Title
Inferential Errors in Social Learning and Markets

Author
Gagnon-Bartsch, Tristan Michael

Publication Date
2014

Inferential Errors in Social Learning and Markets

by

Tristan Michael Gagnon-Bartsch

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Matthew Rabin, Chair
Professor Shachar Kariv
Professor Edward Augenblick

Spring 2014

Inferential Errors in Social Learning and Markets

Abstract

Inferential Errors in Social Learning and Markets

by

Tristan Michael Gagnon-Bartsch

Doctor of Philosophy in Economics

University of California, Berkeley

Professor Matthew Rabin, Chair

This dissertation explores economic implications of misinferring from others' behavior. The first two chapters study misinference in models of social learning. They explore in turn two distinct inferential errors: (1) taste projection—the tendency for people to overestimate how similar others' tastes are to their own, and (2) redundancy neglect—people fail to realize that those acting before them also infer from the behavior of predecessors. The final chapter draws out the implications of taste projection in auctions.

More specifically, within social-learning environments, Chapter 1 explores the implications of "taste projection": agents overestimate how common is their own taste. Agents with heterogeneous tastes learn about unknown (taste-dependent) payoffs of actions from the privately-informed choices of predecessors. For instance, investors with varied risk preferences learn which prospect minimizes risk and which maximizes expected return. Inference requires agents to assess if surprising action frequencies are likely provoked by uncommon tastes or contrary private information. Taste projectors miscalculate these odds. In settings where rational agents correctly learn their optimal choice, projection stops some or all from ever learning the right choice. Long-run beliefs and behavior are determined by a player's taste, the degree of all players' biases, and the nature of uncertainty. First, when each thinks her taste is most common, society comes to believe a single action is best for all, irrespective of whether this is true. Second, when the bias is weaker, social beliefs and behavior perpetually cycle—history never provides a clear message about the optimal choice. Third, when quality is highly uncertain, popularity due to taste is systematically over-attributed to quality. Finally, this form of biased learning can exacerbate and perpetuate a false-consensus effect: if people neglect differences in perceptions when learning about the distribution of tastes from others' choices, then a small initial bias eventually leads all types to think their own taste is most common. For contrast, I also characterize rational learning among Bayesian agents with taste-dependent beliefs over the distribution of preferences.

Across a range of social-learning settings, Chapter 2 follows Eyster and Rabin (2010) in studying the implications of agents who neglect the redundancy in information when

learning from others.[1] Players naively think each predecessor's action reflects solely that person's private information. We explore new implications that arise in environments richer than ER's canonical binary-state setting. Whereas in both classical learning models and ER society will with positive probability come to believe the true state, we characterize a set of states that agents will always come to disbelieve even when true. Typically when the truth lies in this set, society will "unlearn": an early generation learns the truth, but society's beliefs move away and converge to wrong beliefs. Society only remains confident in those hypotheses such that the behavior observed when people are fully confident in the hypothesis most closely resembles the behavior we'd see by privately informed agents if that hypothesis were true. We provide specific implications of these principles. First, in cases where options such as restaurants or stocks have independent quality, people form polarized beliefs—they come to believe that the best option is the best it could be and that all lesser options are the worst they could be. Second, in an investment setting, polarized perceptions lead investors to allocate all their wealth to a single prospect, generating a welfare loss through under-diversification. Third, agents generally overestimate the extent of private information in the economy.

Chapter 3 explores how taste projection affects bidding in auctions.[2] We consider auctions for a good with both private- and common-value elements. We model projection by assuming bidders with higher private values perceive a distribution of valuations that first-order stochastically dominates the perception of those with lesser private values. Those with low private values perceive a distribution shifted to the left whereas those with high private values perceive one shifted to the right. We draw out the implications of this assumption in first- and second-price sealed-bid auctions and English auctions. When the good has only private value, projection leads players to misperceive the extent of competition. This induces overbidding, on average, in first-price auctions, but has no effect in second-price or English auctions. If the good also has some common-value component, players draw inference about others' signals from their equilibrium bids. No matter the auction format, projection leads to distorted inference that reduces efficiency. The probability the player with the highest value receives the good is decreasing in the extent of projection.

[1] This chapter is co-authored with Matthew Rabin.

[2] This chapter will eventually be incorporated into a larger project joint with Marco Pagnozzi and Antonio Rosato. In light of this, I write in first person plural.

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation benefited greatly from helpful discussions with many professors, colleagues, and friends.[3] I'd like to express my appreciation and gratitude. First and foremost, I thank Matthew Rabin for invaluable advice and indefatigable support and encouragement. He made writing this thing fun. He continues to be extremely generous with the opportunities and time he provides. I thank Shachar Kariv for his sage advice on navigating through graduate school and academia. I've had the privilege to learn from him ever since I was an undergrad here at Berkeley. Speaking of undergrad, I must thank Bryan Graham for his support during those years. It's safe to say I would've never won an NSF fellowship if not for him. For helpful discussions and suggestions on the content of this dissertation, I thank David Ahn, Ned Augenblick, Nick Barberis, Stefano DellaVigna, Erik Eyster, David Hirshleifer, Shachar Kariv, Marc Kaufmann, Jussi Keppo, Brian Knight, Botond Kőszegi, Kristof Madarasz, Takeshi Murooka, Omar Nayeem, Ted O'Donoghue, Marco Pagnozzi, Antonio Rosato, Matthew Rabin, Josh Schwartzstein, Adam Szeidl, Xiaoyu Xia, and various seminar audiences at UC Berkeley. Chapter 2 is coauthored with Matthew Rabin. Chapter 3 has led to a larger project in progress with Marco Pagnozzi and Antonio Rosato. I thank the National Science Foundation Graduate Research Fellowship for generous financial support.

Of course, I also benefit daily from these same friends, family, colleagues, and professors in domains tangential to any specific results in this dissertation. It's perhaps tradition to acknowledge them here. While I can't express my gratitude enough, they deserve way better than a mere "thank you" written here. This venue—likely hidden deep in the university's archives—isn't worthy for such appreciation. Instead, I hope to make it known to all of them in much more regular and salient ways. That said, thanks! Especially to Cat, my parents, Johann, and my grandmothers—I'm sure they would be proud.

[3] These sets are not disjoint.

Chapter 1

Taste Projection in a Model of Social Learning

1.1 Introduction

How well do people predict others' tastes? Do we accurately predict, say, the population share that favors a Republican presidential candidate or the safety of bonds over stocks? Evidence from many domains—economics, public policy, and social psychology—suggests that our own tastes influence such predictions. While this inference is rational when knowledge of others' preferences is limited, the literature argues in favor of a systematic non-Bayesian bias: people perceive their own taste as more common than it is. For example, people overestimate how many share their views on income redistribution (Cruces, Perez-Truglia, and Tetaz, 2013), and on the preferred political candidate (Delavande and Manski, 2012). While research on this bias—called the "false-consensus effect" or interpersonal projection—has progressed in showing its existence, few studies demonstrate the extent or scope of its implications. When and how does it matter?

This paper argues that "taste projection" has important consequences in social learning. Of course, inference from others' behavior influences many economic decisions—it guides technology adoption, participation in the stock market, voting behavior, and various forms of consumption—but how we interpret the informational content of others' choices depends crucially on our perception of their preferences.[1] Whether we believe an ambiguous tax reform will help the poor or widen inequality after learning others' positions depends on whether we think most favor redistribution. Or, how we judge the quality of a new film or restaurant based on popularity depends on our perception of peers' tastes for the genre or cuisine. To understand how projection distorts such inference, this paper formally mod-

[1] To give but a few examples, within the domain of technology adoption, Conley and Udry (2010) show that social learning drives investment in new crops in Ghana; in a voting context, Knight and Schiff (2010) show it generates momentum in U.S. primary elections; within consumption domains, Cai et al. (2009), Salganik et al. (2006), and Moretti (2010) respectively show its impact on demand for restaurants, music, and movies.

els the bias and draws out its implications within social-learning environments: building on canonical models of observational learning (Banerjee, 1992; Bikhchandani et al., 1992; Smith and Sørensen, 2000), I examine how biased agents, each of whom overestimates the commonness of her own taste or motive, learn from observing the actions—but not directly the information or tastes—of other biased agents.

To outline the model, suppose investors with varied risk preferences wish to learn whether new asset A is riskier (and has higher expected return) than a known alternative, B. A fraction λ prefers the safer asset whereas $1 - \lambda$ prefers the higher-return alternative. From experience with similar securities, investors have private but imperfect information about the relative risk. To acquire additional information before adequate performance data materializes, investors use others' choices. But heterogeneity in tastes complicates inference. Did a predecessor choose A because she's risk averse with private information that A is safe? Or due to precisely opposite preferences and opposite information? Smith and Sørensen (2000) characterize players' long-run beliefs and behavior in such settings provided that λ—the distribution of preferences—is common knowledge. This paper, in contrast, does so assuming agents project tastes: each overestimates how many seek their same objective. The risk averse think $\hat{\lambda} > \lambda$; the return seekers think $\hat{\lambda} < \lambda$.

More generally, in settings where agents differ in their ideal features of an option, I characterize long-run beliefs about the optimal action for each type of player. Inference requires agents to assess if unexpected action frequencies are likely provoked by uncommon tastes or contrary private information. Since agents have inconsistent theories of what provokes actions, those with different tastes develop divergent beliefs. As a consequence, taste projectors never reach long-run agreement. Hence, importantly, biased agents cannot all mutually learn the truth: even in environments where rational agents necessarily learn, taste projection leads some to choose incorrectly in the long run. Long-run beliefs and behavior are determined by a player's taste, the degree of all players' biases, and the nature of uncertainty. When the bias is sufficiently strong, agents herd on a single action X, and all grow confident (some rightly, some wrongly) that X is optimal for their taste. Taste projection explains how confident, but false, beliefs can persist despite sufficient evidence to learn, and why uniform herds may emerge despite differences in tastes.[2] But when the bias is weaker, the model yields a much different prediction: opinions and behavior forever cycle over time, resembling fads. In addition, if agents also learn about others' tastes from actions, the bias in perceived taste distributions can be intensified: agents conclude more share their taste than anticipated. That is, naive learning can exacerbate and perpetuate a false-consensus bias.

Section 1.2 formalizes the model, which adds taste projection to an observational-learning setting based on Smith and Sørensen (2000). A sequence of agents, N acting per period,

[2] The rational-herding literature shows that when learning from others, society may forever choose suboptimal actions. Importantly, as noted in Eyster and Rabin (2010), in any setting where an incorrect herd may arise, rational agents never grow confident in the state of the world. The more likely is an incorrect herd, the less confident is society in the the long run. As such, the rational-herding literature does not explain how society may often develop *confident* beliefs in some false hypothesis.

choose between two actions, A and B. An action's payoff derives from its commonly-valued quality, q, and a heterogeneously valued attribute, z. Players' tastes for z are distributed along a line; an agent prefers z closest to her "location", θ (e.g., Hotelling, 1929). For instance, q may be the quality of a restaurant or film, while z is the cuisine or genre. Or, among investments, z and q respectively measure risk and transaction costs. Agents learn about (q, z) from private signals and the complete history of predecessors' choices. To crisply identify the effects of projection, I focus on environments where rational agents learn the state.[3]

To model taste projection, I assume agents mispredict the distribution of others' tastes, θ. A θ-type perceives a distribution that first-order stochastically dominates the perception of any player with taste left of θ, but is *dominated* by the perceptions of those right of θ. The more right-leaning is one's taste, the higher is her estimate of those with right-leaning tastes. For example, Anni and Benny, who are respectively risk averse and risk neutral, disagree on the share of investors who are more risk averse than Anni; Anni thinks 75%, Benny thinks just 25%. Additionally, I assume agents are naive about this bias; they neglect that players with different tastes have divergent perceptions of population preferences. An agent best responds to beliefs formed via Bayes rule using her misspecified model, which assumes common knowledge that tastes are distributed according to her perception.[4] Anni wrongly assumes Benny agrees that 75% are risk averse.

Sections 3 and 4 begin with the case where the quality is known: the only uncertainty is over horizontal location. Section 3 first develops preliminaries about individual decision-making as a function of private beliefs and the public history of actions. Following Smith and Sørensen (2000), there are two states of the world—A is to the left or right of B. Players on the same side of the left-right taste spectrum prefer the same action and form identical beliefs from the history; those on opposite sides form divergent beliefs. Importantly, I show that an agent's perceived fraction of those with right-leaning preferences, denoted $\hat{\lambda}$, dictates how she uses new observations. If she underestimates the variance in tastes—say, $\hat{\lambda} = 0.9$ when in truth $\lambda = 0.75$—then she perceives actions as more precise signals of underlying private information; her beliefs overreact relative to rational beliefs. If she overestimates the variance in tastes—say, $\hat{\lambda} = 0.6$—then she perceives actions as less informative, and her beliefs underreact. If she mispredicts the majority taste—say, $\hat{\lambda} = 0.4$—then her beliefs move opposite the rational belief after any observation.

Section 4 studies asymptotic properties of learning. Since agents mispredict how beliefs

[3] I assume private signals have unbounded informativeness. In this environment, it is well understood that bounded informativeness generates information cascades: the information contained in the history of play eventually swamps the information contained in the most informative signal. The setting also precludes "confounded learning", discovered by Smith and Sørensen (2000), where players converge to an uncertain stationary belief. This outcomes arises only when the quality difference is sufficiently large. Appendix 1.B addresses when confounded learning may occur, and how this possibility alters my basic results.

[4] Agents think others' draw inference using a common model of the world, and that any differences in beliefs are purely a result of private information. Indeed, in a panel survey of private investors, Egan, Merkle, and Weber (2012) show that people overestimate how many share their beliefs about returns.

map to action frequencies, convergence to stationary beliefs is not guaranteed, nor is convergence to fully-incorrect beliefs ruled out. I derive conditions on agents' perceptions that determine whether a candidate equilibrium belief is stochastically stable: near equilibrium, is the unexpected frequency of actions interpreted as evidence for or *against* that belief?[5] A belief is stable only if all players observe a greater share choosing their anticipated majority action than expected. From this, I show it's impossible for all taste projectors to converge on identical long-run beliefs: some agents necessarily fail to learn. To understand how and why learning fails, I closely analyze two classes of projection: (1) "strong"—where each type thinks her preference is most common—and "weak"—where all agree on the majority taste.

When each thinks her taste is most common, agents herd on a single action X, and all grow confident (some rightly, some wrongly) that X is optimal for their taste. Beliefs are polarized according to taste: the risk averse grow confident asset A is safe; the return seeking think it's risky.[6] Quite simply, since each thinks her taste is most common, absent strong contrary signals, she wants to follow the herd. This result explains how confident, but false, beliefs can persist despite sufficient evidence to learn, and why herds may emerge despite heterogeneity in tastes.[7] When many act per round ($N \to \infty$), dynamics are deterministic: the minority necessarily learns incorrectly, and *all* choose the option optimal for the majority taste. As an implication, the adoption of new technologies or welfare programs beneficial only to a minority fails when people learn from others' take-up decisions. Instead, society inefficiently over-adopts practices optimal for the majority, implying that observational learning is not only inefficient, but can be socially harmful.[8]

"Uniform" herding is not a general consequence. When taste projectors (correctly) agree on the majority preference, they never settle on a fixed belief, let alone herd. Society's opinion of the optimal action perpetually cycles, offering an explanation for "fads" in settings where rational behavior must converge.[9] When society is confident that A is safe, 75% (the share of risk averse agents) choose A. But risk-averse investors expect a higher frequency, say 90%.

[5] Gagnon-Bartsch and Rabin (2013) study a similar issue of stability in a model of biased social learning where players neglect the redundancy in behavior.

[6] Agents' beliefs display a strong form of polarization, where they grow fully confident in alternative hypotheses. Other researchers studying how disagreement may persist in learning settings, like Andreoni and Mylovanov (2012), demonstrate a much weaker form of polarization where agents with common preferences disagree (but not confidently) on the optimal action. Rational models fail to explain *confident* disagreement.

[7] Sorensen (2006), who studies observational learning in the selection of health-care plans, demonstrates that observing others leads to uniformity in choice, despite heterogeneity in individuals' optimal plans. After the study, many switch away from the "herd" plan. Similarly, some medical practices are widely believed to be beneficial to all, when in fact efficacy depends on heterogeneous characteristics of patients.

[8] Although rational observational learning can cause incorrect herds, (e.g., Bikchandani, et al., 1992) it is necessarily welfare improving, on average, relative to following private information. Eyster and Rabin (2010, 2013) show how a different form of naive learning can be socially harmful.

[9] Rational behavior fails to converge to a single action when players observe only immediate predecessors (Çelen and Kariv, 2004). Acemoglu, Como, Fagnani and Ozdaglar (2012) show that opinions may persistently fluctuate when learning in a network if some agents are "stubborn" and never update their beliefs. Such models help explain, for instance, persistent fluctuations in political opinion (e.g., Kramer, 1971; and Cohen, 2003).

Their best explanation for such low investment is that other risk-averse agents have strong signals that B is safer. As the risk averse lose faith that A is safe, the share investing in A falls so that even the risk seekers, who expect only 60% to choose A, start doubting A is safe. Beliefs reverse: investors think B is safe. But since this logic will repeat itself, beliefs perpetually oscillate.[10] When the minority's perceptions are sufficiently more biased than those in the majority, a long spell where all believe A is safer is followed by a longer spell where all think B is safer, and so on. Beliefs spend roughly equal time favoring each state, and all players are worse off, on average, by observing others than if they simply followed their private information.

Sections 1.5 and 1.6 expand the environment with additional dimensions of uncertainty, and explore how projection leads to mislearning about quality differences and population tastes, respectively. Intuitively, agents use these additional dimensions to best explain otherwise anomalous herds. Section 5 allows for highly-uncertain quality, so that all players may prefer the same action. For instance, a restaurant with astounding quality is preferred by all diners, despite taste in cuisine. With two types, no matter the true quality difference, society necessarily concludes it's large enough so that all prefer the same choice. Projection leads agents to herd on X, which they subsequently explain by assuming X has superior quality. In essence, taste-based popularity is systemically over-attributed to quality. This systematic misconstrual of "vertical" and "horizontal" components of preference may help explain the notoriously slow adoption of new agricultural technologies in environments where their productivities vary across farms: low take-up up is mistaken for *global* ineffectiveness rather than *selective* efficacy.[11] Additionally, even when horizontal locations of A and B are known—for instance, film goers know A is an action film, and B a romance—agents still systematically mislearn quality. Fans of the romance attribute moderate popularity to limited quality rather than admitting few enjoy such films, while action fans adopt a higher perception of B's quality in order to explain higher-than-expected attendance. Essentially, those with the most positive view of B's quality are those who prefer the attributes of A.

Section 1.6 more realistically assumes uncertainty over the distribution of tastes, so agents revise their models of others' preferences as they observe actions. Does learning about tastes ameliorate errors in learning about payoffs? With uncertainty, an agent uses her taste as information, so those with different tastes start with different priors. Still assuming agents are naive—they neglect heterogeneity in priors—I show that learning needn't correct mislearning about payoffs and can intensify the bias in perceived taste distributions: all agents confidently conclude that most share their taste. As such, naivete still generates herds. When A is chosen most often, risk averse infer that A is likely safe and $\lambda > \frac{1}{2}$, but to a risk-neutral agent, this indicates that A is likely risky and $\lambda < \frac{1}{2}$. Absent strong

[10] While tempting to think that non-convergence results from the coarseness of the state space—each option takes one of two locations, implying agents expect to observe one of only two long-run frequencies—one can generate examples with non-convergence in a continuous state space.

[11] Munshi (2003) shows that the adoption rates of hybrid "high-yield" crops in India greatly depend on how variable is output with respect to inputs. Strands of hybrid rice with productivity sensitive to the mix of inputs on the farm have very slow adoption rates.

contrary information, each type best responds with A. When a herd arises, an agent's best explanation is that *all* investors share her taste.[12]

Throughout the paper, I contrast naive-learning results with those following from *rational* taste-dependent perceptions that arise from uncertainty over the distribution.[13] Rational beliefs always converge and never grow fully polarized. However, learning may still fail. With positive probability, agents converge to an interior belief where they cannot discern, say, if a high frequency of A's follows because A is safe and most are risk averse, or because A is risky and most seek high return. Smith and Sørensen (2000) show that under perfect information about the distribution, such beliefs exist only when quality differences are sufficiently large. In contrast, I show that with imperfect information, they *always* exist. This extension provides a simple and natural explanation for persistent disagreement.[14] At a confounding belief, people with different tastes disagree on payoffs: relative to a risk-seeking agent, a risk-averse agent thinks it's more likely that most are risk averse and that A is safe.

I conclude in Section 3.5 by putting both taste projection and social "mislearning" in broader context. I discuss why and how taste projection can distort inference in more general social-learning environments where agents can directly communicate beliefs or payoffs. Mislearning results from taste projectors' incorrect theories about others' payoff functions and how they form beliefs. Diners with "sophisticated" tastes may report mediocre payoffs from a meal that typical diners find remarkable. Typical diners are misled if they underestimate how often they glean advice from "sophisticates". I also discuss settings where agents have biased perceptions of the type distribution distinct from projection, such as a false sense of uniqueness; the tools developed in this paper directly apply. I conclude the paper by highlighting the shortcomings of this model and suggesting avenues for future research.

1.1.1 Relation to Previous Research

This paper contributes to a growing literature incorporating informational biases into economics, and, more specifically, social learning, in order to explain how false or divergent beliefs may persist.[15] Ellison and Fudenberg (1993), who were among the first to study

[12] Exploring false-consensus perceptions in a social network, Flynn and Wiltermuth (2011) find that individuals with higher betweenness centrality—more exposure to others in the network—had higher (and more incorrect) estimates of how common was their taste.

[13] This model, analyzed in the Appendix, is identical to Section 1.6 aside from the assumption of full rationality.

[14] Alternative explanations include uncertainty over the distribution of private information, as explored in Acemoglu, Chernozhukov, and Yildiz (2007 and 2009).

[15] One strand of this literature studies the consequences of probabilistic errors—like over-inferring from small samples (Rabin, 2002; Rabin and Vayanos 2010) or under-appreciating properties of statistical processes, like mean reversion (Barberis, Shleifer, and Vishny, 1998). A distinct strand studies agents that neglect the information content of others' behavior, providing explanations for the winner's curse and excessive trading in asset markets (Eyster and Rabin, 2005; Eyster, Rabin and Vayanos, 2013). In this paper, agents have incorrect beliefs about the distribution of tastes—at the root, a probabilistic error. But since this leads to inaccurate perceptions of others' information, agents additionally misinfer from others' behavior.

biased social learning among agents with heterogeneous tastes, explore the efficiency of "rule-of-thumb" learning in a setting with observable payoffs, where agents with heterogeneous tastes simply choose whichever action performed best of those observed. While I assume fully-Bayesian learning within a misspecified model, their naive learning rule is akin to projection where each player thinks *all* share her taste. Similarly, they show that their rule never leads to exact long-run efficiency, but efficiency improves as tastes become less heterogeneous. Bohren (2010) studies a variant of the canonical Bikhchandani et al. (1992) where only a fraction of players observe the history, and players mispredict this fraction. As here, various degrees of misprediction can lead to both stable incorrect herds and persistent fluctuations in beliefs. The focus, however, is on a commonly-held misprediction, where I emphasize the interaction of misperceptions that differ across types of agents. Further, the inferential error studied by Bohren (2010) has a much different motivation, as it captures players' ignorance of the redundancy in social behavior. This form of redundancy neglect has been studied elsewhere in the literature, namely by DeMarzo, Vayanos and Zwiebel (2003), Eyster and Rabin (2010, 2013) and Gagnon-Bartsch and Rabin (2014), who also show how biased observational learning generates confident, yet false, beliefs. Finally, the basic error I analyze is closely related to information projection, explored in Madarasz (2012). He assumes agents overestimate the likelihood that people have the same private information as themselves and draws out the implications of this error in a variety of principal-agent problems.

From a broader perspective, this paper studies learning among agents with both non-common priors and inconsistent beliefs about others' priors. While a large literature studies the implications of non-common priors, most notably as explanations for speculative trade (e.g., Harrison and Kreps, 1978; and Morris, 1996), warranted caution on modeling non-common priors has been advised. As subjective heterogeneous priors can justify any outcome ex post, Morris (1995) argues that we should allow non-common priors only when we can identify a source for the disagreement and precisely model these differences. This paper proposes a disciplined way of incorporating non-common priors: an agent's own taste systematically dictates her beliefs about others' tastes.[16] Further, the literature on non-common priors typically assumes people have correct beliefs about the distribution of these priors—people simply "agree to disagree." My key departure from this literature is that I instead characterize learning among people who neglect disagreement, and wrongly believe in a commonly-shared interpretation of public information.

[16] Models of overconfidence (e.g., Scheinkman and Xiong, 2003), where individuals disagree on the information content of particular signals, are similar attempts to incorporate non-common priors in a structured fashion.

1.2 Basic Setting and Formalization of Naive Taste Projection

This section describes the basic decision environment (Subsection 1.2.1) and proposes a model of taste projection (Subsection 3.2.2). Subsection 1.2.3 defines a solution concept in the presence of projection, pinning down beliefs about others' perceptions and strategies. Subsection 1.2.4 discusses some immediate implications of these assumptions: (1) players with different tastes draw a distinct inference from any history of play, but (2) each player wrongly thinks all draw the *same* inference.

1.2.1 Setting

Actions and States. There are two options $\{A, B\} =: \mathcal{X}$; each $X \in \mathcal{X}$ has *quality* $q^X \in \mathcal{Q}^X \subset \mathbb{R}$, and *location* $z^X \in \mathcal{Z}^X \subset \mathbb{R}$. As in standard spatial-differentiation models, each players prefers higher "vertical" quality, but preference over "horizontal" location depends on her type.[17] A player's payoff from X is entirely determined by this point in the "characteristic space", $(q^X, z^X) \in \mathbb{R}^2$. Using Downs' (1957) model of political competition as an example, q may measure the competence or integrity of a political candidate, while z indicates how liberal or conservative she is. Or, q is the skill of a chef or a writer, and z is his cuisine or genre. Or, q is the transaction cost of an investment, and z measures its risk: assets to the left are riskier—but have higher expected return—than those to the right. The collection of each options' characteristics, $\omega = ((q^A, q^B), (z^A, z^B))$ comprises the state of the world which agents aim to learn; $\omega \in \Omega$ has common prior $\pi_1 \in \Delta(\Omega)$.

To make clear how taste projection can lead learning astray, I focus on the simplest such environment: there are only two possible location profiles, $(z^A, z^B) \in \{(-1, 1), (1, -1)\}$. That is, A is either to the left of B, $(z^A, z^B) = (-1, 1)$, or to the right of B, $(z^A, z^B) = (1, -1)$.[18] To keep notation simple, I write the decision-relevant information of a state $\omega = ((q^A, q^B), (z^A, z^B))$ using only two dimensions. Let $\zeta \in \{L, R\}$ denote the "location state", where $\zeta = L$ if and only if A is left of B, $(z^A, z^B) = (-1, 1)$. And, as will become clear, an agent's choice depends on the *difference* in quality, $\Delta_q := q^A - q^B$, so, I write $\omega = (\zeta, \Delta_q)$. Let $\mathcal{D} := \{\Delta_1, ..., \Delta_D\}$ be the set of possible quality differences; the state space is $\Omega = \{L, R\} \times \mathcal{D}$.

Preferences. Preference over horizontal location depends on one's *preference type*, or "taste", $\theta \in \Theta \subset \mathbb{R}$. θ denotes an individual's most preferred location. For instance, θ may measure risk aversion, reflecting a person's optimal level of portfolio risk, or political ideology. Preferences are represented by a von Neumann-Morgenstern utility function separable in

[17] Hotelling (1929) is the classic example of a location model, and Downs' (1957) model of political competition extends it to a two-dimensional characteristic space, as I similarly do here.

[18] While admittedly restrictive, this binary-state assumption is common in the literature. Smith and Sørensen (2000), who study rational learning in a similar setting, also focus on two feasible location profiles, noting that additional states come at "significant algebraic cost."

quality and location:

$$u(X, \theta) = q^X - k(z^X - \theta)^2, \tag{1.1}$$

where $k > 0$ is a commonly-known preference parameter.[19] To simplify exposition, suppose Θ is a grid: $\Theta = \{\theta : \theta = \pm j\delta,\ j = 1, ..., J\}$ for some $\delta > 0$ and $J \in \mathbb{N}$. Types are i.i.d. across players with c.d.f. G.[20]

For the moment, I assume beliefs over the distribution of tastes, G, are degenerate. In later sections, I allow uncertainty over G—people may be uncertain whether the majority is risk averse or risk neutral—so an agent's perceived distribution of tastes *rationally* depends on her own type θ. I allow for this generalization to contrast learning in the presence of rational taste-dependent distributional beliefs with learning under biased taste projection, where a player's belief about the taste distribution depends on taste beyond rational Bayesian inference.

Discussing inference is simplified by classifying individuals according to their preferred location $z \in \{-1, 1\}$. Ignoring quality differences, all types $\theta < 0$ strictly prefer a left-located option while $\theta > 0$ strictly prefer a right-located option, leading to the following definition:

Definition 1. *Preference types are dichotomously categorized as follows:*

1. $\theta < 0$ *is referred to as a* left *type.*
2. $\theta > 0$ *is referred to as a* right *type.*

The measure of right types, $\lambda := 1 - G(0)$, is a critical statistic of G for drawing inference from predecessors' actions. Without loss of generality, assume $\lambda > \frac{1}{2}$; right types are the majority.

Assumption 1. (Right types comprise the majority.) $\lambda = 1 - G(0) > \frac{1}{2}$.

Players and Timing. In every period $t \in \mathbb{N}$, a new set of $N \geq 1$ players is drawn according to taste distribution G, and each takes an action $X \in \mathcal{X}$. Players are labeled nt; t is the period in which she acts, and $n \in \{1, 2, ..., N\}$. Let $X_t = (X_{1t}, ..., X_{Nt})$ denote the profile of actions taken in period t. Since all N players in t act independently conditional on the history of play, the number of A's taken in t, denoted by $a_t \in \{0, 1, ..., N\}$, is a sufficient statistic for X_t. Hence, let $h_t = (a_1, ..., a_{t-1})$ denote the history of the game up to time t, where $h_1 = \emptyset$.

Beliefs. Before acting, Player nt observes her preference type θ_{nt}, a private i.i.d. signal $s_{nt} \in \mathcal{S}$ about the state, and the complete ordered history of actions, h_t.[21] Her choice, based on the combination of this information, partially reveals her private signal to followers. For

[19] The assumption that attribute z has value equal to the squared distance from one's location is without loss of generality. Results are identical if $(z^X - \theta)^2$ is replaced by any metric $d(z^X, \theta)$.

[20] As here, I often refer to preference types simply as "types". Below, I endow players with private signals, hence a complete description of a player's type is her signal and preference type. I will be explicit whenever I refer to the complete notion of type as to avoid confusion.

[21] Specific details of the signal structure are provided in the following section.

each $\omega \in \Omega$, let $\pi_t(\omega)$ denote the belief in ω drawn solely from history h_t and the prior; I call this the *public belief* in t.[22] The sequence of public beliefs $\langle \pi_t \rangle$ is this paper's key object of analysis.

Finally, let Γ denote this game form, and let $\Gamma(G)$ denote it explicitly as a function of the taste distribution (keeping all other aspects fixed). The next section makes the purpose of this notation clear: if one misperceives the taste distribution as $\widehat{G} \neq G$, but has an otherwise correct model of the game, her *perceived game* is $\Gamma(\widehat{G})$.

1.2.2 Taste Projection

This subsection reviews the literature motivating my main assumption of taste projection, and provides a simple formulation of this bias, which consists of two key assumptions: (1) an agent's perceived preference distribution depends on her own taste, and (2) she neglects that others' perceptions depend on their tastes.

1.2.2.1 Evidence

The notion that people systematically misptredict others' tastes is supported by several strands of research. A large literature in social psychology studies inter-personal projection—the idea that people's own habits, values, and behavioral responses bias their estimates of how common are such habits, values, and actions in the general population.[23] Early work, including Ross, Greene, and House (1977)—who coin the term "false-consensus effect"—find positive correlation between subjects' own preference responses and their estimates of others' responses. Subjects in Ross, Greene, and House (1977) gave their own (binary) response to a question and predicted the fraction of subjects who answered similarly. (E.g., "Are you politically left of center?"; "Do you prefer basketball over football?"; "Will there be women in the supreme court in the next decade?"; "Do you prefer Italian movies over French?") Out

[22] I use the term "public belief" to conform to existing literature. "Public" in this context does not mean the belief is common across society—taste projection and the solution concept introduced below naturally imply that different taste types draw different inference from h_t. Instead, "public" refers to the source of the belief; that is, publicly observable behavior. The phrase "public belief" originates from a literature on rational learning in which both of the points above are valid: when there is no aggregate uncertainty over tastes, rational players do draw identical inference from h_t in a Bayesian Nash equilibrium.

[23] For example, US citizens display taste-dependent responses when predicting how many support their government's use of torture. During the Bush administration, politicians advocating the use of controversial interrogation methods often alluded to polls indicating the methods had wide public support. Gronke et al. (2010) collected a more comprehensive data set that both falsifies these claims and demonstrates the correlation between one's own opinion on torture and their prediction of others' opinions. Survey participants stated their opinion on how frequently torture should be used—either never, rarely, sometimes or often—and estimated the percent of people who chose each of those options. Each row in the table below shows the predictions of a particular response type. For instance, the first row is comprised of estimates of people who stated "never"—people who chose never, on average, guessed that 5% of people chose "often", etc. The last row shows the true percentages of responses.

of 34 questions, responses to 32 were consistent with taste projection: those who answer "yes" to a question overestimate how many others answer "yes" relative to those who answer "no". Many similar studies followed, documenting this correlation across a wide range of domains, including preferences over political candidates and ideology, perceptions of the income distribution and preferences for redistribution, and risk preferences.[24]

Each of these studies, however, simply document correlation between a subject's own taste and her prediction. Is such correlation necessarily indicative of an error? If there is uncertainty about others' tastes, the answer is no. As first noted by Dawes (1989), with uncertainty, a Bayesian should use her own taste as information, resulting in *rational* type-dependent estimates that appear consistent with a "false-consensus" bias.

Motivated by this critique to demonstrate a systematic error, Krueger and Clement (1994) and others provide evidence that this "bias" remains even when subjects have information about other' preferences. They find that subjects use their own preference information more so than that of anonymous others when making population predictions, inconsistent with Bayesian rationality.[25] In incentivized settings, Engelmann and Strobel (2012) verify that a truly-false-consensus bias remains so long as subjects must exert a small amount of effort to get information on others' choices; when this information is not freely available or made salient, people rely too heavily on their own choice when predicting the choices of others. So long as attending to others' tastes comes at some cost, this result suggests that people can hold incorrect type-dependent beliefs about population preferences even in settings with ample opportunity to observe others—where the "Dawes critique" should have little bearing.[26]

Taste-Dependent Predictions

		Prediction			
		Never	Rarely	Sometimes	Often
Response	Never	31%	25%	39%	5%
	Rarely	7%	51%	35%	7%
	Sometimes	3%	26%	67%	4%
	Often	4%	12%	48%	36%
	True	30%	24%	29%	17%

[24] Marks and Miller (1987) review 45 different studies documenting the false-consensus effect published over the decade following Ross, Greene, and House (1977). Mullen, Atkins, Champion, Edwards, Hardy, Story, and Vanderlok (1985) find robust evidence of this correlation in a meta-study of 115 tests. Evidence of type-dependent misprediction has been found in a variety of domains. For instance, Brown (1982) and Rouhana, O'Dwyer and Vaso (1997) find type-dependent perceptions of political preference. Cruces, et al. (2013) find type-dependent misprediction of the income distribution in Argentina, and demonstrate that this leads to misprediction of population preferences for income redistribution. Faro and Rottenstreich (2006) find correlation between subjects' own risk preference and their perception of others' risk preferences.

[25] Krueger and Clement (1994) deduce that when estimating the percent of subjects that endorse some action or preference, subjects use their own response nearly twice as much as the response of an anonymous other. A rational Bayesian should, of course, use these two responses equally.

[26] Using data from the American Life Panel, Delavande and Manski (2012) show that perceptions of others'

Relatedly, economists have argued *intra*-personal projection bias—exaggerating the degree to which future preferences resemble current preferences—influences behavior.[27] To the extent that preferences of contemporaneous others are similarly difficult to predict, we should expect the logic of intrapersonal projection bias to suggest *inter*personal-projection. An intuition for intrapersonal projection is that we "mentally trade places" with our future selves, and in doing so, project our current preference states. But this exact logic applies when empathizing with another. Indeed, Van Boven and Loewenstein (2003) show that the same transient preference states shown to warp subjects' perceptions of own future preferences also distort predictions of *others'* preferences. Subjects' predictions of whether thirst or hunger would be more bothersome to hypothetical hikers lost without food or water were biased in the direction of subjects' own exercise-induced thirst. More economically relevant, Van Boven, Dunning and Loewenstein (2000, 2003) show that sellers who experience an endowment effect project their high valuation of a good onto the valuations of potential buyers, causing sellers to set inefficiently high prices.

1.2.2.2 Perceived Distributions: Biased First-Order Beliefs Over Tastes

I model taste projection by assuming an individual's preference type θ influences her perceived distribution of types. In truth $\theta \sim G(\cdot)$. Denote a θ type's perception of $G(\cdot)$ by $\widehat{G}(\cdot \mid \theta)$. Consistent with the false-consensus effect, I assume right-leaning types think right types are relatively more common, while left-leaning types think the opposite. I capture this intuition with the following assumption:

Assumption 2. (Stochastically Dominating Perceptions.) *$\widehat{G}(\theta \mid \theta')$ weakly first-order stochastically dominates $\widehat{G}(\theta \mid \theta'')$ if and only if $\theta' > \theta''$. That is, whenever $\theta' > \theta''$, $\widehat{G}(\theta \mid \theta') \leq \widehat{G}(\theta \mid \theta'')$ for all $\theta \in \Theta$.*[28]

The more right-leaning is an agent's taste, the higher is her estimate of those with right-leaning tastes. People with conservative political views overestimate the share of people who prefer the conservative candidate. Or, those with high risk aversion overestimate the share seeking safe investment strategies. The perceived measure of right types, which is a key

candidate preferences in the 2008 U.S. presidential election and 2010 congressional election were consistent with the false-consensus effect even after the release of poll results. While this finding may indicate additional statistical biases (e.g., failure to appreciate the Law of Large Numbers—see Benjamin, Raymond, and Rabin, 2013), it shows that taste-dependent perceptions can persist despite opportunity to learn about others' tastes.

[27] For empirical studies see Busse, Pope, Pope, and Silva-Risso (2012), Simonsohn, (2010), and Conlin, O'Donoghue, and Vogelsang (2007). For example, Busse, et al. shows that projection bias affects demand and prices in large, high-stakes markets for cars and houses. Loewenstein, O'Donoghue, and Rabin (2003) provide a general overview of the evidence and draw out implications of a formal theoretical model.

[28] I assume *weak* domination to allow different θ's to hold identical perceptions. If $\theta' > \theta''$ then the two types need not have different perceptions of the distribution of tastes; however, if their perceptions do differ, it must be the case that $\widehat{G}(\theta \mid \theta')$ *strictly* first-order stochastically dominates $\widehat{G}(\theta \mid \theta'')$: $\widehat{G}(\theta \mid \theta') \leq \widehat{G}(\theta \mid \theta'')$ for all $\theta \in \Theta$ with strict inequality for some θ. Let $\succsim$ and $\succ$ denote weak and strict first-order stochastic dominance, respectively.

statistic of $\widehat{G}$ in drawing inference from actions, is denoted by $\hat{\lambda}(\theta)$; importantly, dominance implies $\hat{\lambda}(\theta)$ is weakly increasing in θ. Note that I model the bias in perception relative to *others'* perceptions.[29]

I also assume that perceptions have full support. This ensures that no player ever observes "off-equilibrium-path" actions arising from preference types she assumed did not exist.

Assumption 3. (Identical Supports.) *For all* $\theta \in \Theta$, $\text{supp}\big(\widehat{G}(\cdot \mid \theta)\big) = \text{supp}(G)$.

All individuals agree on the possible tastes others' may have, but may disagree on the likelihood of these tastes.

Most basic results follow from Assumptions 12 and 3, but for sake of clear intuition, I often focus extensively on a particular form of error I call *choice-dependent* projection: one's perceived distribution depends only on her preferred location, not on the intensity of this preference. All left types share a common misperception of G—call it $\widehat{G}_l$—as do right types—$\widehat{G}_r$.[30] Formally:

Definition 2. *Players suffer* choice-dependent projection *if they have the following misperceptions of G:*

$$\widehat{G}(\theta \mid \theta') = \begin{cases} \widehat{G}_l(\theta) & \textit{if} \quad \theta' < 0 \\ \\ \widehat{G}_r(\theta) & \textit{if} \quad \theta' > 0 \end{cases} \tag{1.2}$$

where $\widehat{G}_r \succ G \succ \widehat{G}_l$.

Under choice-dependent projection, there are essentially just two types—left and right. Left types think the measure of right types is $\hat{\lambda}^l := 1 - \widehat{G}_r(0)$, while right types perceive it as $\hat{\lambda}^r := 1 - \widehat{G}_l(0)$. Stochastically-dominating perceptions (i.e., $\widehat{G}_r \succ G \succ \widehat{G}_l$) implies $\hat{\lambda}^l < \lambda < \hat{\lambda}^r$: left types underestimate the measure of right types, but right types *over*estimate it.[31]

[29] Assumptions 12 and 3 characterize how perceptions compare *across* types. How do perceptions relate to the truth? Section 1.A in the appendix gives a simple paramaterization of perceived distributions where each depends only on the true distribution and a single bias parameter, $\beta \in [0, 1]$. That model, which specifies perceptions relative to the truth, satisfies Assumption 12: perceptions relative to others' exhibit dominance.

[30] The term "choice dependent" follows from the fact that under this class of misperceptions, when the characteristics of the options are known, people overestimate the share of others that would choose the same option as themselves. However, they don't necessarily overestimate the share of people with their identical taste parameter, θ. Hence we can think of one's preferred choice or behavior as the object of projection rather than the underlying intensity of that choice.

[31] This model of taste projection makes assumptions directly on perceived distributions of tastes, and maintains that players understand how taste θ translates into decision utility. Alternatively, following Loewenstein, O'Donoghue and Rabin's (2003) model of intrapersonal projection, we could assume a player with taste θ mispredicts the utility of a player with taste $\tilde{\theta}$, which ultimately leads to a misperception of the measure of players that prefer different actions. This footnote briefly demonstrates the equivalence of these two approaches. Suppose that it is known that A is on the right and B on the left. A taste-type θ whose own location-dependent utility from consuming A is $u(A, \theta) = -k(1-\theta)^2$ mispredicts a $\tilde{\theta}$-type's location-dependent utility from A as $\hat{u}^{\theta}(A, \tilde{\theta}) = -\alpha k(1-\theta)^2 - (1-\alpha)kd(1-\tilde{\theta})^2$ where $\alpha \in [0, 1]$ parameterizes the

Within the regime of choice-dependent projection ($\hat{\lambda}^l < \lambda < \hat{\lambda}^r$), two classes of errors, which I now differentiate and define, lead to very different learning properties.

Definition 3. *Suppose players suffer choice-dependent projection where left and right types believe the measure of right types is $\hat{\lambda}^l$ and $\hat{\lambda}^r$, respectively.*

1. $(\hat{\lambda}^l, \hat{\lambda}^r)$ *satisfy* strong taste projection *if* $\hat{\lambda}^l < \frac{1}{2} < \lambda < \hat{\lambda}^r$.
2. $(\hat{\lambda}^l, \hat{\lambda}^r)$ *satisfy* weak taste projection *if* $\frac{1}{2} < \hat{\lambda}^l < \lambda < \hat{\lambda}^r$.

Strong and weak taste projection differ in whether people agree on the majority preference. Strong projection implies that types disagree on the majority preference: each type thinks her own taste is most common. A weak false consensus implies that all players correctly acknowledge that right types comprise the majority.

