to infinity for any $\rho > 0$, it follows, that, for any $\varepsilon > 0$, there exists $m(\rho)$ such that $q_{m,-i} \leq \varepsilon$ if $m > m(\rho)$. It follows that there exists $\overline{m}$ such that

$$\theta^\rho_{i,\overline{m}+1} = \frac{(1 - q^\rho_{\overline{m},-i})c_T}{(1-\rho)\,(q^\rho_{\overline{m},-i} - z^\rho_{\overline{m},-i})} \geq x. \tag{19}$$

Furthermore, the same holds a fortiori for a weakly dominant sequence c' .

Proof of Proposition 4. To simplify notation, let $r = (x+n)^{-1}$ corresponding to the range of types.

1. First, I show that equilibrium is in cutoff strategies. Note that the projected version of player $-i$ has a dominant strategy and enters iff $\min(\theta_i, \theta_{-i}) > 0$. Let z_{-i} be the equilibrium probability, given some strategy $\sigma_{-i} \in S_{-i}$ of real player $-i$, that real $-i$ enters. For any real type $\theta_i > 0$, the expected utility difference between entering versus staying out is then

$$\rho(rx(\theta_i - f(\theta_i)) + \textstyle\int_{-n}^{0} rg(\theta_i, \theta_{-i})d\theta_{-i}) + \\ (1-\rho)(z_{-i}(\theta_i - f(\theta_i)) + (1 - z_{-i})E[g(\theta_i, \theta_{-i}) \mid \sigma_{-i} = out]). \tag{20}$$

Differentiating the expression in Eq.(20) with respect to θ_i, it follows that this difference is strictly increasing in θ_i for any given σ_{-i} since $f_1 < 1$ and $g_1 \geq 0$, for $\theta_i > 0$. Hence, equilibrium

must be in cutoff strategies.

2. Consider then the best-response functions of the real players. The function determining player i's cutoff is $\beta^\rho(\theta_{-i}) : [0, \theta_{\max}] \to [0, \theta_{\max}]$. It is defined only on the positive domain and range since negative types stay out in equilibrium. Note that $\beta^\rho(\theta_{-i})$ is continuous in $\theta_{-i} > 0$ and Eq.(20) is continuously differentiable in θ_{-i}. The implicit function theorem implies that the slope of $\beta^\rho(\theta_{-i})$, evaluated at some point $(\widehat{\theta}_i, \widehat{\theta}_{-i})$, is

$$\frac{\overbrace{(1-\rho)(\widehat{\theta}_i - f(\widehat{\theta}_i) - g(\widehat{\theta}_i, \widehat{\theta}_{-i}) - \int_{-n}^{\widehat{\theta}_{-i}} r g_2(\widehat{\theta}_i, \theta_{-i}) d\theta_{-i})}^{I} *}{\underbrace{[\rho(x(1 - f'(\widehat{\theta}_i)) + \int_{-n}^{0} r g_1(\widehat{\theta}_i, \theta_{-i}) d\theta_{-i}) + (1 - \rho((\Pr(\theta_{-i} > \widehat{\theta}_{-i})(1 - f'(\widehat{\theta}_i)) + \int_{-n}^{\widehat{\theta}_{-i}} r g_1(\widehat{\theta}_i, \theta_{-i}) d\theta_{-i})]^{-1}}_{II}}. \tag{21}$$

Term II is strictly positive. Term I is strictly negative if investments are substitutes, and strictly positive if investments are complements and $g_2 = 0$.

3. By the intermediate value theorem, a symmetric equilibrium must exist since $\beta^\rho(\theta_{-i})$ is continuous and monotone, with $\beta^\rho(0) \leq \theta_{\max}$ and $\beta^\rho(\theta_{\max}) \leq \theta_{\max}$, and the players' best-response functions are mirror images on the 45-degree line. Consider substitute investments. Since $\beta^\rho(\theta_{-i})$ is strictly decreasing, there is a unique symmetric equilibrium. Consider complement investments. Here, all equilibria must be symmetric. This is true since, given that $\beta^\rho(\theta_{-i})$ is strictly increasing, $\widehat{\theta}_i = \beta^\rho(\widehat{\theta}_{-i}) > \beta^\rho(\widehat{\theta}_i) = \widehat{\theta}_{-i}$ cannot hold.