1.2.2.3 Naivete: Biased Second-Order Beliefs Over Tastes

I assume that a taste projector is *naive* about her bias: she neglects that those with different tastes have alternative perceptions of the distribution. She thinks *all* agents share a common perception.

Assumption 4. (Naivete.) *For any $\theta' \in \Theta$, a θ-type believes $\widehat{G}(\cdot \mid \theta') = \widehat{G}(\cdot \mid \theta)$.*

Naivete pins down second-order beliefs—beliefs about others' perceived distributions—and implies they are incorrect. An individual with taste θ perceives the game as $\Gamma(\widehat{G}(\cdot \mid \theta))$; she assumes the structure of Γ—including $\widehat{G}(\cdot \mid \theta)$—is common knowledge to all players. For instance, a risk-averse agent who thinks 90% of investors are risk averse naively assumes that a risk-neutral agents thinks the same. In essence, agents imagine they are playing a game with common priors, when in fact priors are heterogeneous. Naivete is the key assumption that differentiates "taste projection" from a model with rational taste-dependent distributional beliefs. Rational agents know precisely the map between an agent's type and her belief about the distribution.

extent of the bias: θ's perception of $\tilde{\theta}$'s utility is a linear combination of $\tilde{\theta}$'s true utility and θ's own utility—θ projects her own valuation onto $\tilde{\theta}$'s. It follows that a θ-type's perception of the measure of individuals who prefer A to B in terms of location is the measure of $\tilde{\theta}$ such that

$$\tilde{\theta} > -\frac{\alpha}{1-\alpha}\theta. \tag{1.3}$$

The true measure is that of the set of types that satisfy Equation 1.3 when the right-hand side of is set to zero. But when $\theta > 0$—the agent has right-leaning preferences—the right-hand side of Equation 1.3 is negative. She thus overestimates the share of players that prefer the right-positioned option. Similarly, when $\theta < 0$, the left-type θ *under*estimates the fraction that prefer right-located options. Hence, projection of utility leads to the same qualitative result that people overestimate the share of payers that prefer their desired action that I directly assume here.

Further, naivete departs from much of the literature on non-common priors, which assumes individuals have rational expectations about the *distribution* of heterogeneous beliefs across players.[32] Here, however, players assume the distribution of beliefs (about G) is degenerate on their own perception. As such, within the particular application of observational learning, this paper provides a first step in analyzing the implications of neglecting heterogeneity in beliefs.[33]

1.2.3 Naive Quasi-Bayesian Best Response

This subsection completes the model by specifying a solution concept, called *Naive Quasi-Bayesian Best Response*, formalizing how agents draw inference from past play. It incorporates naivete into the standard notion of Bayesian Nash equilibrium.

Aside from naive taste projection—which implies incorrect first- and second-order beliefs about G—each player nt upholds basic epistemic conditions governing play in a Bayesian Nash equilibrium of her perceived game, $\Gamma(\widehat{G}(\cdot \mid \theta_{nt}))$. First, each player is "quasi-Bayes rational" in that she maximizes expected payoffs given beliefs, which are formed by engaging in putatively correct Bayesian updating using her (false) model of the world.[34] Second, each assumes common knowledge of Bayes rationality within her perceived game. Thus, players correctly predict others' strategies—the map $\sigma : \Theta \times \mathcal{S} \rightarrow \mathcal{X}$ from one's preference θ and private signal to an action.[35] But since they fail to account for others' discrepant models, they systematically mispredict other types' *beliefs*.

To summarize, each player nt best responds given beliefs she draws from Bayesian inference—using her misspecified model $\Gamma(\widehat{G}(\cdot \mid \theta_{nt}))$—from the history of play h_t assuming: (correctly) that all predecessors follow the strategy σ, and (incorrectly) that all predecessors draw the same inference as herself from any all histories h_τ, $\tau < t$.[36] The model of non-

[32] For instance, see Harrison and Kreps (1979) or Morris (1996).

[33] Little work has been done in this area, however there are many domains where this form of neglect seems plausible and worthy of further exploration. Nisbett and Ross (1980), when discussing how people fail to allow for uncertainties in others' perceptions, make the following point emphasizing the need to address naivete: "The real source of difficulty does not lie in the fact that human beings subjectively define the situations they face, nor even in the fact that they do so in variable and unpredictable ways. Rather, the problem lies in their failure to recognize and make adequate inferential allowance for this variability and unpredictability."

[34] This modeling technique—assuming people are "quasi-Bayesian"—is often used in a growing literature in economics studying the implications of systematic biases on inference. While pioneered by Barberis, Shleifer and Vishny (1998) to study biased inference in asset markets, it has since been adopted, to name a few, by Rabin (2002), Rabin and Vayanos (2010) to study inference by believers in the "law of small numbers", Madarasz (2012) to study information projection, and Benjamin, Rabin and Raymond (2012) to study inference by non-believers in the law of large numbers.

[35] Here, the strategy σ is in fact the rational Bayesian-equilibrium strategy.

[36] Note that the social-learning game studied in this paper, which is dominance solvable, requires only a weak solution concept of best response rather than equilibrium. For this reason, I make no additional assumptions relating to the equilibrium condition of consistent beliefs about strategies—whether players believe others hold correct beliefs about their strategy.

rational play simply comprises a particular theory of how players form the incorrect beliefs against which they optimize.

1.2.4 Basic Implications of Taste Projection and Discussion

A central implication of taste projection is that players who differ in their perceptions of G draw different inference from the same history of play. Let π_t^θ be the inference drawn by a θ-type from h_t. It's clear that for any t, $\pi_t^\theta = \pi_t^{\theta'}$ if and only if $\widehat{G}(\cdot \mid \theta) = \widehat{G}(\cdot \mid \theta')$. π_t^θ is θ's perception of the public belief at time t, but contrary to its name it's not commonly shared by all players observing h_t. However, naivete implies each agent projects her mode of inference, and hence *thinks* her "public" belief is commonly shared. Simply put, agents unknowingly draw distinct beliefs from behavior.

There are two errors a naive projector commits when arriving at π_t^θ: she commits the "individual" error of updating using the wrong type distribution, and the "social" error of neglecting how others draw inference. Consider inference after observing action $X_1 = A$ in $t = 1$. With a common prior π_1, an observer knows player 1's initial belief. However, so long as $\widehat{G}(\cdot \mid \theta) \neq G$, the inference drawn by a θ-type, π_2^θ, is incorrect. In $t = 2$, the observer additionally commits the "social" error—she assumes Player 2 also thinks π_2^θ after observing $X_1 = A$. But this is only true if Player 2 happens to have to taste θ. The observer has the wrong theory of the player 2's public belief, and thus mislearns about Player 2's private infromation.[37]

Importantly, aside from Section 1.6, I assume players don't revise their model of $\widehat{G}$ and maintain their belief that others are rational. I do so for two reasons. First, it seems reasonable that if one is unaware of her own mistake, then she likely fails to conclude that others make this same mistake. Agents likely fail to realize over time that others use misspecified models, let alone decipher these models. Thus a fixed perception is a reasonable starting point for analyzing this error. Second, it is standard within models of non-common priors to assume players have rational expectations about the distribution of beliefs. A primary goal of this paper is to understand the consequences when this assumption is relaxed—that is, when agents neglect heterogeneity in perceptions.

1.3 Learning about Horizontal Differentiation: Preliminaries

In this section and the next, I analyze learning about the horizontal locations of A and B when the quality difference is known. The unknown payoff state is simply $\omega = \zeta \in \{L, R\}$:

[37] Nisbett and Ross (1980) fittingly point out: "One of the most important consequences of this state of affairs is that when people make incorrect inferences about situational details, or fail to recognize that the same situation can be construed in different ways by different people, they are likely to draw erroneous conclusions about individuals whose behavior they learn about or observe." Here, neglecting the fact that others hold different perceptions of the taste distribution leads to erroneous conclusions about others' beliefs.

the only uncertainty is whether A is located to the left ($\omega = L$) or right ($\omega = R$) of B. For example, suppose the transaction costs, q, of two investments are known, but their risk is not: assets to the left are riskier—but have higher expected return—than those to the right. Agents with high values of θ prefer the safer asset. Or, consider learning about the characteristics of two jobs beyond observable starting wages (q). Jobs to the "left" offer greater flexibility, and those to the right offer greater opportunity for promotions and bonuses; θ represents an agents' taste for flexibility.[38]

The remainder of this section derives players' choice and inference rules. Implications of projection on individual inference are discussed in Subsection 1.3.2, while the implications on long-run learning are the focus of Section 1.4.

Private Information. Before acting, each player nt observes her preference type θ_{nt}, and a private signal s_{nt} about ω from which she computes via Bayes' rule her private belief p_{nt} that $\omega = R$. Following Smith and Sørensen (2000), I work directly with the distribution of private beliefs. Conditional on ω, private beliefs are i.i.d. across individuals with c.d.f. F_ω; F_L and F_R are differentiable, and mutually absolutely continuous with common support $\text{supp}(F)$, so that no signal perfectly reveals the state of the world. Additionally, the distributions satisfy the following two assumptions:

Assumption 5. (Monotone Likelihood Ratio Property (MLRP).) *Let f_ω denote the density of private beliefs in state ω. $f_R(p)/f_L(p)$ is increasing in p.*

Assumption 6. (Unbounded Private Beliefs.) *For each ω,* $\text{co}(\text{supp}(F_\omega)) = [0, 1]$.

Assumption 5 simply implies private beliefs in favor of ω are relatively more likely whenever ω is true. MLRP implies private-belief distributions exhibit first-order stochastic dominance: for all $p \in (0, 1)$, $F_R(p) \leq F_L(p)$. Assumption 6 implies private beliefs are unbounded: from any non-degenerate prior π and for any $\bar{r} \in (0, 1)$, a player moves with positive probability to beliefs at most $\bar{r}$ and with positive probability to beliefs at least $\bar{r}$. Hence, players receive signals ranging from nearly fully revealing, to uninformative, to (rarely) nearly fully misleading. The "unbounded" signal structure provides a sharp rational benchmark, as it allows rational agents to learn ω.[39]

[38] Holding the average quality of the two jobs fixed, the assumption of negatively-correlated characteristics seems reasonable—the same job will not specialize in both flexible work hours and pecuniary perks. The risk-aversion example also fits this negatively-correlated characteristic paradigm, as the market naturally generates higher expected return on riskier assets. Additional applications include learning about new technologies with known prices but unknown productivities that depend on θ—for example, hybrid seeds with output that depends on soil or other input characteristics, θ. Farmers prefer the seed type that's most productive on their plot.

[39] An understanding has emerged that unbounded private beliefs lead to the successful aggregation of information in a variety of models and contexts. Aside from Smith and Sørensen (2000), Acemoglu, Dahleh, Lobel, and Ozdaglar (2011) and Smith and Sørensen (2008), respectively, show that unbounded beliefs lead to learning in a large class of networks and sampling regimes. Mossel, Sly and Tamuz (2012) show that unbounded beliefs lead to learning in a setting with repeated interactions.

Public Information and Individual Decision-Making. Before acting, each Player nt observes the history h_t, and computes public belief π_t, the probability of $\omega = R$ conditional on h_t.[40] From private belief p and public belief π, a player forms posterior r that $\omega = R$ via Bayes' rule: $r(p, \pi) = p\pi/[p\pi + (1-p)(1-\pi)]$. Players maximize expected utility given their beliefs: type θ takes action A if and only if

$$r\left[q^A - k(1-\theta)^2\right] + (1-r)\left[q^A - k(-1-\theta)^2\right] \geq r\left[q^B - k(-1-\theta)^2\right] + (1-r)\left[q^B - k(1-\theta)^2\right].^{41} \quad (1.4)$$

Rearranging yields the following decision rule:

Lemma 1. *Player nt with private belief p and public belief π has the following decision rule:*

1. *If $\theta_{nt} < 0$, then $X_{nt} = A \Leftrightarrow r(p, \pi) \leq \bar{r}(\theta)$*
2. *If $\theta_{nt} > 0$, then $X_{nt} = A \Leftrightarrow r(p, \pi) \geq \bar{r}(\theta)$*

where

$$\bar{r}(\theta) := \frac{1}{2} + \frac{\Delta_q}{2k\Delta_d(\theta)}, \quad \text{and} \quad \Delta_d(\theta) := (1-\theta)^2 - (-1-\theta)^2 = -4\theta. \quad (1.5)$$

$\Delta_d(\theta)$ is the difference between a θ type's (squared) distance from 1 and -1. The decision rule is intuitive. If $\Delta_q < 0$, then a "right" type ($\theta > 0$) has a cutoff $\bar{r}(\theta) > \frac{1}{2}$: she must be quite certain that A is to the right of B in order to forgo the quality advantage of B and choose A. In terms of the investment example, when B has enticingly low transaction costs, a risk-averse investor must be fairly confident that A is safer if she's to choose A over B.

Observers learn about Player nt's private information p_{nt} from her action, X_{nt}. Observers invert her strategy (Lemma 1) to form cutoffs on p_{nt}: conditional on θ_{nt}, X_{nt} reveals if her private signal was above or below this cutoff.[42] I now derive these private-belief cutoffs critical for drawing inference.

First, there may be some players who's action reveals no private information: when $\Delta_q > 0$, for values of θ near 0—agents with weak preferences over location—the posterior-belief cutoff $\bar{r}(\theta) \notin (0, 1)$. That is, the quality advantage of A outweighs the benefit of choosing B no matter their belief about ω, and hence they always choose A. I call such players *passive*, while a players who's beliefs influence her choice is *active*. The following lemma identifies the set of passive players.

Lemma 2. *Suppose $\Delta_q \geq 0$.*

[40] To be clear, the public belief, π, is drawn solely from the history of play while the private belief, p, is derived solely from one's private signal.

[41] The implicit assumption that A is chosen when indifferent is without loss of generality. Given continuous signals, indifference is a zero-probability event.

[42] While the solution concept implies that a naive projector correctly knows others' strategies, she mispredicts their private-belief thresholds—she neglects that other types have divergent perceptions of the public belief. This error, which I highlight in the next subsection, is one of the two ways in which a naive projector mislearns from others' actions.

1. *The set of active left types is* $\Theta^l = \{\theta \in \Theta \mid k\theta < -\Delta_q/4k\}$*.*

2. *The set of active right types is* $\Theta^r = \{\theta \in \Theta \mid \theta > \Delta_q/4k\}$*.*

3. *The set of passive types is* $\Theta^p = \{\theta \in \Theta \mid -\Delta_q/4k \leq \theta \leq \Delta_q/4k\}$*.*

So that learning takes place, I assume there exist some active right and left types.

Assumption 7. *There exists a positive measure of both active left types and active right types:* $\sum_{\theta\in\Theta^l} g(\theta) > 0$*, and* $\sum_{\theta\in\Theta^r} g(\theta) > 0$*.*

I present the private-belief cutoffs in terms of the *public likelihood ratio*, $\ell := (1-\pi)/\pi$, which is the inverse of the relative likelihood of state R; the lower is ℓ, the more likely is $\omega = R$. Since a player's posterior as a function of ℓ is $r(p,\ell) = p/[p+(1-p)\ell]$, it follows from Lemma 1 that agents' decisions reflect the following cutoff rule on private beliefs:

Lemma 3. *Suppose* $\Delta_q \geq 0$*. Player* nt *with private belief* p *and public likelihood ratio* ℓ *has the following decision rule:*

1. *If* $\theta_{nt} \in \Theta^l$*, then* $X_{nt} = A \Leftrightarrow p \leq p(\ell,\theta)$

2. *If* $\theta_{nt} \in \Theta^r$*, then* $X_{nt} = A \Leftrightarrow p \leq p(\ell,\theta)$

3. *If* $\theta_{nt} \in \Theta^p$*, then* $X_{nt} = A$ *for all* $p \in (0,1)$

where

$$p(\ell,\theta) := \frac{\ell}{v(\theta)+\ell}, \qquad \text{and} \quad v(\theta) := \frac{1-\bar{r}(\theta)}{\bar{r}(\theta)} = \frac{4k\theta+\Delta_q}{4k\theta-\Delta_q}. \tag{1.6}$$

The *private-belief threshold* $p(\ell,\theta)$ is the private belief that renders an active θ type indifferent between A and B given public likelihood ratio ℓ. Intuitively, a "left" type must have a sufficiently strong private belief that A is located to the left in order to choose A—her p must be sufficiently low—while a "right" type must have a sufficiently strong private belief that A is located to the right—her p must be sufficiently high. $v(\theta)$ is the likelihood ratio at which a θ type is indifferent. $v(\theta)$ is decreasing in θ on both Θ^l and Θ^r: $\overline{\theta}^l := \max\Theta^l$ must be most convinced of $\omega = R$ in order to be indifferent, while $\underline{\theta}^r := \min\Theta^r$ must be most convinced of $\omega = L$ to be indifferent.

The private-belief thresholds are very simple when each option has the same quality, $\Delta_q = 0$. First, from Equation 1.6, $p(\ell,\theta) = \ell/(1+\ell) = 1-\pi$ is independent of θ. Each player simply chooses the action that is most likely closest to her. Second, there are no passive players: $\Theta^l = \{\theta \in \Theta \mid \theta < 0\}$ and $\Theta^r = \{\theta \in \Theta \mid \theta > 0\}$. These two implications greatly simplify the inference problem analyzed in the following subsection, and since fixing $\Delta_q = 0$ has little impact qualitatively, I assume $\Delta_q = 0$ for the remainder of Sections 1.3 and 1.4, unless I specifically mention otherwise. There is, however, one important loss of generality that comes with this assumption, which I address in Section 1.3.1.

Two-Type Example. At several points in the paper, I consider the model with only two types: $\Theta = \{-1, 1\}$. In such cases, I synonymously use $\theta = l, r$ ("left" and "right") in place of $\theta = -1, 1$. $\lambda = \Pr(\theta_{nt} = r) > \frac{1}{2}$ is the fraction of right types. $\hat{\lambda}^l$ and $\hat{\lambda}^r$ are left and right types perceptions of λ. Players are active so long as $4k > \Delta_q$. Table 1.1 shows the payoff matrix.

TYPE-l

ω/X	A	B
L	q^A	$q^B - 4k$
R	$q^A - 4k$	q^B

TYPE-r

ω/X	A	B
L	$q^A - 2k$	q^B
R	q^A	$q^B - 2k$

Table 1.1: State-dependent payoffs for left ($\theta = l$) and right ($\theta = r$) types.

1.3.1 Belief Dynamics

This subsection describes the stochastic processes of public likelihood ratios $\langle \ell_t^\theta \rangle$. From Section 1.2.4, types with distinct perceptions of G draw different inference from history h_t; consequently, their public beliefs follow distinct processes. Let $\boldsymbol{\ell}_t \in \mathbb{R}_+^{|\Theta|}$ be the vector of each type's public likelihood ratio in t, ordered from least to greatest θ. Let ℓ_t^θ denote a generic element of $\boldsymbol{\ell}_t$. When there are just two distinct perceptions, as is so with choice-dependent projection (Definition 2), I write $\boldsymbol{\ell}_t = (\ell_t^l, \ell_t^r)$, where $\ell_t; := \ell_t(\theta < 0)$ is a left type's inference from h_t, and, $\ell_t^r := \ell_t(\theta > 0)$ is a right type's.

Each process $\langle \ell_t^\theta \rangle$ is described by the initial value $\ell_1^\theta = 1$ (recall players beginning with common prior $\pi_1 = 1/2$) and Bayesian transition equation

$$\ell_{t+1}^\theta = \frac{\widehat{\Pr}_\theta(a_t \mid \ell_t^\theta, L)}{\widehat{\Pr}_\theta(a_t \mid \ell_t^\theta, R)} \ell_t, \tag{1.7}$$

where $\widehat{\Pr}_\theta(a_t \mid \ell_t^\theta, \omega)$ is the probability of observing a_t people choose A in state ω according to type-θ's incorrect model, in which all players in t share public belief ℓ_t^θ, and tastes have distribution $\widehat{G}(\cdot \mid \theta)$. For notational purposes, let $\psi_\theta(a \mid \ell, \omega) := \widehat{\Pr}_\theta(a \mid \ell, \omega)$. Since behavior of each player in t is independent conditional on h_t,

$$\psi_\theta(a \mid \ell, \omega) = \binom{N}{a} \alpha_\theta(\ell, \omega)^a \big[1 - \alpha_\theta(\ell, \omega)\big]^{N-a}, \tag{1.8}$$

where $\alpha_\theta(\ell, \omega) := \widehat{\Pr}_\theta(X_{nt} = A \mid \ell, \omega)$ is a θ-type's perceived probability that a random player chooses A given ℓ and ω; that is, conditional on h_t, the perceived distribution of actions *within* a period is Binomial(N,$\alpha_\theta(\ell, \omega)$).

Note that

$$\alpha_\theta(\ell, \omega) = \sum_{\tilde{\theta} \in \Theta^l} F_\omega\big(p(\ell, \tilde{\theta})\big) \hat{g}(\tilde{\theta} \mid \theta) + \sum_{\tilde{\theta} \in \Theta^r} \big[1 - F_\omega\big(p(\ell, \tilde{\theta})\big)\big] \hat{g}(\tilde{\theta} \mid \theta) + \sum_{\tilde{\theta} \in \Theta^p} \hat{g}(\tilde{\theta} \mid \theta). \tag{1.9}$$

From left to right, the terms in Equation 1.9 are the probabilities that Player nt: (1) is an active left type and receives private information low enough to provoke action A, $p < p(\ell,\theta)$; (2) is an active right type and receives private information high enough to provoke action A, $p > p(\ell,\theta)$; and (3) is passive, and necessarily chooses A. Equation 1.9 simplifies greatly when $\Delta_q = 0$, so that belief thresholds are independent of θ and all types are active. Equation 1.9 simplifies to

$$\begin{aligned}\alpha_\theta(\ell,\omega) &= \left(\sum_{\tilde\theta<0} \hat g(\tilde\theta \mid \theta)\right) F_\omega\big(p(\ell)\big) + \left(\sum_{\tilde\theta>0} \hat g(\tilde\theta \mid \theta)\right)\big[1 - F_\omega\big(p(\ell)\big)\big] \\ &= \big[1-\hat\lambda(\theta)\big]F_\omega\big(p(\ell)\big) + \hat\lambda(\theta)\big[1-F_\omega\big(p(\ell)\big)\big]. \qquad (1.10)\end{aligned}$$

The first (second) term of Equation 1.10 is just the perceived measure of left (right) types times the perceived probability that a left (right) type takes A. In contrast, the true probability that of $X_{nt} = A$ in state ω, denoted $\alpha(\boldsymbol{\ell},\omega)$, depends on the current beliefs of *all* types, $\boldsymbol{\ell}$:

$$\alpha(\boldsymbol{\ell},\omega) = \sum_{\tilde\theta\in\Theta^l} g(\tilde\theta)F_\omega\big(p(\ell_t^{\tilde\theta})\big) + \sum_{\tilde\theta\in\Theta^r} g(\tilde\theta)\big[1-F_\omega\big(p(\ell_t^{\tilde\theta})\big)\big]. \qquad (1.11)$$

Comparing Equations 1.10 and 1.11 makes clear the two errors a naive taste-projector commits when learning from actions: she (a) mispredicts the frequency of types, $\hat g$, and (b) wrongly thinks all types share her public belief ℓ, so she miscalculates other types' cutoffs $p(\ell)$, and thus mispredicts the probability that other types take A.

Piecing these probabilities together, the transition function for type-θ's beliefs given observation a and current belief ℓ (Equation 1.7) is simply

$$\varphi_\theta(a,\ell) = \psi_\theta(a \mid \ell, L)/\psi_\theta(a \mid \ell, R)\ell = \Psi_\theta(a,\ell)\ell. \qquad (1.12)$$

where the ratio Ψ_θ is the likelihood of observing a in L relative to R given θ's model of the world. Next period's public likelihood ratio is a multiplicative factor Ψ_θ of today's public likelihood ratio.

Confounded Learning. The assumption that $\Delta_q = 0$, which I make for the remainder of this section and 1.4, comes at some loss of generality. Smith and Sørensen (2000) show that rational observational learning with heterogeneous preferences may fail even when private beliefs are unbounded. Specifically, there may exist an interior steady-state belief $\hat\ell$, which they call a "confounding belief", such that $\varphi(a,\ell) = \ell$ for any $a \in \{0,...,N\}$; each possible observation is equally likely in $\omega = L$ and $\omega = R$. If beliefs converge to this interior point, which happens with positive probability whenever $\hat\ell$ exists, then agents never learn. In this environment, a confounding belief exists only if $|\Delta_q|$ is sufficiently large. Hence, assuming $\Delta_q = 0$ rules out this possibility. However, $\Delta_q = 0$ is not a knife-edge case; the non-existence of confounding beliefs is robust.

Lemma 4. *Fixing all components of the game Γ aside from Δ_q, there is a robust (open, non-empty) set of quality differences Δ_q for which there exist no confounding beliefs.*

As a function of the perceived distributions of tastes, one can find a $\tilde{\Delta}_q > 0$ such that for all $\Delta_q \in (-\tilde{\Delta}_q, \tilde{\Delta}_q)$, no confounding belief exists.

Appendix 1.B discusses confounded learning in more detail, and derives bounds on Δ_q such that no confounding belief exists. Further, it explores how the basic results derived under the assumption of $\Delta_q = 0$ change when a confounding belief exists. If so, the logic under $\Delta_q = 0$ still holds, and results are identical aside from the possibility of convergence to to confounding belief. Consequently, results indicating the possibility, or impossibility, of society reaching some *confident* belief are unchanged by the presence of a confounding belief.

1.3.2 Effect of Taste Projection on Individual Inference: Comparative Statics on $\hat{\lambda}(\theta)$

This subsection analyzes comparative statics of $\hat{\lambda}(\theta)$ on the belief-transition equation $\ell_{t+1} = \varphi_\theta(a, \ell_t)$, demonstrating how taste projection distorts inference. The results established here play a key role in understanding the long-run dynamics studied in Section 1.4.

First, one's current belief and perception of tastes, $\hat{\lambda}(\theta)$, dictates the interpretation of an observation $a_t \in \{0, ..., N\}$; a_t is evidence in favor of $\omega = R$ whenever $\ell^\theta_{t+1} = \varphi_\theta(a_t, \ell^\theta_t) < \ell^\theta_t$.

Lemma 5. *For each $\theta \in \Theta$ and perceived public likelihood ratio $\ell^\theta_t \in \mathbb{R}_+$, there exists a value $\kappa\big(\ell^\theta_t, \theta\big) \in (0,1)$ such that observation a_t is interpreted as evidence in favor of $\omega = R$ if and only if*

$$\begin{gathered} \frac{a_t}{N} > \kappa\big(\ell^\theta_t, \theta\big) \quad \text{and} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ \textit{or} \\ \frac{a_t}{N} < \kappa\big(\ell^\theta_t, \theta\big) \quad \text{and} \quad \hat{\lambda}(\theta) < \frac{1}{2}, \end{gathered} \tag{1.13}$$

where

$$\kappa(\ell, \theta) = \frac{1}{1 + \log\left(\frac{\alpha_\theta(\ell, L)}{\alpha_\theta(\ell, R)}\right) / \log\left(\frac{1-\alpha_\theta(\ell, R)}{1-\alpha_\theta(\ell, L)}\right)}. \tag{1.14}$$

An investor who thinks most are risk averse must observe sufficiently many choose A in order to interpret a_t as evidence that A is safer than B, but one who thinks most are risk seeking must see a sufficiently *few* choose A. Figure 1.1 depicts $\kappa(\pi, \theta)$ for 4 values of $\hat{\lambda}(\theta)$.[43]

The limit values of $\kappa(\pi, \theta)$—values near $\pi = 0$ and $\pi = 1$—are critical for determining whether a confident belief is "stable": if players grows confident, what they subsequently observe maintains this confidence. For instance, when $\hat{\lambda}(\theta) > \frac{1}{2}$, $\lim_{\pi^\theta \to 1} \kappa(\pi, \theta) = \hat{\lambda}(\theta)$, so if a θ-type grows nearly confident that $\omega = R$ ($\pi \approx 1$), then she must observe at least $\hat{\lambda}(\theta)$ A's (in frequency) for beliefs to continue growing toward 1. But if fraction $\lambda < \hat{\lambda}(\theta)$ is observed—the true fraction that prefer A in $\omega = R$—then π^θ actually moves downward

[43] This example of $\kappa(\pi, \theta)$ assumes private beliefs have conditional pdfs $f_R(p) = 2p$ and $f_L(p) = 2(1 - p)$.

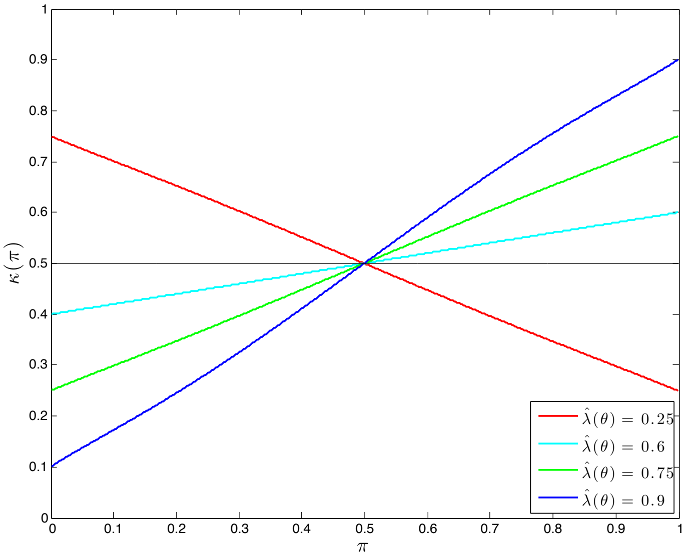

Figure 1.1: *Examples of $\kappa(\pi, \theta)$. Each curve represents the minimal fraction of A's (y-axis) a θ-type must observe in period t in order for a_t to be interpreted as evidence for $\omega = R$ given period-t belief π (x-axis). The curves differ only in the value of $\hat{\lambda}(\theta)$.*

from 1; observing exactly what a rational agent expects to see in $\omega = R$ *reduces* the biased agent's confidence in $\omega = R$. This fact is critical for understanding when some constellation of beliefs across types is stable, and is developed further in Section 1.4.

Figure 1.2 demonstrates this effect on beliefs. Suppose $N = 100$ and $a_t = 75$ is observed. The various curves show the effect of a_t on beliefs as a function of the current public belief (x-axis) for various values of $\hat{\lambda}(\theta)$. The y-axis is the (negative) change in the log-likelihood ratio: $\log \ell_{t+1}^{\theta} - \log \ell_t^{\theta}$. If this value is positive, the agent perceives a_t as evidence for $\omega = R$; if it is negative, then a_t supports $\omega = L$.

Another key feature of Lemma 5, evident from Figure 1.2, is what I call the *perceived-majority effect*: two agents may draw precisely opposite interpretations from the same observed behavior.

Proposition 1. (Perceived-Majority Effect.) *For any $\ell_t^{\theta} \in \mathbb{R}_+$, if $a_t/N > \hat{\lambda}(\theta)$, then $\ell_{t+1}^{\theta} < \ell_t^{\theta}$ if and only if $\hat{\lambda}(\theta) > \frac{1}{2}$. Similarly, if $a_t/N < 1 - \hat{\lambda}(\theta)$, then $\ell_{t+1}^{\theta} > \ell_t^{\theta}$ if and only if $\hat{\lambda}(\theta) > \frac{1}{2}$.*

Corollary 1. (Single-File Majority Effect.) *Suppose $N = 1$. If $X_t = A$ then $\ell_{t+1}^{\theta} < \ell_t^{\theta}$ if and only if $\hat{\lambda}(\theta) > \frac{1}{2}$. Similarly, if $X_t = B$ then $\ell_{t+1}^{\theta} > \ell_t^{\theta}$ if and only if $\hat{\lambda}(\theta) > \frac{1}{2}$.*

Proposition 1 states that when a sufficiently large proportion of agents choose X in t, people who disagree on the majority preference will disagree on the interpretation of this evidence. If 75% of investors buy A in period t, then one who thinks 60% are risk averse concludes

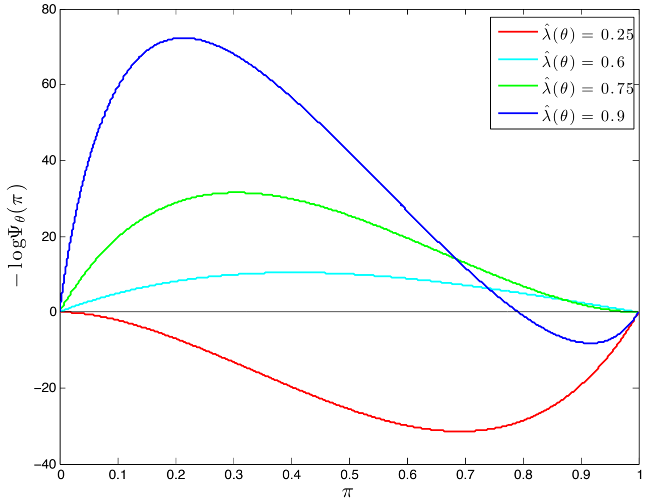

Figure 1.2: *Negative change the in public log-likelihood ratio, $-\log\Psi_\theta(\pi)$, as a function of the current belief, π, after observing action $a_t/N = .75$ for various values of $\hat{\lambda}(\theta)$. A θ-type interprets a_t as evidence for $\omega = R$ if and only if $-\log\Psi_\theta(\pi) > 0$.*

A is likely safe, but another who thinks 40% are risk averse thinks A is likely risky. The corollary, which assumes agents act in single file ($N = 1$), is even more straightforward: an individual always interprets A as evidence for $\omega = R$ if and only if she believes that the majority of players are right types. This result has very different implications depending on whether people suffer strong or weak projection (Definition 3). If it's weak—all agree on the majority taste—then all always agree on the interpretation of an A choice. But, if it's strong, then left and right types always disagree: two individuals can observe the same evidence but disparately conclude that it supports opposite hypotheses.

The preceding result determines how $\hat{\lambda}(\theta)$ influences the direction in which beliefs move. The next result, which I call the *variance effect*, addresses the magnitude of changes in beliefs.

Proposition 2. (Variance Effect.) *Suppose $N = 1$. For any $\ell_t^\theta \in \mathbb{R}_+$ and $X_t \in \{A, B\}$, $|\ell_{t+1}^\theta - \ell_t^\theta|$ strictly increasing in $\hat{\lambda}(\theta)$ on $\left[\frac{1}{2}, 1\right]$ and strictly decreasing in $\hat{\lambda}(\theta)$ on $\left[0, \frac{1}{2}\right]$*

The perceived variance in whether a player is a right or left type is $\hat{\lambda}(\theta)\left[1 - \hat{\lambda}(\theta)\right]$, which is maximized at $\hat{\lambda}(\theta) = \frac{1}{2}$ and decreasing as $\hat{\lambda}(\theta)$ moves away from $\frac{1}{2}$ in either direction. Hence, Proposition 2 implies that as one's perceived variance in types decreases, her beliefs change by a greater amount after any observation. $\hat{\lambda}(\theta)$ dictates the (perceived) informativeness of actions.As perceived variance decreases, a player becomes more confident about the tastes of those whom she observes—observed choices are more precise signals of the decision maker's private information. If she overestimates the likelihood that predecessors are right types, then observing A, say, is interpreted as overly strong evidence that A is optimal for right types. This result has important implications in the case of weak projection. Fixing an initial belief, the right-type belief changes by a greater amount than the rational belief after any action—beliefs are volatile, and over-responsive. The left-type belief, however, changes by a smaller amount than is rational—they are relatively conservative, and under-responsive. In terms of an example, a very risk-averse investor (a right type) reacts too strongly to a predecessor's choice, as she's too confident that it reflects her own best investment strategy. But a risk-neutral investor (a left type), is too skeptical of the evidence—if she thinks her processor was roughly equally likely risk averse or risk neutral, then his choice tells her relatively little about her own optimal strategy.

Figure 1.3 demonstrates both the "single-file" majority and variance effects. The plot shows tomorrow's belief, π^θ_{t+1}, supposing A is observed today, as a function of today's belief, π^θ_t. The various curves assume different values of $\hat{\lambda}(\theta)$. Comparing $\hat{\lambda}(\theta) = 0.25$ to the other cases highlights the perceived-majority effect: unlike when $\hat{\lambda}(\theta) > 1/2$, A causes tomorrow's belief to move below today's. Comparing the curves with $\hat{\lambda}(\theta) > 1/2$ demonstrates the variance effect: the magnitude of changes in beliefs increases with $\hat{\lambda}(\theta)$.

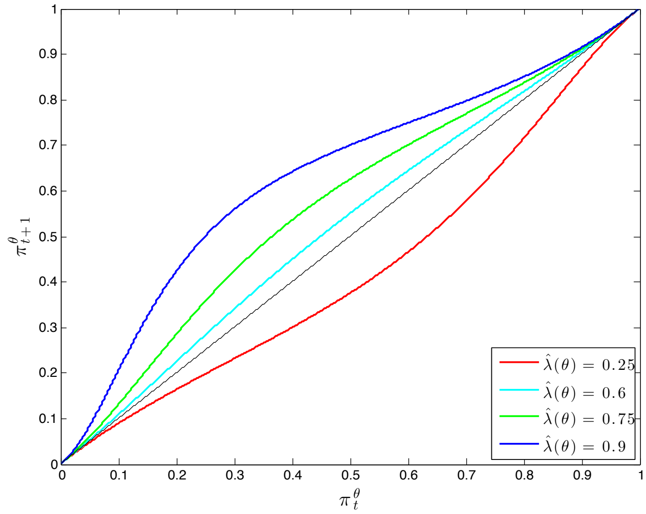

Figure 1.3: *Next-period's public belief π^θ_{t+1} as a function of the current public belief, π^θ_t, assuming choice A is observed in t. The $45°$-line is plotted for reference.*

1.4 Learning About Horizontal Differentiation: Long-Run Beliefs

Building on the setup of Section 1.3 where agents learn about horizontal locations, this section investigates the effect of taste projection on long-run beliefs and behavior. I show that when the bias is strong, taste projection always leads to inefficient herds and fully-confident beliefs. But when it is weak, it leads to cyclical behavior and persistently fluctuating beliefs. Subsection 1.4.1 introduces the possible learning outcomes under projection, and 1.4.2 derives conditions on players' perceptions of λ that determine which equilibrium beliefs are stochastically stable: if beliefs reach the neighborhood of a fixed point, does the (unexpected) resulting behavior confirm or contradict these beliefs? This analysis, conducted directly on the primitive $\hat{\boldsymbol{\lambda}}$—the vector of each types' perception, ordered from least to greatest θ—assumes a general model of perceptions, where each of an arbitrary finite number of types may hold a distinct perception $\hat{\lambda}(\theta)$; I only assume $\hat{\lambda}(\theta)$ is increasing in θ. To build intuition for the particular way in which learning fails as a function of $\hat{\boldsymbol{\lambda}}$, and to explore comparative statics, Subsections 1.4.3 and 1.4.4 assume a simple two-type setting ("choice-dependent" projection), where left and right types have distinct perceptions, $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}^l, \hat{\lambda}^r)$, and beliefs, $\boldsymbol{\ell}_t = (\ell_t^l, \ell_t^r)$. Subsection 1.4.5 discusses how these "two-type" results generalize to cases with many perceptions.

1.4.1 Potential Learning Outcomes

I first introduce terminology for the various learning outcomes that can occur. Learning among θ-types is (1) *complete* if π_t^θ converge almost surely to the truth; (2) *incorrect* if π_t^θ converges to certainty in a false state; (3) *incomplete* if π_t^θ does not converge to certainty in any state. Learning fails for type-θ if it is incorrect or incomplete. Finally, I say *universal learning* is complete if learning is complete for all $\theta \in \Theta$. Otherwise, universal learning fails. Without loss of generality, I assume $\omega = R$—action A is located on the right—so complete learning for type-θ entails $\Pr(\lim_{t\to\infty} \pi_t^\theta = 1) = 1$, or, in terms of the public likelihood ratio, $\ell_t^\theta \xrightarrow{a.s.} 0$.[44]

As a benchmark, given the assumption of no quality differences, if people are fully-rational ($\hat{\lambda}(\theta) = \lambda$ for all $\theta \in \Theta$), then they necessarily come to learn the true state in the long run.

Proposition 3. *If $\hat{\lambda}(\theta) = \lambda$ for all $\theta \in \Theta$, then learning is complete: $\pi_t^\theta \xrightarrow{a.s.} 1$ for all $\theta \in \Theta$.*

This result—first derived in Smith and Sørensen (2000)—follows from the martingale character of rational public beliefs. Provided $\hat{\lambda}(\theta) = \lambda$ for each θ, ℓ_t^θ is identical across θ in all t, and $\langle \ell_t^\theta \rangle$ is a martingale conditional on state $\omega = R$. It follows from the Martingale Convergence Theorem that $\langle \ell_t^\theta \rangle$ converges almost surely to some random variable $\ell_\infty^\theta := \lim_{t\to\infty} \ell_t^\theta$.

Yet with projection, public beliefs do not form a martingale:

Lemma 6. *The likelihood-ratio processes $\langle \ell_t^\theta \rangle$ is a martingale conditional on state R if and only if $\hat{\lambda}(\theta) = \lambda$ for all $\theta \in \Theta$.*

As long as $\hat{\lambda}(\theta) \neq \lambda$ for *some* $\theta \in \Theta$, all players mispredict the distribution of actions in t. The perceived probability of outcome a_t in $\omega = R$ according to *any* θ's model of the world, $\psi_\theta\big(a_t \mid \ell_t^\theta, R\big)$, is generically not equal to the true probability of observing a_t in

[44] While much of the analysis is in terms of the public likelihood ratio, for sake of intuition, I present some results in terms of the sequence of beliefs, π_t^θ.

$\omega = R$, which depends on *all* types' beliefs, $\boldsymbol{\ell}_t$. Why? First, if $\hat{\lambda}(\theta) \neq \lambda$, then a θ-type directly mispredicts the distribution of actions in t via her misprediction of the distribution of preferences. Second, even if $\hat{\lambda}(\theta) = \lambda$ but $\hat{\lambda}(\theta') \neq \lambda$, then a θ-type mispredicts the *beliefs* of θ'-types.

Lemma 6 implies we cannot rely on standard martingale methods to study the limit properties of the joint-belief process $\langle \boldsymbol{\ell}_t \rangle$. To proceed, I first identify the set $\mathscr{L} \subset \mathbb{R}_+^{|\Theta|}$ of potential limit points such that *if* biased agents converges to stationary beliefs, they must lie in $\mathscr{L}$. I call $\mathscr{L}$ the set of "candidate equilibria".[45] Second, I evaluate whether these candidate equilibria are stochastically stable.

I now show that $\mathscr{L}$ is the set of *confident beliefs*: $\boldsymbol{\ell}$ such that for each θ, $\ell^\theta \in \{0, \infty\}$. Conditional on state $\omega = R$, the process of actions and beliefs $\langle a_t, \boldsymbol{\ell}_t \rangle$ is a discrete-time Markov process on $\{0, ..., N\} \times \mathbb{R}_+^{|\Theta|}$ with transitions along the θ-dimension given by

$$\ell_{t+1}^\theta = \varphi_\theta\big(a, \ell_t^\theta\big) \text{ with probability } \psi(a, \boldsymbol{\ell}_t) \quad (a = 0, ..., N), \tag{1.15}$$

where $\varphi_\theta(a, \ell)$ is the belief-transition function introduced in Section 1.3.1 (Equation 1.12) and $\psi(a, \boldsymbol{\ell})$ is the true probability of observing a at $\boldsymbol{\ell}$. Granted stationary limits exist, Theorems B.1 and B.2 of Smith and Sørensen (2000) determine ℓ_∞^θ for a such a Markovian belief process with state-dependent transitions. Since private beliefs are continuously distributed and the transition functions $\varphi_\theta(a, \cdot)$ are continuous for all a, it follows from their result that any $\hat{\ell}^\theta \in \operatorname{supp}\big(\ell_\infty^\theta\big)$ is a fixed point of the Markov process: for each component $\hat{\ell}^\theta$ of $\hat{\boldsymbol{\ell}} \in \operatorname{supp}(\boldsymbol{\ell}_\infty)$ and all $a \in \{0, ..., N\}$,

$$\hat{\ell}^\theta = \varphi_\theta\big(a, \hat{\ell}^\theta\big). \tag{1.16}$$

The following lemma shows that the only fixed points of process 1.15—the only possible stationary beliefs—are confident beliefs. This follows from the assumption of unbounded private beliefs—learning never stops at uncertain beliefs.[46]

Lemma 7. *Suppose that there exists a real, nonnegative random variable ℓ_∞^θ such that $\ell_t^\theta \xrightarrow{a.s.} \ell_\infty^\theta$. Then* $\operatorname{supp}\big(\ell_\infty^\theta\big) \subseteq \{0, \infty\}$.

Lemma 7 implies that $\mathscr{L} = \{0, \infty\}^{|\Theta|}$. For sake of presenting key results in terms of *beliefs* $\pi \in [0, 1]$, rather than likelihood ratios $\ell \in \mathbb{R}_+$, let Π be the set of candidate equilibria in belief space. From Lemma 7, any long-run stationary belief lies in $\Pi := \{0, 1\}^{|\Theta|}$.

We have now identified our candidate long-run stationary equilibria: Π. But to which of these equilibria will society converge? The next subsection (1.4.2) shows that agents perceptions of population preferences, $\hat{\boldsymbol{\lambda}}$, dictate which, if any, of these beliefs are asymptotically stable. The subsections to follow apply the stability criterion developed in 1.4.2 to two regimes of projection—strong and weak.

[45] The term equilibrium, in this context, refers to a profile of beliefs $\boldsymbol{\ell}$ which is a fixed point of the beliefs process.

[46] Recall that $\Delta_q = 0$ implies no confounding outcomes exist.

1.4.2 Stability of Confident Beliefs

This section derives, as a function of mispredictions $\hat{\boldsymbol{\lambda}}$, a condition specifying when a fixed point of the belief process is locally stable: if the process enters a neighborhood of the fixed point, then with positive probability it remains in that neighborhood forever after. Section 1.4.2.1 derives sufficient conditions on the Markov process (1.15) for a fixed point to be either stable or unstable. From these abstract conditions, Section 1.4.2.2 derives a stability criterion based directly on the primitives of the model: each agent's perception of others' tastes, $\hat{\lambda}(\theta)$.

1.4.2.1 Local Stability of the Joint-Belief Process

Let $\hat{\boldsymbol{\ell}} \in \mathscr{L}$ denote a fixed point of process 1.15 with generic element $\hat{\ell}^\theta$. Formally, stability is defined as follows.

Definition 4. *Fixed point $\hat{\boldsymbol{\ell}}$ is stable if for any open ball about $\hat{\boldsymbol{\ell}}$, $\mathscr{N}(\hat{\boldsymbol{\ell}})$, there is a positive probability that $\boldsymbol{\ell}_t \in \mathscr{N}(\hat{\boldsymbol{\ell}})$ for all $t \in \mathbb{N}$ provided $\boldsymbol{\ell}_1 \in \mathscr{N}(\hat{\boldsymbol{\ell}})$.*

The conditions for stability follow from the logic of stability theory for linear systems. Although the belief process is nonlinear, near the fixed point we can approximate the process by its first-order Taylor series expansion, and stability is assessed locally by applying standard linear-system criteria to this "linearized" approximation.

More formally, near fixed point $\hat{\boldsymbol{\ell}}$, each type's belief process $\langle \ell_t^\theta \rangle$ is well approximated by the following stochastic difference equation: starting at $(a_t, \ell_t^\theta - \hat{\ell}^\theta)$, the continuation is $\left(a_{t+1}, \frac{\partial}{\partial \ell}\varphi_\theta(a_t, \hat{\ell})(\ell_t^\theta - \hat{\ell}^\theta)\right)$ with chance $\psi(a_t, \hat{\boldsymbol{\ell}})$. That is, the continuation is well approximated by the first-order Taylor expansion of $\varphi_\theta\big(a_t, \ell_t^\theta\big)$ about fixed point $\hat{\ell}^\theta$. Now, for any linear process $\langle y_t \rangle$, where $y_{t+1} = b_a y_t$ with chance p_a for $a = 0, 1, ..., N$, we can write

$$y_t = b_0^{I_0(t)} \times ... \times b_N^{I_N(t)} y_1 \tag{1.17}$$

where $I_a(t)$ counts the realization of a's in the first $t-1$ steps. Since $I_a(t)/t \to p_a$ almost surely by the Strong Law of Large Numbers, the product $\chi := b_0^{p_0} \times ... \times b_N^{p_N}$ fixes the long-run stability of the stochastic system $\langle y_t \rangle$ near 0:

$$\lim_{t\to\infty} y_t = \lim_{t\to\infty} \left(b_0^{p_0} \times ... \times b_N^{p_N}\right)^t y_1 = \lim_{t\to\infty} \chi^t y_1. \tag{1.18}$$

Clearly from Equation 1.18, the linear process converges to the fixed point 0 if and only if the product $\chi < 1$. The analog of χ for the linearized belief process in the neighborhood of $\hat{\boldsymbol{\ell}}$ is

$$\chi_\theta(\hat{\boldsymbol{\ell}}) := \prod_{a=0}^{N} \left(\frac{\partial}{\partial \ell}\varphi_\theta\big(a, \hat{\ell}^\theta\big)\right)^{\psi(a,\boldsymbol{\ell})} \tag{1.19}$$

Accordingly, $\chi_\theta(\hat{\boldsymbol{\ell}})$—which I call the *stability coefficient* of type-θ's beliefs near $\hat{\boldsymbol{\ell}}$—determines the local stability of the original nonlinear system (1.15) near $\hat{\ell}$.

Lemma 8. *Suppose $\hat{\boldsymbol{\ell}} \in \mathscr{L}$. $\hat{\boldsymbol{\ell}}$ is stable if $\chi_\theta(\hat{\boldsymbol{\ell}}) < 1$ for all $\theta \in \Theta$, and unstable if for any $\theta \in \Theta$, $\chi_\theta(\hat{\boldsymbol{\ell}}) > 1$.*

Lemma 8 and the preceding discussion is simply an extension of Smith and Sørensen's (2000) Theorem 4, which establishes this stability criterion for an arbitrary Markov process as in 1.15 so long as continuation functions $\varphi_\theta(a, \cdot)$ and transition probability functions $\psi(a, \cdot)$ are C^1 (once continuously differentiable). While they use this condition to show stability of interior fixed points of the rational learning process, I use it to demonstrate both the instability of correct beliefs and the stability of false beliefs within the biased learning model.

1.4.2.2 Characterization of Confident Equilibria

This subsection derives from Lemma 8 a stability criterion based directly on the primitives of the model—people's perceptions of others' tastes. This key proposition shows that we can asses the stability of an equilibrium belief simply by comparing what people *expect* to observe at that belief with what they actually observe.[47] This requires some final pieces of notation. Let $\widehat{\mathscr{F}_\theta} : \mathcal{X} \times \mathbb{R}_+ \to [0, 1]$ be a θ-type's perceived probability of observing action X given ℓ^θ, and let $\mathscr{F} : \mathcal{X} \times \mathbb{R}_+^{|\Theta|} \to [0, 1]$ be the true probability of observing action X given belief profile $\boldsymbol{\ell}$. Additionally, let $M_\theta : \mathbb{R}_+ \to \mathcal{X}$ denote the the expected majority action according to θ's model at ℓ:

$$M_\theta(\ell) = \arg\max_{X \in \mathcal{X}} \widehat{\mathscr{F}_\theta}(X, \ell) \tag{1.20}$$

Finally, the main stability result follows.

Proposition 4. *Let $\hat{\boldsymbol{\ell}} \in \mathscr{L}$ be a fixed point of the joint-belief process.*

1. *$\hat{\boldsymbol{\ell}}$ is a stable if for all $\theta \in \Theta$, $\widehat{\mathscr{F}_\theta}\Big(M\big(\hat{\ell}^\theta\big), \hat{\ell}^\theta\Big) < \mathscr{F}\Big(M\big(\hat{\ell}^\theta\big), \hat{\boldsymbol{\ell}}\Big)$.*

2. *$\hat{\boldsymbol{\ell}}$ is unstable if for any $\theta \in \Theta$, $\widehat{\mathscr{F}_\theta}\Big(M\big(\hat{\ell}^\theta\big), \hat{\ell}^\theta\Big) > \mathscr{F}\Big(M\big(\hat{\ell}^\theta\big), \hat{\boldsymbol{\ell}}\Big)$.*

Note that $\mathscr{F}(A, \hat{\boldsymbol{\ell}})$ is the long-run frequency of action A if all players beliefs are fixed at $\hat{\boldsymbol{\ell}}$. So, Proposition 4 states that, given long-run behavior $\mathscr{F}(A, \hat{\boldsymbol{\ell}})$, stationary-equilibrium belief $\hat{\boldsymbol{\ell}}$ is stable if all players observe a greater share choosing their anticipated majority

[47] Gagnon-Bartsch and Rabin (2014) study a similar issue of stability in a model of biased social learning in which players draw inference from the history of play, but wrongly assume the behavior of each person they observe reflects solely that person's private information. In some settings, the behavior of a generation confident in the true state can lead observers to beliefs far from the truth: confident, correct beliefs are unstable. The "inferential naiveté" bias in Gagnon-Bartsch and Rabin (2014) was first studied in a more standard environment by Eyster and Rabin (2010), and a similar error where people neglect the redundancy in information when learning socially is analyzed by DeMarzo, Vayanos and Zwiebel (2003).

action than expected; it is unstable if any player observes *fewer* than expected choosing her anticipated majority action.

For example, suppose a risk-averse agent believes most seek the safer asset ($\hat{\lambda}^r > \frac{1}{2}$), and grows nearly confident that A is safe. Then she must observe a long-run frequency of A at least equal to $\hat{\lambda}^r$ in order for her to remain confident, if not, she necessarily becomes less confident over time. Essentially, observing a larger majority share than expected only corroborates a player's hypothesis about which action is optimal for the majority taste. The concept is similar self-confirming equilibrium (e.g., Fudenberg and Levine, 1993): an incorrect belief may be stable so long as the behavior of those best responding to that belief reinforces the false hypothesis. Even if the investor *wrongly* concludes that A is safe, so long as more people choose it than she anticipates, she'll continue to believe it is safe.

An implication of Proposition 4 is that agents cannot all hold identical beliefs in the long run:

Proposition 5. *For any degree of taste projection, long-run agreement across types does not arise. That is, for each $\hat{\pi} \in \{0,1\}$, $\Pr(\lim_{t\to\infty} \pi_t^\theta = \hat{\pi} \ \forall\ \theta) = 0$.*

An immediate, but important, corollary is that not all agents can learn the truth.

Corollary 2. *For any degree of taste projection, universal learning fails.*

Since rational agents necessarily learn the truth in this setting, Corollary 2 demonstrates a discontinuity of rational learning. Adding any degree of taste projection implies some agents necessarily fail to learn. The basic intuition is that if beliefs grow close to the truth—A is optimal for the majority taste—society would observe roughly $\lambda > \frac{1}{2}$ choose A. But people with the majority taste—who overestimate their commonness—expect to observe a frequency of A's strictly greater than λ. By Proposition 4, their beliefs necessarily become less confident over time.

The mere fact that some agents necessarily fail to learn tells us little about what agents *do* come to believe, or their long-run behavior. Within a simple two-type setting, the next two subsections use Proposition 4 to answer these questions, which crucially depends on the extent of players' biases. Section 1.D of the Appendix applies Proposition 4 to two additional settings: (1) subsection 1.D.1 shows that universal learning still fails among biased agents when an arbitrary fraction of society is fully rational, and (2) subsection 1.D.2 shows how learning may fail when agents suffer alternative distributional errors distinct from projection, such as a false sense of uniqueness.

1.4.3 Strong Taste Projection

In this subsection (1.4.3) and the next (1.4.4), suppose agents suffer "choice-dependent" projection (Definition 2). Hence, there are just two distinct perceptions of λ, $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}^l, \hat{\lambda}^r)$, and two distinct belief sequences, $\boldsymbol{\ell}_t = (\ell_t^l, \ell^\theta)$. Subsections 1.4.3 and 1.4.4 apply Proposition 4 respectively to the cases of strong and weak projection (Definition 3), identifying $\Pi^*(\hat{\lambda}^l, \hat{\lambda}^r)$—the set of stable equilibrium beliefs given misperceptions $\hat{\lambda}^l, \hat{\lambda}^r$—and showing in each case specifically how and why learning fails.

When each agent thinks her own taste is most common ("strong" taste projection), inevitably all herd on a single action X, and each grows confident that X is optimal for her taste. Beliefs of different types grow polarized. For example, two types of investors—risk-averse ("right types") and risk neutral ("left types")—each think their type is most common; say, $\lambda^r = 0.8$, and $\lambda^l = 0.4$, when in truth, $\lambda = 0.6$. In a large market—many act per period—agents have very different expectations about first-period purchases: risk averse expect roughly 80% buy A if A is safe, while risk seeking expect roughly 40% buy A if safe. Assuming A is safe, they observe roughly 60%. Recall that Proposition 1 states that agents with opposite tastes may interpret the same observation as evidence of opposite states. Here, risk averse perceive A as evidence that A is safe, but risk seeking think it indicates A is risky. After one round, risk averse are quite confident that A is safe, while risk seeking are quite confident that A is risky. With these polarized beliefs, nearly *all* investors in the next round best respond by buying A, which only further corroborates the investors' divergent beliefs. All risk seekers grow confident in the incorrect state. A numerical version of this example, demonstrating precise values of beliefs, is provided in Section 1.4.3.1. Formally, we have the following result.

Proposition 6. *Assume a strong taste projection.* $\Pi^*(\hat{\lambda}^l, \hat{\lambda}^r) = \{(0,1),(1,0)\}$*:* π_t^r *converges almost surely to either 0 or 1, each outcome arising with positive probability when N is finite. If* π_t^r *converges to 0 (1), then* π_t^l *almost surely converges to 1 (0).*

The intuition is simple, and precisely that displayed in the example above. Eventually, some action, say A, earns a majority following, and, absent strong contrary signals, all individuals believe it's optimal to follow the herd; agents flock to A. As such, the action frequency of A grows in expectation, reinforcing an observer's belief that A suits the more common taste—and hence *her* taste. Players don't realize that those with different tastes develop opposing beliefs, and thus don't expect *all* to choose A. After such a history, left types expect right types to choose B, and vice versa. But all best respond to this surprising history with A. The equilibrium is essentially self confirming: behavior following from polarized beliefs reinforces, and never contradicts, false beliefs.

When $N < \infty$, the number of players acting per period is finite, either action may be most popular in early periods. Hence, the action on which players inevitably herd is ex-ante random.[48] Society suffers a form of "social" confirmation bias, where initial evidence dominates long-run beliefs. Since players don't expect a herd, the surprisingly uniform behavior quickly

[48] In simulations of the model with signal densities $f_R(p) = 2p$ and $f_L(p) = 2(1-p)$ and parameters $\lambda = 0.75$, $\hat{\lambda}^r = 0.9$, $\hat{\lambda}^l = 0.4$ and $N = 1$ (agents move in single file), the majority type learns correctly roughly 80% of the time.

moves naive agents to confident beliefs.[49] The numerical example in Section 1.4.3.1highlights this point. While either herd is possible, the majority type learns correctly more often when λ or N increases. Higher values of these variables increase the likelihood that the action optimal for the majority taste, A, is most popular in early periods. As $N \to \infty$, agents almost surely herd on A.

The logic above explains that polarized beliefs $\hat{\boldsymbol{\pi}} = (0,1)$, and $\hat{\boldsymbol{\pi}} = (1,0)$ are stable, and Proposition 4 rules out the stability of $\hat{\boldsymbol{\pi}} = (0,0)$ and $\hat{\boldsymbol{\pi}} = (1,1)$.[50] When society is nearly certain that $\omega = R$, then fraction λ chooses A. To left types, who think they're most common, this observation suggests $\omega = L$, reducing confidence that $\omega = R$. Biased beliefs do not form a martingale; instead, taste projection imparts predictable drift on beliefs. Near the truth $\hat{\boldsymbol{\ell}} = (0,0)$, both ℓ_t^l and ℓ_t^r are strict submartingales—they drift away from the truth.

Lemma 9. *Assume strong taste projection. There exists a neighborhood about the truth* $\hat{\boldsymbol{\ell}} = (0,0)$ *such that:*

1. ℓ_t^r *is a strict submartingale:* $\mathbb{E}[\ell_{t+1}^r \mid \ell_t^l, \ell_t^r] > \ell_t^r$
2. ℓ_t^l *is a strict submartingale:* $\mathbb{E}[\ell_{t+1}^l \mid \ell_t^l, \ell_t^r] > \ell_t^l$.

Relating back to the stability condition in Proposition 4, near $\hat{\boldsymbol{\ell}} = (0,0)$ each player sees fewer than expected choose the action she thought would be most popular. Figure 1.4 shows the drift in beliefs for all regions of the joint-belief space.[51] Beliefs drift *away* from each fixed point where people agree, $\hat{\boldsymbol{\pi}} = (0,0)$ and $\hat{\boldsymbol{\pi}} = (1,1)$, but drift *toward* confident disagreement.

Strong taste projection leads to an extremely strong form of herding. With heterogeneous preferences, a "herd" is typically defined (e.g., Smith and Sørensen, 2000) by players of each type acting identically—rational "herds" do not preclude heterogeneity in behavior. With strong projection, however, players of *every* type act identically, an outcome I call *uniform* herding. Heterogeneity in long-run behavior is eliminated by strong taste projection. Agents inefficiently over adopt the popular action. Sorensen (2006) shows an interesting example of this in the domain of health-care plans. Workers within an academic department learn about others' choices before making a selection, but employees differ in their preferred

[49] See Rabin and Schrag (1999) for a discussion of confirmatory bias in individual learning settings. Eyster and Rabin (2010) also show how biased social learning causes society to grow too confident too quickly in which ever state initial evidence supports.

[50] This is also a direct consequence of Proposition 5. For sake of explaining the intuition for Proposition 5, I walk through the intuition for why different types fail to agree in the long run.

[51] As shown in Figure 1.4, there are four regions with distinct martingale properties. The label $(+,-)$, for example, implies that ℓ_t^l is a submartingale and ℓ_t^r is a supermartingale when restricted to the indicated region of $\mathbb{R}_+^2$. In general, there exists a function $L_l : \mathbb{R}_+ \times [0,1] \to \mathbb{R}_+$ such that if $\hat{\lambda}^l > 1/2$, then $\mathbb{E}[\ell_{t+1}^l \mid \boldsymbol{\ell}_t] > \ell_t^l \Leftrightarrow \ell_t^r > L_l(\ell_t^l, \hat{\lambda}^l)$, and if $\hat{\lambda}^l < 1/2$, then $\mathbb{E}[\ell_{t+1}^l \mid \boldsymbol{\ell}_t] > \ell_t^l \Leftrightarrow \ell_t^r < L_l(\ell_t^l, \hat{\lambda}^l)$. Similarly, there exists a function $L_r : \mathbb{R}_+ \times [0,1] \to \mathbb{R}_+$ such that if $\hat{\lambda}^r > 1/2$, then $\mathbb{E}[\ell_{t+1}^r \mid \boldsymbol{\ell}_t] > \ell_t^r \Leftrightarrow \ell_t^l < L_r(\ell_t^r, \hat{\lambda}^r)$, and if $\hat{\lambda}^r < 1/2$, then $\mathbb{E}[\ell_{t+1}^r \mid \boldsymbol{\ell}_t] > \ell_t^r \Leftrightarrow \ell_t^l > L_r(\ell_t^r, \hat{\lambda}^r)$. Both L_l and L_r are monotonic in ℓ^θ and intersect exactly once. Figure 1.4 (and also Figure 1.5) show L_θ in units of probabilities rather than likelihood ratios. That is, the figures plot $P_\theta(\pi) := L_\theta\big(\pi/(1-\pi), \hat{\lambda}(\theta)\big)/\big[1 + L_\theta\big(\pi/(1-\pi), \hat{\lambda}(\theta)\big)\big]$.

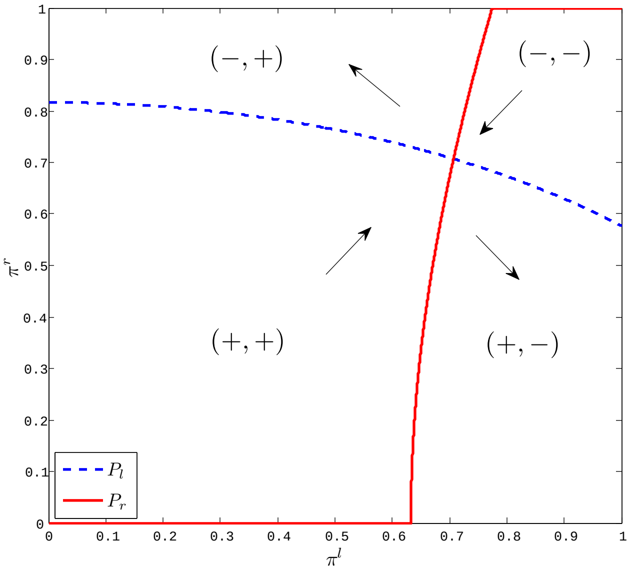

Figure 1.4: *Belief "phase diagram" for Strong Taste Projection.*

plan characteristics. Despite this, employees tend to herd on a particular plan.[52] Many employees later switch, reflecting the heterogeneity in the optimal match. In the extensions considered in Sections 1.5 and 1.6 agents explain the uniform herd using additional dimensions of uncertainty: respectively, it (wrongly) indicates large quality differences or very little heterogeneity in tastes.

As a result of uniform herding, observing others can be socially harmful. That is, public information reduces welfare, on average. Recall, a share $\nu \in \{1-\lambda, \lambda\}$ correctly learns, while $1-\nu$ chooses the inferior option. For sufficiently precise private signals, individuals are necessarily worse off by observing others than if they simply followed private information. Supposing 50-50 priors, an agent choosing solely on private information does so correctly with probability $\rho := 1 - F_R(.5)$. So long as

$$\begin{aligned}\rho > \frac{\lambda\mathbb{E}\big[u(A,\theta)-u(B,\theta)\mid\theta>0\big]}{\lambda\mathbb{E}\big[u(A,\theta)-u(B,\theta)\mid\theta>0\big]+(1-\lambda)\mathbb{E}\big[u(B,\theta)-u(A,\theta)\mid\theta<0\big]}\\ =\frac{\lambda\mathbb{E}\big[\theta\mid\theta>0\big]}{\lambda\mathbb{E}\big[\theta\mid\theta>0\big]-(1-\lambda)\mathbb{E}\big[\theta\mid\theta<0\big]},\end{aligned} \tag{1.21}$$

observing others reduces social welfare. With only two types, $\theta \in \{-1, 1\}$, condition 1.21 reduces to $\rho > \lambda$: it's more likely that an agent has a correct signal than that a random other shares her taste. The welfare loss is asymmetric in that within each game, it falls entirely

[52] Employees only observe choices of others' within their department. Interestingly, the "herd" plan varies across departments.

on agents with a particular taste. For large N, the inefficiency falls entirely on those in the minority.

Uniform herding, and hence polarized beliefs, are robust consequences of strong projection under a variety of perturbations to the canonical environment. First, both arise when more than two actions are available. So long as each thinks her taste is most common, net of private information, she finds it optimal to follow whichever action is most popular. Second, uniform herding is robust to non-Bayesian inference within the misspecified model. For instance, if agents neglect that others infer from the history and think others' actions rely solely on their private information (as in Eyster and Rabin, 2010), then the logic for uniform herding is unchanged—agents still have incentive to follow the herd. Non-Bayesian "agreeing to disagree" also leads to this conclusion. Suppose each individual believes her taste is most common, and that these conflicting perceptions are common knowledge. Right types are confident they are correct to think they are most common, and think left types have incorrect perceptions; left types think precisely the opposite. A uniform herd still emerges. In contrast to when players are naive, players are aware that some fraction of players err by following the herd. However, each thinks it's the *other* type who plays incorrectly.

1.4.3.1 Example

Building on the investment story that starts this subsection (1.4.3), this numerical example provides values of beliefs and the likelihood people choose the herd action, A. Specifically, it demonstrates how quickly biased agents grow confident in opposite hypotheses. Suppose that $\lambda = 0.6$, but $\hat{\lambda}^l = 0.4$ and $\hat{\lambda}^r = 0.8$, and 100 people act each period ($N = 100$). I calculate (expected) beliefs in periods 2 and 3 as a function of the initial choices in period 1, and compare the belief path of biased players, with that of rational players. I assume private beliefs have conditional densities $f_R(p) = 2p$ and $f_L(p) = 2(1-p)$, and that $\omega = R$.

Table 1.2 shows the evolution of behavior and beliefs for rational agents, and Table 1.3 does so for biased agents of each type. Each row of the table shows dynamics for a different initial observation, $a_1 \in \{45, 55, 65\}$.[53] The second column specifies beliefs upon observing a_1, and the third column shows the likelihood that a θ-type takes A given these beliefs. The fourth column is the expected observation in $t = 2$ given investors' updated beliefs, π_2^θ. The final column gives beliefs assuming the expected behavior in column 4 is observed.

Most striking is the last column of Table 1.3—the speed at which a herd emerges. Recall from Proposition 1, if $\hat{\lambda}(\theta) > \frac{1}{2}$, then no matter her prior, a θ-type interprets $a > \hat{\lambda}(\theta)N$ as unambiguous evidence for $\omega = R$. Likewise, if $\hat{\lambda}(\theta) < \frac{1}{2}$, $a > (1 - \hat{\lambda}(\theta))N$ is unambiguous evidence for $\omega = L$. In this example, if $a_2 > 80$, all the risk averse view a_2 as evidence that A is safe, and all the risk seeking view it as evidence that A is risky. What's the likelihood of observing such strong evidence for one state or the other? Under biased learning, if $a_1 = 55$, then $\Pr(a_2 > 80) = 0.9995$—99.95% of the time, investors observe evidence that necessarily

[53] Given the signal structure, $E[a_1 \mid R] = 55$ and $\mathbb{E}[a_1 \mid L] = 45$, so these initial conditions match exactly what rational players expect to see for some ω. Additionally, given $\omega = R$, the probability of such initial values are $\Pr(a_1 = 45) = 0.0107$, $\Pr(a_1 = 55) = 0.0800$, and $\Pr(a_1 = 65) = 0.0107$.

Rational Inference

a_1	π_2	$\Pr(X_{n2} = A \mid \theta_{n2} = \theta)$	$\mathbb{E}[a_2]$	π_3
45	0.1185	$\Pr(X_{n2} = A \mid \theta_{n2} = l) = 0.7770$ $\Pr(X_{n2} = A \mid \theta_{n2} = r) = 0.2230$	0.4446	0.1615
55	0.8815	$\Pr(X_{n2} = A \mid \theta_{n2} = l) = 0.0140$ $\Pr(X_{n2} = A \mid \theta_{n2} = r) = 0.9860$	0.7430	0.9140
65	0.9976	$\Pr(X_{n2} = A \mid \theta_{n2} = l) = 0$ $\Pr(X_{n2} = A \mid \theta_{n2} = r) = 1$	0.6000	0.9976

Table 1.2: Expected evolution of rational beliefs and behavior.

Biased Inference

a_1	π_2^θ	$\Pr(X_{n2} = A \mid \theta_{n2} = \theta)$	$\mathbb{E}[a_2]$	π_3^θ
45	$\pi_2^l = 0.8815$	$\Pr(X_{n2} = A \mid \theta_{n2} = l) = 0.0140$	0.0081	$\pi_3^l = 0.9999$
	$\pi_2^r = 0.0020$	$\Pr(X_{n2} = A \mid \theta_{n2} = r) = 0.0041$		$\pi_3^r = 0.0015$
55	$\pi_2^l = 0.1185$	$\Pr(X_{n2} = A \mid \theta_{n2} = l) = 0.7770$	0.9108	$\pi_3^l = 0.0004$
	$\pi_2^r = 0.9980$	$\Pr(X_{n2} = A \mid \theta_{n2} = r) = 1$		$\pi_3^r = 0.9983$
65	$\pi_2^l = 0.0024$	$\Pr(X_{n2} = A \mid \theta_{n2} = l) = 0.9952$	0.9981	$\pi_3^l = 0.0021$
	$\pi_2^r = 1$	$\Pr(X_{n2} = A \mid \theta_{n2} = r) = 1$		$\pi_3^r = 1$

Table 1.3: Expected evolution of biased beliefs and behavior.

pushes them toward polarized, divergent beliefs. In contrast, under rational learning, such evidence is *extremely* rare: $\Pr(a_2 > 80) = 4.5678 \times 10^{-6}$. Evidence that rational people find unambiguously in favor of $\omega = R$—$a_2 > 60$—is still relatively rare: $\Pr(a_2 > 60) = 0.4394$. Since behavior under rational play is quite heterogeneous, rational beliefs grow confident more slowly.

1.4.4 Weak Taste Projection

While "strong" taste projection presents a clear herding logic, uniform herding is not a general consequence of taste projection. Under the weak form of the bias, where people exhibit projection but still agree on the majority preference, agents never settle on a fixed belief, let alone herd on a single action.

Proposition 7. *Assume weak taste projection.* $\Pi^*(\hat{\lambda}^l, \hat{\lambda}^r) = \emptyset$*: There exists no stable fixed point for* ℓ_t^l *or* ℓ_t^r*.*

Since no belief is stable, beliefs of each type almost surely fail to converge to a fixed value. Agents never observe a pattern of behavior consistent with any hypothesis in their model of the world. As such, beliefs perpetually oscillate from favoring one state to the other. With respect to her own model, an agent's belief process forms a martingale: she anticipates that her opinion will eventually settle down on the truth. However, whenever it begins to settle down, she observes new, "shocking" evidence (with respect to her model) that pushes her back toward uncertainty. The agent is continually surprised, as she sees the most popular action repeatedly swing from A to B.

What is the logic for non-convergence? It relates to the "variance effect" discussed in Section 1.3.2. Since right types overestimate their frequency, $\hat{\lambda}^r > \lambda$, they underestimate the variance in tastes; they think actions reveal more private information than is so. In particular, when a right type observes a "contrarian" action—one that deviates from the most likely choice—she overweights the possibility that it's due to a player who shares her taste, but has strong information contrary to the current public opinion.[54] Importantly, contrarian actions are overattributed to private information rather than taste. When society is nearly confident of the truth, $\omega = R$, people observe a frequency of A near λ. But a right type expects a frequency near $\hat{\lambda}^r > \lambda$—she observes roughly $\hat{\lambda}^r - \lambda$ more contrarian choices (in frequency) than anticipated. And each of these choices pushes her belief toward $\omega = L$ more so than is rational. It follows that ℓ^r_t is a submartingale when nearly certain of the truth: beliefs drift toward less confident beliefs.

On the other hand, ℓ^l_t are a supermartingale near $\ell^\theta = 0$—left-type beliefs move toward $\ell = 0$ in expectation. A left type observes more As than anticipated, reinforcing her belief in $\omega = R$.

Lemma 10. *Assume weak taste projection. There exists a neighborhood about the truth* $\hat{\boldsymbol{\ell}} = (0, 0)$ *such that:*

1. ℓ^r_t *is a submartingale:* $\mathbb{E}[\ell^r_{t+1} \mid \ell^l_t, \ell^r_t] > \ell^r_t$
2. ℓ^l_t *is a supermartingale:* $\mathbb{E}[\ell^l_{t+1} \mid \ell^l_t, \ell^r_t] < \ell^l_t$.

Despite Lemma 10, ℓ^l_t must eventually move away from 0. For a contradiction, suppose ℓ^l_t remained near 0 for all t. As right-type beliefs move toward greater uncertainty, the frequency at which right types choose B increases, and the observed frequency of B increases in expectation. Since the only fixed points of $\langle \ell^r_t \rangle$ are 0 and ∞, ℓ^r_t diverges to infinity, implying the frequency of B converges to 1. But a left type is aware she is in the minority; an arbitrarily long herd on B must cause her to eventually think $\omega = L$. This logic makes clear that while right-type beliefs move from favoring $\omega = R$ to $\omega = L$, the resulting behavior causes left-type beliefs to follow. Once all agree that $\omega = L$ is most likely, the logic repeats, sending right-type beliefs back toward uncertainty. No matter which state society agrees on, no action ever gains as much support as the majority anticipates—the majority never grow confident their optimal action.

[54] In this setting, a *contrarian* action is defined relative to an individual's belief: action X_{nt} is contrarian if it's the action least likely observed according to an observer with belief ℓ^θ_t.

Figure 1.5 shows the expected drift in biased beliefs for all regions of the joint-belief space. Beliefs drift away from each fixed point. But they do so in a particular way: play near each potential equilibrium reinforces the beliefs of some types, while deteriorating the beliefs of others.

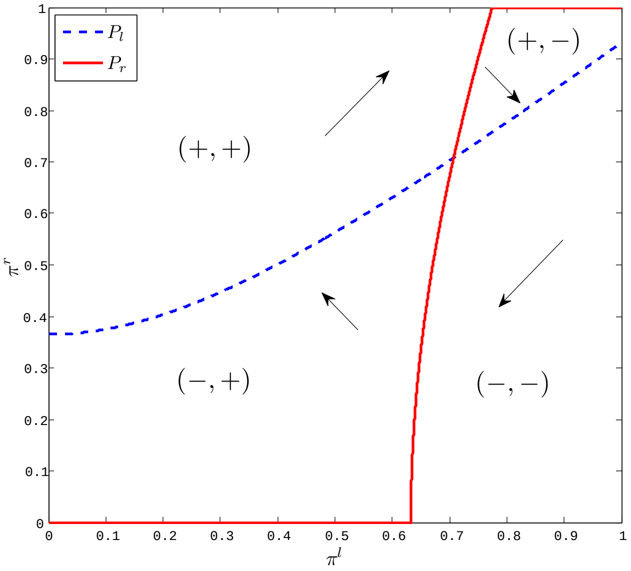

Figure 1.5: *Belief "phase diagram" for "weak" taste projection.*

Weak projection generates persistent opinion fluctuations. Given the pattern of drift shown in Figure 1.5) and that there exists no stable limit point, each type's beliefs enter of the four regions of belief space infinitely often. As a consequence, for any degree of weak projection and each $\theta = l, r$, $\langle \pi_t^\theta \rangle$ crosses 0.5 infinitely often. Society forever alternates between supporting $\omega = R$—where most people choose A—and supporting $\omega = L$—where most choose B. Behavior resembles "fad" behavior, where spells in which A is most popular are followed by those in which B is most popular. Although ubiquitous, such behavior is not well explained by rational learning models in settings with strongly connected networks or "unbounded" private information. For example, Çelen and Kariv (2004) show that if rational players observe only immediate predecessors and receive private signals with bounded informativeness, then fads can occur. Acemoglu, Como, Fagnani and Ozdaglar (2012) suggest an alternative model for persistent oscillations in the public opinion. They explore learning in a network where some agents are "stubborn" and never update their beliefs. Acemoglu, Como, Fagnani, and Ozdaglar (2012) suggest that such models help explain persistent fluctuations in political opinion, documented by Kramer (1971) and Cohen (2003). In my model, the public is surprised how little support a policy receives, rationalizing that if the policy was in fact optimal for the majority, it would garner more support. But when society changes its mind, the alternative policy also fails to earn sufficient support. People perpetually misinterpret the "surprising" amount of heterogeneity in choice.

As with strong projection, weak projection can harm social welfare. This occurs whenever beliefs spend a significant proportion of time below $\frac{1}{2}$. To determine when this occurs, we must consider the long-run distribution of beliefs. The distribution depends on the relative magnitudes of the mispredictions, $\hat{\lambda}^l$ and $\hat{\lambda}^r$. Near $\hat{\boldsymbol{\ell}} = (0,0)$, $\hat{\lambda}^r$ dictates how quickly ℓ_t^r moves away from 0, while $\hat{\lambda}^l$ dictates how quickly ℓ_t^l moves *toward* 0. The dynamics are most interesting when $\hat{\lambda}^r$ is sufficiently close to λ. In particular, let

$$\bar{\lambda}^r(\hat{\lambda}^l, \lambda) = 1 - \left(\frac{1-\lambda}{\lambda}\right)\hat{\lambda}^l, \tag{1.22}$$

and consider the case when $\hat{\lambda}^r < \bar{\lambda}^r(\hat{\lambda}^l, \lambda)$. Simulations confirm that each belief process oscillates between increasingly confident beliefs in the two states.

Mispredictions of λ such that condition 1.22 holds—shown in Figure 1.6—lead to "cyclical beliefs." With parameters in this set, Figure 1.7 depicts a simulated sample path of $\log \ell_t^\theta$: beliefs of each type spend longer and longer near each confident fixed point before transitioning to the next. Additionally, Figure 1.8 shows this same path, but in "phase" space. The unstable orbit increases in its distance from neutral beliefs, $\boldsymbol{\ell} = (0,0)$. Notice from Figure 1.7, at most points in time, players are mutually confident in the same state. When confident in $\omega = R$, all choose optimally—all right types, a fraction λ, take A. But when confident in L, all choose incorrectly—all left types, a fraction $1-\lambda$ take A. Figure 1.9 displays this finding: the frequency at which players choose A oscillates between λ and $1-\lambda$. Finally, Figure 1.10 shows a sample path of $\log \ell_t^\theta$ for parameters that fail condition 1.22: the beliefs take on a much more stationary limit distribution.

When society exhibits "cyclical beliefs", expected welfare falls below the autarkic level. Roughly 50% of the time, each agent holds nearly confident, but false, beliefs and chooses the incorrect action. But when relying solely on private infromation, agents necessarily choose correctly more than 50% of the time. Observing others makes society worse off, on average.

1.4.5 Biased Learning Under General Taste Projection

The previous subsections draw out the implications of taste projection within the domain of "choice-dependent" projection. This imposed that all players on a particular side of the taste spectrum share the same perception of λ. The assumption essentially implies only two types—left and right. Here, I briefly characterize learning when misperceptions may differ across an arbitrary finite number of types. The only assumption on perceptions is Assumption 12: those with high θ hold perceptions that first-order stochastically dominate the perceptions of those with lower θ, so that $\hat{\lambda}(\theta)$ is monotonically increasing in θ.

I focus on conditions on perceptions that define when stable equilibria exist. Let W denote the share of types that wrongly think left types comprise the majority. Define $\tilde{\theta} = \arg\max_\theta \hat{\lambda}(\theta)$ subject to $\hat{\lambda}(\theta) < 1/2$; $\tilde{\theta}$ is the right-most type who believes left types comprise the majority. If $\tilde{\theta}$ exists, then $W = G(\tilde{\theta})$, otherwise $W = 0$. Let $\underline{\theta} = \min \Theta$ and $\overline{\theta} = \max \Theta$.

Proposition 8. *A stable equilibrium exists if and only if*

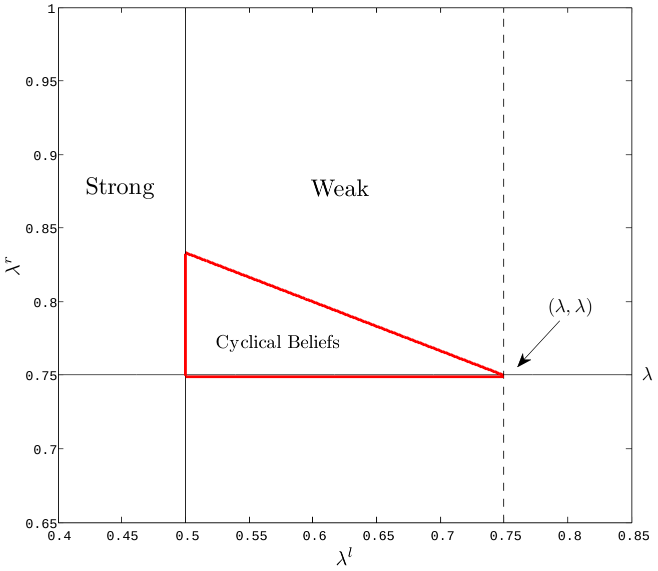

Figure 1.6: *Set of Weak-Projection parameters leading to cyclical beliefs.*

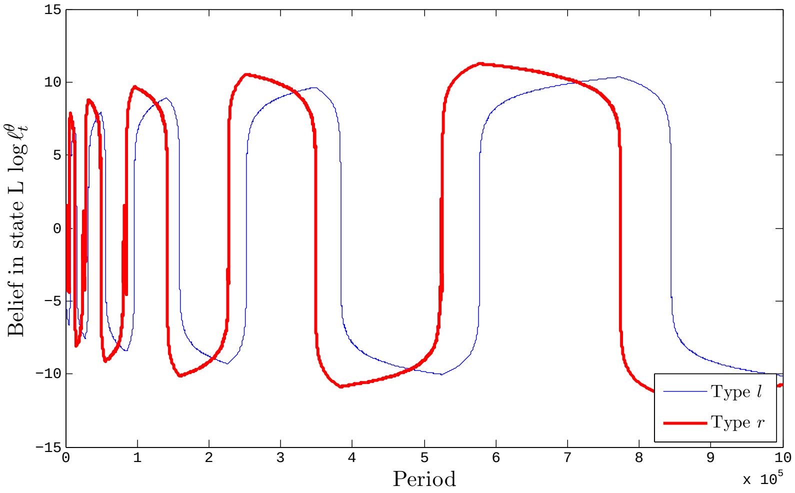

Figure 1.7: *Sample path of log-likelihood ratios for* $\lambda = 0.75$, $\hat{\lambda}^l = 0.55$, *and* $\hat{\lambda}^r = 0.8$.

1. $\tilde{\theta} < 0$ *and* $W + \lambda > \max\left\{1 - \hat{\lambda}(\underline{\theta}), \hat{\lambda}(\overline{\theta})\right\}$

2. $\tilde{\theta} > 0$ *and* $2 - (W + \lambda) > \max\left\{1 - \hat{\lambda}(\underline{\theta}), \hat{\lambda}(\overline{\theta})\right\}$

The left-hand side of each condition in Proposition 8 is the measure of agents who believe it is optimal to follow the majority action. The right-hand side is the most biased perception of the extent of the majority held by any agent. So long as *all* agents observe more people than they anticipated choosing a single action, then the equilibrium is stable. In any

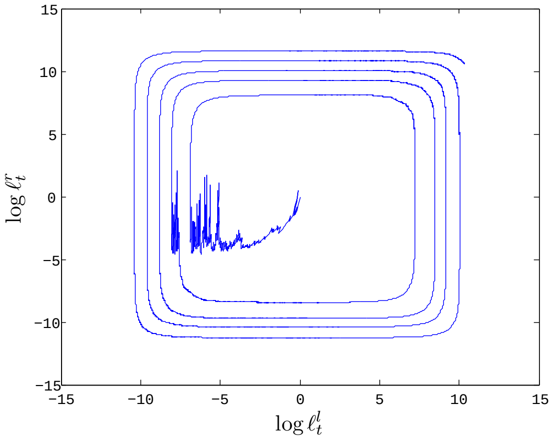

Figure 1.8: *Sample path of log-likelihood ratios for* $\lambda = 0.75$*,* $\hat{\lambda}^l = 0.55$*, and* $\hat{\lambda}^r = 0.8$ *in phase space.*

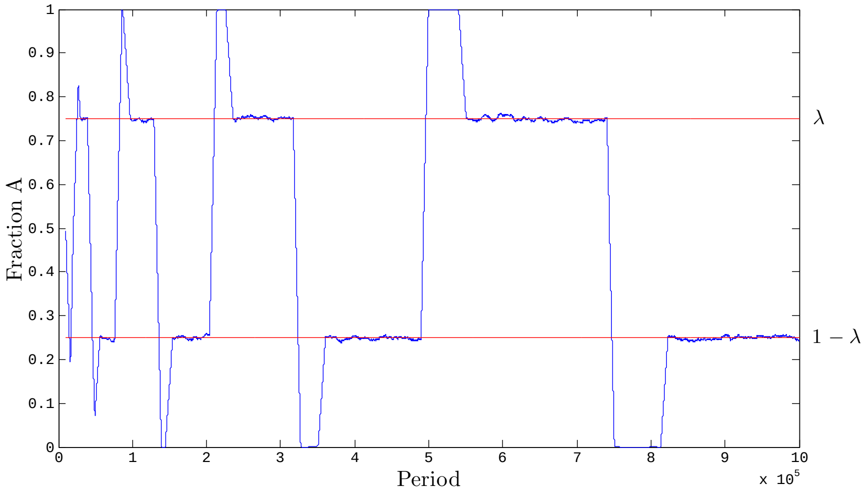

Figure 1.9: *Sample path of log-likelihood ratios for* $\lambda = 0.75$*,* $\hat{\lambda}^l = 0.55$*, and* $\hat{\lambda}^r = 0.9$*.*

stable equilibrium, it is always the extreme types (those far from indifferent) who (rightly or wrongly) follow the majority action. They are the types who most overestimate how many share their taste. It is those with weak preference over location who concede that they have less-common preferences and choose the minority action. Turning to equilibrium beliefs, $\tilde{\theta}$ represents a turning point in beliefs: all types to one side of $\tilde{\theta}$ agree on the state, while those on opposite sides disagree.