4. Consider the comparative static with respect to ρ. Consider cutoffs $(\theta_i^\rho, \theta_{-i}^\rho)$ that constitute a $\rho - IPE$ for a given ρ. Rewriting the equilibrium condition, using Eq.(20), one gets that

$$\overbrace{\rho[\int_0^{\theta_{-i}^\rho} r(\theta_i^\rho - f(\theta_i^\rho) - g(\theta_i^\rho, \theta_{-i})) d\theta_{-i}]}^{V} + \tag{22}$$

$$\Pr(\theta_{-i} > \theta_{-i}^\rho)(\theta_i^\rho - f(\theta_i^\rho)) + \int_{-n}^{\theta_{-i}^\rho} r g(\theta_i^\rho, \theta_{-i}) d\theta_{-i}] = 0, \tag{23}$$

Note again that the LHS is increasing in θ_i^ρ. In addition, if investments are substitutes, Term V is negative. Holding $(\theta_i^\rho, \theta_{-i}^\rho)$ fixed, the LHS of Eq.(22) is decreasing in ρ. Hence, for a fixed θ_{-i}^ρ, an increase in ρ must be compensated by an increase in θ_i^ρ; an increase in ρ shifts the decreasing best-response function up. Hence, the symmetric equilibrium must increase in ρ.

If investments are complements, Term V is positive. Holding $(\theta_i^\rho, \theta_{-i}^\rho)$ fixed, the LHS of Eq.(22) is increasing in ρ. Furthermore, since $\theta_{-i}^+(\theta_i) = 0$ for any $\theta_i > 0$, $\beta^\rho(0)$ is independent of ρ. An increase in ρ, thus, shifts $\beta^\rho(\theta_{-i})$ down for all $\theta_{-i} > 0$. 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 exists, is increasing in ρ since $\beta^\rho(\theta_{-i})$ is continuous and monotone increasing in θ_{-i}.

6. Suppose that $\theta_i > 0$. Since $g(\theta_i, \theta_{-i}) < 0$ if $\min\{\theta_i, \theta_{-i}\} < 0$, and $g(0, \theta_{-i}) = 0$, it must be the case that $\theta^\rho_{-i} > \theta^+_{-i}(\theta_i) = 0$. Hence, underestimation follows from the proof of Proposition 3. Note that Eq. (16) still holds; hence, false antagonism also follows from the proof of Proposition 3 .

Proof of Proposition 5. If $\rho = 0$, since the benefit of checking is strictly decreasing in c, the receiver adopts a cutoff checking strategy for checking. The indifferent type is $c^0 = p^0/(1+p^0)^2$. Since without checking $y = 1/(1+p^0)$, and the indifference condition is

$$0 = \frac{1}{1+p^0}(\frac{1}{1+p^0} - 1)^2 + \frac{p^0}{1+p^0}(\frac{1}{1+p^0})^2$$

Let $c_{\max} = 1/4$. Since $B > 0$, $p^0(c^0)$ is uniquely determined by c^0 solving $c^0 = \min\{F^{-1}(B), c_{\max}\}$.

Proof of Proposition 6. Suppose that $\rho > 0$. Let $p^+(c)$ be the projected sender's lying probability given receiver type c. This strategy $p^+(c)$ must smoothly increase in c. To see this, consider $c'' > c'$, but suppose that $p^+(c'') < p^+(c')$. Since p^ρ does not depend on c, type c'' now would have a strictly lower incentive to check than type c', but then $p^+(c'') = 1$, a contradiction. Hence, $p^+(c)$ must be increasing. This also implies that checking frequency must be weakly monotone decreasing in c. As a consequence, there cannot be a discontinuous jump in $p^+(c)$ at some $\widehat{c}$. Such a jump would imply the existence of a $\tau > 0$ such that type $\widehat{c} + \tau$ checked strictly more often than a type $\widehat{c} - \tau$, contradicting monotonicity. Hence $p^+(c)$ must smoothly increase on some $[c^\rho_1, c^\rho_3]$ with $c^\rho_1 < c^\rho_3$ and $p^+(c^\rho_1) = 0$ and $p^+(c^\rho_3) = 1$. Each $c \in [c^\rho_1, c^\rho_3]$ must play a mixed checking strategy to ensure that $p^+(c) \in (0,1)$ for $c \in (c^\rho_1, c^\rho_3)$. There then exists $c^\rho_2 \in (c^\rho_1, c^\rho_3]$ such that $p^+(c^\rho_2) = p^\rho$. Hence, if $c \in (c^\rho_1, c^\rho_2)$, $E_\theta[y^*] > \frac{1}{2}$ and if $c > c^\rho_2$, $E_\theta[y^*] \leq \frac{1}{2}$.