Proposition 8 generalizes the findings of "strong" and "weak" projection to a broad class of taste-dependent perceptions, $\hat{\boldsymbol{\lambda}}$. Strong projection implies *all* agents choose identically. Here, relative to the efficient outcome, any stable equilibrium requires that too many agents adopt the popular action. "Over-adoption" of the majority choice is the general implication

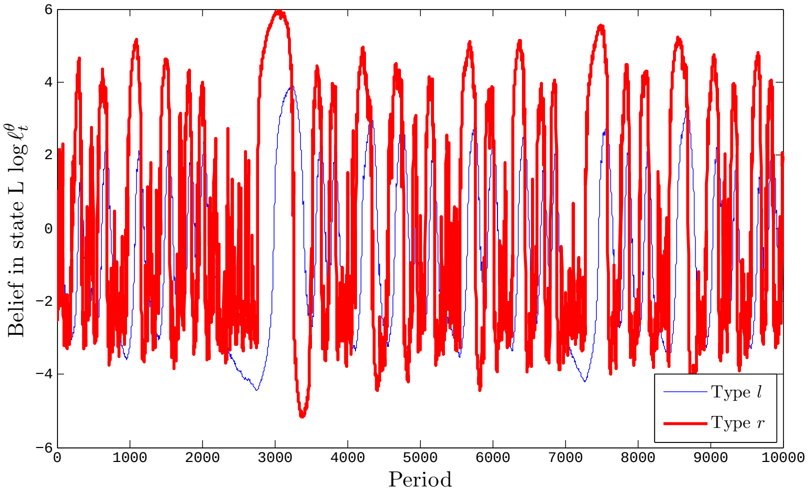

Figure 1.10: *Sample path of log-likelihood ratios for* $\lambda = 0.75$*,* $\hat{\lambda}^l = 0.55$*, and* $\hat{\lambda}^r = 0.9$*.*

of a stable projection equilibrium; strong projection demonstrates a particularly limit case in which *all* choose a single action. Additionally, so long as each type correctly recognizes the majority preference, $\hat{\lambda}(\theta) > \frac{1}{2}$ for all $\theta \in \Theta$, then results match those of the "weak" projection case: there exist no stable equilibrium beliefs. As such, the two-type examples of "strong" and "weak" accurately capture the essence of learning with projection, albeit in extreme fashion.

1.5 Learning About Quality

While the previous sections analyze an environment where the commonly-valued quality of A and B are known, there are natural settings with uncertainty about horizontal location *and* quality. This section assumes quality differences are highly uncertain and potentially large enough that all players prefer a single option. While in Sections 1.4 and 1.4, beliefs about location dictated choice, here quality concerns may dominate decisions. How might taste projection distort inference about quality?

To build on examples above, consider learning about the prospects of investing in foreign real estate. Countries vary drastically on the fixed costs associated with acquiring real estate—both in taxes and fees charged by local intermediaries that buy the property. These fees may be difficult to assess prior to going forth with the investment, so agents glean information from past choices. If they find a country has very low fees, all may prefer this investment irrespective of risk preferences. I also discuss an example where farmers learn which seed is most productive on their plot—productivity depends on inputs, like soil type, but it may be that some seed is universally most productive, irrespective of inputs.

Perceived quality is always mislearned. With two types, if people suffer strong taste projection, then society necessarily comes to believe in the largest possible quality difference. Differences in vertical quality are always weakly exaggerated. This mislearning both arises from, and perpetuates, uniform herding, where all agents choose the action with the perceived quality advantage. With a continuum of types, if long-run behavior is stable and heterogeneous—there exist some types that prefer different actions—then the equilibrium must display a particular form of mislearning about quality differences: people underestimate the quality of the option close to their location relative to those far from it.

1.5.1 Preliminaries

States. There are now two dimensions of uncertainty associated with each action: (1) quality, and (2) location—whether A is to the left or right of B. Recall from Section 1.2, the unknown state is characterized by $(\zeta, \Delta_q) \in \{L, R\} \times \mathcal{D}$. Let the minimal and maximal quality differences be $\underline{\Delta} := \min \mathcal{D}$ and $\overline{\Delta} := \max \mathcal{D}$. Since $u(q^X, z^X) = q^X - k(z^X - \theta)^2$, all agents prefer A over B if $\Delta_q > \widehat{\Delta} := 4k\overline{\theta}$, where $\overline{\theta} = \max \Theta$. Likewise, all prefer B if $\Delta_q < -\widehat{\Delta}$. I call a state *aligned* when the quality of an option trumps any concern for horizontal differentiation, so all prefer the same action. Such states occur so long as $\overline{\Delta} > \widehat{\Delta}$ or $\underline{\Delta} < -\widehat{\Delta}$. If society grows confident in some state ω, players expect to observe uniform behavior if and only if ω is aligned; otherwise, agents anticipate heterogeneous behavior.

Private Information and Public Beliefs. Despite multi-dimensional uncertainty, I assume a simple uni-dimensional signal structure. Players receive signals informing them which action is optimal for their own taste. For each θ, let $\Omega_\theta \subset \Omega$ denote the set of states in which it is optimal for a θ-type to take action A. Hence, each θ-type receives signals indicating whether $\omega \in \Omega_\theta$; her private belief that $\omega \in \Omega_\theta$ has distribution F_A whenever $\omega \in \Omega_\theta$, and otherwise has distribution F_B. F_A and F_B meet precisely the same assumptions as F_H and F_L, respectively (Assumptions 5 and 6). I assume this signal structure simply for ease of exposition: agents follow decision rules analogous to those derived in Section 1.3. The structure still allows rational agents to learn the optimal action, and I emphasize below that it does not drive any incorrect-learning results.

While private information alone leads to coarse inference over Ω, this signal structure implies agents achieve finer inference when observing others—agents can discern which action is optimal for each type. Let $\pi_t^\theta(\zeta, \Delta_q)$ denote a θ-type's perception of the public belief that $\omega = (\zeta, \Delta_q)$ in t; let $\pi_1(\omega)$ denote the common prior for ω.

1.5.2 Biased Learning with Two Types

Suppose only two types, $\Theta = \{-1, 1\}$—a left type ($\theta = -1$) and a right type ($\theta = 1$)—and the probability any agent is a right type is $\lambda := \Pr(\theta_{nt} = 1)$.[55] In this setting, if people suffer

[55] For continuity with previous sections, I use the superscript l to denote perceptions held by $\theta = -1$ and r for those of $\theta = 1$.

strong taste projection, then society necessarily comes to believe in a state in which quality dominates preferences. Differences in vertical quality are always weakly exaggerated.

With two types, observing actions leads players to hold public beliefs over a partition of Ω comprised of four subsets—players can only determine which action is optimal for each type. These subsets are

$$\begin{aligned}
\Omega^{AA} &= \{\omega = (\zeta, \Delta_q) \mid \Delta_q \geq \widehat{\Delta}\}, \\
\Omega^{AB} &= \{\omega = (\zeta, \Delta_q) \mid \Delta_q \in (-\widehat{\Delta}, \widehat{\Delta})\ ,\ z^A = L\}, \\
\Omega^{BA} &= \{\omega = (\zeta, \Delta_q) \mid \Delta_q \in (-\widehat{\Delta}, \widehat{\Delta})\ ,\ z^A = R\}, \text{ and} \\
\Omega^{BB} &= \{\omega = (\zeta, \Delta_q) \mid \Delta_q \leq -\widehat{\Delta}\}.
\end{aligned}$$

The set $\Omega^{XX'}$ contains all the states where it is optimal for left types to chose X and right types to choose X'. For instance, Ω^{AB}—the set of states where A is on the left and actions have intermediate quality values—is the set of states where left types prefer A, and right prefer B. Let $\pi_t^{\theta}(XX')$ denote a θ-type's belief that the state is in set $\Omega^{XX'}$ after observing history h_t.

The following proposition characterizes long-run beliefs, and demonstrates that left and right types never agree on the location state, but always agree that one action has superior quality.

Proposition 9. *Suppose players suffer strong taste projection and $\overline{\Delta} > \widehat{\Delta}$ and $\underline{\Delta} < -\widehat{\Delta}$.*

1. *Any confident joint belief with $\pi^r(AB) = \pi^l(AB) = 1$ or $\pi^r(BA) = \pi^l(BA) = 1$ is unstable: agents never agree on the location state in the long run.*

2. *Suppose the number of players per period is arbitrarily large, $N \to \infty$, and agents observe only those in the previous period. If in truth $\omega \in \Omega^{BA} \cup \Omega^{AA}$, then for each θ, $\pi_t^{\theta}(AA) \to 1$. Otherwise, $\pi_t^{\theta}(BB) \to 1$ for each θ. Agents necessarily converge to an aligned state.*

Part 1 of Proposition 9 follows from the stability criteria established in Proposition 4. The logic is identical to learning under strong taste projection absent quality differences (Section 1.4.3)—agents never agree on the location state, and instead form fully-polarized beliefs over ζ. Whenever the majority chooses A, left types come to believe $\zeta = L$, while right types conclude $\zeta = R$. As usual, all agents believe A is optimal; a uniform herd on A emerges.

Part 2 of Proposition 9 is a consequence of how agents explain this uniform herd. Variable quality allows agents to make sense of the herd: when all choose A, it must be that A simply has superior quality, $\Delta_q > \widehat{\Delta}$. The observation structure, where the number of agents each period is infinite but players observe only the behavior of the previous generation, is assumed simply to make precise claims about limit beliefs—limit beliefs are deterministic.

This result implies society necessarily concludes that quality differentiation trumps horizontal differentiation. And this is independent of priors: even as the prior of an aligned state becomes arbitrarily small, $\Pr(\omega \in \Omega^{AA} \cup \Omega^{BB}) \to 0$, people still conclude some option has such a large quality advantage that all should choose it irrespective of preferences. This has important consequences in markets with niche goods. Even when some product or technology is optimal for a minority of consumers, low demand is misinterpreted as a signal of poor quality.[56]

Consider an example where farmers learn whether to adopt new hybrid rice (A) over the status-quo crop (B). Suppose farmers fall into two categories, those with high-salinity soil ($\theta = r$) and those with low ($\theta = l$). It is known that the hybrid seed is sensitive to salinity, but the direction is unknown. The status-quo crop is insensitive. Further, farmers are uncertain about the potential yield of the new seed. It's possible that even when sowed on suboptimal soil, the hybrid may trump the alternative. Suppose in truth that this is not the case; the new seed is only beneficial for low-salinity farms, which account for 40% of the region's farms. Before investing in the new crop, farmers cultivate a small test plot—they have noisy signals if the new seed is a good match with their farm. Initial adoption in $t = 1$ is based on this private information. In $t = 2$, farmers use both private information and the fraction of neighbors that previously adopted, say roughly 40%. If both low- and high-salinity farms perceive themselves as the majority, then both types find the initial demand too weak to adopt. The next period, new farmers learn that none of those from the previous generation adopted the new seed. The only reasonable conclusion is that the yield is inferior to the status quo, irrespective of variation across farms: A is suboptimal for *all* farmers.

With more general signal structures, society always overestimates quality differences, even when it is commonly known that agents *never* all prefer the same choice; that is, in cases where quality is necessarily in the intermediate range $\Delta_q \in (-\widehat{\Delta}, \widehat{\Delta})$ so $\Omega^{AA} = \Omega^{BB} = \emptyset$. This follows immediately from the logic of Proposition 9, Part 1. If agents receive independent private information about the quality and location of each option, then Proposition 9, Part 1 remains true. But beliefs can be stable when left and right types disagree on location. If this is so, all agents choose an identical action. While agents understand that no possible state leads to such behavior, they must decide which state best explains it. So long as signals about quality follow the monotone likelihood-ratio property so that increasing Δ_q increases the chance an agent chooses A, then a herd on A is best explained by $\Delta_q = \overline{\Delta}$. Society comes to believe A attains the largest possible quality advantage.

The model above rules out learning from own past experience by assuming each agent makes a single choice. While allowing for multiple choices over time complicates the inference problem, it does not necessarily imply that taste projection has no effect on learning. So long as feedback from experience is stochastic and the number of predecessors any agent observes is large relative to the number of choices she makes, then the (misinterpreted) public information can dominate personal experience. If society wrongly concludes that large

[56] An older literature in industrial organization attempts to explain how social learning may deteriorate the market share of niche goods. See McFadden and Train (1996).

crowds at a restaurant imply high-quality food and service, when in reality the restaurant has mediocre quality but serves the majority's preferred style of cuisine, then a diner may continue visiting the restaurant despite many bad experiences. The diner is convinced that in this information-rich environment, the only way such a herd could persist is if quality is high; she concedes that she simply has bad luck.

1.5.3 Biased Learning with Many Types

With many types, there may exist long-run stable equilibria in which society learns the true location state and behavior remains heterogeneous. While agents may hold correct beliefs along the horizontal dimension, such equilibria entail an interesting form of mislearning along the quality dimension. Instead of universally concluding one action has superior quality, agents disagree on quality in a particular way. If all agents are confident that action A best suits right-leaning tastes ($\zeta = R$), then, relative to left types, right types conclude A has *low* quality. Those with innate taste for an option develop a relatively pessimistic view of its quality.

The logic is simple. Consider agents learning about two films newly released at the local cinema. They rightfully conclude that film A is an action film, but B is a romance. Upon observing box-office sales, action fans—who overestimate the share of action fans—find the attendance to A surprisingly low, while romance aficionados find it surprisingly high. Action fans attribute this to quality: the action movie must have limited quality if so many are passing in favor of the romance film. But romance fans think precisely the opposite.

The previous subsection and this one reiterate an important point: taste projectors must disagree on some dimension in order to explain observed behavior. If they agree on quality, then they must disagree on location, and vice versa.

To show this result formally, I assume choice-dependent projection (Definition 2), and assume a continuum of types: $\Theta = [\underline{\theta}, \overline{\theta}]$. Although there are many types, there are only two distinct perceptions of the taste distribution: $\widehat{G}_l$ held by $\theta < 0$ and $\widehat{G}_r$ held by $\theta > 0$, and right-type perceptions dominate left in the sense of FOSD. I argue that the logic of the equilibria discussed here extends to the case where each type holds a distinct perception in an intuitive way. I also assume the number of players each period is large so that the fraction choosing A each round, denoted $\alpha_t = a_t/N$, is a deterministic function of beliefs and the state.

Suppose all agents agree on the location state ζ, and with out loss of generality, $\zeta = R$. Let Δ_q^l and Δ_q^r denote perceived quality differences held by left and right types, respectively. Does there exist (Δ_q^l, Δ_q^r) that jointly rationalizes fraction α choosing A each period? Types that choose A are those with $\theta > \hat{\theta}$, where $\hat{\theta}$ denotes the agent indifferent between A and B. In state (R, Δ_q), the marginal type is $\hat{\theta} = -\Delta_q/4k$. Let $\hat{\theta}^l$ and $\hat{\theta}^r$ denote each type's perceived marginal agent. Equilibrium requires $\alpha = 1 - \widehat{G}_l(\hat{\theta}^l)$ and $\alpha = 1 - \widehat{G}_r(\hat{\theta}^r)$, which in turn implies

$$\begin{aligned}\Delta_q^l &= -4k\widehat{G}_l^{-1}(1-\alpha),\\ \Delta_q^r &= -4k\widehat{G}_r^{-1}(1-\alpha).\end{aligned}\tag{1.23}$$

Since $\widehat{G}_r(x) \leq \widehat{G}_l(x)$, it follows that $\Delta_q^r \leq \Delta_q^l$. Summarizing in a proposition, we have the following.

Proposition 10. *Suppose a continuum of types suffer choice-dependent projection. If agents agree that A is optimal for right-leaning tastes ($\zeta = R$), then right-leaning agents have a lower perception of A's quality than do left-leaning agents: $\Delta_q^r \leq \Delta_q^l$.*

In general, with perceptions that can vary for each θ, the equilibrium requirement is $\alpha = 1 - \widehat{G}(\hat{\theta}(\theta) \mid \theta)$ for all θ, where $\widehat{G}(\cdot \mid \theta)$ is a θ-type's perceived distribution and $\hat{\theta}(\theta)$ is her perception of the marginal agent. This condition does not necessarily hold—existence requires the degree to which $\widehat{G}(\cdot \mid \theta)$ varies across θ to be small.[57] However, in any such equilibrium, it's clear that perceptions of the relative quality of A are decreasing in type. $\alpha = 1 - \widehat{G}(\hat{\theta}(\theta) \mid \theta)$ implies $\hat{\theta} = \widehat{G}^{-1}(1-\alpha \mid \theta)$, and using $\hat{\theta} = -\Delta_q(\theta)/4k$ yields

$$\Delta_q(\theta) = -4k\widehat{G}^{-1}(1-\alpha \mid \theta),\tag{1.24}$$

where $\Delta_q(\theta)$ denotes a θ-type's perception of the quality difference. By first-order stochastic dominance (Assumption 12), $\widehat{G}^{-1}(1-\alpha \mid \theta)$ is increasing in θ, so $\Delta_q(\theta)$ is decreasing in θ.

1.6 Learning About Preferences

This section explores learning about horizontal differentiation, as in Sections 1.3 and 1.4, among agents who revise their models of others' preferences after observing actions. In Sections 1.3 and 1.4, agents have fixed perceptions: they believe the distribution of tastes (which they mispredict) is perfectly known by all agents.[58] Was this restrictive assumption responsible for errors in long-run learning? This section considers a more realistic model where all agents perceive some uncertainty over the distribution, and learn about others' tastes through their actions. When the true taste distribution lies in the support, does updating their models ameliorate agents' mislearning of payoffs? If agents are naive—they neglect that different types start at different priors—then the answer is no. Specifically, agents with different tastes *rationally* form divergent priors over the distribution. A naive agent only errs by wrongly assuming all others share her prior, and, hence, she develops incorrect beliefs about what other types infer. It is not heterogeneous priors, per se, that

[57] General conditions on the collection of perceptions that guarantee such an equilibrium are beyond the scope of this paper.

[58] While this sounds dogmatic, this assumption forms the premise of many Bayesian games, including the canonical model of Smith and Sørensen (2000).

leads agents astray, but rather that they neglect others' discrepant beliefs. I show that a particular class of priors can cause agents to become fully biased in their perceptions of others tastes: they wrongly conclude society shares a common preference.

Subsection 1.6.1 extends the model and defines taste projection in a setting with uncertainty. For the sake of demonstrating how naivete can generate incorrect learning even when agents put positive weight on the true environment, I consider the most simple variant of the model. Within this setting, Subsection 1.6.2 explores properties of biased long-run learning.

1.6.1 Extension of the Model

Consider the model of Section 1.3 and 1.4 with no uncertainty over quality nor quality differences, $\Delta_q = 0$. Further, assume only two types, $\Theta = \{-1, 1\}$—a left type ($\theta = -1$) and a right type ($\theta = 1$); the probability any agent is a right type is $\lambda := \Pr(\theta_{nt} = 1)$. Learning about the taste distribution entails learning a single parameter, λ.

Public and Private Beliefs. Suppose that λ is a random draw from distribution μ_0 on $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_K\}$ with $\underline{\lambda} = \min \Lambda$ and $\overline{\lambda} = \max \Lambda$. The state space is now $\{L, R\} \times \Lambda$, consisting of payoff states, $\omega \in \{L, R\}$, and distribution states, λ_k. Let $\pi_t^\theta(\omega, \lambda_k)$ denote a θ-type's public belief that the state is (ω, λ_k) after observing h_t. Without loss of generality, suppose the state is (R, λ^*) for some $\lambda^* \in \Lambda$, and let $\ell_t^\theta(\omega, \lambda_k)$ denote the likelihood ratio of (ω, λ_k) relative to the truth (R, λ^*). Correct learning entails $\ell_t^\theta(\omega, \lambda_k) \to 0$ for all $(\omega, \lambda_k) \neq (R, \lambda^*)$. Finally, let the conditional distributions of private beliefs, F_ω, meet Assumptions 5 and 6 from Section 1.3.

Priors. Assume $\Pr(\omega = R) = 1/2$. In truth, λ has distribution μ_0. Importantly, with uncertainty over population preferences, one's taste is information about λ. Learning θ causes an agent to revise her prior μ_0. I model taste projection as a biased perception of revised priors over λ. Specifically, in period 0, each player learns her taste θ, and uses it as a proper Bayesian to update prior $\mu_0(\lambda_k)$ to posterior $\mu^\theta(\lambda_k) = \Pr(\lambda_k \mid \theta)$. But naive agents neglect that players who receive conflicting signals (i.e., tastes) arrive at different priors: a θ-type thinks *all* players share her prior μ^θ regardless of their tastes.[59] This is the only way in which a θ-type's model is misspecified: she has a perfectly rational theory of how λ is distributed, but an incorrect theory of what others think. Explicitly, left and right types derive priors μ^l, and μ^r, respectively, from μ_0 where[60]

$$\begin{aligned}
\mu^r(\lambda_k) &:= \Pr(\lambda_k \mid \theta = r) = \frac{\lambda_k \mu_0(\lambda_k)}{\sum_i \lambda_i \mu_0(\lambda_i)} = \left(\frac{\lambda_k}{\mathbb{E}[\lambda]}\right) \mu_0(\lambda_k), \\
\mu^l(\lambda_k) &:= \Pr(\lambda_k \mid \theta = l) = \frac{(1 - \lambda_k)\mu_0(\lambda_k)}{\sum_i (1 - \lambda_i)\mu_0(\lambda_i)} = \left(\frac{1 - \lambda_k}{1 - \mathbb{E}[\lambda]}\right) \mu_0(\lambda_k). \qquad (1.25)
\end{aligned}$$

[59] This assumption is similar to Madarasz's (2012) model of "information projection". A θ-type forms beliefs as if her private-taste signal was publicly observed by all agents. But she also projects ignorance: she neglects that other agents may receive contradictory information.

[60] For continuity with previous sections, I use the superscript l for $\theta = -1$ and r for $\theta = 1$.

I make the admittedly strong assumption that an agent thinks others' priors match exactly her own.[61] It will become clear that this is stronger than necessary, and I assume this only because the error is particularly simple. All that is necessary in order for agents to mislearn with positive probability is that an agent (unknowingly) infers too much (relative to a Bayesian) from her private taste, which means that agents underappreciate the extent to which priors differ across types.

Decision Making and Updating. While beliefs about λ dictate the interpretation of actions, an individual's decision depends only on her belief about whether A is to the right. This belief, the marginal probability of $\omega = R$, is denoted by $\pi_t^\theta := \sum_k \pi_t^\theta(R, \lambda_k)$, and let $\ell^\theta = (1 - \pi^\theta)/\pi^\theta$ denote the likelihood ratio that $\omega = L$ relative to $\omega = R$; agents follow the same decision rule as in Section 1.3 (Lemma 3). Since a naive agents thinks all share a common prior, she thinks all types share her public belief $\pi_t^\theta(\omega, \lambda_k)$ in each state, for all t.

After observation $a \in \{0, 1, ..., N\}$ in period t, type-θ beliefs update according to

$$\ell_{t+1}^\theta(\omega, \lambda_k) = \ell_t^\theta(\omega, \lambda_k) \frac{\psi_\theta(a \mid \ell_t^\theta, \omega, \lambda_k)}{\psi_\theta(a \mid \ell_t^\theta, R, \lambda^*)}, \tag{1.26}$$

where $\psi_\theta(a \mid \ell^\theta, \omega, \lambda_k)$ is the probability of observing a in state (ω, λ_k) given ℓ_t^θ according to type-θ. The key difference between rational and naive updating is that a rational player has correct second-order beliefs—she knows that left and right types have different beliefs—so, in a rational model, ψ depends on both ℓ_t^l and ℓ_t^r.

1.6.2 Biased Long-Run Learning

This subsection shows that incorrect second-order beliefs over λ arising from naivete can generate polarized beliefs about location and tastes. For some priors, agents disagree on the interpretation of actions, causing left types to grow confident that $\omega = L$ while right types grow certain that $\omega = R$. With polarized beliefs about payoffs, a uniform herd develops—all agents take the same action. To explain this herd, agents' perceptions of others' tastes also become polarized: each interprets the herd as an indication that her taste is maximally common. I provide sufficient conditions on priors guaranteeing that such an outcome occurs with positive probability.

In general, uniform herding can occur whenever an arbitrarily long herd on some action, say A, is polarizing—left and right types always (unknowingly) disagree on the interpretation of A no matter how often it is played. The herd on A leads left types to believe $\omega = L$, and right types to believe $\omega = R$.

A simple test determines whether agents may, with positive probability, converge to precisely opposite beliefs. Within a given environment, suppose people act in single file.

[61] This notion of naivete is entirely consistent with the earlier definition in Assumption 13. A more general definition of naivete that extends to settings with uncertainty is that all players think each agent shares her prior over the taste distribution. Then Assumption 13 follows from this definition in settings with no uncertainty—where the prior is degenerate—as in previous sections.

Let $\pi_t^\theta(h_t)$ be a θ-type's belief in $\omega = R$ entering period t following history h_t. Let h_t^A be a history of length $t-1$ consisting of all A's. Then fully-polarized beliefs occur in this environment if for all $t \in \mathbb{N}$, $\pi_{t+1}^r(A, h_t^A) > \pi_t^r(h_t^A)$ and $\pi_{t+1}^l(A, h_t^A) < \pi_t^l(h_t^A)$. Right-type beliefs are monotonically increasing along paths of all A's and left-type beliefs are monotonically decreasing along such paths.

Although rational beliefs never satisfy this condition—eventually agents must agree on the interpretation of A—they may be polarized by *finite* sequences of A's. For example, suppose the fraction of risk-averse agents is $\lambda \in \Lambda = \{\frac{1}{4}, \frac{3}{4}\}$. Suppose rational people agree on prior $\mu_0(3/4) = 0.6$. Then, after learning one's own risk preference, a risk-averse agent thinks $\lambda = 3/4$ with approximately 0.82 chance. But a risk-neutral agent thinks $\lambda = 1/4$ with chance 2/3. That is, initially, before observing any actions, each type *rationally* believes her taste is most common. If the first player chooses A, each agent reasons that the first player likely shared her taste. Thus, a risk-neutral agent believes A is likely risky, while a risk-averse believes A is likely safe. Observing A polarizes *rational* agents' beliefs over the riskiness of A.

But so long as agents are rational, A cannot forever polarize beliefs. After a sufficiently long string of A's, an additional A must move rational players' beliefs in the same direction. Why? Agents have correct second-order beliefs: they know exactly what people of opposite tastes believe. A rational agent cannot simultaneously grow confident of some hypothesis while fully aware that another rational agent is confident of another after observing precisely the same information.[62] In the example above, observing a second A tells everybody very little—all know that each type likely chooses A irrespective of private information. After a long enough sequence of A's, people eventually rely on the original prior μ_0 to draw conclusions, not their taste dependent prior. In the example above, all people eventually agree that a long sequence of A's is strong evidence for $(R, 3/4)$. Beliefs of the two types eventually grow close.

This logic needn't hold for biased agents. In the example above, a risk-averse agent thinks all share her initial belief that $\Pr(\lambda = 3/4) = 0.82$. She neglects the fact that risk-neutral agents—who think $\Pr(\lambda = 3/4) = 1/3$—initially disagree on the interpretation of A. As more A's are played, beliefs converge toward opposite payoff states. However, agents are not aware that other types are learning differently: risk averse think all agree that the sequence is indicative of $(R, 3/4)$, while risk neutral think all agree that $(L, 1/4)$ is most likely.

In general, if actions can have a lasting polarizing effect on the two types, then with positive probability agents with different tastes converge to confident beliefs in opposite payoff states, and a uniform herd results. Why? If the game starts with a long sequence of A's, which occurs with positive probability, then agents' beliefs grow polarized. Starting from this "initial condition", where people unknowingly disagree on the state, most continue to choose A—risk neutral are confident A is risky, and risk averse are confident it's safe. Crucially, these polar-opposite beliefs are stable: they lead all agents to play A with high probability,

[62] Acemoglu, Chernozhukov, and Yildiz (2007, 2009) show that rational agents may "agree to disagree" on the interpretation of an infinite sequence of evidence, however, they never fully disagree.

which only reinforces these beliefs. Thus, so long as beliefs can reach a neighborhood of the polar-opposite beliefs—which occurs with positive probability whenever A has a lasting polarizing effect on agents—then society may forever remain at these polar beliefs.

1.6.2.1 Two-Point Distributions

I first demonstrate mislearning in the simple case where, like the example above, λ takes one of two values. Suppose $\Lambda = \{\underline{\lambda}, \overline{\lambda}\}$ with $\underline{\lambda} < \overline{\lambda}$. The following lemma establishes what a player comes to believe after observing an arbitrarily long herd on A as a function of her prior, assuming she believes all people share this prior. In this setting with $|\Lambda| = 2$, let μ^θ denote a θ-type's perceived probability that $\lambda = \overline{\lambda}$.

Lemma 11. *Suppose $\Lambda = \{\underline{\lambda}, \overline{\lambda}\}$. For any $\underline{\lambda} < \frac{1}{2} < \overline{\lambda}$, there exists a value $\hat{\mu}(\underline{\lambda}, \overline{\lambda}) \in (0,1)$ such that $\mu^\theta < \hat{\mu}(\underline{\lambda}, \overline{\lambda})$ implies $\lim_{t\to\infty} \pi_t^\theta(h_t^A) = 0$ and $\mu^\theta > \hat{\mu}(\underline{\lambda}, \overline{\lambda})$ implies $\lim_{t\to\infty} \pi_t^\theta(h_t^A) = 1$.*

Lemma 11 implies that if agents are initially sufficiently confident that $\lambda = \overline{\lambda}$, then a herd on A indicates $(R, \overline{\lambda})$. But if μ^θ is low, the herd indicates $(L, \underline{\lambda})$. Hence, whenever agents have priors that fall on opposite sides of $\hat{\mu}(\overline{\lambda}, \underline{\lambda})$, the two types disagree on the interpretation of an arbitrarily long herd. However, if $\underline{\lambda} > \frac{1}{2}$ or $\underline{\lambda} < \frac{1}{2}$, so that both $\underline{\lambda}$ and $\overline{\lambda}$ lie on the same side of $\frac{1}{2}$, then the two types always agree on the interpretation of a herd. The logic of Lemma 11 implies the following mislearning result.

Proposition 11. *Suppose $\mu^l < \hat{\mu}(\underline{\lambda}, \overline{\lambda}) < \mu^r$. With positive probability, $\pi_t^r(R, \underline{\lambda}) \to 1$ and $\pi_t^l(L, \overline{\lambda}) \to 1$.*

Agents grow fully polarized along both dimensions on which they learn: they disagree on the payoff state, and each type of agent thinks *all* others share her taste. In the next subsection, I show that this same phenomenon occurs for more general distributions of λ, and discuss the intuition and significance of these results.

1.6.2.2 General Distributions

I now discuss when this logic holds for any general support $\Lambda \subseteq [0,1]$. While it has not been shown formally, I believe the following conjecture provides a sufficient condition for polarized long-run beliefs.

> *Suppose $\Lambda = [0,1]$. Suppose type-dependent prior μ^l is strictly decreasing on Λ, and μ^r is strictly increasing on Λ. If the number of players each round is arbitrarily large, $N \to \infty$, then $\pi_t^r(R, \overline{\lambda}) \to 1$ and $\pi_t^l(L, \underline{\lambda}) \to 1$. Actions converge on option A.*

The intuition is as follows. Suppose the truth is (R, λ^*) with $\lambda^* > \frac{1}{2}$. First period actions a_1 collapse beliefs onto the truth and (L, λ') for some $\lambda' < \frac{1}{2}$. Type-r believes (R, λ^*) is most likely, and type-l believes (L, λ') is most likely. In period 2, net of private information, each type believes A is optimal. And, since agents thinks their beliefs are commonly shared, each expects a player with taste different than her own to choose B. Agents neglect the fact that *all* have incentive to choose A. Thus, a_2 exceeds what any player expects to see in either state. Given monotonic priors, the most likely explanation for this unexpectedly-high outcome within a right-type's model is that $\lambda > \lambda^*$. Within a left-type's model, the most likely explanation is $\lambda < \lambda'$. That is, a_2 polarizes the agent's beliefs about λ: a right type's estimate moves toward 1, while a left-type's estimate moves toward 0. Increased polarization implies still more choose A in round 3—$a_3 > a_2$, and polarization increases further. In general, $a_{t+1} > a_t$ for all t, and $a_t/N \to 1$. In the long-run, all choose A. Type-r thinks $(\omega, \lambda) = (R, 1)$ and type-l believes in $(\omega, \lambda) = (L, 0)$.

Similar to Section 1.5, where agents explain a uniform herd by assuming one option has superior quality, agents here explain the herd by assuming common preferences. A risk-averse agent's best explanation for why all invest in A is that A is safe and all have preferences similar to her own. A risk-neutral agent concludes precisely the opposite. The equilibrium is essentially self confirming: agent's incorrect beliefs never generate evidence inconsistent with these beliefs.

The fact that perceptions of population tastes become fully polarized emphasizes that learning can exacerbate taste projection when agents aren't aware of their initial bias. Agents move from a mild error—they assume others share their uncertain beliefs about λ—to a strong error—they are confident that all share their taste. That is, naive learning can generate a strong false-consensus bias. This result highlights the role of neglecting others' discrepant beliefs in learning settings. Aside from wrong theories of others' beliefs, agents have precisely correct models of the world. Even so, ignoring heterogeneity in beliefs can lead society far from the truth.

1.7 Discussion and Conclusion

This paper demonstrates the implications of taste projection on social learning. I clearly demonstrate how one's interpretation of others' behavior depends on the lens through which it is observed—those with different perceptions of tastes often develop drastically different beliefs about the state of the world. And in many cases, this discrepancy in beliefs can lead behavior far from the optimum. The results of this paper help explain three important phenomenon inconsistent with rational learning models. First, taste projection offers an explanation for why uniform behavior may arise despite diverse preferences. Second, it shows how society can develop and maintain confident but false beliefs even when observing an arbitrarily large sample of privately-informed behavior. Third, false-consensus errors can arise from naive learning: when people ignore differences in prior beliefs, otherwise rational learning leads agents to think their own taste is most common.

While the formal model focuses exclusively on observational learning, I conjecture that taste projection has important consequences in other natural social-learning environments.

For instance, consider a setting where agents directly share their experiences. Players observe the action and *payoff* of all predecessors, consistent with word-of-mouth learning (e.g., Banerjee and Fudenberg, 2004) or learning from online reviews. Projection still leads learning astray. Suppose restaurant X generates stochastic outcomes y, which provide a θ-type with utility $u(y, \theta)$. And suppose an observer sees a large collection of payoffs from random sample of the population. With correct knowledge of the distribution of θ, a rational observer could back out the distribution of Y from the sample of payoffs. But a taste projector, who has wrong beliefs about the distribution of θ, develops a distorted perception of the underlying distribution of outcomes, Y. If some unsophisticated diners earn high payoffs from average-quality meals, "foodies" who think high payoffs come only from exceptional meals will be mislead by the shining reviews of those with limited taste, and vice versa.

From a broad perspective, a novel feature of this paper is the assumption that agents within a non-common-prior environment neglect heterogeneity in beliefs. Of course, this paper focuses on the very specific case of social learning, but it naturally provokes curiosity about how similar forms of naivete alter the results of well-known non-common-prior models like Harrison and Kreps (1978), Morris (1996), and Scheinkman and Xiong (2003). What do speculative traders come to believe about returns when they neglect disagreement? Beyond taste projection, there are other reasons to expect disagreement neglect. For example, Malmendier and Nagel (2011) find that market conditions experienced early in life shape investors expectations about stock-market returns. It seems natural that investors under-appreciate the influence of experience on perceptions, and may wrongly conclude that investors from different generations hold similar perceptions. How do conflicting expectations interact in the market, and how does this interaction shape the perceptions of the current young generation? And how does this naive learning process play out in the long run? These questions are left open for future research.

Appendix

1.A A Simple Model Of Taste Projection

This section presents a simple model of taste projection, providing a specific parameterization for the general bias outlined in Section 3.2.2. The model specifies an agent's perceived taste distribution as a function of her own taste, the true distribution, and a single bias parameter, $\beta \in [0, 1]$.

Consider the model of preferences introduced in Section 1.2.1 with finite type space Θ. Suppose θ are distributed according to c.d.f. G with mass function g. As formalized in Section 3.2.2,type θ perceives the taste distribution as $\widehat{G}(\cdot \mid \theta)$, and let $\hat{g}(\cdot \mid \theta)$ be the associated mass function. The perceived mass function is given by

$$\hat{g}(\tilde{\theta} \mid \theta) = (1 - \beta)^{|\tilde{\theta} - \theta|} g(\tilde{\theta}) / \Phi_\theta \tag{A.1}$$

where $\beta \in [0, 1]$ is the degree of the bias, and Φ_θ is a normalization constant,

$$\Phi_\theta := \sum_{\tilde{\theta} \in \Theta} (1 - \beta)^{|\tilde{\theta} - \theta|} g(\tilde{\theta}). \tag{A.2}$$

Essentially, for a player with taste θ, the true mass function is scaled by a weighting function $w_\theta(\tilde{\theta}) = (1 - \beta)^{|\tilde{\theta} - \theta|}$ that gives higher weight to types close to her own. β controls how quickly the weighting function decreases when moving away from one's own type. Note that $\beta = 0$ corresponds to rational perceptions: $\hat{g}(\cdot \mid \theta) = g(\cdot)$. At the other extreme, $\beta = 1$ implies full projection—the player believes all share her taste: $\hat{g}(\theta \mid \theta) = 1$ and $\hat{g}(\tilde{\theta} \mid \theta) = 0$ for all $\tilde{\theta}$. This simple parameterization of taste projection meets the general conditions assumed in Section 3.2.2: (1) first-order stochastic dominance in θ (Assumption 12), and (2) full support (Assumption 3).[63]

The following figures give a sense of perceived distributions under this parameterization. In each, $\Theta = \{j\delta \mid j = 0, ..., 100\}$ and $\delta = 0.01$. Figures 1.A.1 and 1.A.2 show the biased pmfs and cfs, respectively, for various θ-types supposing the true distribution is (approximately) U$[0, 1]$. Figures 1.A.3 and 1.A.4 do the same when the true distribution is (approximately) Beta$(8, 8)$.

[63] Trivially, Assumption 3 holds so long as $\beta \neq 1$.

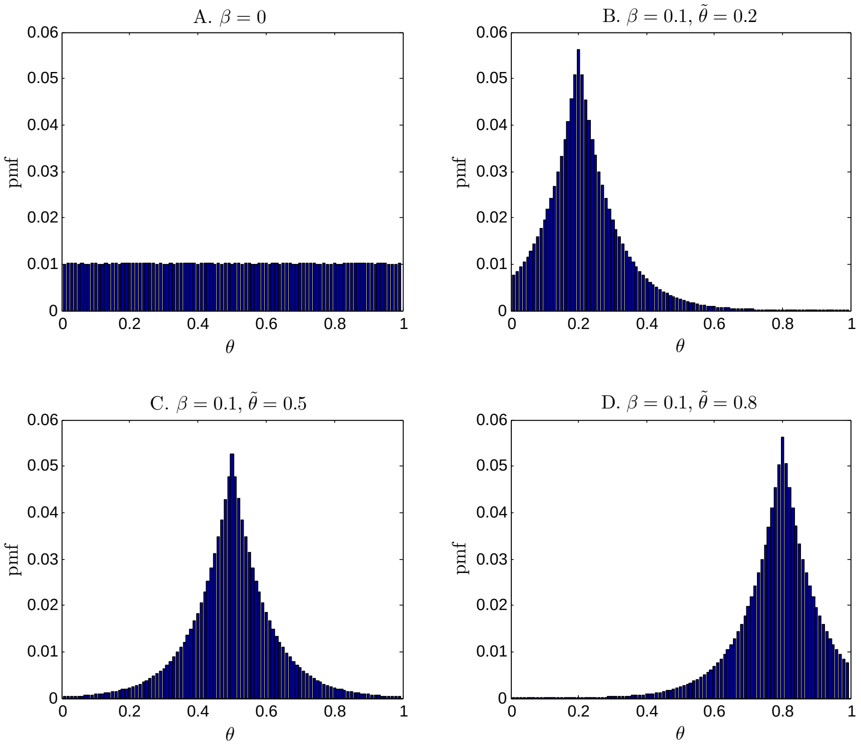

Figure 1.A.1: *Biased pmfs for* $\beta = 0.1$ *and* $\theta \sim \mathrm{U}[0, 1]$.

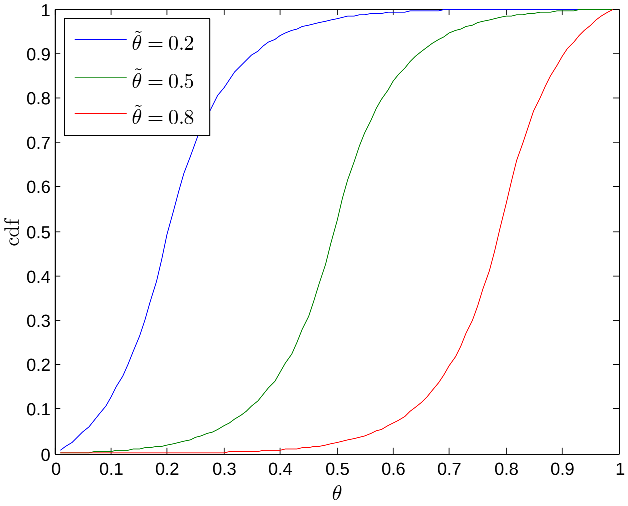

Figure 1.A.2: *Biased cdfs for* $\beta = 0.1$ *and* $\theta \sim \mathrm{U}[0, 1]$.

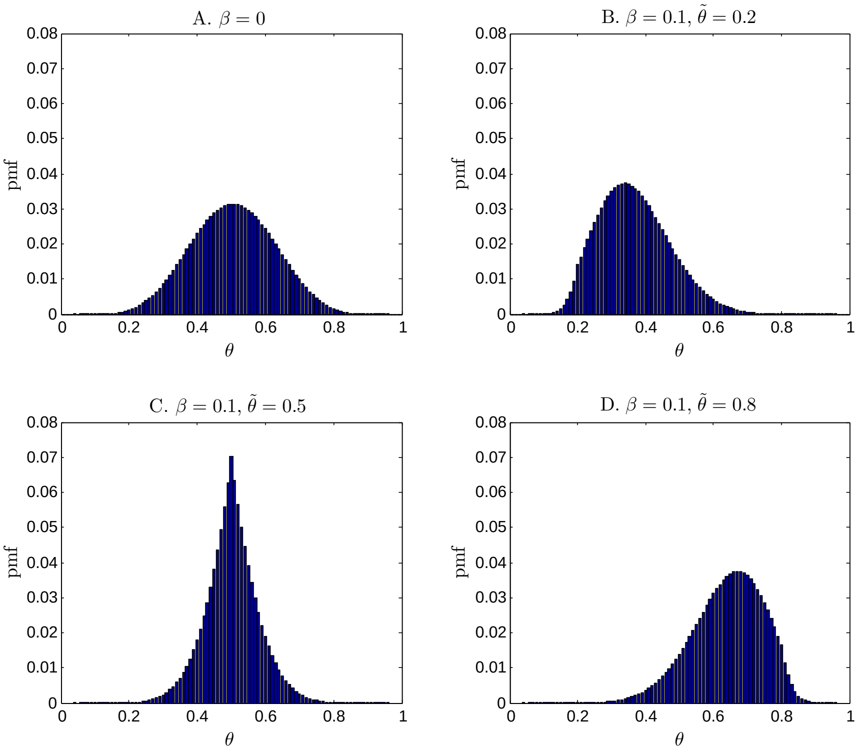

Figure 1.A.3: *Biased pmfs for* $\beta = 0.1$ *and* $\theta \sim \text{Beta}(8, 8)$.

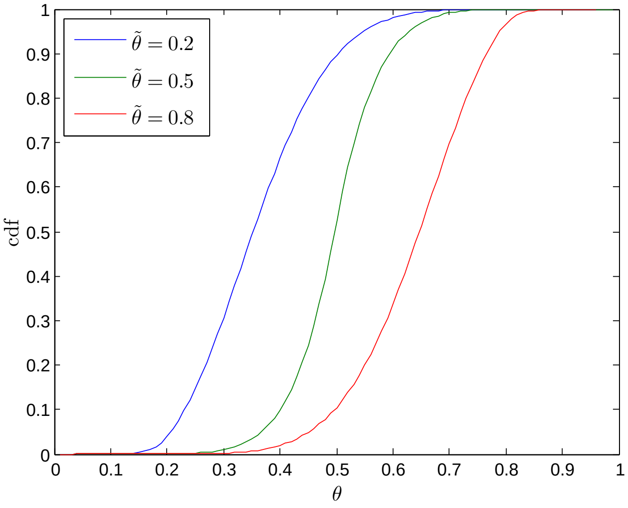

Figure 1.A.4: *Biased cdfs for* $\beta = 0.1$ *and* $\theta \sim \text{Beta}(8, 8)$.

1.B Smith and Sørensen's Confounded Learning

Consider the model of Sections 1.3 and 1.4 where Δ_q is known. This section demonstrates that confounding beliefs only exist when $|\Delta_q|$ is sufficiently large, and show how their existence changes the basic results derived in above. Smith and Sørensen (2000) show that in this setting, observational learning with heterogeneous preferences may lead to "confounded learning". With rational agents, there may exist an interior steady-state belief, $\hat{\pi}$, such that if public beliefs reach this value, then learning stops. Beliefs remain at $\hat{\pi}$. The steady state is such that the probability of any observation a is equal in both states R and L. Observing a when public beliefs are at the steady state reveals no new information. In terms of updating process defined above, $\hat{\pi}$ is the value that satisfies $\psi(a \mid \hat{\ell}, R) = \psi(a \mid \hat{\ell}, L)$ where $\hat{\ell} = (1-\hat{\pi})/\hat{\pi}$. Smith and Sørensen (2000) show that under rational play, if such a confounding belief exists, long-run beliefs converge to this value with positive probability.

1.B.1 Existence of Confounding Beliefs

Lemma A.1. *Let $\bar{\theta}^l = \max_\theta \Theta^l$. Then no confounding beliefs exist if*

$$\Delta_q < k\Delta_d(\overline{\theta}^l)(1-\xi^\theta)/(1+\xi^\theta),$$

where

$$\xi^\theta := \min\left\{\sqrt{\frac{\sum_{\theta' \in \Theta^l} \hat{g}(\theta' \mid \theta)}{\sum_{\theta' \in \Theta^r} \hat{g}(\theta' \mid \theta)}}, \sqrt{\frac{\sum_{\theta' \in \Theta^r} \hat{g}(\theta' \mid \theta)}{\sum_{\theta' \in \Theta^l} \hat{g}(\theta' \mid \theta)}}\right\} < 1.$$

1.C Rational Learning with Aggregate Preference Uncertainty

This section characterizes long-run learning among rational agents with taste-dependent distributional beliefs, which arise from uncertainty over the taste distribution (as in Section 1.6). For instance, investors are uncertain if others are primarily risk averse or risk neutral, so an agent's own preference is information.

Rational learning contrasts sharply with learning under naive projection. Namely, rational beliefs always converge, and people with different tastes never reach fully-polarized beliefs—they never grow confident in different states. The various failures in learning that arise with naive projection—incorrect learning, fully-polarized beliefs, and perpetually fluctuating beliefs—are thus not a sole consequence of taste-dependent distributional beliefs. Rather, they result from *ignorance* regarding others' taste-dependent beliefs—from thinking others' think like oneself.

However, rational learning in this setting is not complete. Depending on the sample path, rational agents either fully learn or converge to an interior fixed point. Disagreement may exist in a long-run equilibrium, but in such cases, society remains uncertain: two agents with different tastes never grow confident in two distinct hypotheses. Interestingly, when there is uncertainty over the type distribution, confounded learning always arises with positive probability. This contrasts the standard Smith and Sørensen (2000) model, where it arises only if quality differences, $|\Delta_q|$, are sufficiently large.

1.C.1 Rational Long-Run Learning

The model I consider here is identical to the model in Section 1.6 with the following exception: agents are fully-rational, so second-order beliefs are correct. Each player knows precisely the priors of all others.

Rational learning may still fail in an important way. In particular, confounding beliefs exist for any quality difference Δ_q. Let $\pi_t^\theta = \sum_k \pi_n^t(R, \lambda_k)$ denote marginal probability of preference state $\omega = R$. I now define "confounding beliefs".

Definition A.1. *Let the pair $\hat{\boldsymbol{\pi}}^l$ and $\hat{\boldsymbol{\pi}}^r$ be public beliefs held by types l and r, respectively. The pair $(\hat{\boldsymbol{\pi}}^l,\hat{\boldsymbol{\pi}}^r)$ are* confounding beliefs *if for all states given positive weight—$\zeta, \zeta' \in \{L, R\}$ and $\lambda_k, \lambda_j \in \Lambda$ such that $\hat{\pi}^\theta(\zeta, \lambda_k), \pi^\theta(\zeta', \lambda_j) > 0$,*

$$\Pr(a \mid \hat{\pi}^l, \hat{\pi}^r, \zeta, \lambda_k) = \Pr(a \mid \hat{\pi}^l, \hat{\pi}^r, \zeta', \lambda_j)$$

for any $a \in \{0, 1, ..., N\}$.

The next proposition shows that such belief profiles generically exist when there is uncertainty about λ.

Proposition A.1. *For any Λ with $|\Lambda| \geq 2$ and any non-degenerate prior $\mu_0 \in \Delta(\Lambda)$, there exists at least one pair of confounding beliefs $(\hat{\boldsymbol{\pi}}^l,\hat{\boldsymbol{\pi}}^r)$ satisfying Definition A.1.*

However, to show learning is complete, it must be the case that beliefs converge with positive probability to such a profile. The next proposition establishes this.

Proposition A.2. *At least one pair of confounding beliefs is locally stochastically stable: a confounding outcome occurs with positive probability. However, the probability of correct learning goes to 1 as $\pi_1 \to 1$; for each θ, $\Pr(\pi_t^\theta(R, \lambda^*) \to 1) = 1$ as $\pi_1 \to 1$.*

This result is similar to that of Jackson and Kalai (1997). In a model of "recurring games" with both type uncertainty and payoff uncertainty, behavior doesn't converge to Bayesian Nash equilibrium of the stage game with *known* type distributions whenever payoffs depend on type. Here we see such non-convergence. However, players still learn with positive probability. Uncertainty doesn't imply society *necessarily* fails to learn.

Rational learning with uncertainty about tastes provides a simple and natural explanation for persistent disagreement. At a confounding belief, people with different tastes disagree on payoffs: relative to a risk-seeking agent, a risk-averse agent thinks it's more likely that most are risk averse and that A is safe. Despite continually observing behavior, players persistently disagree. Why? At a confounding belief, new observations reveal no new information. Hence beliefs across types depend on priors, which are necessarily taste specific. There are alternative explanations for how individuals who observe the same evidence disagree in the long run. Such models include uncertainty over the distribution of private information, as explored in Acemoglu, Chernozhukov, and Yildiz (2007 and 2009), or public signals about a single dimension of uncertainty despite an environment with many dimensions of uncertainty

(Andreoni and Mylovanov, 2012). In all cases, so long as players are rational, disagreement is never "fully" polarized. As I've argued, full polarization—where agents grow confident in alternative hypotheses, can occur under taste projection.

1.D Extensions

1.D.1 Can Some Rational Agents Correct Biased Learning?

Can the existence of fully-rational agents—those who know λ, but also know how others misperceive λ—correct biased learning among naive agents? Following the analysis in Sections 1.4.3 and 1.4.4, I address this question assuming agents suffer choice-dependent. The answer is no; no matter how many individuals are rational, biased learning remains incomplete.

To see this, suppose a fraction $\gamma \in (0,1)$ are fully rational. And suppose rationality is independent of taste. A rational player knows λ and $\hat{\lambda}(\theta)$ for each θ—they know exactly how taste maps to perception. Let $\langle \pi_t^c \rangle$ denote the belief process among rational—"correct"—agents. Since all correct players agree on the model of the world, beliefs are independent of taste. First, learning is complete among rational players.

Proposition A.3. *For any $\gamma \in (0,1)$, λ, and $\hat{\lambda}(\theta)$, $\pi_t^c \to 1$ a.s.*

This follows immediately from the fact that $\langle \ell_t^c \rangle$—the process of rational likelihood ratios—forms a conditional martingale, and hence must converge to a finite fixed point of the process. The only such fixed point is $\ell^c = 0$.

Since rational agents learn, all rational right types choose A in the long run. Hence in any candidate stable equilibrium $\hat{\boldsymbol{\pi}} = (\hat{\pi}^l, \hat{\pi}^h)$, the frequency of A is $\gamma\lambda + (1-\gamma)[(1-\lambda)(1-\hat{\pi}^l) + \lambda\hat{\pi}^h]$. We can simply invoke Proposition 4 to determine whether the equilibrium is stable. From logic identical to Proposition 5, biased beliefs never converge to a point of agreement, and hence never converge to the truth. While rational agents never lead biased agents to the truth, the next two propositions (which follow from direct applications of Proposition 4) demonstrate the limited impact of rational agents on biased learning outcomes in both the strong and weak case.

Proposition A.4. *Assume strong taste projection. For all $\gamma \in (0,1)$, universal learning fails. Furthermore, there exist $\underline{\gamma}$ and $\overline{\gamma}$, $\underline{\gamma} < \overline{\gamma}$, such that*

1. *If $\gamma < \underline{\gamma}$, then naive beliefs $\hat{\boldsymbol{\pi}} = (0,1)$ and $\hat{\boldsymbol{\pi}} = (1,0)$ are stable.*
2. *If $\gamma \in (\underline{\gamma}, \overline{\gamma})$, then only $\hat{\boldsymbol{\pi}} = (0,1)$ is stable.*
3. *If $\gamma > \overline{\gamma}$, then no naive beliefs are stable.*

In Proposition A.4, the values $\underline{\gamma}$ and $\overline{\gamma}$ are

$$\underline{\gamma} = \min\left\{\frac{\hat{\lambda}^l}{\lambda}, \frac{1-\hat{\lambda}^r}{\lambda}\right\} \quad \overline{\gamma} = \min\left\{\frac{\hat{\lambda}^l}{1-\lambda}, \frac{1-\hat{\lambda}^r}{1-\lambda}\right\}$$

It follows that a sufficiently large proportion of rational agents, $\gamma > \overline{\gamma}$, can break a uniform herd. For example, when $\lambda = 0.6$, but $(\hat{\lambda}^l, \hat{\lambda}^r) = (0.4, 0.8)$, then $\overline{\gamma} = \frac{1}{2}$.

Proposition A.5. *Assume weak taste projection. For all $\gamma \in (0,1)$, universal learning fails. Furthermore, there exists $\tilde{\gamma} \in (0,1)$ such that*

1. *$\gamma \in (0, \tilde{\gamma})$ implies both $\langle \ell_t^l \rangle$ and $\langle \ell_t^r \rangle$ are non-convergent.*
2. *$\gamma \in (\tilde{\gamma}, 1)$ implies only $\langle \ell_t^r \rangle$ is non-convergent.*

The value $\tilde{\gamma}$ satisfies

$$\hat{\lambda}^l = \tilde{\gamma}\lambda + (1-\tilde{\gamma})\mathbb{E}[X_{nt} = A \mid \theta_{nt} = r, \ell_t^l = 0\,\forall t]$$

$\tilde{\gamma}$ is the value such that the expected frequency of A in the long-run exactly matches a low-type's expected frequency. Since this value depends on the limit distribution of right-type behavior—which only converges in distribution—$\tilde{\gamma}$ depends on the distribution of signals. For example, if $f_R(p) = 2p$, $f_L(p) = 2(1-p)$, $\lambda = 0.75$, $\hat{\lambda}^l = 0.6$ and $\hat{\lambda}^r = 0.9$, then $\tilde{\gamma} \approx 0.2$. So a rather modest fraction of rational players ensures that low-type beliefs converge. The existence of rational players can ensure that the frequency of A never falls below $\hat{\lambda}^l$, even when right-type beliefs favor $\omega = L$.

1.D.2 Alternative Forms of Misprediction

This section considers alternative distributional errors distinct from projection. For instance, people might perceive a false sense of uniqueness. The analysis of limit beliefs contained in Sections 1.4.1 and 1.4.2 was independent of assumptions placed on $\hat{\boldsymbol{\lambda}}$. Hence, those results can be applied to $\hat{\boldsymbol{\lambda}}$ exhibiting any particular pattern of error.

Following simply from Proposition 4, which tells us when a confident equilibrium belief is stable, we have the following general result for any form of misprediction of type proportion $\hat{\lambda}$:

Proposition A.6. *As $N \to \infty$, economy-wide learning is complete if and only if for all $\theta \in \Theta$, $\hat{\lambda}(\theta) \in \left(\frac{1}{2}, \lambda\right]$.*

When all individuals mutually underestimate the share of people that have the majority preference, then the truth is asymptotically stable. Near an equilibrium, people observe more people taking the majority action than they anticipated, only strengthening their beliefs. However, this logic implies learning may backfire in settings with small N: people may grow confident in a false state of the world. As N grows large, however, the probability of incorrect learning goes to 0.

In all other scenarios not discussed in this paper, some—and possibly all—types hold non-convergent beliefs. A particular example of interest is when people suffer a "false-uniqueness" bias: each type thinks her type is least common.[64] In such a case, it's intuitive that action frequencies evolve in a cyclical fashion. As some option gains popularity, say A, an individual of *either* type believes B best suits her tastes. Her reasoning is that she has the minority preference, thus the less popular option is most likely optimal. But since *all* people follow this reasoning, B will eventually become the majority choice. At this point, individuals will admit they must have been wrong, once again believing A must be optimal for their preference. Under the false-uniqueness bias, followers avoid the majority action, causing society's most prevalent choice to oscillate over time. This contrasts sharply with the intuition of the strong false-consensus bias: there, followers flock to the majority action, increasing the frequency at which it is chosen over time.

1.E Proofs

Proof of Lemma 1.

Proof. From Equation 1.4 and using definitions $\Delta_q := q^A - q^B$ and $\Delta_d(\theta) := (1-\theta)^2 - (-1-\theta)^2 = -4\theta$, it follows that an individual chooses A if and only if

$$r\Delta_d(\theta) \leq \frac{\Delta_d(\theta)}{2} + \frac{\Delta_q}{2k}. \tag{A.3}$$

Dividing through by $\Delta_d(\theta)$ and noting that $\Delta_d(\theta) > 0 \Leftrightarrow \theta < 0$ yields the decision rule. □

Proof of Lemma 2.

Proof. Type-θ is actively learning if $\bar{r}(\theta) \in [0,1]$ where $\bar{r}(\theta) = \frac{1}{2} + \frac{\Delta_q}{2k\Delta_d(\theta)}$. $\bar{r}(\theta) \notin [0,1] \Leftrightarrow \frac{\Delta_q}{\Delta_d(\theta)} > k$ or $\frac{\Delta_q}{\Delta_d(\theta)} < -k$. It follows that passive types comprise the set $\Theta^p := \{\theta \in \Theta \mid -\Delta_q < k\Delta_d(\theta) < \Delta_q\}$. Active left types comprise the set $\Theta^l := \Theta \cap (-\infty, 0) \setminus \Theta^p$. Given that $\theta < 0 \Rightarrow \Delta_d(\theta) > 0$, $\Theta^l = \{\theta \in \Theta \mid k\Delta_d(\theta) \geq \Delta_q\}$. Similarly, since $\theta > 0 \Rightarrow \Delta_d(\theta) < 0$, $\Theta^r := \Theta \cap [0, \infty) \setminus \Theta^p = \{\theta \in \Theta \mid k\Delta_d(\theta) \leq -\Delta_q\}$ □

Proof of Lemma 3.

Proof. For $\theta \notin \Theta^p$, the cutoff rule followa immediately from rewriting the posterior r in Lemma 1 as $r = p/(p + (1-p)\ell)$ and solving for a threshold on p. By definition, $\theta \in \Theta^p$ choose A for all $p \in (0,1)$. □

[64] Wallace (1996) puts it well: "everybody is identical in their unspoken belief that way deep down they are different from everyone else."

Proof of Lemma 4.

Proof. See Lemma A.1. □

Proof of Lemma 5.

Proof. Fix $\theta \in \Theta$ and $\ell_t^\theta \in \mathbb{R}_+$. Suppose $a_t/N > \hat{\lambda}(\theta)$. From Equation 1.12, $\ell_{t+1}^\theta < \ell_t^\theta \Leftrightarrow \psi(a_t \mid \ell_t^\theta, L) < \psi(a_t \mid \ell_t^\theta, R) \Leftrightarrow \alpha_\theta(\ell_t^\theta, \omega)^a \big[1 - \alpha_\theta(\ell_t^\theta, L)\big]^{N-a} < \alpha_\theta(\ell_t^\theta, R)^a \big[1 - \alpha_\theta(\ell_t^\theta, R)\big]^{N-a}$,

$$\begin{aligned} &\Leftrightarrow \quad \left(\frac{\alpha_\theta(\ell_t^\theta, L)\big[1 - \alpha_\theta(\ell_t^\theta, R)\big]}{\big[1 - \alpha_\theta(\ell_t^\theta, L)\big]\alpha_\theta(\ell_t^\theta, R)}\right)^a \left(\frac{1 - \alpha_\theta(\ell_t^\theta, R)}{1 - \alpha_\theta(\ell_t^\theta, L)}\right)^N < 1 \\ &\Leftrightarrow \quad a \log\left(\frac{\alpha_\theta(\ell_t^\theta, L)\big[1 - \alpha_\theta(\ell_t^\theta, R)\big]}{\big[1 - \alpha_\theta(\ell_t^\theta, L)\big]\alpha_\theta(\ell_t^\theta, R)}\right) + N \log\left(\frac{1 - \alpha_\theta(\ell_t^\theta, R)}{1 - \alpha_\theta(\ell_t^\theta, L)}\right) < 0. \end{aligned} \tag{A.4}$$

If $\left(\frac{\alpha_\theta(\ell_t^\theta, L)\big[1 - \alpha_\theta(\ell_t^\theta, R)\big]}{\big[1 - \alpha_\theta(\ell_t^\theta, L)\big]\alpha_\theta(\ell_t^\theta, R)}\right) > 1$, then inequality A.4 holds iff

$$a/N < \frac{1}{1 + \log\left(\frac{\alpha_\theta(\ell_t^\theta, L)}{\alpha_\theta(\ell_t^\theta, R)}\right) / \log\left(\frac{1 - \alpha_\theta(\ell_t^\theta, R)}{1 - \alpha_\theta(\ell_t^\theta, L)}\right)} \equiv \kappa(\ell_t^\theta, \theta).$$

Otherwise, A.4 holds iff $a/N > \kappa(\ell_t^\theta, \theta)$. Finally, note that $\left(\frac{\alpha_\theta(\ell_t^\theta, L)\big[1 - \alpha_\theta(\ell_t^\theta, R)\big]}{\big[1 - \alpha_\theta(\ell_t^\theta, L)\big]\alpha_\theta(\ell_t^\theta, R)}\right) > 1 \Leftrightarrow \alpha_\theta(\ell_t^\theta, L) > \alpha_\theta(\ell_t^\theta, R) \Leftrightarrow \hat{\lambda}(\theta) + [1 - 2\hat{\lambda}(\theta)]F_L(p(\ell_t^\theta)) > \hat{\lambda}(\theta) + [1 - 2\hat{\lambda}(\theta)]F_R(p(\ell_t^\theta)) \Leftrightarrow \hat{\lambda}(\theta) < 1/2$, since $F_R(p(\ell_t^\theta)) < F_L(p(\ell_t^\theta))$ by Assumption 5.

□

Proof of Proposition 1.

Proof. Fix $\theta \in \Theta$ $\ell_t^\theta \in \mathbb{R}_+$. Let $\underline{m} = \min\{1 - \hat{\lambda}(\theta), \hat{\lambda}(\theta)\}$ and $\overline{m} = \max\{1 - \hat{\lambda}(\theta), \hat{\lambda}(\theta)\}$. To proceed, I show that for all $\ell_t^\theta \in \mathbb{R}_+$, $\kappa(\ell_t^\theta, \theta) \in [\underline{m}, \overline{m}]$. Since $\kappa(\ell_t^\theta, \theta)$ is monotonic in ℓ_t^θ, we must consider $\lim_{\ell \to 0} \kappa(\ell, \theta)$ and $\lim_{\ell \to \infty} \kappa(\ell, \theta)$. First, note that $\lim_{\ell \to 0} \alpha_\theta(\ell, \omega) = \hat{\lambda}(\theta)$ and $\lim_{\ell \to \infty} \alpha_\theta(\ell, \omega) = 1 - \hat{\lambda}(\theta)$. Thus, we must use L'Hoptial's rule to evaluate the limits:

$$\frac{\partial}{\partial \ell} \log\left(\frac{\alpha_\theta(\ell, L)}{\alpha_\theta(\ell, R)}\right) = \left(\frac{\alpha_\theta(\ell, L)}{\alpha_\theta(\ell, R)}\right)^{-1} \frac{\alpha_\theta(\ell, R)\frac{\partial}{\partial \ell}\alpha_\theta(\ell, L) - \alpha_\theta(\ell, L)\frac{\partial}{\partial \ell}\alpha_\theta(\ell, R)}{\alpha_\theta(\ell, R)^2},$$

$$\begin{aligned} &\frac{\partial}{\partial \ell} \log\left(\frac{1 - \alpha_\theta(\ell, R)}{1 - \alpha_\theta(\ell, L)}\right) = \\ &\quad \left(\frac{1 - \alpha_\theta(\ell, R)}{1 - \alpha_\theta(\ell, L)}\right)^{-1} \frac{\frac{\partial}{\partial \ell}\alpha_\theta(\ell, L) - \frac{\partial}{\partial \ell}\alpha_\theta(\ell, R) + \alpha_\theta(\ell, L)\frac{\partial}{\partial \ell}\alpha_\theta(\ell, R) - \alpha_\theta(\ell, R)\frac{\partial}{\partial \ell}\alpha_\theta(\ell, L)}{[1 - \alpha_\theta(\ell, L)]^2}, \end{aligned}$$

and, since Equation 1.10 implies

$$\frac{\partial}{\partial \ell}\alpha_\theta(\ell,\omega) = \big[1-2\hat{\lambda}(\theta)\big] f_\omega\big(p(\ell)\big)\frac{\partial}{\partial \ell}p(\ell),$$

it follows that

$$\lim_{\ell \to 0} \frac{\frac{\partial}{\partial \ell}\log\left(\frac{\alpha_\theta(\ell,L)}{\alpha_\theta(\ell,R)}\right)}{\frac{\partial}{\partial \ell}\log\left(\frac{1-\alpha_\theta(\ell,R)}{1-\alpha_\theta(\ell,L)}\right)} = \frac{1-\hat{\lambda}(\theta)}{\hat{\lambda}(\theta)},$$

and

$$\lim_{\ell \to \infty} \frac{\frac{\partial}{\partial \ell}\log\left(\frac{\alpha_\theta(\ell,L)}{\alpha_\theta(\ell,R)}\right)}{\frac{\partial}{\partial \ell}\log\left(\frac{1-\alpha_\theta(\ell,R)}{1-\alpha_\theta(\ell,L)}\right)} = \frac{\hat{\lambda}(\theta)}{1-\hat{\lambda}(\theta)}.$$

As such, $\lim_{\ell\to 0}\kappa(\ell,\theta) = \hat{\lambda}(\theta)$ and $\lim_{\ell\to\infty}\kappa(\ell,\theta) = 1-\hat{\lambda}(\theta)$, and $\kappa(\ell,\theta) \in [\underline{m},\overline{m}]$ for all $\ell \in \mathbb{R}_+$. Suppose $a_t/N > \hat{\lambda}(\theta)$. If $\hat{\lambda}(\theta) > 1/2$, then $a_t/N > \overline{m} > \kappa(\ell_t^\theta,\theta)$ and Lemma 5 implies $\ell_{t+1}^\theta < \ell_t^\theta$. Otherwise, Lemma 5 implies $\ell_{t+1}^\theta > \ell_t^\theta$. Now suppose $a_t/N < 1-\hat{\lambda}(\theta)$. Similarly, if $\hat{\lambda}(\theta) > 1/2$, then $a_t/N < \underline{m} < \kappa(\ell_t^\theta,\theta)$ and Lemma 5 implies $\ell_{t+1}^\theta > \ell_t^\theta$. Otherwise, if $\hat{\lambda}(\theta) < 1/2$, Lemma 5 implies $\ell_{t+1}^\theta < \ell_t^\theta$.

□

Proof of Corollary 1.

Proof. Fix an arbitrary $\theta \in \Theta$ and suppose she has likelihood ratio $\ell_t^\theta \in \mathbb{R}_+$. Given observation X_t at public belief ℓ_t^θ, Equation 1.12 implies $\ell_{t+1}^\theta > \ell_t^\theta \Leftrightarrow \Psi_\theta(X_t,\ell_t^\theta) > 1 \Leftrightarrow \psi(X_t \mid \ell_t^\theta, L) > \psi(X_t \mid \ell_t^\theta, R)$. Suppose $N=1$ and $X_t = A$ and let $\bar{p} := p(\ell_t^\theta)$ denote θ's private-belief threshold in t.. Then Equation 1.10 implies $\psi(X_t \mid \ell_t^\theta, L) > \psi(X_t \mid \ell_t^\theta, R)$ if and only if

$$\big[1-\hat{\lambda}(\theta)\big]F_L(\bar{p}) + \hat{\lambda}(\theta)\big[1-F_L(\bar{p})\big] > \big[1-\hat{\lambda}(\theta)\big]F_R(\bar{p}) + \hat{\lambda}(\theta)\big[1-F_R(\bar{p})\big],$$

which holds if and only if

$$\big[1-2\hat{\lambda}(\theta)\big]F_L(\bar{p}) > \big[1-2\hat{\lambda}(\theta)\big]F_R(\bar{p}). \tag{A.5}$$

By Assumption 5, $F_L(\bar{p}) > F_R(\bar{p})$ for all $\ell_t^\theta \in \mathbb{R}_+$, so Relation A.5 holds if and only if $\hat{\lambda}(\theta) < \frac{1}{2}$. □

Proof of Proposition 2.

Proof. Fix an arbitrary $\theta \in \Theta$ and suppose she has likelihood ratio $\ell_t^\theta \in \mathbb{R}_+$. Assuming $N=1$ and $X_t = A$, Proposition 1 $\Psi_\theta(X_t = A, \ell_t^\theta) > 1 \Leftrightarrow \hat{\lambda}(\theta) < \frac{1}{2}$. First consider $\hat{\lambda}(\theta) \in (\frac{1}{2}, 1)$ so $\Psi_\theta(X_t = A, \ell_t^\theta) < 1$. We want to show $|\ell_{t+1}^\theta - \ell_t^\theta|$ is increasing in $\hat{\lambda}(\theta)$ on this domain. Note that $|\ell^\theta_{t+1} - \ell^\theta_t| = \ell^\theta_t |\Psi_\theta(A, \ell^\theta_t) - 1|$, which is increasing in $\hat{\lambda}(\theta) \Leftrightarrow \Psi_\theta(A, \ell^\theta_t)$ is decreasing in $\hat{\lambda}(\theta)$. Let $\bar{p} := p(\ell^\theta_t)$ denote θ's private-belief threshold in t. Note

$$\begin{aligned}\Psi_\theta(A, \ell^\theta_t) &= \frac{\big[1 - \hat{\lambda}(\theta)\big] F_L(\bar{p}) + \hat{\lambda}(\theta)\big[1 - F_L(\bar{p})\big]}{\big[1 - \hat{\lambda}(\theta)\big] F_R(\bar{p}) + \hat{\lambda}(\theta)\big[1 - F_R(\bar{p})\big]} \\ &= \frac{\hat{\lambda}(\theta)\big[1 - 2F_R(\bar{p})\big] + F_R(\bar{p})}{\hat{\lambda}(\theta)\big[1 - 2F_L(\bar{p})\big] + F_R(\bar{p})} \qquad \text{(A.6)}\end{aligned}$$

so $\frac{\partial}{\partial \hat{\lambda}(\theta)} \Psi_\theta(A, \ell^\theta_t) < 0$ if and only if

$$\begin{aligned}&\hat{\lambda}(\theta)\big[1 - 2F_R(\bar{p})\big]\big[1 - 2F_L(\bar{p})\big] + F_R(\bar{p})\big[1 - 2F_L(\bar{p})\big] > \\ &\qquad\qquad \hat{\lambda}(\theta)\big[1 - 2F_L(\bar{p})\big]\big[1 - 2F_R(\bar{p})\big] + F_L(\bar{p})\big[1 - 2F_R(\bar{p})\big],\end{aligned}$$

which holds if and only if $F_L(\bar{p}) > F_R(\bar{p})$, which is true for all $\ell^\theta_t \in \mathbb{R}_+$. Next, suppose that $\hat{\lambda}(\theta) \in (0 < \frac{1}{2})$ so $\Psi_\theta(X_t = A, \ell^\theta_t) > 1$. We want to show that $|\ell^\theta_{t+1} - \ell^\theta_t| = \ell^\theta_t |\Psi_\theta(A, \ell^\theta_t) - 1|$ is decreasing in $\hat{\lambda}(\theta)$ on this domain. This is true iff $\Psi_\theta(A, \ell^\theta_t)$ is decreasing in $\hat{\lambda}(\theta)$, which was shown in the case above. The logic is identical for $X_t = B$, but uses the fact that $\Psi_\theta(X_t, \ell^\theta_t) > 1 \Leftrightarrow \hat{\lambda}(\theta) > \frac{1}{2}$. □

Proof of Proposition 3.

Proof. If $\hat{\lambda}^l = \hat{\lambda}^r$, then play and beliefs correspond to the true Bayesian equilibrium and for all $t \in \mathbb{N}$, $\pi^\theta_t = \pi^{\theta'}_t$ for all $\theta, \theta' \in \Theta$. This equilibrium is studied in Smith and Sørensen (2000) and this result follows directly from their Theorem 5. Intuition is as follows: By Lemma 6, $\langle \ell^\theta_t \rangle$ forms a conditional martingale on $\omega = R$. By the Martingale Convergence Theorem, it must converge almost surely to some stationary limit. By Lemma 7, the only stationary limit points are $\ell \in \{0, \infty\}$. But rational beliefs never converge to fully-incorrect beliefs, so it must be that $\ell^\theta_t \to 0$ a.s. □

Proof of Lemma 6.

Proof. Fix an arbitrary $\theta \in \Theta$ and suppose $\omega = R$. Note that

$$\mathbb{E}[\ell^\theta_{t+1} \mid \boldsymbol{\ell}_t] = \sum_{a_t=0}^{N} \psi(a_t \mid \boldsymbol{\ell}_t, R) \Psi_\theta\big(a_t, \ell^\theta_t\big) \ell^\theta_t \qquad \text{(A.7)}$$

Thus in order for $\langle \ell^\theta_t \rangle$ to form a Martingale conditional on R, we would need

$$\sum_{a_t=0}^{N} \psi(a_t \mid \boldsymbol{\ell}_t, R) \Psi_\theta\big(a_t, \ell^\theta_t\big) = 1. \qquad \text{(A.8)}$$

for all $\ell^\theta_t \in \mathbb{R}_+$. But note

$$\begin{aligned}\sum_{a_t=0}^{N} \psi(a_t \mid \boldsymbol{\ell}_t, R)\Psi_\theta(a_t, \ell_t^\theta) &= \sum_{a_t=0}^{N} \psi(a_t \mid \boldsymbol{\ell}_t, R)\frac{\psi_\theta\big(a_t \mid \ell_t^\theta, L\big)}{\psi_\theta\big(a_t \mid \ell_t^\theta, L\big)} \\ &= \sum_{a_t=0}^{N} \frac{\psi(a_t \mid \boldsymbol{\ell}_t, R)}{\psi_\theta\big(a_t \mid \ell_t^\theta, R\big)}\psi_\theta\big(a_t \mid \ell_t^\theta, L\big) \qquad \text{(A.9)}\end{aligned}$$

Trivially, by the Law of Total Probability, $\sum_{a_t=0}^{N} \psi_\theta(a_t \mid \ell_t^\theta, L) = 1$. Hence in order for Equation A.8 to hold generically, we require $\psi(a_t \mid \boldsymbol{\ell}_t, R) = \psi_\theta(A_t \mid \ell_t^\theta, R)$ for all $a_t \in \{0, 1, ..., N\}$ in each $t \in \mathbb{N}$, which is only true if $\hat{\lambda}(\theta) = \lambda$ and for each $\theta, \theta' \in \Theta$, $\ell_t^\theta = \ell_t^{\theta'}$ in each $t \in \mathbb{N}$. But $\ell_t^\theta = \ell_t^{\theta'}$ in each $t \in \mathbb{N} \Leftrightarrow \hat{\lambda}(\theta) = \hat{\lambda}(\theta')$. Hence, the martingale condition holds if and only if $\hat{\lambda}(\theta) = \lambda$ for all $\theta \in \Theta$.

□

Proof of Lemma 7.

Proof. This is a direct application of Theorem B.1 and B.2 of S&S. They show that any limit point must be a steady-state of the process. That is, if $\ell^\theta \in \text{supp}\big(\ell_\infty^\theta\big)$, then it must be that $\varphi(X, \ell^\theta) = \ell^\theta$. For all $\theta \in \Theta$, the only beliefs that satisfy this condition are $\pi^\theta \in \{0, 1\}$. □

Proof of Lemma 8.

Proof. Adapted from Theorem C.1 of Smith and Sørensen (2000). □

Proof of Proposition 4.

Proof. Let $\hat{\boldsymbol{\ell}}$ be a fixed point of the joint belief process 1.15. From Lemma 8, $\hat{\boldsymbol{\ell}}$ is stable if $\chi_\theta(\hat{\boldsymbol{\ell}}) < 1$ for all $\theta \in \Theta$, and unstable if $\chi_\theta(\hat{\boldsymbol{\ell}}) > 1$ for some θ. I determine when this condition holds as a function of $\hat{\boldsymbol{\lambda}}$, which dictates the action frequency each type expects at fixed point $\hat{\boldsymbol{\ell}}$. At $\hat{\boldsymbol{\ell}}$, a θ-type believes all share confident belief $\hat{\ell}^\theta$, and thus expects A with frequency $\alpha_\theta\big(\hat{\ell}^\theta, \omega\big)$; the true frequency is $\alpha(\hat{\boldsymbol{\ell}})$. To determine whether this unexpected frequency reinforces each θ's beliefs, we must calculate $\chi_\theta(\hat{\boldsymbol{\ell}}) = \prod_{a=0}^{N} \Big(\frac{\partial}{\partial \ell}\varphi_\theta\big(a, \hat{\ell}^\theta\big)\Big)^{\psi(a, \boldsymbol{\ell})}$ for each θ.

Step 1: Calculate $\frac{\partial}{\partial \ell}\varphi_\theta(a, \ell)$.

Recall $\varphi_\theta(a, \ell) = \Psi_\theta(a, \ell)\ell$, where $\Psi_\theta(a, \ell) = \psi_\theta(a \mid \ell, L)/\psi_\theta(a \mid \ell, R)$. From the definition of $\psi_\theta(a \mid \ell, \omega)$ in Equation 1.8, it follows that

$$\begin{aligned}\frac{\partial}{\partial \ell}\psi_\theta(a \mid \ell, \omega) &= \binom{N}{a}\bigg(a\alpha_\theta(\ell, \omega)^{a-1}\big[1 - \alpha_\theta(\ell, \omega)\big]^{N-a}\frac{\partial}{\partial \ell}\alpha_\theta(\ell, \omega) \\ &\quad -(N-a)\alpha_\theta(\ell, \omega)^{a}\big[1 - \alpha_\theta(\ell, \omega)\big]^{N-a-1}\frac{\partial}{\partial \ell}\alpha_\theta(\ell, \omega)\bigg) \\ &= \frac{\partial}{\partial \ell}\alpha_\theta(\ell, \omega)\left(a\frac{\psi_\theta(a \mid \ell, \omega)}{\alpha_\theta(\ell, \omega)} - (N-a)\frac{\psi_\theta(a \mid \ell, \omega)}{1 - \alpha_\theta(\ell, \omega)}\right). \qquad \text{(A.10)}\end{aligned}$$

From Equation 1.10 it follows that

$$\frac{\partial}{\partial \ell}\alpha_\theta(\ell,\omega) = \big[1-2\hat{\lambda}(\theta)\big] f_\omega\big(p(\ell)\big)\frac{\partial}{\partial \ell}p(\ell).$$

Plugging this into Equation A.10 and using the fact $p(\ell) = \ell/(1+\ell) \Rightarrow \frac{\partial}{\partial \ell}p(\ell) = 1/(1+\ell)^2$ yields

$$\frac{\partial}{\partial \ell}\psi_\theta(a \mid \ell,\omega) = \frac{\big[1-2\hat{\lambda}(\theta)\big]}{(1+\ell)^2}\psi_\theta(a \mid \ell,\omega) f_\omega\big(p(\ell)\big)\left(\frac{a - N\alpha_\theta(\ell,\omega)}{\alpha_\theta(\ell,\omega)\big[1-\alpha_\theta(\ell,\omega)\big]}\right). \tag{A.11}$$

From the definition of $\Psi_\theta(a,\ell)$, we have

$$\frac{\partial}{\partial \ell}\Psi_\theta(a,\ell) = \frac{\frac{\partial}{\partial \ell}\psi_\theta(a \mid \ell, L)}{\psi_\theta(a \mid \ell, R)} - \Psi_\theta(a,\ell)\frac{\frac{\partial}{\partial \ell}\psi_\theta(a \mid \ell, R)}{\psi_\theta(a \mid \ell, R)}, \tag{A.12}$$

so Equation A.10 implies

$$\begin{aligned}\frac{\partial}{\partial \ell}\Psi_\theta(a,\ell) = \Psi_\theta(a,\ell)\Bigg\{\frac{\big[1-2\hat{\lambda}(\theta)\big]}{(1+\ell)^2}\Bigg[& f_L\big(p(\ell)\big)\left(\frac{a - N\alpha_\theta(\ell,L)}{\alpha_\theta(\ell,L)\big[1-\alpha_\theta(\ell,L)\big]}\right) \\ & - f_R\big(p(\ell)\big)\left(\frac{a - N\alpha_\theta(\ell,R)}{\alpha_\theta(\ell,R)\big[1-\alpha_\theta(\ell,R)\big]}\right)\Bigg]\Bigg\}.\end{aligned} \tag{A.13}$$

Finally, $\frac{\partial}{\partial \ell}\varphi_\theta(a,\ell) = \Psi_\theta(a,\ell) + \ell\frac{\partial}{\partial \ell}\Psi_\theta(a,\ell)$, so Equation A.13 implies

$$\begin{aligned}\frac{\partial}{\partial \ell}\varphi_\theta(a,\ell) = \Psi_\theta(a,\ell)\Bigg\{1 + \frac{\big[1-2\hat{\lambda}(\theta)\big]\ell}{(1+\ell)^2}\Bigg[& f_L\big(p(\ell)\big)\left(\frac{a - N\alpha_\theta(\ell,L)}{\alpha_\theta(\ell,L)\big[1-\alpha_\theta(\ell,L)\big]}\right) \\ & - f_R\big(p(\ell)\big)\left(\frac{a - N\alpha_\theta(\ell,R)}{\alpha_\theta(\ell,R)\big[1-\alpha_\theta(\ell,R)\big]}\right)\Bigg]\Bigg\}.\end{aligned} \tag{A.14}$$

Step 2: Evaluation of $\chi_\theta(\hat{\boldsymbol{\ell}})$.

While we want to assess whether $\chi_\theta(\hat{\boldsymbol{\ell}})$ exceeds 1 at the candidate equilibrium belief, the fact that fixed points are confident beliefs adds a complication to this approach. If each component of $\hat{\boldsymbol{\ell}}$ is 0 or ∞,then $\chi_\theta(\hat{\boldsymbol{\ell}}) = 1$ for all $\theta \in \Theta$. I now show this.

It is clear from Equation A.14 that if $\ell \in \{0,\infty\}$, then $\frac{\partial}{\partial \ell}\varphi_\theta(a,\ell) = \Psi_\theta(a,\ell)$. Furthermore, it is easy to show that $\Psi_\theta(a,0) = \Psi_\theta(a,\infty) = 1$: if θ is confident in ω, then her perceived probability of outcome a is identical in each $\omega \in \{L,R\}$, so $\psi_\theta(a \mid 0,L) = \psi_\theta(a \mid 0,R)$ and $\psi_\theta(a \mid \infty,L) = \psi_\theta(a \mid \infty,R)$. Formally, consider $\hat{\ell}^\theta = 0$. The private belief threshold is $p(\hat{\ell}^\theta) = 0$, so the perceived probability that a random player takes A in ω is $\alpha_\theta(0,\omega) = \big[1-$

$\hat{\lambda}(\theta)\big]F_\omega(0) + \hat{\lambda}(\theta)\big[1 - F_\omega(0)\big] = \hat{\lambda}(\theta)$. If instead $\hat{\ell}^\theta = \infty$, then $p\big(\hat{\ell}^\theta\big) = 1$ and $\alpha_\theta(\infty, \omega) = 1 - \hat{\lambda}(\theta)$. In either case, $\alpha_\theta\big(\hat{\ell}^\theta, \omega\big)$ is independent of ω, so it follows immediately from Equation 1.8 that $\psi_\theta\big(a \mid \hat{\ell}^\theta, \omega\big)$ is also independent of ω. Hence $\Psi_\theta\big(a, \hat{\ell}^\theta\big) = \psi_\theta\big(a \mid \hat{\ell}^\theta, L\big)/\psi_\theta\big(a \mid \hat{\ell}^\theta, R\big) = 1$. So for any $\hat{\boldsymbol{\pi}} \in \Pi$ and corresponding likelihood ratios $\hat{\boldsymbol{\ell}}$,

$$\frac{\partial}{\partial \ell}\varphi_\theta\big(a, \hat{\ell}^\theta\big)\bigg|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} = 1. \tag{A.15}$$

It follows from Equation 1.19 that $\chi_\theta(\hat{\boldsymbol{\ell}}) = 1$, which tells us nothing about the stability of the process in the neighborhood of $\hat{\boldsymbol{\ell}}$. To address this, note that $\chi_\theta(\cdot)$ is differentiable with respect to any ℓ^θ in the neighborhood of any $\hat{\boldsymbol{\ell}}$. So, stability is determined by whether $\lim_{\ell^\theta \to \hat{\ell}^\theta} \chi_\theta(\ell^\theta, \boldsymbol{\ell}^{-\theta}) = 1$ from below or above. If it's from below, then $\chi_\theta(\boldsymbol{\ell}) < 1$ at all points $\boldsymbol{\ell}$ in the neighborhood of $\hat{\boldsymbol{\ell}}$. So any linear approximation of the system within this neighborhood converges toward the fixed point, implying stability. But if $\chi_\theta(\boldsymbol{\ell})$ approaches 1 from above, $\chi_\theta(\boldsymbol{\ell}) > 1$ at all points $\boldsymbol{\ell}$ in the neighborhood of $\hat{\boldsymbol{\ell}}$, implying the fixed point is *not* stable. Hence the sign of the derivative of $\chi_\theta(\boldsymbol{\ell})$ with respect to $\hat{\ell}^\theta$ determines stability analogously to Lemma 8: $\hat{\boldsymbol{\ell}}$ is stable if $\frac{\partial}{\partial}\chi_\theta(\hat{\boldsymbol{\ell}}) < 0$ for all $\theta \in \Theta$, and unstable if $\frac{\partial}{\partial \ell}\chi_\theta(\hat{\boldsymbol{\ell}}) > 0$ for some $\theta \in \Theta$.