Lemma 1 *The cutoff c^ρ_1 is weakly decreasing and the cutoff c^ρ_3 is weakly increasing in ρ.*

Proof of Lemma 1. I proceed by contradiction. Suppose that $\rho' > \rho$, but $c^{\rho'}_3 < c^\rho_3$. Since $p^+(c^\rho_3) = p^+(c^{\rho'}_3) = 1$, it must be that $p^{\rho'} < p^\rho$ since c^ρ_3 is increasing in p^ρ and in ρ separately. This implies that $c^{\rho'}_1 < c^\rho_1$ must also hold since these are also increasing in p^ρ. The sender, however, now has a strictly lower incentive to lie under ρ than under ρ' implying that $p^{\rho'} > p^\rho$ must hold; a contradiction. Hence, c^ρ_3 is weakly increasing in ρ.

Suppose that $\rho' > \rho$, but $c^{\rho'}_1 > c^\rho_1$. Since $p^+(c^\rho_1) = p^+(c^{\rho'}_1) = 0$ and $c^{\rho'}_3 \geq c^\rho_3$ by the previous argument, it must be that $p^{\rho'} \leq p^\rho$, which then implies that $c^{\rho'}_1 \leq c^\rho_1$, a contradiction .

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

$$B = F(c^\rho_1)/(1 - F(c^\rho_3) + F(c^\rho_1)). \tag{24}$$

An increase in B increases the LHS of Eq.(24). Holding ρ constant, since c_3^ρ moves in the same direction as c_1^ρ in p^ρ, an increase in B must increase c_3^ρ. Since $c_3^\rho \leq c_{\max}$, and $F(c_{\max}) < 1$, if B is sufficiently high, the equality can no longer hold; instead, $c_2^\rho = c_3^\rho$ binds and $p^\rho = 1$. This establishes the existence of $\overline{B}(\rho, F)$. Given Lemma 1, $\overline{B}(\rho, F)$ must decrease in ρ, because c_3^ρ increases in ρ.

To show the existence of a $\overline{F}(\rho, B)$, rewrite the sender's interior incentive condition as

$$B = F(c_3^\rho)B + F(c_1^\rho)(1 - B). \tag{25}$$

Consider now an increase in F in the sense of fosd. Holding c_1^ρ and c_3^ρ constant, the RHS of Eq.(25) decreases; hence, c_3^ρ must increase in F. Since $F(c_{\max}) < 1$, such an increase is always bounded which implies that an $\overline{F}(\rho, B)$ must exist. The second part follows again from the fact that c_3^ρ is weakly increasing in ρ .

Proof of Proposition 8. The $\rho = 0$ case is immediate. Suppose that $\rho > 0$. As long as uniform credulity holds, c_1^ρ and c_3^ρ do not depend on B or F. Consider now an increase in $B > \overline{B}(\rho, F)$. For each $c \in [c_1^\rho, c_3^\rho]$, the receiver's investment, conditional on a positive recommendation and not checking, is $y^{\rho,+}(c) = \frac{1}{1+\overline{p}^\rho(c)}$, where $\overline{p}^\rho(c)$ is given by the solution to $c = \frac{\overline{p}^\rho(c)}{(1+\overline{p}^\rho(c))^2}$, or, equivalently, by

$$\overline{p}^\rho(c) = -\frac{1}{2c}\left(4c - 2c + \sqrt{1-4c} - 1\right) \tag{26}$$

Hence, given uniform credulity, the expected payoff of any type $c \in [c_1^\rho, c_3^\rho]$ is

$$\begin{aligned} E[u^\rho \mid c] &= B(-c) + (1-B)(-\frac{1}{2}(1 - y^{\rho,+}(c))^2 - \frac{1}{2}y^{\rho,+}(c)^2) \\ &= (2B-1)(c_{\max} - c) - c_{\max}, \end{aligned}$$

where the second equality follows when expressing $y^{\rho,+}(c)$ as function of c substituting in Eq. (26). It follows that the expected utility of a type $c \in [c_1^\rho, c_3^\rho]$ is decreasing in B. Finally, the behavior and the payoff of a type $c < c_1^\rho$ is not changing in B, as long as uniform credulity holds, and the same is true for $c > c_3^\rho$. Hence, receiver welfare is increasing in B.