To proceed, I determine when $\frac{\partial}{\partial}\chi_\theta(\hat{\boldsymbol{\ell}}) \lesseqgtr 0$ for an arbitrary θ-type at each of the possible limit points, $\hat{\ell}^\theta = 0$ and $\hat{\ell}^\theta = \infty$, respectively.

Step 3: Stability of ℓ_t^θ near $\hat{\ell}^\theta = 0$.

Suppose $\hat{\pi}(\theta) = 1 \Rightarrow \hat{\ell}^\theta = 0$. Note that $\frac{\partial}{\partial \ell^\theta}\chi_\theta(\boldsymbol{\ell}) > 0 \Leftrightarrow \frac{\partial}{\partial \ell^\theta}\log\chi_\theta(\boldsymbol{\ell}) > 0$. Notice

$$\begin{aligned}
\frac{\partial}{\partial \ell^\theta}\log\chi_\theta(\boldsymbol{\ell})\bigg|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} &= \sum_{a=0}^{N}\psi(a, \hat{\boldsymbol{\ell}})\left(\frac{\partial}{\partial \ell}\varphi_\theta(a, 0)\right)^{-1}\left(\frac{\partial^2}{\partial \ell^2}\varphi_\theta\big(a, \ell^\theta\big)\bigg|_{\hat{\ell}^\theta=0}\right) \\
&\quad + \sum_{a=0}^{N}\left(\frac{\partial}{\partial \hat{\ell}^\theta}\psi(a, \boldsymbol{\ell})\bigg|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}}\right)\log\left(\frac{\partial}{\partial \ell}\varphi_\theta(a, 0)\right) \\
&= \sum_{a=0}^{N}\psi(a, \hat{\boldsymbol{\ell}})\left(\frac{\partial^2}{\partial \ell^2}\varphi_\theta\big(a, \ell^\theta\big)\bigg|_{\hat{\ell}^\theta=0}\right)
\end{aligned} \tag{A.16}$$

where the final equality follows from $\frac{\partial}{\partial \ell}\varphi_\theta(a, 0) = 1$ (as shown above in Equation A.15). Since

$$\frac{\partial^2}{\partial \ell^2}\varphi_\theta(a, \ell) = \frac{\partial^2}{\partial \ell^2}\left\{\Psi_\theta(a, \ell)\ell\right\} = 2\frac{\partial}{\partial \ell}\Psi_\theta(a, \ell) + \ell\frac{\partial^2}{\partial \ell^2}\Psi_\theta(a, \ell), \tag{A.17}$$

Equation A.16 reduces to

$$\frac{\partial}{\partial \ell^\theta}\log\chi_\theta(\boldsymbol{\ell})\bigg|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} = \sum_{a=0}^{N}2\psi(a, \hat{\boldsymbol{\ell}})\frac{\partial}{\partial \ell}\Psi_\theta(a, 0) \tag{A.18}$$

From Equation A.13 and using the fact that $p(0) = 0 \Rightarrow \alpha_\theta(0, \omega) = \hat{\lambda}(\theta)$ and $\Psi_\theta(a, 0) = 1$,

$$\frac{\partial}{\partial \ell}\Psi_\theta(a,0) = \big[1 - 2\hat{\lambda}(\theta)\big]\big[f_L(0) - f_R(0)\big]\left(\frac{a - N\hat{\lambda}(\theta)}{\hat{\lambda}(\theta)\big[1-\hat{\lambda}(\theta)\big]}\right), \tag{A.19}$$

so Equation A.18 implies

$$\begin{aligned}\frac{\partial}{\partial \ell^\theta}\log\chi_\theta(\boldsymbol{\ell})\bigg|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} &= \frac{2\big[1 - 2\hat{\lambda}(\theta)\big]\big[f_L(0) - f_R(0)\big]}{\hat{\lambda}(\theta)\big[1-\hat{\lambda}(\theta)\big]}\sum_{a=0}^{N}\psi(a, \hat{\boldsymbol{\ell}})\big(a - N\hat{\lambda}(\theta)\big) \\ &= \frac{2\big[1 - 2\hat{\lambda}(\theta)\big]\big[f_L(0) - f_R(0)\big]}{\hat{\lambda}(\theta)\big[1-\hat{\lambda}(\theta)\big]}\Big(N\alpha(\hat{\boldsymbol{\ell}}) - N\hat{\lambda}(\theta)\Big). \end{aligned}\tag{A.20}$$

where the second equality follows from the fact that $\sum_{a=0}^{N}\psi(a,\hat{\boldsymbol{\ell}}) = 1$ and $\sum_{a=0}^{N} a\psi(a,\hat{\boldsymbol{\ell}})$ is simply the expected value of a Binomial(N,$\alpha(\hat{\boldsymbol{\ell}})$) random variable, so $\sum_{a=0}^{N} a\psi(a,\hat{\boldsymbol{\ell}}) = N\alpha(\hat{\boldsymbol{\ell}})$. Since $f_L(0) - f_H(0) > 0$, Equation A.20 implies the following result:

$$\frac{\partial}{\partial \ell^\theta}\chi_\theta(\boldsymbol{\ell})\bigg|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} < 0 \Leftrightarrow \begin{cases} \hat{\lambda}(\theta) < \alpha(\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ \hat{\lambda}(\theta) > \alpha(\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}. \end{cases}\tag{A.21}$$

Step 4: Stability of ℓ_t^θ near $\hat{\ell}^\theta = \infty$:

Recall that ℓ_t^θ is the likelihood ratio of state L relative to state R, hence $\ell_t^\theta = \infty$ indicates confidence in state L. This is equivalent to the likelihood ratio of state R relative to state L—the inverse of ℓ_t^θ—equal to 0. Denote the inverse likelihood ratio by $\rotatebox{-60}{$\ell$}_t^\theta := (\ell_t^\theta)^{-1}$. In order to follow the logic of the case in Step 3, which determined stability of $\hat{\ell}^\theta = 0$, I assess the stability of $\hat{\ell}^\theta = \infty$ by determining the stability of the *inverse* likelihood ratio $\rotatebox{-60}{$\ell$}$ at 0. The stability coefficient of interest is now that of the inverse likelihood ratio:

$$\tilde{\chi}_\theta(\hat{\rotatebox{-60}{ℓ}}) = \prod_{a=0}^{N}\left(\frac{\partial}{\partial \rotatebox{-60}{ℓ}}\tilde{\varphi}_\theta\big(a, \hat{\rotatebox{-60}{ℓ}}^\theta\big)\right)^{\tilde{\psi}(a,\rotatebox{-60}{ℓ})}\tag{A.22}$$

where $\tilde{\varphi}_\theta(a, \rotatebox{-60}{ℓ})$ is the transition equation for the process $\langle \rotatebox{-60}{$\ell$}_t^\theta\rangle$:

$$\tilde{\varphi}_\theta(a, \rotatebox{-60}{ℓ}) = \tilde{\Psi}_\theta(a, \rotatebox{-60}{ℓ})\,\rotatebox{-60}{ℓ},$$

with

$$\tilde{\Psi}_\theta(a, \rotatebox{-60}{ℓ}) = \frac{\tilde{\psi}_\theta(a \mid \rotatebox{-60}{ℓ}, R)}{\tilde{\psi}_\theta(a \mid \rotatebox{-60}{ℓ}, L)}.$$

$\tilde{\psi}_\theta(a \mid \rotatebox{-60}{ℓ}, \omega)$ is the direct analog of $\psi_\theta(a \mid \ell, \omega)$: it is the probability of observing a at belief $\rotatebox{-60}{$\ell$}$ in state ω according to type-θ's theory of tastes.

As above, $\tilde{\chi}_\theta(\hat{\boldsymbol{\mathscr{r}}}) = 1$ if $\hat{\boldsymbol{\mathscr{r}}}^\theta = 0$, so we must calculate the derivative or $\tilde{\chi}_\theta(\hat{\boldsymbol{\mathscr{r}}})$ with respect to $\boldsymbol{\mathscr{r}}^\theta$ and evaluate the sign at 0. As above, the fixed point is stable the sign is negative, and unstable when positive. Identical calculations to those in Step 3 yield

$$\frac{\partial}{\partial \boldsymbol{\mathscr{r}}^\theta} \log \tilde{\chi}_\theta(\boldsymbol{\mathscr{r}})\bigg|_{\boldsymbol{\mathscr{r}}=\hat{\boldsymbol{\mathscr{r}}}} = \sum_{a=0}^{N} 2\tilde{\psi}(a, \hat{\boldsymbol{\mathscr{r}}}) \frac{\partial}{\partial \boldsymbol{\mathscr{r}}} \tilde{\Psi}_\theta\big(a, \boldsymbol{\mathscr{r}}\big)\bigg|_{\boldsymbol{\mathscr{r}}=\hat{\boldsymbol{\mathscr{r}}}}. \tag{A.23}$$

Note that

$$\frac{\partial}{\partial \boldsymbol{\mathscr{r}}} \tilde{\Psi}_\theta(a, \boldsymbol{\mathscr{r}}) = \tilde{\Psi}(a, \boldsymbol{\mathscr{r}}) \Bigg\{ \frac{\big[1 - 2\hat{\lambda}(\theta)\big]}{(1 + \boldsymbol{\mathscr{r}})^2} \Bigg[f_L\big(p(\boldsymbol{\mathscr{r}})\big) \left(\frac{a - N\alpha_\theta(\boldsymbol{\mathscr{r}}, L)}{\alpha_\theta(\boldsymbol{\mathscr{r}}, L)\big[1 - \alpha_\theta(\boldsymbol{\mathscr{r}}, L)\big]} \right)$$
$$f_R\big(p(\boldsymbol{\mathscr{r}})\big) \left(\frac{a - N\alpha_\theta(\boldsymbol{\mathscr{r}}, R)}{\alpha_\theta(\boldsymbol{\mathscr{r}}, R)\big[1 - \alpha_\theta(\boldsymbol{\mathscr{r}}, R)\big]} \right) \Bigg] \Bigg\}. \tag{A.24}$$

At $\boldsymbol{\mathscr{r}} = 0$, $p(\boldsymbol{\mathscr{r}}) = 1$ and $\alpha_\theta(\boldsymbol{\mathscr{r}}, \omega) = 1 - \hat{\lambda}(\theta)$, so when $\hat{\boldsymbol{\mathscr{r}}}^\theta = 0$,

$$\frac{\partial}{\partial \boldsymbol{\mathscr{r}}^\theta} \tilde{\Psi}_\theta\big(a, \boldsymbol{\mathscr{r}}^\theta\big)\bigg|_{\boldsymbol{\mathscr{r}}=\hat{\boldsymbol{\mathscr{r}}}} = \big[1 - 2\hat{\lambda}(\theta)\big]\big[f_L(1) - f_R(1)\big] \left(\frac{a - N\big(1 - \hat{\lambda}(\theta)\big)}{\hat{\lambda}(\theta)\big[1 - \hat{\lambda}(\theta)\big]} \right). \tag{A.25}$$

Plugging into Equation A.23,

$$\frac{\partial}{\partial \boldsymbol{\mathscr{r}}^\theta} \log \tilde{\chi}_\theta(\boldsymbol{\mathscr{r}})\bigg|_{\boldsymbol{\mathscr{r}}=\hat{\boldsymbol{\mathscr{r}}}} = \frac{2\big[1 - 2\hat{\lambda}(\theta)\big]\big[f_L(1) - f_R(1)\big]}{\hat{\lambda}(\theta)\big[1 - \hat{\lambda}(\theta)\big]} \Big(N\alpha(\hat{\boldsymbol{\mathscr{r}}}) - N\big(1 - \hat{\lambda}(\theta)\big)\Big). \tag{A.26}$$

Since $f_R(1) > f_L(1)$, we have the following result:

$$\frac{\partial}{\partial \boldsymbol{\mathscr{r}}^\theta} \tilde{\chi}_\theta(\boldsymbol{\mathscr{r}})\bigg|_{\boldsymbol{\mathscr{r}}=\hat{\boldsymbol{\mathscr{r}}}} < 0 \Leftrightarrow \begin{cases} 1 - \hat{\lambda}(\theta) > \alpha(\hat{\boldsymbol{\mathscr{r}}}) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ 1 - \hat{\lambda}(\theta) < \alpha(\hat{\boldsymbol{\mathscr{r}}}) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2} \end{cases} \tag{A.27}$$

Step 5. Linking stability to expected action frequencies.

Finally, I write the stability conditions derived in Steps 3 and 4—Results A.21 and A.27—in terms of the expected and true action frequencies at $\hat{\boldsymbol{\ell}}$. First, note that

$$\widehat{\mathscr{F}_\theta}\big(M_\theta(\hat{\ell}), \hat{\ell}\big) = \begin{cases} \hat{\lambda}(\theta) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ 1 - \hat{\lambda}(\theta) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}. \end{cases} \tag{A.28}$$

Second, note that by definition, $\alpha(\hat{\boldsymbol{\ell}}) = \mathscr{F}(A, \hat{\boldsymbol{\ell}})$ and $1 - \alpha(\hat{\boldsymbol{\ell}}) = \mathscr{F}(B, \hat{\boldsymbol{\ell}})$. Plugging these identities into Results A.21 and A.27 respectively yield

$$\frac{\partial}{\partial \ell^\theta} \chi_\theta(\boldsymbol{\ell})\bigg|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} < 0 \Leftrightarrow \begin{cases} \widehat{\mathscr{F}_\theta}\big(M_\theta(0), 0\big) < \alpha(\hat{\boldsymbol{\ell}}) = \mathscr{F}(A, \hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ \widehat{\mathscr{F}_\theta}\big(M_\theta(0), 0\big) < 1 - \alpha(\hat{\boldsymbol{\ell}}) = \mathscr{F}(B, \hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}, \end{cases} \tag{A.29}$$

and

$$\frac{\partial}{\partial \boldsymbol{r}^{\theta}}\tilde{\chi}_{\theta}(\boldsymbol{r})\bigg|_{\boldsymbol{r}=\hat{\boldsymbol{r}}} < 0 \Leftrightarrow \left\{ \begin{array}{ll} \widehat{\mathscr{F}_{\theta}}\big(M_{\theta}(\infty),\infty\big) < 1-\alpha(\hat{\boldsymbol{\ell}}) = \mathscr{F}(B,\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ \widehat{\mathscr{F}_{\theta}}\big(M_{\theta}(\infty),\infty\big) < \alpha(\hat{\boldsymbol{\ell}}) = \mathscr{F}(A,\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}. \end{array} \right. \tag{A.30}$$

Finally, we can rewrite the $\mathscr{F}(X,\hat{\boldsymbol{\ell}})$ terms on the right-hand side of the expressions above in terms of a θ-type's expected majority action at $\hat{\boldsymbol{\ell}}$. Note

$$M_{\theta}(0) = \left\{ \begin{array}{lll} A & \text{if} & \hat{\lambda}(\theta) > \frac{1}{2} \\ B & \text{if} & \hat{\lambda}(\theta) < \frac{1}{2}, \end{array} \right. \quad \text{and} \quad M_{\theta}(\infty) = \left\{ \begin{array}{lll} B & \text{if} & \hat{\lambda}(\theta) > \frac{1}{2} \\ A & \text{if} & \hat{\lambda}(\theta) < \frac{1}{2}. \end{array} \right. \tag{A.31}$$

Appropriately incorporating these identities into A.32 and A.33 finally yields the following stability conditions:

$$\frac{\partial}{\partial \ell^{\theta}}\chi_{\theta}(\boldsymbol{\ell})\bigg|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} < 0 \Leftrightarrow \left\{ \begin{array}{ll} \widehat{\mathscr{F}_{\theta}}\big(M_{\theta}(0),0\big) < \mathscr{F}\big(M_{\theta}(0),\hat{\boldsymbol{\ell}}\big) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ \widehat{\mathscr{F}_{\theta}}\big(M_{\theta}(0),0\big) < \mathscr{F}(M_{\theta}(0),\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}, \end{array} \right. \tag{A.32}$$

and

$$\frac{\partial}{\partial \boldsymbol{r}^{\theta}}\tilde{\chi}_{\theta}(\boldsymbol{r})\bigg|_{\boldsymbol{r}=\hat{\boldsymbol{r}}} < 0 \Leftrightarrow \left\{ \begin{array}{ll} \widehat{\mathscr{F}_{\theta}}\big(M_{\theta}(\infty),\infty\big) < \mathscr{F}(M_{\theta}(\infty),\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ \widehat{\mathscr{F}_{\theta}}\big(M_{\theta}(\infty),\infty\big) < \mathscr{F}(M_{\theta}(\infty),\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}. \end{array} \right. \tag{A.33}$$

Hence, in all cases—$\hat{\ell} \in \{0,\infty\}$ and $\hat{\lambda}(\theta) \lessgtr \frac{1}{2}$—that the stability condition holds for a θ type if and only if

$$\widehat{\mathscr{F}_{\theta}}\Big(M\big(\hat{\ell}(\theta)\big),\hat{\ell}(\theta)\Big) < \mathscr{F}\Big(M\big(\hat{\ell}(\theta)\big),\hat{\boldsymbol{\ell}}\Big), \tag{A.34}$$

completing the proof.

□

Proof of Proposition 5.

Proof. Suppose $\hat{\boldsymbol{\ell}} \in \mathscr{L}$ is such that $\hat{\ell}^{\theta} = 0$ for all $\theta \in \Theta$. I show that this point of long-run agreement is necessarily unstable; the proof for the alternative case where $\hat{\ell}^{\theta} = \infty$ for all $\theta \in \Theta$, which follows analogously, is omitted.

Instability of asymptotic agreement is established along the lines of Proposition 4. However, to demonstrate the robustness of this result, I extend the proof of Proportion 4 to allow for known quality differences. Without loss of generality, assume $\Delta_q \geq 0$. The logic is identical: $\hat{\boldsymbol{\ell}}$ is unstable if $\frac{\partial}{\partial \ell^{\theta}}\chi_{\theta}(\boldsymbol{\ell})\big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} > 0$ for some $\theta \in \Theta$. The only aspect of that proof that we must change is the function $\alpha_\theta(\ell^\theta, \omega)$ (now given by Equation 1.9). Since $\Delta_q \neq 0$, $\frac{\partial}{\partial \ell} p(0,\theta) = 1/v(\theta)$. Hence

$$\frac{\partial}{\partial \ell}\alpha_\theta(0,\omega) = f_\omega(0)\left[\sum_{\tilde\theta\in\Theta^l}\frac{\hat g(\tilde\theta \mid \theta)}{v(\tilde\theta)} - \sum_{\tilde\theta\in\Theta^r}\frac{\hat g(\tilde\theta \mid \theta)}{v(\tilde\theta)}\right]. \tag{A.35}$$

We can now rely on many of the derivations in Proposition 4. It follows that

$$\begin{aligned}\frac{\partial}{\partial \ell^\theta}\log\chi_\theta(\boldsymbol{\ell})\Big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} &= \sum_{a=0}^{N} 2\psi(a,\hat{\boldsymbol{\ell}})\frac{\partial}{\partial \ell}\Psi_\theta(a,0)\\ = \frac{2\big[f_L(0)-f_R(0)\big]}{\alpha_\theta(0,\omega)\big[1-\alpha_\theta(0,\omega)\big]}&\left[\sum_{\tilde\theta\in\Theta^l}\frac{\hat g(\tilde\theta \mid \theta)}{v(\tilde\theta)} - \sum_{\tilde\theta\in\Theta^r}\frac{\hat g(\tilde\theta \mid \theta)}{v(\tilde\theta)}\right]\sum_{a=0}^{N}\psi(a,\hat{\boldsymbol{\ell}})(a-N\alpha_\theta(0,\omega)). \quad \text{(A.36)}\end{aligned}$$

The first equality follows from Equation A.18. To arrive at the second equality, first plug $\frac{\partial}{\partial \ell}\alpha_\theta(0,\omega)$ from A.35 into the expression for $\frac{\partial}{\partial \ell}\psi_\theta(a \mid \ell,\omega)$ in Equation A.10, then plug the result into Equation A.12, and evaluate the expression at $\ell^\theta = 0$. Given $\Delta_q \geq 0$, all right and passive players take A at $\hat{\boldsymbol{\ell}}$, so $\alpha_\theta(0,L) = \alpha_\theta(0,R) = 1 - \sum_{\tilde\theta\in\Theta^l}\hat g(\tilde\theta \mid \theta)$—$\theta$'s perceived measure of all types other than active left types. Since $\sum_{a=0}^{N}\psi(a,\hat{\boldsymbol{\ell}})a = \mathbb{E}[\tilde a]$ assuming $\tilde a \sim \text{Binomial}(N,\alpha(\hat{\boldsymbol{\ell}}))$, and since $f_L(0) > f_R(0)$, it follows from Equation A.36 that $\frac{\partial}{\partial \ell^\theta}\log\chi_\theta(\boldsymbol{\ell})\big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} > 0$ if and only if

$$\left[\sum_{\tilde\theta\in\Theta^l}\frac{\hat g(\tilde\theta \mid \theta)}{v(\tilde\theta)} - \sum_{\tilde\theta\in\Theta^r}\frac{\hat g(\tilde\theta \mid \theta)}{v(\tilde\theta)}\right]\big[\alpha(\hat{\boldsymbol{\ell}}) - \alpha_\theta(0,\omega)\big] > 0. \tag{A.37}$$

I now argue that, generically, Condition A.37 must hold for some $\theta \in \Theta$. First, for any $\theta \in \Theta^r$, $\alpha_\theta(0,\omega) > \alpha(\hat{\boldsymbol{\ell}})$. If not, this implies that an active right type *over*estimates the share of active left types, providing a contradiction. Similarly, for any $\theta \in \Theta^l$, $\alpha_\theta(0,\omega) < \alpha(\hat{\boldsymbol{\ell}})$. Next, define $V(\theta) := \sum_{\tilde\theta\in\Theta^l}\frac{\hat g(\tilde\theta|\theta)}{v(\tilde\theta)} - \sum_{\tilde\theta\in\Theta^r}\frac{\hat g(\tilde\theta|\theta)}{v(\tilde\theta)}$. The only way for condition A.37 to fail at all θ is if $V(\theta) < 0$ for all $\theta \in \Theta^l$, and $V(\theta) > 0$ for all $\theta \in \Theta^r$. For a contradiction, suppose this is true. Recall $v(\theta) = (4k\theta + \Delta_q)/(4k\theta - \Delta_q)$. From the definition of Θ^r in Lemma 2, $1/v(\theta)$ is increasing on Θ^r. Because $\widehat{G}(\tilde\theta \mid \theta)$ first-order stochastically dominates $\widehat{G}(\tilde\theta \mid \theta')$ whenever $\theta > \theta'$, $\sum_{\tilde\theta\in\Theta^r}\frac{\hat g(\tilde\theta|\theta)}{v(\tilde\theta)}$ is increasing in θ. Hence, for large enough θ, $V(\theta) < 0$. Similarly, for small enough θ, $V(\theta) > 0$. Thus Condition A.37 must fail for some θ, implying a vector of beliefs such that all agents agree on the state is necessarily unstable.

□

Proof of Proposition 6.

Proof. Proposition 4 determines Π^*. From Proposition 5, we know $(0,0) \notin \Pi^*$ and $(1,1) \notin \Pi^*$. But $\hat{\boldsymbol{\pi}} = (0,1)$ and $\hat{\boldsymbol{\pi}} = (1,0)$ satisfy the stability requirement of Proposition 4: each type observes more taking her anticipated majority action than expected. We need only show that beliefs reach a neighborhood of these stable limit points. Suppose $\langle \ell^l_t, \ell^r_t \rangle$ reaches the north-west quadrant of belief space (see Figure 1.4), which we define by all points $\boldsymbol{\ell}_t$ such that $\ell^r_t > L_l(\ell^l_t)$ and $\ell^l_t < L_r(\ell^r_t)$ (see footnote 51). Call this set L_{NW}. Restricted to L_{NW}, each $\langle \ell^l_t \rangle$ and $\langle 1/\ell^r_t \rangle$ are non-negative supermartingales, and thus, by the Martingale Convergence Theorem, converge. Since 0 is a stable limit point of each of these processes, they either both converge to 0 (which occurs with positive probability) or exit L_{NW} in finite time. Similarly, consider the south-east quadrant defined by all points $\boldsymbol{\ell}_t$ such that $\ell^r_t < L_l(\ell^l_t)$ and $\ell^l_t > L_r(\ell^r_t)$. Call this space L_{SE}. Restricted to L_{SE}, each $\langle \ell^r_t \rangle$ and $\langle 1/\ell^l_t \rangle$ are non-negative supermartingales, and thus converge. Hence, if process $\langle \ell^l_t, \ell^r_t \rangle$ enters L_{SE}, it either converges to $(\infty, 0)$ (which occurs with positive probability) or exits. Further more, since no stable limit points exist outside of $L_{NW} \cup L_{SE}$, the process must enter $L_{NW} \cup L_{SE}$ infinitely often. Thus, eventually, the process converges to one of the two stationary points. □

Proof of Lemma 9.

Proof. Since

$$\mathbb{E}[\ell^\theta_{t+1} \mid \boldsymbol{\ell}_t] = \sum_{a_t=0}^{N} \psi(a_t \mid \boldsymbol{\ell}_t, R)\Psi_\theta\big(a_t, \ell^\theta_t\big)\ell^\theta_t, \tag{A.38}$$

$\mathbb{E}[\ell^\theta_{t+1} \mid \boldsymbol{\ell}_t] > \ell^\theta_t \Leftrightarrow \xi_\theta(\ell^l_t, \ell^r_t) \equiv \sum_{a_t=0}^{N} \psi(a_t \mid \boldsymbol{\ell}_t, R)\Psi_\theta\big(a_t, \ell^\theta_t\big) > 1$. We want to assess whether this holds for each θ in a neighborhood of $\boldsymbol{\ell} = \mathbf{0} = (0,0)$. Since $\mathbf{0}$ is a fixed point of the belief process for each θ, $\xi_\theta(0,0) = 1$. Hence we consider the (first-order) Taylor-Series expansion of $\xi_\theta(\ell^l_t, \ell^r_t)$ near $\mathbf{0}$. Note that

$$\begin{aligned}\xi_\theta(\epsilon,\epsilon) \approx \xi_\theta(0,0) + \sum_{a=0}^{N} \psi(a \mid \mathbf{0}, R)\frac{\partial}{\partial \ell^\theta}\Psi_\theta\big(a,0\big) \\ + \epsilon\left(\sum_{a=0}^{N}\left(\frac{\partial}{\partial \ell^l}\psi(a \mid \mathbf{0}, R) + \frac{\partial}{\partial \ell^r}\psi(a \mid \mathbf{0}, R)\right)\Psi_\theta\big(a,0\big)\right).\end{aligned} \tag{A.39}$$

From Equation A.40,

$$\frac{\partial}{\partial \ell^\theta}\psi(a \mid \mathbf{0}, R) = (1-2\lambda)\psi(a \mid \mathbf{0}, R) f_R(0)\left(\frac{a - N\lambda}{\lambda(1-\lambda)}\right), \tag{A.40}$$

and since $\Psi_\theta\big(a,0\big) = 1$,

$$\sum_{a=0}^{N} \frac{\partial}{\partial \ell^\theta}\psi(a \mid \mathbf{0}, R)\Psi_\theta\big(a,0\big) = (1-2\lambda) f_R(0)\sum_{a=0}^{N}\psi(a \mid \mathbf{0}, R)\left(\frac{a - N\lambda}{\lambda(1-\lambda)}\right),$$

which equals $(1-2\lambda)f_R(0)\mathbb{E}[a-N\lambda]/[\lambda(1-\lambda)]$ where the expectation is with respect to $a \sim \text{Binomial}(N,\lambda)$. Thus, $\mathbb{E}[a-N\lambda]=0$. Substituting this result into Equation A.39 yields

$$\xi_\theta(\epsilon,\epsilon) \approx 1 + \sum_{a=0}^{N} \psi(a \mid \mathbf{0}, R)\frac{\partial}{\partial \ell^\theta}\Psi_\theta\big(a,0\big).$$

Finally, recall that $\mathbb{E}[\ell^\theta_{t+1} \mid \boldsymbol{\ell}_t = (\epsilon,\epsilon)] > \ell^\theta_t = \epsilon \Leftrightarrow \xi_\theta(\epsilon,\epsilon) > 1 \Leftrightarrow \sum_{a=0}^{N}\psi(a \mid \mathbf{0}, R)\frac{\partial}{\partial \ell^\theta}\Psi_\theta\big(a,0\big) > 0$. From Equation A.19,

$$\frac{\partial}{\partial \ell}\Psi_\theta(a,0) = \big[1-2\hat{\lambda}(\theta)\big]\big[f_L(0)-f_R(0)\big]\left(\frac{a-N\hat{\lambda}(\theta)}{\hat{\lambda}(\theta)\big[1-\hat{\lambda}(\theta)\big]}\right),$$

so

$$\sum_{a=0}^{N}\psi(a \mid \mathbf{0}, R)\frac{\partial}{\partial \ell^\theta}\Psi_\theta\big(a,0\big) = N\frac{\big[1-2\hat{\lambda}(\theta)\big]\big[f_L(0)-f_R(0)\big]}{\hat{\lambda}(\theta)\big[1-\hat{\lambda}(\theta)\big]}(\lambda-\hat{\lambda}(\theta)),$$

which exceeds 0 if and only if $\big[1-2\hat{\lambda}(\theta)\big]\big[\lambda-\hat{\lambda}(\theta)\big] > 0$. With Strong projection, $\hat{\lambda}^r > \lambda > 1/2$, so $[1-2\hat{\lambda}^r][\lambda-\hat{\lambda}^r] > 0$. Hence, ℓ^r_t is locally a submartingale in the neighborhood of $\boldsymbol{\ell}=(0,0)$. Likewise, $\hat{\lambda}^l < 1/2$, so $[1-2\hat{\lambda}^l][\lambda-\hat{\lambda}^l] > 0$. Hence, ℓ^l_t is locally a submartingale in the neighborhood of $\boldsymbol{\ell}=(0,0)$. □

Proof of Proposition 7.

Proof. We must show that $\langle\boldsymbol{\ell}_t\rangle$ is unstable at each $\hat{\boldsymbol{\ell}}$. First consider a limit point in which types agree, $\hat{\boldsymbol{\ell}}=(0,0)$. At this belief, the observed frequency of A converges to λ, while right types anticipate $\hat{\lambda}^r > \lambda$. By Proposition 4, ℓ^r_t is unstable near 0. ℓ^l_t must also be unstable near 0: by Lemma 10, there exists an $\epsilon > 0$ such that ℓ^r_t is submartingale so long as $\ell^l_t < \epsilon$. If $\ell^l_t < \epsilon$ for all t, then ℓ^r_t diverges to ∞ and the frequency of A converges to 1, which necessarily implies $\ell^l_t \to \infty$, a contradiction. The analogous argument holds at *any* potential limit point $\hat{\boldsymbol{\ell}}$: for some $\theta \in \{l,r\}$, ℓ^θ_t is immediately unstable by Proposition 4, and the martingale property of the unstable ℓ^θ_t, which moves away from $\hat{\ell}^\theta$ in expectation, implies $\ell^{\theta'}_t$ $\theta' \neq \theta$ necessarily exits a neighborhood about $\hat{\ell}^{\theta'}$, contradicting stability of $\ell^{\theta'}_t$. □

Proof of Lemma 10.

Proof. The proof of Lemma 9 shows that $\mathbb{E}[\ell^\theta_{t+1} \mid \boldsymbol{\ell}_t = (\epsilon,\epsilon)] > \ell^\theta_t = \epsilon \Leftrightarrow \big[1-2\hat{\lambda}(\theta)\big]\big[\lambda-\hat{\lambda}(\theta)\big] > 0$. This holds for $\hat{\lambda}^r > \lambda > 1/2$, but fails for $\hat{\lambda}^l \in (1/2,\lambda)$. Hence ℓ^r_t is locally a submartingale in the neighborhood of $\boldsymbol{\ell}=(0,0)$ whereas ℓ^l_t is locally a supermartingale in the neighborhood of $\boldsymbol{\ell}=(0,0)$. □

Proof of Proposition 8.

Proof. This follows from a direct application of Proposition 4. In any stable equilibrium, all players who think their taste matches the majority taste must take the majority action, X. In Case 1 ($\tilde{\theta} < 0$), all right types (measure λ) and all left types with $\hat{\lambda}(\theta) < 1/2$ (measure $G(\tilde{\theta})$) take the majority action. By Proposition 4, this outcome is stable if and only if no type expects to observe a share greater than $G(\tilde{\theta}) + \lambda$ take X at their respective equilibrium beliefs. This is true so long as $G(\tilde{\theta}) + \lambda > \max\left\{1 - \hat{\lambda}(\underline{\theta}), \hat{\lambda}(\overline{\theta})\right\}$. In Case 2 ($\tilde{\theta} > 0$), some right types think they are in the minority. Now all left types (measure $1-\lambda$) and right types with $\hat{\lambda}(\theta) > 1/2$ (measure $1 - G(\tilde{\theta})$) take X. Hence, by Proposition 4, this outcome is stable if and only if $(1-\lambda) + 1 - G(\tilde{\theta}) = 2 - (\lambda - -G(\tilde{\theta})) > \max\left\{1 - \hat{\lambda}(\underline{\theta}), \hat{\lambda}(\overline{\theta})\right\}$. □

Proof of Proposition 9.

Proof. As N grows large, for any θ, there exists some (X, X') such that $\pi_2^\theta(X, X')$ is arbitrarily close to 1. The only case in which this does not imply that a_2/N is arbitrarily close to 0 or 1—nearly all players take the same action—is when either $\pi_2^l(B, A) \approx 1$ and $\pi_2^r(B, A) \approx 1$ or $\pi_2^l(A, B) \approx 1$ and $\pi_2^r(A, B) \approx 1$. That is, we do not observe a (nearly) uniform herd in period 2 whenever both types grow confident in a state where it is optimal for players with opposing tastes to take different actions. I focus on the case where $\pi^\theta(B, A)$ is arbitrarily close to 1 for each θ.[65] So $a_2/N \approx \lambda$. More precisely, by the Strong Law of Large Numbers, there exists some $\epsilon(N) > 0$ such that $a_2/N = \lambda - \epsilon(N)$, where $\epsilon(N) \to 0$ as $N \to \infty$. Now we evaluate the perceived likelihood ratio of observing $a_2/N \approx \lambda - \epsilon_N$ in state (B, A) with (B, A) for a right type. Notice that a right type expects to observe $a_2/N = \hat{\lambda}^r - \hat{\epsilon}(N)$ for some $\hat{\epsilon}(N) > 0$ such that $\hat{\epsilon}(N) \to 0$ as $N \to \infty$. So this likelihood ratio is

$$\mathcal{L}^r = \left[\left(\frac{\widehat{\Pr}^\theta\left(X_{n2} = A \,\middle|\, \omega \in \Omega^{BA}\right)}{\widehat{\Pr}^\theta\left(X_{n2} = A \,\middle|\, \omega \in \Omega^{AB}\right)}\right)^{a_2/N} \left(\frac{1 - \widehat{\Pr}^\theta\left(X_{n2} = A \,\middle|\, \omega \in \Omega^{BA}\right)}{1 - \widehat{\Pr}^\theta\left(X_{n2} = A \,\middle|\, \omega \in \Omega^{AB}\right)}\right)^{1 - a_2/N}\right]^N \tag{A.41}$$

$$\mathcal{L}^r = \left(\frac{\hat{\lambda}^r - \hat{\epsilon}(N)}{1 - \hat{\lambda}^r + \hat{\epsilon}(N)}\right)^{\lambda - \epsilon_N} \left(\frac{1 - \hat{\lambda}^r + \hat{\epsilon}(N)}{\hat{\lambda}^r - \hat{\epsilon}(N)}\right)^{1 - \lambda + \epsilon_N} = \left(\frac{\hat{\lambda}^r - \hat{\epsilon}(N)}{1 - \hat{\lambda}^r + \hat{\epsilon}(N)}\right)^{2\lambda - 1 - 2\epsilon_N} \tag{A.42}$$

Note that $(\mathcal{L}^r)^{1/N} > 1$ if and only if both $\hat{\lambda}^r > \frac{1}{2} + \hat{\epsilon}(N)$ and $\lambda > \frac{1}{2} + \epsilon(N)$. Since $\hat{\lambda}^r > \lambda > \frac{1}{2}$, this holds for sufficiently large N. So $(\mathcal{L}^r)^{1/N} > 1$ implies $\mathcal{L}^r \to \infty$ as $N \to \infty$. So right types in period 3 are arbitrarily confident that A is their optimal choice.

Left types, however, draw the opposite inference. As above,

[65] Proving the alternative case in which all types are arbitrarily confident in (A, B) is essentially identical.

$$\mathcal{L}^l = \left(\frac{\hat{\lambda}^l - \hat{\epsilon}^l(N)}{1-\hat{\lambda}^l + \hat{\epsilon}^l(N)}\right)^{2\lambda-1-2\epsilon(N)}, \tag{A.43}$$

so $(\mathcal{L}^l)^{1/N} > 1$ if and only if both $\hat{\lambda}^l > \frac{1}{2}+\hat{\epsilon}^l(N)$ and $\lambda > \frac{1}{2}+\epsilon(N)$. Since $\hat{\lambda}^l < \frac{1}{2}$, this fails to hold for sufficiently large N. So $(\mathcal{L}^l)^{1/N} < 1$ implies $\mathcal{L}^l \to 0$ as $N \to \infty$. Hence left types in $t=3$ grow arbitrarily confident that A is their optimal choice. Thus all players enter $t=3$ arbitrarily confident that A is their optimal choice. Only those in $t=3$ with strong contrary signals take B, but the measure of such players goes to 0 as $N \to \infty$. Hence $a_3/N \to 1$ as $N \to \infty$. Once $a_3/N \approx 1$ is observed, players remain confident that A is optimal for all types. As all $\omega \in \Omega^{AA}$ are absorbing states, beliefs remain confident that $\omega \in \Omega^{AA}$ for all future periods.

□

Proof of Proposition 11.

Proof. Let $\underline{\omega} := (L, \underline{\lambda})$ and $\overline{\omega} := (R, \overline{\lambda})$. Suppose the history up to time t is a herd on A: $h_t = h_t^A$. For any finite t, this occurs with positive probability. By Lemma 11, for large t, this initial history moves both $\pi^l(\underline{\omega})$ and $\pi^r(\overline{\omega})$ close to 1. Hence, given arbitrary neighborhoods about beliefs degenerate on states $\underline{\omega}$ and $\overline{\omega}$, denoted $\mathscr{N}(\underline{\omega})$ and $\mathscr{N}(\overline{\omega})$, respectively, with positive probability, $\boldsymbol{\pi}_t^l \in \mathscr{N}(\underline{\omega})$ and $\boldsymbol{\pi}_t^r(\overline{\omega}) \in \mathscr{N}(\overline{\omega})$ for some finite t. Now we must simply show that the joint-belief process is stochastically stable within these neighborhoods. I build on the stability arguments of Proposition 4, extending the logic to larger state spaces (the state space considered in Proposition 4 is binary). As above, I work with likelihood ratios. Only for the purpose of this proof, I define left-type likelihood ratios relative to state $\underline{\omega}$, but right-type's relative to $\overline{\omega}$; let $\ell_t^l(\omega) := \pi^l(\omega)/\pi^l(\underline{\omega})$ and $\ell_t^r(\omega) := \pi^r(\omega)/\pi^r(\overline{\omega})$. Let $\boldsymbol{\ell}_t^l = (\ell_t^l(L,\overline{\lambda}), \ell_t^l(R,\underline{\lambda}), \ell_t^l(R,\overline{\lambda}))$ and $\boldsymbol{\ell}_t^r = (\ell_t^r(L,\underline{\lambda}), \ell_t^r(L,\overline{\lambda}), \ell_t^r(R,\underline{\lambda}))$. With these definitions, $\boldsymbol{\pi}_t^l \in \mathscr{N}(\underline{\omega})$ and $\boldsymbol{\pi}_t^r \in \mathscr{N}(\overline{\omega}) \Leftrightarrow$ for each $\theta = l, r$, $\boldsymbol{\ell}_t^\theta$ is in a neighborhood about the origin, $\mathbf{0} \in \mathbb{R}_+^3$.

Step 2: Linearized System Like Proposition 4, I show the stability of the linear approximation of the system near fixed points $\hat{\boldsymbol{\ell}}^l = \mathbf{0}$ and $\hat{\boldsymbol{\ell}}^r = \mathbf{0}$. The system is multi-dimensional; let $\boldsymbol{\ell}_{t+1}^\theta = \varphi(a, \boldsymbol{\ell}_t^\theta)$ define the transition function for a θ-type's vector of beliefs, and each element evolves according to $\ell_{t+1}^\theta(\omega) = \varphi_\theta(a, \boldsymbol{\ell}_t^\theta, \omega) := \ell_t^\theta(\omega)\psi_\theta(a \mid \boldsymbol{\ell}_t^\theta, \omega)/\psi_\theta(a \mid \boldsymbol{\ell}_t^\theta, \omega^*)$ where $\omega^* = \overline{\omega}$ if $\theta = r$, and $\omega^* = \underline{\omega}$ if $\theta = l$.

For each θ, the system is approximated by the Jacobian of $\varphi_\theta(a, \boldsymbol{\ell}^\theta)$ at $\hat{\boldsymbol{\ell}}^\theta = \mathbf{0}$. Note that the (ω', ω) term of the Jacobian (the derivative of the $\ell_t^\theta(\omega')$ transition function with respect to belief $\ell_t^\theta(\omega)$) is

$$\frac{\partial}{\partial \ell(\omega)}\varphi(a, \boldsymbol{\ell}, \omega') = \ell(\omega')\frac{\partial}{\partial \ell(\omega)}\left(\frac{\psi_\theta(a \mid \boldsymbol{\ell}, \omega')}{\psi_\theta(a \mid \boldsymbol{\ell}, \omega^*)}\right) + \frac{\partial \ell(\omega')}{\partial \ell(\omega)}\left(\frac{\psi_\theta(a \mid \boldsymbol{\ell}, \omega')}{\psi_\theta(a \mid \boldsymbol{\ell}, \omega^*)}\right) \tag{A.44}$$

which, evaluated at $\boldsymbol{\ell} = \mathbf{0}$, is 0 when $\omega' \neq \omega$—off-diagonal terms of the Jacobian are 0. Hence, the approximate system is diagonal: to a first-order approximation, the likelihood ratio of ω' has no effect on the evolution of the likelihood ratio of $\omega \neq \omega'$. As such, the fixed point is stable if each dimension satisfies the uni-dimensional stability criterion developed in Proposition 4. Accordingly, the remainder of this proof follows the same steps as Proposition 4, but within this modified environment; for brevity, the arguments here are terse—some analogous derivations in 4 are referenced for details.

From Proposition 4, $\boldsymbol{\ell}_t^\theta$ will remain in the neighborhood of $\mathbf{0}$ so long as for each θ, the "stability coefficient" (Equation 1.19) for each ω and $a \in \{0, 1, ..., N\}$ is less than one at $\hat{\boldsymbol{\ell}}^l = \mathbf{0}$, $\hat{\boldsymbol{\ell}}^r = \mathbf{0}$:

$$\chi_\theta(\hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r, \omega)\bigg|_{(\hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r) = (\mathbf{0}, \mathbf{0})} < 1, \tag{A.45}$$

where

$$\chi_\theta(\hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r, \omega) = \prod_{a=0}^{N} \left(\frac{\partial}{\partial \ell(\omega)} \varphi_\theta\big(a, \boldsymbol{\ell}^\theta, \omega\big) \right)^{\psi(a, \hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r)}, \tag{A.46}$$

and $\psi(a, \hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r)$ is the true probability of observation a at beliefs $\hat{\boldsymbol{\ell}}^l$, $\hat{\boldsymbol{\ell}}^r$. Note $\psi(a, \mathbf{0}, \mathbf{0}) = 1 \Leftrightarrow a = N$, and 0 otherwise; all agents play A at these beliefs. So, $\chi_\theta(\mathbf{0}, \mathbf{0}, \omega) < 1 \Leftrightarrow \frac{\partial}{\partial \ell(\omega)} \varphi_\theta\big(N, \mathbf{0}, \omega\big) < 1$. From A.44, for any ω, $\frac{\partial}{\partial \ell(\omega)} \varphi_\theta\big(N, \boldsymbol{\ell}^\theta, \omega\big) = \psi_\theta(N \mid \mathbf{0}, \omega)/\psi_\theta(N \mid \mathbf{0}, \omega^*) = \alpha_\theta(\mathbf{0}, \omega)/\alpha_\theta(\mathbf{0}, \omega^*)$, where $\alpha_\theta(\boldsymbol{\ell}^\theta, \omega)$ is the probability a random player chooses A at beliefs $\boldsymbol{\ell}^\theta$ according to a θ-type. (At $\boldsymbol{\ell}^l = \mathbf{0}$, $\boldsymbol{\ell}^r = \mathbf{0}$, left types think all left types choose A, and right types think all right types choose A.) First consider $\theta = l$, so $\omega^* = \underline{\omega} = (L, \underline{\lambda})$, and $\alpha_l(\mathbf{0}, \omega^*) = 1 - \underline{\lambda}$. If $\omega = (\zeta, \overline{\lambda})$ for either $\zeta \in \{L, R\}$, then $\alpha_l(\mathbf{0}, \omega)/\alpha_l(\mathbf{0}, \omega^*) = (1 - \overline{\lambda})/(1 - \underline{\lambda}) < 1$ since $\underline{\lambda} < \overline{\lambda}$, so $\chi_l(\mathbf{0}, \mathbf{0}, \omega) < 1$. For $\omega = (R, \underline{\lambda})$, $\alpha_l(\mathbf{0}, \omega)/\alpha_l(\mathbf{0}, \omega^*) = (1 - \underline{\lambda})/(1 - \underline{\lambda}) = 1$, and the stability test is inconclusive. Before turning to the inconclusive case, consider $\theta = r$: $\omega^* = \overline{\omega} = (R, \overline{\lambda})$, and $\alpha_r(\mathbf{0}, \omega^*) = \overline{\lambda}$. If $\omega = (\zeta, \underline{\lambda})$ for either $\zeta \in \{L, R\}$, then $\alpha_r(\mathbf{0}, \omega)/\alpha_r(\mathbf{0}, \omega^*) = \underline{\lambda}/\overline{\lambda} < 1$, so $\chi_r(\mathbf{0}, \mathbf{0}, \omega) < 1$. For $\omega = (L, \underline{\lambda})$, $\alpha_r(\mathbf{0}, \omega)/\alpha_r(\mathbf{0}, \omega^*) = (1 - \underline{\lambda})/(1 - \underline{\lambda}) = 1$. So, for each type, we've established stability along each dimension except for one.

To deal with the "inconclusive" cases where $\chi_\theta(\mathbf{0}, \mathbf{0}, \omega) = 1$, I follow Proposition 4, and show that $\frac{\partial}{\partial \ell^\theta(\omega)} \chi_\theta(\mathbf{0}, \mathbf{0}, \omega) < 0$—the stability coefficient is less than one at all points in the neighborhood of the fixed-point (excluding the fixed point itself). Analogous to Equation A.19,

$$\frac{\partial}{\partial \ell^\theta(\omega)} \log \chi_\theta(\hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r, \omega)\bigg|_{(\hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r) = (\mathbf{0}, \mathbf{0})} = 2 \sum_{z=0}^{N} \psi(a, \mathbf{0}, \mathbf{0}) \frac{\partial}{\partial \ell^\theta(\omega)} \left(\frac{\psi_\theta(a \mid \mathbf{0}, \omega)}{\psi_\theta(a \mid \mathbf{0}, \omega^*)} \right). \tag{A.47}$$

For $\omega = (\zeta, \lambda)$ and $\omega^* = (\zeta^*, \lambda^*)$, analogous to Equation A.13

$$\frac{\partial}{\partial \ell^\theta(\omega)} \left(\frac{\psi_\theta(a \mid \boldsymbol{\ell}^\theta, \omega)}{\psi_\theta(a \mid \boldsymbol{\ell}^\theta, \omega^*)} \right) =$$

$$\left(\frac{\psi_\theta(a \mid \boldsymbol{\ell}^\theta, \omega)}{\psi_\theta(a \mid \boldsymbol{\ell}^\theta, \omega^*)} \right) \left\{ \frac{\partial p^\theta(\boldsymbol{\ell}^\theta)}{\partial \ell^\theta(\omega)} \Bigg[[1 - 2\lambda] f_\zeta\big(p^\theta(\boldsymbol{\ell}^\theta)\big) \left(\frac{a - N\alpha_\theta(\boldsymbol{\ell}^\theta, \omega)}{\alpha_\theta(\boldsymbol{\ell}^\theta, \omega)\big[1 - \alpha_\theta(\boldsymbol{\ell}^\theta, \omega)\big]} \right) \right.$$

$$- [1-2\lambda^*] f_{\zeta^*}\big(p^\theta(\boldsymbol{\ell}^\theta)\big) \left(\frac{a - N\alpha_\theta(\boldsymbol{\ell}^\theta,\omega^*)}{\alpha_\theta(\boldsymbol{\ell}^\theta,\omega^*)\big[1-\alpha_\theta(\boldsymbol{\ell}^\theta,\omega^*)\big]} \right) \Bigg] \Bigg\}, \qquad \text{(A.48)}$$

where $p^\theta(\boldsymbol{\ell}^\theta)$ is the probability of location state L according to a θ-type. For each θ, let Σ^θ be the sum of the components of $\boldsymbol{\ell}^\theta$; for $\theta = l$, $p^l(\boldsymbol{\ell}^l) = (1+\ell^l(L,\overline{\lambda}))/(1+\Sigma^l)$, and for $\theta = r$, $p^l(\boldsymbol{\ell}^l) = (\ell^r(L,\underline{\lambda}) + \ell^r(L,\overline{\lambda}))/(1+\Sigma^r)$. Note that $p^l(\mathbf{0}) = 1$ and $p^r(\mathbf{0}) = 0$. Note A.47 is less than 0 so long as A.48 is less than 0 when evaluated at $\boldsymbol{\ell}^l = \mathbf{0}$, $\boldsymbol{\ell}^r = \mathbf{0}$, and $a = N$. Assuming $\lambda = \lambda^*$ (which is always so in any “inconclusive case”), this holds if and only if

$$C^\theta(\omega) := \frac{\partial p^\theta(\mathbf{0})}{\partial \ell^\theta(\omega)}[1-2\lambda^*]\big[1-\alpha_\theta(\mathbf{0},\omega^*)\big]\Big[f_\zeta\big(p^\theta(\mathbf{0})\big) - f_{\zeta^*}\big(p^\theta(\mathbf{0})\big)\Big] < 0. \qquad \text{(A.49)}$$

Hence I need only show show $C^l(R,\underline{\lambda}) < 0$ and $C^r(L,\overline{\lambda}) < 0$. From the definition of p^θ above, $\partial p^l(\mathbf{0})/\partial \ell^l(R,\underline{\lambda}) < 0$, and $\partial p^r(\mathbf{0})/\partial \ell^r(L,\overline{\lambda}) > 0$. So, $\theta = l \Rightarrow \omega^* = (L,\underline{\lambda}) \Rightarrow C^l(R,\underline{\lambda}) < 0 \Leftrightarrow \underline{\lambda}[1-2\underline{\lambda}]\big[f_R(1) - f_L(1)\big] > 0$, which holds since $f_R(1) > f_L(1)$ and $\underline{\lambda} < \frac{1}{2}$. And, $\theta = r \Rightarrow \omega^* = (R,\overline{\lambda}) \Rightarrow C^r(L,\overline{\lambda}) < 0 \Leftrightarrow (1-\overline{\lambda})[1-2\overline{\lambda}]\big[f_L(0) - f_R(0)\big] < 0$, which holds since $f_L(0) > f_R(0)$ and $\overline{\lambda} > \frac{1}{2}$.

□

Proof of Lemma A.1.

Proof. From Equation 1.9,

$$\alpha_\theta(\ell,\omega) = \sum_{\theta'\in\Theta^p} \hat{g}(\theta' \mid \theta) + \sum_{\theta'\in\Theta^l} \hat{g}(\theta' \mid \theta) F_\omega(\ell/(v(\theta')+\ell)) + \sum_{\theta'\in\Theta^r} \hat{g}(\theta' \mid \theta)[1 - F_\omega(\ell/(v(\theta')+\ell))], \qquad \text{(A.50)}$$

so

$$\frac{\partial}{\partial \ell}\alpha_\theta(\ell,\omega) = f_\omega(\ell)\left(\sum_{\theta'\in\Theta^r} \hat{g}(\theta' \mid \theta) v(\theta')/(v(\theta')+\ell)^2 - \sum_{\theta'\in\Theta^l} \hat{g}(\theta' \mid \theta) v(\theta')/(v(\theta')+\ell)^2 \right). \qquad \text{(A.51)}$$

Since $f_L(0) > f_H(0)$, $\frac{\partial}{\partial \ell}\alpha_\theta(0,R) < \frac{\partial}{\partial \ell}\alpha_\theta(0,L)$ when $\sum_{\theta'\in\Theta^r} \hat{g}(\theta' \mid \theta)/v(\theta') > \sum_{\theta'\in\Theta^l} \hat{g}(\theta' \mid \theta)/v(\theta')$. Similarly, since $f_L(1) > f_H(1)$, $\frac{\partial}{\partial \ell}\alpha_\theta(\infty,R) < \frac{\partial}{\partial \ell}\alpha_\theta(\infty,L)$ when $\sum_{\theta'\in\Theta^r} \hat{g}(\theta' \mid \theta)v(\theta') < \sum_{\theta'\in\Theta^l} \hat{g}(\theta' \mid \theta)v(\theta')$. Thus there exists some $\bar{\ell}^\theta$ such that $\alpha_\theta(\bar{\ell}^\theta, L) = \alpha_\theta(\bar{\ell}^\theta, R)$ so long as both of the preceding inequalities hold, or both fail.

Let $\overline{\theta}^l := \max \Theta^l$ and $\underline{\theta}^l := \min \Theta^r$. We want to find conditions on Δ_q such that for any given $\hat{g}(\cdot \mid \theta)$, no such point $\bar{\ell}^\theta$ exists. Hence we need one of the inequalities to hold, and one to fail. Suppose $\sum_{\theta'\in\Theta^r} \hat{g}(\theta' \mid \theta) > \sum_{\theta'\in\Theta^l} \hat{g}(\theta' \mid \theta)$. Since $v(\theta') > v(\theta'')$ for all $\theta' \in \Theta^r$ and $\theta'' \in \Theta^l$, trivially $\sum_{\theta'\in\Theta^r} \hat{g}(\theta' \mid \theta)v(\theta') > \sum_{\theta'\in\Theta^l} \hat{g}(\theta' \mid \theta)v(\theta')$. So no confound exists so long as $\sum_{\theta'\in\Theta^r} \hat{g}(\theta' \mid \theta)/v(\theta') > \sum_{\theta'\in\Theta^l} \hat{g}(\theta' \mid \theta)/v(\theta')$. Since $\underline{\theta}^r = \arg\max_{\Theta^r} v(\theta)$ and $\overline{\theta}^l = \arg\min_{\Theta^l} v(\theta)$,

$$\sum_{\theta'\in\Theta^r} \hat{g}(\theta' \mid \theta)/v(\theta') > (1/v(\underline{\theta}^r)) \sum_{\theta'\in\Theta^r} \hat{g}(\theta' \mid \theta),$$

and

$$(1/v(\overline{\theta}^l)) \sum_{\theta' \in \Theta^l} \hat{g}(\theta' \mid \theta) > \sum_{\theta' \in \Theta^l} \hat{g}(\theta' \mid \theta)/v(\theta').$$

So no confound exists if $v(\overline{\theta}^l)/v(\underline{\theta}^r) > \sum_{\theta' \in \Theta^l} \hat{g}(\theta' \mid \theta)/\sum_{\theta' \in \Theta^r} \hat{g}(\theta' \mid \theta)$. While not necessary, assuming Θ is a grid, $\overline{\theta}^l = -\underline{\theta}^r$, so $v(\overline{\theta}^l) = 1/v(\underline{\theta}^r)$. Hence no confound exists whenever $v(\overline{\theta}^l)^2 > \sum_{\theta' \in \Theta^l} \hat{g}(\theta' \mid \theta)/\sum_{\theta' \in \Theta^r} \hat{g}(\theta' \mid \theta) \Leftrightarrow \Delta_q < k\Delta_d(\overline{\theta}^l)(1-\xi)/(1+\xi)$ where

$$\xi := \sqrt{\frac{\sum_{\theta' \in \Theta^l} \hat{g}(\theta' \mid \theta)}{\sum_{\theta' \in \Theta^r} \hat{g}(\theta' \mid \theta)}} < 1.$$

The case for $\sum_{\theta' \in \Theta^r} \hat{g}(\theta' \mid \theta) < \sum_{\theta' \in \Theta^l} \hat{g}(\theta' \mid \theta)$ follows nearly identically, and we find no confound exists if $\Delta_q < k\Delta_d(\overline{\theta}^l)(1-\xi')/(1+\xi')$ where

$$\xi' := \sqrt{\frac{\sum_{\theta' \in \Theta^r} \hat{g}(\theta' \mid \theta)}{\sum_{\theta' \in \Theta^l} \hat{g}(\theta' \mid \theta)}} < 1.$$

Together, we see that no confound exists for any type's belief process so long as

$$\Delta_q < k\Delta_d(\overline{\theta}^l)(1-\xi^\theta)/(1+\xi^\theta)$$

where

$$\xi^\theta := \min\left\{\sqrt{\frac{\sum_{\theta' \in \Theta^l} \hat{g}(\theta' \mid \theta)}{\sum_{\theta' \in \Theta^r} \hat{g}(\theta' \mid \theta)}}, \sqrt{\frac{\sum_{\theta' \in \Theta^r} \hat{g}(\theta' \mid \theta)}{\sum_{\theta' \in \Theta^l} \hat{g}(\theta' \mid \theta)}}\right\} < 1.$$

□

Proof of Proposition A.1.

Proof. Let $\underline{\lambda}, \overline{\lambda}$ be arbitrary elements of Λ with $\underline{\lambda} < \overline{\lambda}$. I show that there exists a confounding belief that puts positive weight on states $\underline{\omega} := (L, \underline{\lambda})$ and $\overline{\omega} := (R, \overline{\lambda})$, and zero weight on all other states. At this belief, players are nearly certain the state is one of $\overline{\omega}$ or $\underline{\omega}$, but cannot discern which is true. We want to find $\hat{\boldsymbol{\pi}}^l$ and $\hat{\boldsymbol{\pi}}^r$ such that $\Pr(a_t \mid \hat{\pi}^l, \hat{\pi}^r, L, \underline{\lambda}) = \Pr(a_t \mid \hat{\pi}^l, \hat{\pi}^r, R, \overline{\lambda})$, which holds so long as the probability any random player chooses A given these beliefs is equal in each state of the world. Denote this probability $\alpha(\hat{\pi}^l,\hat{\pi}^r,\omega)$. When $\omega = (\zeta, \lambda)$ for $\zeta \in \{L, R\}$, then $\alpha(\hat{\pi}^l,\hat{\pi}^r,\omega) = \lambda\big[1 - F_\zeta(1-\hat{\pi}^r)\big] + (1-\lambda)F_\zeta(1-\hat{\pi}^l)$. I now construct $\hat{\boldsymbol{\pi}}^l$ and $\hat{\boldsymbol{\pi}}^r$ that meet the condition for "confounding" beliefs, above. For each θ, parameterize beliefs by some $p^\theta \in (0,1)$: let $\hat{\pi}^\theta(\overline{\omega}) = p^\theta$, $\hat{\pi}^\theta(\underline{\omega}) = 1 - p^\theta$, and $\hat{\pi}^\theta(\omega) = 0$ for all $\omega \neq \underline{\omega}, \overline{\omega}$. Importantly, we can write both p^l and p^l as a function of some neutral belief p. Note that p^θ is the belief that $\omega = \underline{\omega}$ held by an agent with taste θ after history h. Consider a neutral observer who observes history h, but does not yet know her taste—say her belief that $\omega = \underline{\omega}$ is p. If she then learns her taste is θ, then p^θ must follow from Bayes' rule as a function of p: $p^l(p) = \Pr(\underline{\omega} \mid h, \theta = l) = (1-\underline{\lambda})p/((1-\underline{\lambda})p + (1-\overline{\lambda})(1-p))$ and

$p^r(p) = \Pr(\underline{\omega} \mid h, \theta = r) = \underline{\lambda} p/(\underline{\lambda} p + \overline{\lambda}(1-p))$. Clearly, for each θ, $\lim_{p\to 0} p^\theta(p) = 0$ and $\lim_{p\to 1} p^\theta(p) = 1$. Now consider the condition for confounding beliefs: $\Pr(a_t \mid \hat{\pi}^l, \hat{\pi}^r, L, \underline{\lambda}) = \Pr(a_t \mid \hat{\pi}^l, \hat{\pi}^r, R, \overline{\lambda}) \Leftrightarrow \alpha(\hat{\pi}^l, \hat{\pi}^r, \underline{\omega}) = \alpha(\hat{\pi}^l, \hat{\pi}^r, \overline{\omega}) \Leftrightarrow$

$$\underline{\lambda}\big[1 - F_L(p^r(p))\big] + (1-\underline{\lambda})F_L(p^l(p)) = \overline{\lambda}\big[1 - F_R(p^r(p))\big] + (1-\overline{\lambda})F_R(p^l(p)). \tag{A.52}$$

I now argue that there must exist $p \in (0,1)$ such that Equation A.52 holds. At $p = 0$, the left-hand side is $\underline{\lambda}$, and the is $\overline{\lambda}$. At $p = 1$, the left is $1 - \underline{\lambda}$, and the right is $1 - \overline{\lambda}$. Since $\underline{\lambda} < \overline{\lambda}$, the left-hand side is less than the right at $p = 0$, but greater than than the right at $p = 1$. By continuity, there exists $p \in (0,1)$ so that Equation A.52 holds. Hence, I've constructed a pair of confounding beliefs, $\hat{\boldsymbol{\pi}}^l$ and $\hat{\boldsymbol{\pi}}^r$.

□

Proof of Proposition A.2.

Proof. Let $\underline{\lambda}, \overline{\lambda}$ be arbitrary elements of Λ with $\underline{\lambda} < \overline{\lambda}$, and let $\underline{\omega} := (L, \underline{\lambda})$ and $\overline{\omega} := (R, \overline{\lambda})$. Consider the confounding belief constructed in the proof of Proposition A.1, above. That is, $\hat{\boldsymbol{\pi}}^l$ and $\hat{\boldsymbol{\pi}}^r$ such that, for each θ, $\hat{\pi}^\theta(\overline{\omega}) = p^\theta$, $\hat{\pi}^\theta(\underline{\omega}) = 1 - p^\theta$, and $\hat{pi}^\theta(\omega) = 0$, where $p^l(p) = \Pr(\underline{\omega} \mid h, \theta = l) = (1-\underline{\lambda})p/((1-\underline{\lambda})p + (1-\overline{\lambda})(1-p))$ and $p^r(p) = \Pr(\underline{\omega} \mid h, \theta = r) = \underline{\lambda} p/(\underline{\lambda} p + \overline{\lambda}(1-p))$, and p is the value that solves Equation A.52. I show that the neutral belief process—the belief of a player who does not know her taste—is stochastically stable in the neighborhood of p. If this is so, then taste dependent beliefs converge with positive probability to the confounding belief identified above. Let the neutral likelihood ratio of state $\underline{\omega}$ relative to $\overline{\omega}$ after history h_t be denoted by ℓ^n_t. Let $\psi(a \mid \ell^n_t, \omega)$ be the probability of observation $a \in \{0, 1, ..., N\}$ in state ω given neutral belief ℓ^n_t. Fix $\omega = \overline{\omega}$. Then process $\langle \ell^n_t \rangle$ evolves according to $\ell^n_{t+1} = \ell^n_t \psi(a \mid \ell^n_t, \underline{\omega})/\psi(a \mid \ell^n_t, \overline{\omega}) := \varphi(a, \ell^n_t)$ with transition probability $\psi(a \mid \ell^n_t, \overline{\omega})$. We want to show this process is stable in the neighborhood of $\hat{\ell}^n := p/(1-p)$, where p generates the confounding belief, given above. By definition of the confounding belief, $\hat{\ell}^n$ is a fixed point of the neutral-belief Markov process: $\hat{\ell}^n = \varphi(a, \hat{\ell}^n)$ for any a. We can use Lemma 8 to assess whether $\hat{\ell}^n$ is stable. That is, it must be that $\chi(\hat{\ell}^n) < 1$, where $\chi(\hat{\ell}^n) = \prod_{a=0}^N \left(\frac{\partial}{\partial \ell}\varphi(a, \hat{\ell}^n)\right)^{\psi(a|\hat{\ell}^n, \overline{\omega})}$. If this Markov process is also a martingale, then $\chi(\hat{\ell}^n) < 1$ (see Smith and Sørensen (2000), Theorem 4). Clearly, $\langle \ell^n_t \rangle$ forms a martingale conditional on $\omega = \overline{\omega}$: $\mathbb{E}[\ell^n_{t+1} \mid \ell^n_t] = \sum_{a=0}^N \psi(a \mid \ell^n_t, \overline{\omega})\varphi(a, \ell^t_n) = \ell^t_n \sum_{a=0}^N \psi(a \mid \ell^n_t \underline{\omega}) = \ell^n_t$. □

Chapter 2

Naive Social Learning, Mislearning, and Unlearning

2.1 Introduction

People make inferences from the behavior of others when making economic and social decisions. In the canonical example, diners might use the crowd gathered at a restaurant to infer its quality. Or, investors may use others' portfolio choices to glean information about an asset's expected payoff. A large literature, building from Banerjee (1992), Bikchandani, Hirshleifer, and Welch (1992), and Smith and Sørensen (2000), has fleshed out how such inference by fully rational people might lead them to imitate others rather than follow their private information, and how in some particular settings such imitation can lead society to herd with high probability on actions that people know with high probability might be wrong. In such settings, several recent papers (DeMarzo, Vayanos, and Zwiebel, 2003; Eyster and Rabin, 2010, 2013; Bohren, 2013) draw out implications of naive agents who fail to appreciate the redundancy in information when learning from one another. Eyster and Rabin (2010) (henceforth, "ER") propose a simple model of inferential naivete where each agent neglects that her predecessors are themselves learning from those before them: she assumes each observed action reflects solely that predecessor's private signal. For instance, a naive diner thinks each person crowding at a restaurant is there on the basis of independent private information. She fails to realize that the last in line likely joined because of his inference from those already in the queue. Since naive observers wrongly think that each member of the crowd has positive private information about the restaurant, more naive agents line up at the restaurant, which in turn causes later observers to grow even more confident that it has high quality. ER shows that naive agents inevitably grow unwarrentedly confident in some state of the world. And in information-rich environments where rational observers surely learn the truth, with positive probability naive observers grow confident in some wrong hypothesis.

Moving beyond this false confidence result, this paper explores new implications of ER's model of naive inference that emerge in an array of environments richer than those previously studied. Our central insight is that naivete sharply defines the set of hypotheses society may come to believe in. In many settings, there exist states of the world that people always come to disbelieve, even if true. While ER shows that in very simple environments society mislearns with positive probability, we show that in more general settings, there exist states where people mislearn with probability 1. That is, people *necessarily* grow confident in some false hypothesis. As a corollary, we show that society can "unlearn": in some states of the world, even though early generations perfectly learn the truth, society inevitably assign arbitrarily high weight to some false state. Additionally, in many settings naivete predicts the nature of those hypotheses agents come to believe. For instance, when agents care not only about ranking payoffs of actions but wish to learn the size of payoff differences, naivete predicts polarization in perceived payoffs: people think the payoff difference between the best option and all alternatives is as large as possible. And in settings with uncertainty about the distribution of private information, it predicts that people overestimate the precision of private signals.

The environment studied ER precludes these results. They assume a canonical social-learning environment with perfect negative correlation in payoffs: if one restaurant is good, the other must be bad. Classical rational-choice social-learning models rule out the possibility that actions are equally good (or bad) solely for analytical ease, and it's still unknown the extent to which these simplified settings create a loss of generality. Leveraging the tractability of the naive model, we consider more general—and more natural—settings where agents learn the size of payoff differences across actions, which can range from large down to zero.[1] In doing so, we highlight additional implications of naive learning otherwise hidden by the simple environments previously considered. For instance, when agents have common preferences, so long as payoffs are not perfectly correlated, then there are necessarily some payoff states that naive agents always come to disbelieve, even if true. While ER illustrate that particular calibrations of their model lead naive agents to incorrectly learn the state 11% of the time, in some states within richer models naive agents mislearn *100%* of the time.

Section 2 presents our basic model. Every period, a new generation of players take actions. Payoffs are initially unknown, but each player learns from conditionally independent private signals and predecessors' behavior. Following ER, we assume players naively infer from past behavior: each wrongly infers *as if* others learn nothing from predecessors; she thinks their actions reflect solely their private information and the common prior. While we discuss how our results extend in an array of environments, we focus on a setting with two key features: (1) each generation observes only the behavior of the preceding generation, and (2) the number of players in each generation is arbitrarily large. The first assumption simplifies exposition. The second stacks the deck against long-run mislearning. Since an infinite population of privately informed agents act independently in the first period—so naivety has no implications—the law of large numbers implies the second generation of either rational or naive agents immediately learns the state.

[1] The environments we consider are not developed to generate novel or pathological results, but are natural generalizations of those previously considered.

While Generation 2 correctly infers the state, naive inference can lead Generation 3 astray. They wrongly assume Generation 2 acts "autarkically"—relying only on noisy private signals—neglecting that Generation 2 has perfect information. How do they interpret Generation 2's confident behavior through this "autarkic" lens? Since confident behavior needn't match the action distribution predicted by autarkic play in *any* state, naive Generation 3 puts probability 1 on the state that *best* explains this behavior.[2] If the true state does so, then Generation 3 learns correctly, as do all remaining generations. But the "best explanation" need not be the correct one. When the autarkic interpretation of optimal confident behavior suggests some state other than the truth, Generation 3 unlearns. Indeed, if learning doesn't settle down by Generation 3, then beliefs can never settle down on the truth. If social beliefs converge on a state, it must be a "fixed point" of this process: when interpreted through the autarkic lens, the behavior of those confident in the state is best explained by that same state.

For clarity, consider the following example. Consumers wish to learn the quality of a durable good offered by Firms 1 and 2. Firm m's quality q_m reflects, say, the probability that its good remains operational at least a year after purchase. Suppose $q_1 \in \{.4, .8\}$, $q_2 \in \{.1, .6\}$, and qualities are independent. Signal distributions generate the following map from the state (q_1, q_2) to the percent of first-generation consumers who buy 1: $(.8, .1) \rightarrow 90\%$, $(.4, .1) \rightarrow 70\%$, $(.8, .6) \rightarrow 60\%$, and $(.4, .6) \rightarrow 40\%$. Suppose in truth $(q_1, q_2) = (.8, .6)$, so 60% select 1 in $t = 1$. Generation 2 correctly infers that 1 has a quality advantage, so *all* buy 1. Generation 3 is puzzled—no state predicts uniform behavior in autarky. Generation 3's best explanation is the state *most likely* to produce this observation.[3] Out of all states, $(.8, .1)$ maximizes the likelihood that a consumer buys 1 based on private signals alone; Generation 3 puts probability 1 on this state. By this logic, Generation 3 puts probability 1 on $(.8, .1)$ whenever it observes a herd on 1—whenever $(q_1, q_2) \in \{(.8, .1), (.4, .1), (.8, .6)\}$. Consumers inevitably believe in one of two states, $(.8, .1)$ or $(.4, .6)$, no matter what is true.

Section 2.3 discusses general implications of naive inference on long-run beliefs. First, we provide a key lemma characterizing the set of states on which naive agents can settle, Ω^*. Ω^* consists of those hypotheses such that the distribution behavior observed when people are fully confident in the hypothesis most closely resembles the behavior we'd see by privately informed agents if that hypothesis were true. Formally, we show that "closeness" is naturally defined by the cross-entropy distance between the realized action distribution and that predicted by autarkic play. Sections 2.3.2 and 2.3.3 go on to show several applications of this characterization, which reveal the extent to which naive inference limits the conclusions society is able to draw. For instance, we show that Ω^* may be a singleton—society draws a unique conclusion no matter the true state—or it may be empty—beliefs forever cycle, and society never settles on a singular conclusion.

[2] Formally, Generation 3 settles on the state ω that minimizes the cross-entropy distance between the observed action distribution and the action distribution predicted by autarkic play in ω. Details are provided in Section 3.

[3] Naive agents attribute any discrepancy between observed and predicted action distributions to sampling variation.

Section 2.3.2 presents a simple implication of naive inference in settings where players with common preferences choose among options with payoffs independent of one another. Consider, for instance, investors learning the payoffs to independent stocks, or diners assessing the quality of various restaurants. Naive perceptions of payoffs inevitably grow "polarized": people think the best option is as good as it can be, while all lesser options are the worst they could be. Once a herd starts, people think they observe (infinitely) many independent signals indicating that the herd action is better than all alternatives. Under natural assumptions on the signal structure, this observation suggests a state with polarized payoffs.[4] Intuitively, the essence of herding is that people converge on all taking an action based on shared information that would be less universally chosen if people were acting solely on their private information. Naivety leads people to rationalize such uniform behavior by polarized beliefs that the chosen action is as superior as possible to those unchosen. Although ER's setting precludes this result by assuming only 2 states, naivete sharply restricts the constellation of payoffs society may deem true.

Section 2.4 demonstrates how naivete harms welfare by exploring this notion of polarization within an allocation problem where investors must choose how to split wealth between a risky and safe asset. Investors observe private signals about the risky asset's average return, μ, and draw inference about μ from predecessors' allocations. Although, in truth, the first-period allocation resolves uncertainty, naive traders continually draw inference from allocations as if they reveal new information. When μ beats expectations, neglecting redundancy causes perceptions of μ to increase over time. Loosely, investors infer from the first-period split that μ beats expectations; the amount allocated to the risky asset in the second period revises upward in response to this news. But since later investors neglect that predecessors learn from past investments, they wrongly attribute this revised allocation to new positive information. Investment in the risky asset increase yet again. Eventually, investors allocate all wealth to the risky asset. The same logic holds when when μ falls sufficiently short of expectations: allocation to the risky asset decreases over time. Investors inevitably invest all wealth in a single asset, implying inefficient under-diversification. Furthermore, it's not necessarily so that investors allocate all wealth to the asset with the higher payoff: for some realizations of the risky asset's payoff, investors eventually perceive the dominated asset as superior.

Section 2.5 explores naive inference about the distribution of information in the economy. Preceding sections follow the standard social-learning literature in assuming players know the distribution of private signals conditional on the payoff-relevant state. We relax this assumption in two ways. In Section 2.5.1, the precision of private information is uncertain. Since naive observers expect variation in actions proportional to the variation in private information, a herd indicates that signals have the highest possible precision. Naive observers become convinced that others have perfect private information about the payoff state.

Section 2.5.2 characterizes learning in an environment with *aggregate* uncertainty—after

[4] We assume signals indicating high quality are increasingly likely as quality increases. Formally, for all qualities $q > q'$, signal densities $f(s \mid q)$ and $f(s \mid q')$ satisfy the monotone likelihood ratio property.

combining all information in the economy, a rational agent is still uncertain about payoffs. In this setting, a naive player rightfully anticipates that she'll remain uncertain in the long run. Much to her surprise, however, she inevitably grows confident in some (perhaps false) payoff state. For example, investors learn about the probability that an asset will yield positive return. Naive individuals conclude that the most popular asset will payoff for sure and that the remaining assets never will. In contrast, a rational observer correctly learns the probability that each asset pays off, but remains uncertain about the payoff *realization*. Relative to rational inference, naive inference implies overconfidence about the payoff state.

We conclude in Section 3.5 by putting these principles of naive learning in broader context. First, we argue that our results hold when agents can observe the complete history of play or when only a finite number act each period. Second, we discuss how naive inference generates mislearning in more general social-learning environments where agents can learn from their own past experiences, or where some observe different predecessors than others. Finally, we attempt to make sense of the fact that agents in our model may continually observe events to which they assign zero probability ex-ante. Often there are natural ways to extend the state space so that agents can explain the long-run distribution of actions.

2.2 Model

This section provides our general model. Section 2.2.1 describes our social-learning environment. Within this environment, Section 2.2.2 formally defines naive inference. Section 2.2.3 discusses the environment and results of Eyster and Rabin (2010) to make clear the differences between our setting and theirs.

2.2.1 Social-learning Environment

In every period $t = 1, 2, 3, \ldots$, a new set of N players enters, and each simultaneously takes an action $X_{nt} \in \mathcal{A}$.[5] Each player is labeled nt, where t is the period in which she acts and n is an index ranging from 1 to N. To ease exposition, we assume the action space is discrete; $\mathcal{A} \equiv \{A_1, \ldots, A_M\}$ for some finite $M \geq 2$.[6] For each $m = 1, \ldots, M$, let $a_t(m)$ denote the fraction of players in t who take action A_m:

$$a_t(m) = \frac{1}{N} \sum_{n=1}^{N} \nVdash \{X_{nt} = A_m\}.$$

Vector $\mathbf{a}_t = (a_t(1), \ldots, a_t(M))$ is the action distribution played in period t. Let $h_t = (\mathbf{a}_1, \ldots, \mathbf{a}_{t-1})$ denote the history of the game up to time t, where $h_1 = \emptyset$.

[5] This environment closely resembles Jackson and Kalai's (1997) notion of a "recurring game": each period, a new set of players—randomly drawn according to a time-invariant type distribution—play an identical stage game.

[6] Our model naturally extends to continuous settings. Some of the applications in later sections assume a continuous action space or continuous state space.

Players wish to learn an unknown payoff-relevant state of the world $\omega \in \{\omega_1, \omega_2, ..., \omega_K\} \equiv \Omega$, $K < \infty$. Players share a common prior $\boldsymbol{\pi}_1 \in \Delta(\Omega)$, where $\pi_1(k) > 0$ denotes the probability of state ω_k. We also permit heterogeneous preferences: conditional on ω, the payoff from A_m depends on a player's preference type $\theta \in \{\theta_1, \theta_2, ...\theta_J\} \equiv \Theta$; payoffs are denoted by $u(A_m \mid \omega, \theta)$. Preference types are private information and i.i.d. across players according to a commonly-known probability measure $g \in \Delta(\Theta)$.[7]

Players learn about the state from two information channels: private signals and social observation. Each Player nt is endowed with a private signal $\mathbf{s}_{nt}$ about the state of the world. From this signal and prior $\boldsymbol{\pi}_1$, Player nt computes via Bayes' rule the probability of each state $\omega_k \in \Omega$, denoted $p_{nt}(k)$. We call the belief vector $\boldsymbol{\pi}_{nt} \in \Delta(\Omega)$ derived solely from private information Player nt's *private belief.* Following Smith and Sørensen (2000), we work directly with the distribution of private beliefs, rather than signals.[8] It is common knowledge that in state ω, $\boldsymbol{\pi}_{nt}$ are conditionally independent and identically distributed across players with c.d.f. F_ω. Our only assumption on distributions $\{F_\omega\}$ is an identification assumption, which we make after discussing public information.

Our framework is quite general regarding the extent to which the current generation observes behavior of previous generations. We model observability by specifying for each Player nt an *observation set* O_{nt}, which consists of all predecessors whose actions are observed by Player nt:

Definition 1. *Player nt's observation set O_{nt} is the set of all players observed by nt.*

Definition 2. *The observation structure $\mathcal{O}$ is the collection of all players' observation sets: $\mathcal{O} \equiv \{\{O_{nt}\}_{n=1}^{N}\}_{t=1}^{\infty}$.*

While this framework provides a language to argue that our results are robust to various observation structures, we primarily focus on a particular simple structure. Unless noted otherwise, we impose the following two assumptions. First, we assume players observe only the behavior of the previous period:

Assumption 8. (Recent Observations.) *Players observe only the actions of those acting in the previous generation: for each $t \geq 2$ and all $n = 1, ..., N$, $O_{nt} = \{(n', t-1) \mid n' = 1, ..., N\}$.*

Second, so that observing the most recent generation is sufficient for full learning, we assume each generation is arbitrarily large:

Assumption 9. (Large Populations.) *The set of players acting each period is arbitrarily large: $N \to \infty$.*

[7] As noted in Smith and Sørensen (2000), a model with multiple but observable preference types is informationally equivalent to one with a single preference type.

[8] While we only need to consider private-belief distributions for abstract results, in some of the applications below we explicitly specify the signal distributions that lead to private beliefs.

Assumption 9 simplifies the analysis as it implies deterministic dynamics: by the Law of Large Numbers, action frequencies in each period t are pinned down by the signal distribution and agents' priors entering t.

Together, Assumptions 8 and 9 represent large overlapping generations. Each generation is present for two periods: in the first, they observe the actions of the preceding generation, and in the second, they take actions based on the beliefs formed from this observation. We argue in Section 2.6.1 that our results hold when replacing Assumption 8 with other observation structures, including the most common structure in the literature (e.g., Bikchandani, Hirshleifer, and Welch (1992), or Smith and Sørensen (2000)) in which players observe *all* past actions in order. Assumption 9 is also unnecessary for our results, so long as the total number of predecessors later generations observe grows large. We impose these assumptions for tractability and to starkly highlight how dynamics under naive inference differ from rational dynamics.

Following the literature, we call Player nt's belief formed solely from observed actions her *public belief.* Under Assumption 8, generation t observes the full distribution of actions played in $t-1$, $\mathbf{a}_{t-1}$, and forms public belief $\boldsymbol{\pi}_t \in \Delta(\Omega)$ using Bayes' rule:

$$\pi_t(k) = \frac{\Pr(\mathbf{a}_{t-1} \mid \omega_k)\pi_1(k)}{\sum_{\tilde{k}=1}^{K} \Pr(\mathbf{a}_{t-1} \mid \omega_{\tilde{k}})\pi_1(\tilde{k})}. \tag{2.1}$$

We assume a naive player only errs in her calculation of $\Pr(\mathbf{a}_{t-1} \mid \omega_k)$. She forms beliefs via Bayes rule using her misperception of $\Pr(\mathbf{a}_{t-1} \mid \omega_k)$ and maximizes expected utility with respect to these beliefs. For simplicity, we assume that all players are naive.

In each period $t \geq 2$, players combine public and private information and optimize against these beliefs.[9] From $\boldsymbol{\pi}_{nt}$ and $\boldsymbol{\pi}_t$, Player nt arrives via Bayes rule at posterior belief $\boldsymbol{r}_{nt} \in \Delta(\Omega)$:

$$r_{nt}(k) = \frac{p_{nt}(k)[\pi_t(k)/\pi_1(k)]}{\sum_{\tilde{k}=1}^{K} p_{nt}(\tilde{k})[\pi_t(\tilde{k})/\pi_1(\tilde{k})]}.$$

Finally Player nt chooses an action maximizing expected utility given her beliefs:

$$X_{nt} \in \arg\max_{A \in \mathcal{A}} \sum_{k=1}^{K} r_{nt}(k) u(A \mid \omega_k, \theta_{nt}).^{10}$$

The environment described above defines an observational-learning game,

$$\Gamma = \left\langle \Omega, \Theta, \mathcal{A}, u, \mathcal{O}, \{F_\omega\}, g, \boldsymbol{\pi}_1 \right\rangle.^{11}$$

[9] In Period 1, there is no social observation; actions are based solely on private signals.

[10] When indifferent, we assume players uniformly choose some action from the set of optimal actions at random.

[11] The game form Γ generalizes the canonical social-learning models developed by Banerjee (1992), Bikchandani, Hirshelifer, and Welch (1992), and Smith and Sørensen (2000). Each of these models assume full observability ($\mathcal{O}_{nt} = \{(n', k) \mid n' = 1, ..., N \text{ and } k < t\}$ for all n and t) and a single-file sequence

2.2.2 Naive Social Inference

We now introduce formally how naive players learn from predecessors' behavior. Following Eyster and Rabin (2010), a naive individual thinks that any predecessor's action relies solely on that player's private information. This implies that a naive agent draws inference *as if* she assumes all her predecessors ignored the history of play and hence learned nothing from others' actions. That is, she infers as if all her predecessors acted "in autarky".[12] In Equation 2.1 above, a naive player miscalculates $\Pr(\mathbf{a}_{t-1} \mid \omega)$—the likelihood of observing action profile $\mathbf{a}_{t-1}$ conditional on the state. Why? The true probability depends on the beliefs of Generation $t-1$, which in turn depend on what $t-1$ has observed. While a rational agent knows how $\mathbf{a}_{t-1}$ depends on $\mathbf{a}_{t-2}$ and so on, a naive agent infers as if $\mathbf{a}_{t-1}$ is based entirely on the private information of players in $t-1$. As such, Generation t thinks $\mathbf{a}_{t-1}$ reflects solely the distribution of posteriors given some signal distribution F_ω and prior $\boldsymbol{\pi}_1$, and wrongly neglects that $\mathbf{a}_{t-1}$ contains any information gleaned from earlier generations.[13] Aside from this error in how others use information, naive individuals are otherwise fully

of players ($N = 1$). Banerjee (1992) and Bikchandani, Hirshleifer, and Welch (1992) assume discrete signal distributions and common preferences ($|\Theta| = 1$), and Smith and Sørensen (2000) generalize the model by allowing continuous signals and multiple unobserved preference types. Eyster and Rabin (2010) study naive inference in a binary-state observational-learning environment nearly identical to Smith and Sørensen's (2000) aside from restricting attention to homogeneous preferences and, following Lee (1993), allowing for continuous actions.

[12] This is a literal interpretation of the bias in terms of the model at hand. While it gives an intuition as to how to model the error, it does not match the motivating psychology: we believe people neglect that others *use* public information, not that people wrongly believe others don't have *access* to such information. In our social-learning settings, these two assumptions are isomorphic. Formally, one could model naivete a long these lines by assuming agents best respond within a perceived game which exactly matches Γ aside from a misperception of observation structure $\mathcal{O}$. Denote Player nt's misperception of $\mathcal{O}$ by $\widehat{\mathcal{O}}_{nt} = \{\{\widehat{O}_{\tilde{n}\tilde{t}}\}_{\tilde{n}=1}^{N}\}_{\tilde{t}=1}^{\infty}$. Each Player nt thinks that all players aside from herself face an autarkic observation set (i.e., $O = \emptyset$), while she herself faces the true observation set specified by $\mathcal{O}$. Aside from this misconception of the game form, we assume all individuals are perfectly Bayesian within their wrong model of the game $\widehat{\Gamma}_{nt} = \langle \Omega, \Theta, \mathcal{A}, u, \widehat{\mathcal{O}}_{nt}, \{F_\omega\}, g, \boldsymbol{\pi}_1 \rangle$.