Consider now an increase in F from some level $F > \overline{F}(\rho, B)$. Again the expected utility of types $c \notin [c_1^\rho, c_3^\rho]$ is unaffected. Consider now $c \in [c_1^\rho, c_3^\rho]$. If $B < 0.5$, $E[u^\rho \mid c]$ is strictly decreasing in c on $[c_1^\rho, c_3^\rho]$. Thus, an increase in F, which leaves $F(c_1^\rho) = F((1-\rho)/(2-\rho)^2)$ unaffected, decreases receiver welfare .

Proof of Corollary 4. If $B \geq \overline{B}(\rho, F)$, then $R(\rho, \overline{B}(\rho, F)) > \overline{y}$. Hence, there exists $\overline{\gamma}(\rho, F)$ such that $\overline{\gamma}(\rho, F)[R(\rho, \overline{B}(\rho)) - \overline{y}] > \overline{B}(\rho, F)$.

Proof of Corollary 5. Fix ρ, and consider the set of $\{B, F\}$ such that uniform credulity holds. Here, for any F and B, $c_1^\rho = (1-\rho)/(2-\rho)^2$ must hold. Since $p^+(c)$ is increasing in

c, $y^{\rho,+}(c)$ is maximal for c_1^ρ. Since the probability of checking is 1 for all $c < c_1^\rho$ is constant on $[c_1^\rho, c_3^\rho]$ and is zero for all c_3^ρ, but $y^{\rho,+}(c) = \frac{1}{2}$ for all $c > c_3^\rho$, it follows that the seller's revenue is highest for type c_1^ρ. Furthermore, the probability of checking is increasing in B, the revenue generated by c_1^ρ is highest when B is the smallest. Finally, for uniform credulity to hold, it must be that

$$(1 - F(c_3^\rho))B(\rho, F) - F(c_1^\rho)(1 - B(\rho, F)) > 0.$$

For any given $\varepsilon > 0$, if $B = \varepsilon$ and $F(c_1^\rho) < \frac{\varepsilon^2}{1-\varepsilon}$ and $F(c_3^\rho) = 1 - \varepsilon$, the above inequality is satisfied. Furthermore, holding $B, F(c_1^\rho)$ and $F(c_3^\rho)$ constant, revenue is increasing in $F(c_1^\rho + \varepsilon)$. Since $E[y^{\rho,*}(c)]$ is maximal for c_1^ρ, and is decreasing in B, revenue is increasing in ε^{-1} as long as $\varepsilon > 0$.

Proof of Proposition 9. Both the real and the projected buyer accept any price on the equilibrium path; they also both reject any price greater than $b + x$. The projected seller names a price of $\overline{q} + x$. To show that this is a a ρ projection equilibrium, consider first the real seller.

If $q < \overline{p} - x$, deviating to any price $p < \overline{p}$, leads to a perceived loss. This is true because the payoff from deviating to such a price is bounded by

$$(1 - \rho)\overline{p} + \rho q \leq q + x,$$

as long as $\overline{p} - q \leq x/(1 - \rho)$, which holds for all q. Deviating to some price $p > \overline{p}$ generates an expected payoff of

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

which is lower than $q + x$ because $pe^{-(p-\overline{p})/x} + qe^{-(p-\overline{p})/x} \leq q + x < q + x/(1-\rho)$. If $q > \overline{p} - x$, then naming a price of $p = q + x$ maximizes $pe^{-(p-\overline{p})/x} + q(1 - e^{-(p-\overline{p})/x})$.

Consider the projected seller. Deviating to a price above $\overline{q} + x$ leads to a loss since $\overline{p} \leq \overline{q} + x$. Deviating to a price below $\overline{p}$ leads to a loss if $\overline{p} = \overline{q} + x$. If $\overline{p} = x/(1 - \rho)$, then, $p = \overline{q} + x$ is optimal, since $pe^{-(p-\overline{p})/x} + \overline{q}(1 - e^{-(p-\overline{p})/x})$ is always maximized by $p = \overline{q} + x$.

References

[1] Allen, Jon, Peter Fonagy, and Anthony Baterman. (2008). *Mentalizing in Clinical Practice.* American Psychiatric Publishing, Washington DC.

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

[3] Algan, Yann, and Pierre Cahuc. (2010). "Inherited Trust and Growth." *American Economic Review,* 100(5): 2060–92.

[4] Arrow, Kenneth. (1972). "Gifts and Exchanges." *Philosophy and Public Affairs*, 1: 343–362.

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

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

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

[8] 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.

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

[10] 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.

[11] Danz, David, Kristóf Madarász, and Stephanie Wang. (2014). "The Biases of Others: Anticipating Informational Projection in an Agency Setting." mimeo LSE and U of Pittsburgh.

[12] Dawes, Robyn and Matthew Mulford. (1996). "The False Consensus Effect and Overconfidence: Flaws in Judgment or Flaws in How We Study Judgment?" *Organizational Behavior and Human Decision Processes,* 65(3)*:* 201–211.

[13] DellaVigna, Stefano and Ethan Kaplan. (2007) "The Fox News Effect: Media Bias and Voting." *Quarterly Journal of Economics*, 122: 1187-1234.

[14] Elster, Jon. (2007). *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences* Cambridge University Press.

[15] 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

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

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

[18] 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.

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

[20] Kagel, John. (1995). "Auctions: a Survey of Experimental Research." in *Handbook of Experimental Economics* ed. J. Kagel and S. Roth, Princeton University Press.

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

[22] Katz, Daniel and Floyd Allport. (1931). *Student Attitudes.* Syracuse, N.Y.: The Craftsman Press.

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

[24] Gilovich, Thomas, Victoria Medvec, 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.

[25] Gilovich, Thomas, Victoria Medvec, Kenneth Savitsky. (2000). "The Spotlight Effect in Social Judgment: An Egocentric Bias in Estimates of the Salience of One's Own Actions and Appearance." *Journal of Personality and Social Psychology*, 78(2): 211–22.

[26] Hirschman, Albert. (1970). *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States.* Cambridge, MA: Harvard University Press.

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

[28] Indherst, Roman, and Marco Ottaviani. (2012). "How Not to Pay for Financial Advice." *Journal of Financial Economics,* 105(2): 393-411.

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

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

[31] La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny. (1997). "Trust in Large Organizations." *American Economic Review*, 87: 333–38.

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

[33] Madarász, Kristóf. (2014a). "Projection Equilibrium: Definition and Applications to Social Investment and Persuasion." *Working Paper, LSE.*

[34] Madarász, Kristóf. (2014b). "Bargaining under the Illusion of Transparency." *CEPR* Discussion Paper.

[35] Malmendier, Ulrike, and Devin Shanthikumar. (2007). "Are Small Investors Naive about Incentives?" *Journal of Financial Economics*, 85(2): 457–89.

[36] Mullainathan, Sendhil, Noeth Markus, and Antoinette Schoar. (2012). "The Market for Financial Advice: An Audit Study." *NBER Working Paper*

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

[38] Myerson, Roger. (1981). "Optimal Auction Design." *Mathematics of Operations Research,* 6, 58–73.

[39] Myerson, Roger and Mark Satterthwaite (1983). "Efficient Mechanisms for Bilateral Trading." *Journal of Economic Theory,* 29 (2): 265–281.

[40] Newton, Elizabeth. (1990). "Overconfidence in the communication of intent: Heard and unheard melodies." Unpublished doctoral dissertation, Stanford University.

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

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

[43] 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.

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

[45] Pronin, Emily, Carolyn Puccio, and Lee Ross. (2002). " Understanding Misunderstanding: Social Psychological Perspectives." in *Heuristics and Biases* eds. Thomas Gilovich, Dale Griffin, and Daniel Kahneman, CUP, Cambridge.

[46] Rayo, Luis and Ilya Segal (2010) "Optimal Disclsoure Policy." *Journal of Political Economy*, 118(5): 949-987.

[47] Riley, John. (1989). "Expected Revenue from Open and Sealed Bid Auctions." *Journal of Economic Perspectives*, 3(3): 41-55.

[48] Samuleson, William F. (1984). "Bargaining under Asymmetric Information." *Econometrica*, 995-1006.

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

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

[51] Williamson, Oliver. (1979). "Transaction-cost Economics: the Governance of Contractual Relations." *Journal of Law and Economics*, 22(2): 233–261.

[52] 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.