Definition 3. *A player is* inferentially naive *if she plays a best response to the game* $\widehat{\Gamma}_{nt} = \langle \Omega, \Theta, \mathcal{A}, u, \widehat{\mathcal{O}}_{nt}, \{F_\omega\}, g, \boldsymbol{\pi}_1 \rangle$ *where* $\widehat{\mathcal{O}}_{nt} = \{\{\widehat{O}_{\tilde{n}\tilde{t}}\}_{\tilde{n}=1}^{N}\}_{\tilde{t}=1}^{\infty}$ *is such that*

1. $\widehat{O}_{nt} = O_{nt}$
2. *If* $\tilde{n} \neq n$ *or* $\tilde{t} \neq t$*, then* $\widehat{O}_{\tilde{n}\tilde{t}} = \emptyset$.

Definition 3 implies a naive player has a correct perception of whom she observes, but thinks all others act solely on private information—she believes all others observe no predecessors. This definition of naive inference is isomorphic to, but distinct from, the one originally proposed by Eyster and Rabin (2008). They assume naive players best respond to the belief that all other players are fully cursed in the sense of Eyster and Rabin's (2005) "Cursed Equilibrium".

[13] In essence, naive players fail to realize that past behavior (in $t > 2$) already incorporates all private information. There is nothing new to learn from $\mathbf{a}_t$, $t > 2$. Yet followers use $\mathbf{a}_t$ as if it reflects new independent information. Eyster and Rabin (2013) refer to this as "redundancy neglect". In simple single-file settings, this directly generates over-counting of early signals.

rational in that they form beliefs using Bayes rule within their incorrect model and maximize expected utility with respect to these beliefs.

The primitives of Γ will determine two objects that are key in determining belief dynamics. For each state ω, the *autarkic action distribution* is the expected distribution of actions that occurs in the first generation in both the naive and rational model. It's the distribution that emerges when players in fact act solely on private signals and priors.

Definition 4. (Autarkic Distribution.) *Conditional on ω, $\mathbb{P}_\omega \in \Delta(\mathcal{A})$ is the distribution of actions generated by autarkic play. That is, $\mathbb{P}_\omega(m)$ is probability that a random player takes A_m based solely on her realized private belief $\boldsymbol{\pi}$.*[14]

The *converged action distribution* is the expected distribution of actions that occurs in Generation t when players in t are certain (rightly or wrongly) that the state is ω.

Definition 5. (Converged Distribution.) $\mathbb{T}_\omega \in (\Delta\mathcal{A})$ *is the distribution of actions generated when all players put probability 1 on ω.*[15]

Finally, we make an identifiability assumption that the autarkic distribution in each state is distinct. Thus, as generations grow arbitrarily large, naive agents anticipate that $\mathbf{a}_{t-1}$ perfectly reveals the state ω.

Assumption 10. (Identifiability.) *The collection of private belief distributions $\{F_\omega\}$ is such that for all ω, $\omega' \in \Omega$, there exists $m \in \{1, ..., M\}$ such that $\mathbb{P}_\omega(m) \neq \mathbb{P}_{\omega'}(m)$ whenever $\omega \neq \omega'$. That is, no two distinct states generate identical autarkic action distributions.*

[14] Formally, to derive $\mathbb{P}_\omega$ from primitives, we first define the probability that type θ takes A_m in ω when relying solely on her private belief:

$$\psi_\omega(m,\theta) \equiv \Pr\left(A_m = \arg\max_{X\in\mathcal{A}} \mathbb{E}_{\tilde{\omega}}\left[u(X \mid \theta, \tilde{\omega}) \mid \boldsymbol{\pi}\right] \;\middle|\; \omega\right). \tag{2.2}$$

Recall that our definition of private belief $\boldsymbol{\pi}$ is the updated belief over states given one's private signal and prior $\boldsymbol{\pi}_1$. Hence, while not explicit in the definition of ψ above, ψ depends heavily on prior $\boldsymbol{\pi}_1$. Aggregating across types gives the autarkic frequency of action m:

$$\mathbb{P}_\omega(m) = \sum_{\theta\in\Theta} g(\theta)\psi_\omega(m,\theta). \tag{2.3}$$

[15] Formally, to derive $\mathbb{T}_\omega$ from primitives, denote the set of types who prefer A_m in ω by

$$\Theta_\omega(m) \equiv \left\{\theta \in \Theta \;\middle|\; A_m = \arg\max_{X\in\mathcal{A}} u\left(X \mid \theta, \omega\right)\right\}. \tag{2.4}$$

Within a generation confident in ω, it is only those $\theta \in \Theta_\omega(m)$ who take A_m; thus,

$$\mathbb{T}_\omega(m) = \sum_{\theta\in\Theta} g(\theta)\nVdash\left\{\theta \in \Theta_\omega(m)\right\}. \tag{2.5}$$

Taken together, Assumptions 9 and 10 imply that each naive Generation t is fully confident in some state upon observing their predecessors' behavior, $\mathbf{a}_{t-1}$:[16]

Observation 1. *Assume Assumptions 8, 9, and 10 hold. For each $\omega \in \Omega$, let $\delta(\omega) \in \Delta(\Omega)$ be the degenerate distribution that places probability 1 on state ω. For each $t \geq 2$, $\boldsymbol{\pi}_t \to \delta(\omega)$ as $N \to \infty$ for some $\omega \in \Omega$.*

Conditional on state ω, naive agents in Generation t expect to observe autarkic distribution $\mathbf{a}_{t-1} = \mathbb{P}_\omega$. Intuitively, since they think predecessors use only private signals, they believe play in each round should match (in the limit as $N \to \infty$) the autarkic distribution. However, by Observation 1, Generation t actually observes the behavior of a Generation $t-1$ who's perfectly confident in some state ω'. That is, in truth, they observe $\mathbf{a}_{t-1} = \mathbb{T}_{\omega'}$. As we show in detail in Section 2.3, the interplay between the collection of autarkic and converged distributions precisely determines what naive agents come to believe.

Before turning to learning dynamics in Section 2.3, it's worth noting a few important implications of our setting. First, since Generation 1 does act solely on private information, $\mathbf{a}_1 = \mathbb{P}_\omega$ in state ω. Second, our identifiability assumption implies that as $N \to \infty$, Generation 2, whether naive or rational, perfectly infers the true state ω from $\mathbf{a}_1$. Hence, action frequencies in $t = 2$ converge to $\mathbb{T}_\omega$. It's straightforward that, if all agents are rational, behavior converges by $t = 2$: all future generations will also play $\mathbb{T}_\omega$. The rational Bayesian Nash equilibrium calls for Generation $t \geq 3$ to simply imitate $t-1$, since $\mathbf{a}_{t-1}$ is optimal given all available information. But naive followers may move away from $\mathbb{T}_\omega$, because their inference assumes $\mathbb{T}_\omega$ is the result of autarkic play—they neglect that there is nothing left to learn from $\mathbb{T}_\omega$. Whether and when naive players will converge to $\mathbb{T}_\omega$, and what they converge to if not, is the basic premise of this paper, and studied generally in the next section.

2.2.3 Related Models

For sake of contrast, we briefly review Eyster and Rabin's (2010) model upon which we build and discuss other related approaches in the literature. We extend ER only by applying their model of naive inference to a broader array of settings. ER explores a simple binary-state model with a continuum of actions, common preferences, and one player acting per round. Specifically, $\Omega = \{0, 1\}$, $\mathcal{A} = [0, 1]$, and $u(X|\omega) = -(X - \omega)^2$. With these preferences, a player optimally chooses $X \in [0, 1]$ equal to her belief that $\omega = 1$; actions perfectly reveal an agent's posterior belief. Their main result is that with positive probability, society grows confident in the wrong state. Essentially, naive players treat announced posteriors as independent signals despite the fact that posteriors incorporate predecessors' signals. As

[16] Naive agents are fully confident only in the limit as $N \to \infty$. For finite large N, we can choose an N large enough to achieve any arbitrary level of confidence.

such, players vastly over-count early signals. If early signals are misleading—which happens with positive probability—then players grow confident in the wrong state.[17]

The logic of how early misleading signals can lead society astray is the crux of how naivete plays out in ER's environment. However, because we assume $N \to \infty$, early signals are never misleading in our setting. This is reflected in the result that $\mathbf{a}_2 = \mathbb{T}_\omega$ for the correct ω in period 2. We emphasize how a naive society can still mislearn despite early generations perfectly inferring the state.

Additionally, ER impose perfect-negative correlation in payoffs across states: whenever $X = 1$ is optimal, payoffs are strictly decreasing in X, and learning the payoff of $X = 1$ pins down the payoff of all other actions. In contrast, we consider a more natural array of settings where payoffs aren't perfectly correlated. Unlike ER's setting, knowing that X is superior to X' doesn't necessarily tell us *by how much* X is preferred to X'. In sections to follow, we demonstrate exactly why this distinction matters. Not only will agents grow confident in false states, but when tasked with learning the *size* of payoff differences, naive agents systematically overestimate them. Naive inference imposes restrictions on the constellation of payoffs agents can come to believe in the long run.

Eyster and Rabin's model of naive inference is related to many alternative approaches which attempt to capture the fact that people neglect redundancies in information. DeMarzo, Vayanos, and Zweibel (2003) propose a model of "Persuasion Bias" in which neighbors in a network communicate posterior beliefs. Building on DeGroot's (1974) model of consensus formation, they assume players form posteriors by taking the average of neighbors' beliefs as if they reflected independent signals with known precision. Importantly, players neglect that stated beliefs already incorporate signals previously shared, which generates over-counting of early signals. Our model is also related to Level-k thinking (e.g., Crawford and Iriberri, 2007) in this particular environment. Specifically, our agents act like Level-2 thinkers: they best respond to the belief that others use only private information (Level-1). Additionally, Bohren (2013) studies a model in which only fraction α of players can observe past actions, and those who do have wrong beliefs about α. Our model of naive inference corresponds to the case where $\alpha = 1$, but all players think $\hat{\alpha} = 0$.

Empirically, lab experiments show direct evidence that people neglect redundancies in information when learning from predecessors' actions. Eyster, Rabin, and Weizsäcker (2013) tell subjects the difference in the number of Heads and Tails from 100 flips of a coin. Moving in sequence, each subject must estimate the total difference in Heads and Tails across all predecessors, including herself, and announce this estimate. A Bayesian Nash equilibrium strategy is to add ones observation to the estimate of the previous subject. However, they find that players tend to sum the announcements of *all* predecessors, suggesting that subjects fail to understand that the most recent predecessor's behavior incorporates the information of earlier predecessors. Enke and Zimmermann (2013) find similar redundancy neglect in a

[17] In contrast with the rational model, wrong herds are likely to occur only when those herding are relatively uncertain of the state. Rational herding models do not provide theories of society thinking it knows things it doesn't.

laboratory asset-trading experiment, but do so in environments similar to DeMarzo, Vayanos, and Zweibel (2003) where it is not redundancy of actions that people are neglecting.[18]

2.3 Naive Long-Run Beliefs

This section presents some general implications of naive inference on long-run beliefs. Our central insight is a simple characterization of the set of hypotheses society may come to believe in. In Section 2.3.1, we show that there can exist states of the world that people always come to disbelieve, even if true. Further, we show that society can *unlearn*: in some states of the world, when public beliefs assign arbitrarily high weight to the truth (as in period 2), they inevitably assign arbitrarily high weight to some false state the following period.

Section 2.3.2 applies our characterization to a natural environment in which players have common preferences and actions have payoffs independent of one another. Naive inference leads to polarized beliefs about quality: people conclude the best option is as good as it can be, while all others are as bad as possible. In Section 2.3.3, we go on to show two additional examples which demonstrate the extent to which naive inference restricts the set of hypotheses society can learn. First, this set may be a singleton: society inevitably comes to the same conclusion no matter what is true. Second, the set may be empty, so beliefs continually cycle: each generation is confident in a hypothesis different than that assumed true by the generation before.

2.3.1 Characterization of Stationary Beliefs

This section characterizes the potential limit beliefs of the naive-learning process. In doing so, we present a key lemma that identifies the set of states naive agents come to believe in, denoted Ω^*. Ω^* consists of those hypotheses such that the distribution of behavior observed when people are fully confident in the hypothesis most closely resembles the behavior we'd see by privately informed agents if that hypothesis were true. Formally, we show that "closeness" is naturally defined by the cross-entropy distance between the realized action distribution and that predicted by autarkic play.

To arrive at general convergence principles, we first walk through the inferential logic of the first 3 periods. Within our environment, only 3 periods are necessary to demonstrate the main intuition of how naivete leads inference astray.

Naive beliefs and behavior in the first two periods match those of rational players. Suppose the true state is ω^*. Since the first generation in fact acts in autarky, $\mathbf{a}_1 = \mathbb{P}_{\omega^*}$. Generation 2 (and only Generation 2) is correct in thinking it's predecessors act solely on private signals; by Assumption 10, Generation 2 perfectly learns $\omega = \omega^*$. Accordingly, their behavior matches the converged action distribution given ω^*: $\mathbf{a}_2 = \mathbb{T}_{\omega^*}$.

[18] Eyster and Rabin (2010) discuss how findings from earlier experimental work, like Kübler and Weizsäcker (2004), suggest that people neglect redundancy.

Naivete potentially has an effect starting in period 3: Generation 3 neglects that 2 learns from 1, thinking players in $t = 2$ act solely on private signals. As such, Generation 3 expects $\mathbf{a}_2$ to reflect some autarkic distribution $\mathbb{P}_\omega$. In general though, Generation 3 is surprised by $\mathbf{a}_2 = \mathbb{T}^*_\omega$: there need not exist ω such that $\mathbb{P}_\omega = \mathbb{T}_{\omega^*}$. Any discrepancy between $\mathbf{a}_2$ and the distributions predicted by autarkic play is attributed to sampling variation. Hence, to explain why $\mathbf{a}_2 \neq \mathbb{P}_\omega$ for any ω, Generation 3 concludes that Generation 2 realized a very unlikely constellation of signals. Inference in $t = 3$ entails deciding which $\omega \in \Omega$ is most likely to generate such a signal distribution.

Generation $t = 3$ comes to believe in state $\hat{\omega}$ that best predicts $\mathbf{a}_2$. Formally, $\hat{\omega}$ is the state whose predicted autarkic distribution $\mathbb{P}_{\hat{\omega}}$ is closest to the observed distribution $\mathbb{T}_{\omega^*}$ in terms of "cross-entropy" distance. To see this, recall from Equation 2.1 that the public belief in $t = 3$ that $\omega = \omega_j$ is

$$\pi_3(j) = \frac{\Pr(\mathbf{a}_2 \mid \omega_j)\pi_1(j)}{\sum_{k=1}^{K} \Pr(\mathbf{a}_2 \mid \omega_k)\pi_1(k)}.$$

Naive agents miscalculate $\Pr(\mathbf{a}_2 \mid \omega)$. Under their model, they think that conditional on ω, $N\mathbf{a}_2 \sim \text{Multinomial}(N, \mathbb{P}_\omega)$. That is, the behavior of each predecessor is an independent realization drawn from autarkic action distribution $\mathbb{P}_\omega$. Thus, they perceive

$$\Pr(\mathbf{a}_2 \mid \omega) = C(N, \mathbf{a}_2) \prod_{m=1}^{M} \mathbb{P}_\omega(m)^{Na_2(m)} = C(N, \mathbf{a}_2) \left(\prod_{m=1}^{M} \mathbb{P}_\omega(m)^{\mathbb{T}_{\omega^*}(m)} \right)^N, \tag{2.6}$$

where $C(N, \mathbf{a}_2)$ is a normalization constant independent of ω.[19] Since it plays a recurring role in our analysis, we define the "likelihood" function $\mathcal{L}(\omega \mid \mathbf{a})$ of observing action distribution $\mathbf{a}$ in ω as

$$\mathcal{L}(\omega \mid \mathbf{a}) \equiv \prod_{m=1}^{M} \mathbb{P}_\omega(m)^{a(m)}. \tag{2.7}$$

So, $\mathbb{P}(\mathbf{a}_2 = \mathbb{T}^*_\omega \mid \omega) = C(N, \mathbf{a}_2)\mathcal{L}(\omega \mid \mathbb{T}^*_\omega)^N$. It follows that the naive likelihood ratio between any two states ω_j and ω_k is

$$\frac{\pi_3(j)}{\pi_3(k)} = \left(\frac{\mathcal{L}(\omega_j \mid \mathbb{T}_{\omega^*})}{\mathcal{L}(\omega_k \mid \mathbb{T}_{\omega^*})} \right)^N.$$

Since $N \to \infty$, $\boldsymbol{\pi}_3$ puts all weight on $\hat{\omega}$ ($\boldsymbol{\pi}_3 \to \delta(\hat{\omega})$) that solves

$$\hat{\omega} = \arg\max_{\omega \in \Omega} \mathcal{L}(\omega \mid \mathbb{T}_{\omega^*}) = \arg\max_{\omega \in \Omega} \prod_{m=1}^{M} \mathbb{P}_\omega(m)^{\mathbb{T}_{\omega^*}(m)} \tag{2.8}$$

Simply put, $\hat{\omega}$ is the state most likely to generate $\mathbf{a}_2$ assuming Generation 2 acts solely on private information

[19] If $N\mathbf{a}_t \sim \text{Multinomial}(N, \mathbb{P}_\omega)$, then $C(N, \mathbf{a}_t) = N!/\prod_{m=1}^{M} Na_t(m)!$.

In slightly different terms, we can model inference as a comparison across autarkic distributions given the observed action distribution: Generation 3 grows confident in the state $\hat{\omega}$ whose autarkic distribution $\mathbb{P}_{\hat{\omega}}$ most closely resembles the converged distribution $\mathbb{T}_{\omega^*}$. Our metric for "closeness" is the cross-entropy distance between $\mathbb{P}_{\hat{\omega}}$ and $\mathbb{T}_{\omega^*}$: for any $\mathbb{P}, \mathbb{T} \in \Delta(\mathcal{A})$, the cross entropy is defined as $H(\mathbb{T}, \mathbb{P}) \equiv -\sum_{m=1}^{M} \mathbb{T}(m) \log \mathbb{P}(m)$.[20] This notion of distance follows naturally from the definition of $\hat{\omega}$ in Equation 2.8. State $\hat{\omega}$ equivalently solves

$$\hat{\omega} = \arg\min_{\omega \in \Omega} - \log \mathcal{L}(\omega \mid \mathbb{T}^*_\omega) = \left(- \sum_{m=1}^{M} \mathbb{T}_{\omega^*}(m) \log \mathbb{P}_\omega(m) \right) \tag{2.9}$$

As such, Generation 3 grows confident in state $\hat{\omega}$ which minimizes entropy between the predicted distribution $\mathbb{P}_{\hat{\omega}}$ and the observed distribution $\mathbb{T}_{\omega^*}$.

Having established how a naive Generation 3 wrongly infers from the confident behavior of its predecessors, we extend this logic to describe long-run dynamics. Since each Generation t grows confident from $\mathbf{a}_{t-1}$, we can deterministically characterize belief and action dynamics. To do so, we introduce belief-transition function $\phi : \Omega \to \Omega$ which maps the belief of Generation t to the belief of Generation $t+1$.[21] Formally, suppose Generation t is certain of $\hat{\omega} \in \Omega$, so $\mathbf{a}_t \to \mathbb{T}_{\hat{\omega}}$. Then define

$$\phi(\hat{\omega}) = \arg\min_{\omega \in \Omega} H(\mathbb{T}_{\hat{\omega}}, \mathbb{P}_\omega). \tag{2.10}$$

To describe the belief process, let $\hat{\omega}_t$ satisfy $\lim_{N\to\infty} \boldsymbol{\pi}_t = \delta(\hat{\omega}_t)$. Beliefs thus evolve according to

$$\hat{\omega}_{t+1} = \phi(\hat{\omega}_t), \tag{2.11}$$

starting from initial condition $\hat{\omega}_2 = \omega^*$—Generation 2 correctly infers the true state ω^* from the truly autarkic play of Generation 1.

Naturally, we are interested the limit of this belief process: what hypothesis, if any, will all generations eventually agree upon? Large generations ($N \to \infty$) ensure that beliefs converge to a point belief *within* each generation. To characterize convergence *across* generations, we make the following definition:

Definition 6. *Public-belief process $\langle \boldsymbol{\pi}_t \rangle$ converges in t if there exists τ and ω such that as $N \to \infty$, $\boldsymbol{\pi}_t \to \delta(\omega)$ in all $t \geq \tau$.*

[20] Our reference to H as a metric is colloquial: because it is not symmetric, it is not a proper metric. This measure of distance is well-known in information theory and the literature on model selection. (See, for example, Burnham and Anderson, 1989.) The measure is also used in recent work within economics on learning with incorrect or uncertain models. Examples include Acemoglu, Chernozhukov, and Yildiz (2009), Schwartzstein (2013), and Esponda and Pouzo (2013). An older literature in statistics on Bayesian learning with misspecified models, starting with Berk (1966), takes a similar approach.

[21] Despite being a function that maps how beliefs over Ω evolve, the domain and range of ϕ is Ω rather than $\Delta(\Omega)$. Defining a map over $\Delta(\Omega)$ is unnecessary given that each Generation t is confident of some $\omega \in \Omega$. Thus ϕ denotes the state in which Generation $t+1$ grows certain as a function of the state assumed by Generation t.

It follows immediately that beliefs converge in t to ω only if ω is a fixed point of map ϕ. Formally:

Lemma 1. *If $\langle \boldsymbol{\pi}_t \rangle$ converges in t to some $\omega \in \Omega$, then $\omega \in \Omega^*$, where Ω^* is the set of fixed points of minimum-entropy map ϕ:*

$$\Omega^* \equiv \{\omega \in \Omega \mid \phi(\omega) = \omega\}. \tag{2.12}$$

Lemma 1 is central to our applications and an important result in it's own right. In some sense, Lemma 1 is trivial after establishing transition function ϕ: of course, a limit point of $\hat{\omega}_t$ must be a fixed point of ϕ. Its usefulness lies in identifying the map ϕ that governs long-run stationary beliefs. In applications, once we specify environment Γ, Lemma 1 predicts exactly what naive agents inevitably believe.

More specifically, the necessary condition in Lemma 1 sharply restricts the set of states that naive agents can learn. Ω^* consists of only those hypotheses such that the group behavior observed when people are fully confident in the hypothesis most closely resembles the behavior we'd see by privately informed agents if that hypothesis were true. That is, those states ω where out of all $\mathbb{P}_{\omega'}$, $\mathbb{P}_\omega$ is closest to $\mathbb{T}_\omega$.

It follows that beliefs never settle down on any $\omega \notin \Omega^*$: there can exist states that people never come to believe even if they are true. While in ER's canonical environment society mislearns with positive probability, Lemma 1 implies that in our setting, there exist states where people mislearn with probability 1—people *necessarily* grow confident in some false hypothesis. And this result is independent of priors: even if priors assign arbitrarily low probability to Ω^*, people inevitably conclude one of these extremely unlikely states has been realized.

Assumptions 8 and 9, which imply deterministic dynamics, make the logic behind Lemma 1 straightforward. But the result is much more general. We argue in 2.6.1 that it holds whenever observation sets grow large in t. In short, if beliefs converge in t to ω, then the long-run distribution of behavior must converge to $\mathbb{T}_\omega$.[22]

It's noteworthy that naive inference can cause *unlearning* across generations. Recall that Generation 2 necessarily grows confident in the true state, ω^*. Generation 3 in turn becomes certain of state $\hat{\omega} = \phi(\omega^*)$. Whenever $\omega^* \notin \Omega^*$ so $\hat{\omega} \neq \omega^*$, then society never converges to truth. Agents "unlearn" ω^* between Generations 2 and 3. Even if social beliefs return to ω^* at some later date, they necessarily move away the following period.

This logic highlights a more general result of this environment: if Generation 3 fails to learn the truth ω^*, then beliefs never converge in t to ω^*. Similarly, if society correctly learns, it does so quickly. If $\omega^* \in \Omega^*$, then Generation 3 learns correctly. And so do all remaining periods: $\mathbf{a}_3 = \mathbf{a}_2 = \mathbb{T}_{\omega^*}$, so $t = 4$ draws exactly the same inference exactly as $t = 3$. It

[22] In order for naive agents to remain confident in ω, it must be that $\omega = \arg\min_{\omega' \in \Omega} H(\mathbb{P}_{\omega'}, \mathbb{T}_\omega)$. Essentially, instead of comparing autarkic distributions with per-period action distributions, agents compare them to the long-run distribution. For convergence, the relationship between the autarkic and long-run converged distribution most satisfy exactly the same condition as the one described here between autarkic and short-run action distributions.

follows that $\mathbf{a}_t = \mathbb{T}_{\omega^*}$ for all $t \geq 2$, and all generations put probability 1 on ω^*. Unlike the mislearning and unlearning characterized by Lemma 1, this result is an artifact of our large generations assumption. When N is small—like when agents act in single file—beliefs need not converge to the truth even when the truth lies in Ω^*. As we argue in Section 2.6.1, for each $\omega \in \Omega^*$, beliefs converge with positive probability to $\delta(\omega)$.

The remainder of the paper studies a variety of applications of Lemma 1. While there is little to say about general properties of Ω^* without specifying autarkic distributions, we present one immediate and fairly general implication here. Namely, if agents have common preferences and there are more states than actions, then there are necessarily some states that society never settles down on: Ω^* is a strict subset of Ω.

Proposition 1. *If $|\Omega| > |\mathcal{A}|$ and for all ω and $\theta \neq \theta'$, $u(\cdot \mid \omega, \theta) = u(\cdot \mid \omega, \theta')$, then $\Omega^* \subset \Omega$.*

To understand Proposition 1, note that from Assumption 10, Generation 2 identifies the optimal action. Since they share common tastes, all players in $t = 2$ choose that action (there is a herd). From Assumption 10 again, there is a unique state that best predicts this herd. Since a herd on any action A_m, $m = 1, ..., M$, indicates at most one state, there are at most M fixed points of ϕ. Thus if $|\Omega| > M$, there exist states that naive agents never assume true after observing a herd. This logic is hidden by ER, who assume $|\Omega| \leq |\mathcal{A}|$. But $|\Omega| > |\mathcal{A}|$ naturally applies to many environments: this holds whenever payoffs are not perfectly correlated across states—the payoff of A_m doesn't pin down the payoff of A_j.

2.3.2 Polarization

We now present a key implication of naive inference in settings where players with common preferences choose among options with independent payoffs—the payoff of each option is independent of the payoff of any other options. Naive perceptions of payoffs inevitably grow "polarized": people think one option is as good as it can be, while all other options are as bad as possible. The intuition is that once a herd starts, people think they observe (infinitely) many independent signals indicating that the herd action is better than all alternatives. Under natural assumptions on the signal structure, this suggests a state with polarized payoffs. ER precludes this polarization result by assuming the payoff difference between options is constant in magnitude across states.

Consider a setting where each option $A_m \in \{A_1, ..., A_M\}$ has unknown quality independent of the others. For instance, diners learn about the quality of various restaurants in town, or investors learn about the returns to unrelated assets. The payoff-relevant state $\omega = (q_1, ..., q_M)$ is a vector specifying the quality of each action. Players have homogeneous and monotonic preferences over quality, $u(A_m \mid \omega) = q_m$; the payoff of action m depends only on the quality of m. For each $m = 1, ..., M$, quality q_m is drawn from a compact set Q_m according to known prior π^m with full support over Q_m, and, for any $j \neq m$, q_j and q_m are independent. Hence, any quality profile is a feasible state of the world: $\Omega = \times_{m=1}^{M} Q_m$.

To make sharp claims, we consider a specific but natural class of signal structures that satisfy the standard monotone likelihood ratio property (MLRP). Each player receives a signal $s^m \in S^m \subseteq \mathbb{R}$ about each q_m, and signals are independent across actions—news about q_m provides no information about q_j, $j \neq m$. For each m, s^m has c.d.f. $F^m(\cdot|q_m)$ and density $f^m(\cdot|q_m)$ with full support over S^m.[23]

Assumption 11. (Monotone Likelihood Ratio Property.) *For each m, $f^m(s|q_m)/f^m(s|q'_m)$ is increasing in s if and only if $q_m > q'_m$.*

The MLRP assumption means higher signals unambiguously indicate higher quality. Finally, to rule out trivialities, we assume that in any state, each action is chosen with positive probability in autarky: for each m, priors are such that for all $\omega \in \Omega$, there exists a subset of signals $S^*(m)$ with positive measure yielding $m = \arg\max_j \mathbb{E}[q_j \mid \mathbf{s}]$ whenever $\mathbf{s} \in S^*(m)$. That is, for all ω, autarkic distribution $\mathbb{P}_\omega$ has full support.

Importantly, MLRP implies $\mathbb{P}_\omega(m)$ is increasing in q_m. Holding all other qualities fixed, increasing q_m increases the share of players who choose A_m in autarky. Naturally, if higher quality generates more positive news, then more people autarkicly choose an action more often when its quality increases. Further, it implies $\arg\max_{\omega\in\Omega} \mathbb{P}_\omega(m)$ is the state in which q_m takes its maximum value, and, for all $j \neq m$, q_j takes its minimum value. We define such a state as a "polar state":

Definition 7. *Define"polar state m", denoted by ω^P_m, as follows:*

1. $q_m = \max Q_m$.

2. $q_j = \min Q_j$ *for all* $j \neq m$.

Our assumptions on the signal structure imply the following lemma.

Lemma 2. *Under the* MLRP *signal structure (Assumption 11), $\mathbb{P}_\omega(m)$ is increasing in q_m and decreasing in q_j for all $j \neq m$. Hence $\mathbb{P}_\omega(m)$ is maximized in state ω^P_m.*

Lemma 2 implies that under naive learning, society inevitably grows confident in a polar state. People conclude that one option is the best it could be, while all others are the worst they could be. To see why, suppose $q_1 = \max(q_1, ..., q_M)$. Then $\mathbf{a}_1$ perfectly reveals to Generation 2 that A_1 is optimal. Hence, in $t = 2$, all choose A_1: $\mathbf{a}_2 \to (1, 0, ..., 0)$. Generation 3 comes to believe in the state most likely to induce such a herd under autarkic play. Intuitively, this state must maximize the chance of good news about A_1, but minimize the chance of good news about any other option. As a consequence of Assumption 11, this happens when $q_1 = \max Q_1$, and $q_m = \min Q_m$ for all $m \geq 2$. This is the intuition behind Lemma 2—ω^P_1 is the state that maximizes the likelihood of observing A_1 in autarky. As such, public beliefs in $t = 3$ converge on ω^P_1. Since Generation 3 correctly believes A_1 is optimal, all choose A_1: $\mathbf{a}_3 = \mathbf{a}_2$. By induction, all Generations $t \geq 3$ observe the same

[23] Signals also satisfy the maintained assumptions made in Section 2.2.1: (i) for each m, s^m_{nt}—Player nt's signal about q_m—is conditionally independent and identically distributed across all players, and (2) in the limit as $N \to \infty$, the autarkic distribution perfectly reveals payoff differences across actions.

behavior, and, hence, all conclude $\omega = \omega_1^P$. Long-run naive learning is summarized in the following proposition.

Proposition 2. *In the independent-qualities environment outlined here (Section 2.3.2)*

1. *If* $q_m = \max(q_1, ..., q_M)$, *then naive public beliefs converge on* ω_m^P: $\boldsymbol{\pi}_t \to \delta(\omega_m^P)$ *as* $N \to \infty$.

2. Ω^* *is the set of all polar states.*

The driving force behind Proposition 2 is that people mistake herds caused by social learning as evidence that the solely chosen option has far superior quality. Rational social learning at its best allows society to determine the optimal action by aggregating individuals' private information. With common preferences, this generates a herd on the best option, no matter how small a quality advantage it has over alternatives.[24] In our setting, aggregation and herding occurs by $t = 2$. But all following naive generations neglect that herding is a result of aggregation; instead, they think the herd reflects the underlying distribution of quality across options. To them, the only way to explain why so many choose a single action based purely on private signals is that the quality difference between it and all alternatives is as large as possible.

Part 2 of Proposition 2 demonstrates how naivete restricts the hypotheses society can come to believe. The restriction is most stark in the case where each Q_m is a closed interval, and hence Ω contains an uncountable infinity of states. Despite this, $|\Omega^*| = M$: society necessarily concludes one of M polar states is true.

Quite clearly, naivete necessarily produces exaggerated perceptions of quality differences. Eyster and Rabin's (2010) canonical environment precludes this result. They assume binary actions and binary quality: $\mathcal{A} = \{A_1, A_2\}$, and for $m = 1, 2$, $Q_m = \{0, 1\}$. Crucially, they focus on two states, $(q_1, q_2) \in \{(1, 0), (0, 1)\}$. Thus q_m are not independent, but instead perfectly negatively correlated. Our result shows that if quality is independent—states $(1, 1)$ and $(0, 0)$ occur with positive probability—then, in the long run, people never come to believe $(1, 1)$ or $(0, 0)$, even when those states are true. Society necessarily concludes one action is better than the other.

Despite exaggerating quality differences, in this simple environment society does learn the optimal choice. While naivete has no welfare consequences here, polarization suggests natural welfare implications in richer environments. For instance, Section 2.4 shows that in an investment setting, naivete leads investors form polarized beliefs about returns. This leads to suboptimal under-diversifcation. Additionally, in settings with queuing costs, polarized

[24] When there is no uniquely best option, beliefs still converge on some polar state. However, which polar state society comes to believe is stochastic. For example, suppose $q_1 = q_2 > q_j$ for all $j = 3, ..., M$. First round behavior makes it clear that options 1 and 2 have similarly superior quality. In $t = 1$, most players choose either action 1 or 2, and the fraction of players who choose 1 is roughly the same as the fraction choosing 2. Generation 2 herds on whichever $m \in \{1, 2\}$ that is chosen most often in $t = 1$. Since these action frequencies are identical in expectation (because $q_1 = q_2$), the herd action depends on the particular realization of signals, making it stochastic.

beliefs imply inefficiently high congestion. Since agents assume a polar state, they are less willing—relative to rational players—to switch to the next-best option when congestion costs are high.

The logic of polarization is quite generally, and applies beyond the particular environment specified here. Specifically, polarization arises in any environment where the conclusion of Lemma 2 holds—that is, whenever the autarkic frequency of action m is increasing in q_m and decreasing in q_j for all $j \neq m$. That said, our result is sensitive to some our assumptions, particularly that signals about q^m are independent of all q^j, $j \neq m$. If instead signals are positively correlated—people are more likely to receive good news about A_1 the better is A_2—then, whenever people observe a herd on A_1, naive players conclude that *both* options have high quality.

2.3.3 Additional Examples

The environments considered in this section reveal the extent to which naive inference limits the conclusions society is able to draw. While in the independent-quality setting analyzed above society can only converge to a polar state, the examples here entail more stark restrictions on Ω^*. The first shows that in a setting with heterogeneous preferences, Ω^* is a singleton: society draws a unique conclusion no matter the true state. The second provides a setting where Ω^* is empty: society never settles on a singular conclusion.

2.3.3.1 State-Independent Learning ($|\Omega^*| = 1$)

This example shows how naivete can cause society to reach the same conclusion no matter what is true. This can occur when agents have heterogeneous tastes: when learning which of two preference types some new technology best suits, so long as signals are sufficiently rare, people inevitably conclude it suits the more common taste.

To demonstrate the logic, consider a scenario where farmers learn whether to adopt a new hybrid seed (A) or stick with a well-known variety (B). The new seed A is sensitive to inputs, such as soil type—it's optimal to use A only if the seed matches well with one's plot.[25] Suppose there are two soil types, high salinity ($\theta = H$) and low salinity ($\theta = L$). And, thus, there are two states: A is compatible with type H ($\omega = H$) or with type L ($\omega = L$). Suppose the status-quo crop B grows equivalently for all types, but yields limited output; a farmer prefers B if and only if A is a poor match (i.e., type θ prefers A iff $\theta = \omega$). Finally, suppose some (but not all) farmers are informed about the optimal match. For instance, an NGO educates some farmers whether to use the new seed. Specifically, fraction $\rho \in (0, 1)$ perfectly knows the state, while $1 - \rho$ receives no information. Assume payoffs and priors are such that farmers only switch from B to A when they learn the truth and the seed is a good match; people choose A only if informed.

[25] Munshi (2003) studies learning among farmers in India who choose between hybrid strains of rice or wheat. The output of rice is very sensitive to soil attributes, but wheat grows similarly irrespective of soil type.

To determine what farmers come to believe, we must compare the autarkic and converged action distributions across states. Let $\lambda \equiv \Pr(\theta = H)$ be the fraction of "high-type" farmers; suppose $\lambda > \frac{1}{2}$. In autarky, Player nt chooses A if both informed and $\theta_{nt} = \omega$. Thus $\mathbb{P}_L(A) = \rho(1-\lambda)$ and $\mathbb{P}_H(A) = \rho\lambda$. If a generation is confident of ω, all those with $\theta = \omega$ choose A: $\mathbb{T}_L(A) = 1-\lambda$ and $\mathbb{T}_H(A) = \lambda$. Comparing $\mathbb{P}_\omega$ with $\mathbb{T}_\omega$, social learning naturally leads more to adopt than in autarky: even those without private information might adopt based on information gleaned from predecessors. Naive individuals misattribute this learning-based increase in adoption to preferences. When a large share adopts, they conclude the new seed must be optimal for the majority type. Essentially, people may mistake converged behavior in $\omega = L$ for autarkic play in $\omega = H$.

Suppose in truth $\omega = L$—only the less-common low types should adopt the new seed. In the first period, behavior converges to the true autarkic distribution, and $\rho(1-\lambda)$ adopt. This perfectly reveals $\omega = L$ to Generation 2. In $t = 2$, all of those with $\theta = L$ choose A, yielding adoption rate $1-\lambda$. But Generation 3 expects to see either rate $\rho\lambda$ or $\rho(1-\lambda)$. Given their naive autarkic interpretation, Generation 3 decides which state is most likely to yield rate $1-\lambda$ via sampling variation. When fraction $1-\lambda$ lies "closer" to $\rho\lambda$ than $\rho(1-\lambda)$, they interpret rate $1-\lambda$ as evidence of $\omega = H$.[26] From Equation 2.7, the likelihoods of each state are

$$\mathcal{L}(L \mid 1-\lambda) = \mathbb{P}_L(A)^{1-\lambda}\mathbb{P}_L(B)^{\lambda} = \big[\rho(1-\lambda)\big]^{1-\lambda}\big[1-\rho(1-\lambda)\big]^{\lambda}, \tag{2.13}$$

$$\mathcal{L}(H \mid 1-\lambda) = \mathbb{P}_H(A)^{1-\lambda}\mathbb{P}_H(B)^{\lambda} = \big[\rho\lambda\big]^{1-\lambda}\big[1-\rho\lambda\big]^{\lambda}. \tag{2.14}$$

One can show that unless ρ is sufficiently large, $\mathcal{L}(H \mid 1-\lambda) > \mathcal{L}(L \mid 1-\lambda)$, so $\omega = H$ seems most likely after $1-\lambda$ choose A. And if Generation 3 believes $\omega = H$, all following Generations also infer $\omega = H$. Why? In $t = 3$, all $\theta = H$— fraction λ—adopt. Since $\lambda > \rho\lambda > \lambda(1-\lambda)$, Generation 4 observes more adopt than predicted in *any* state. Thus, they must come to believe the state that predicts the highest adoption rate, $\omega = H$. It follows that in all Generations $t \geq 3$, people are confident that $\omega = H$ and all high types adopt the new seed.

By the same logic, when in fact $\omega = H$, society always learns correctly. Fraction λ chooses A in Generation 2. While this level of adoption is higher than any predicted by Generation 3, their best explanation is the correct one, $\omega = H$. Simply put, when the new technology is best for the majority preference, society always correctly learns. But when it is best for the minority preference, society may wrongly interpret the high adoption rates, relative to autarky, that result from social learning. They intuitively (but wrongly) conclude that high adoption rates indicate that the new technology is best for the majority type. In summary:

Proposition 3. *In the environment outlined here (Section 2.3.3.1):*

1. *If $\omega = H$, then naive public beliefs converge on $\omega = H$: for all $t \geq 3$, $\pi_t \to \delta(H)$ as $N \to \infty$.*

[26] "Closer" in terms of cross-entropy distance: $H(\mathbb{T}_L, \mathbb{P}_H) < H(\mathbb{T}_L, \mathbb{P}_L)$.

2. *If* $\omega = L$, *there exists a value* $\bar{\rho}(\lambda) \in (0, 1)$ *such that naive public beliefs converge on* $\omega = H$ *if and only if* $\rho < \bar{\rho}(\lambda)$.

Hence if $\rho < \bar{\rho}(\lambda)$, *all Generations* $t \geq 3$ *believe* $\omega = H$ *no matter if* $\omega = H$ *or* $\omega = L$.

To give a concrete example, suppose $\lambda = 0.6$ and $\rho = 0.8$. Naive players expect 32% to choose A in $\omega = L$ and 48% do so in $\omega = H$. If $\omega = H$, then $\lambda = 60\%$ choose A in $t = 2$. This more closely resembles the 48% expected in $\omega = H$ than the 32% expected in $\omega = L$. Hence, Generation 3 (and all following generations) conclude $\omega = H$. But if $\omega = L$, then $1 - \lambda = 40\%$ choose A in $t = 2$. Is this more indicative of H or L? Naive players compute the likelihoods: $\mathcal{L}(L \mid .4) = [0.32]^{.4}[0.86]^{.8} = 0.5030 < \mathcal{L}(H \mid .4) = [0.48]^{.4}[0.52]^{.6} = 0.5036$. Hence, Generation 3 wrongly concludes $\omega = H$ is more likely. When $\lambda = 0.6$, $\bar{\rho}(\lambda) = 0.8036$: society is correct in $\omega = L$ only if $\rho > 0.8036$.

2.3.3.2 Non-Convergence ($\Omega^* = \emptyset$)

This example shows that Ω^* may be empty. Every generation grows confident in some hypothesis distinct from that assumed true by the previous generation. Hence, naive beliefs are potentially unstable—they need not converge to any fixed belief over time.

Recall that Ω^* is empty whenever the belief transition function ϕ has no fixed points. We simply construct an example where this is so. Consider a setting with three actions, $\mathcal{A} = \{A, B, C\}$, and three states, $\Omega = \{A, B, C\}$. Players are risk neutral. The payoffs of action $X \in \mathcal{A}$ as a function of state ω are listed in the table below, where we assume $\epsilon > 0$ is arbitrarily close to 0.

X/ω	A	B	C
A	1	0	$1-\epsilon$
B	$1-\epsilon$	1	0
C	0	$1-\epsilon$	1

With perfect information, it's optimal for players to choose $X = \omega$. But if players are fairly confident—but not certain—that $\omega = X$, it may be optimal to take some action $\neg X$. As such, some signal structures are such that the most commonly chosen action in autarky, Y, does *not* match the state. That is, private signals about ω lead a majority of players to rationally take an action different from the perfect-information optimum. To see this, suppose that in $\omega = A$, most people receive a private signal that generates posterior $\pi(A) = 2/3 - \delta/2$, $\pi(B) = 1/3 - \delta/2$, and $\pi(C) = \delta$. In the limit as $\delta \to 0$ and $\epsilon \to 0$, the expected values of A, B, and C, are respectively 2/3, 1, and 1/3. A player with such a signal rationally chooses B. While A is the action most likely to yield the highest possible payoff, it also has the largest downside—she plays it safe and chooses B. Specifically, suppose there are 3 possible signals, $s \in \{a, b, c\}$. The probabilities of each signal realization conditional on ω are listed in the table below.

s/ω	A	B	C
a	$\frac{2}{3}-\frac{\delta}{2}$	$\frac{1}{3}-\frac{\delta}{2}$	δ
b	δ	$\frac{2}{3}-\frac{\delta}{2}$	$\frac{1}{3}-\frac{\delta}{2}$
c	$\frac{1}{3}-\frac{\delta}{2}$	δ	$\frac{2}{3}-\frac{\delta}{2}$

When agents start with a uniform prior over states, it's straightforward that a player with an a signal takes B, a b signal takes C, and an c signal takes A.

This autarkic decision rule leads to cyclical naive beliefs when δ and ϵ are small. Suppose $\omega = A$. In $t = 1$, approximately share $2/3 - \delta/2$ take action B. Among Generation 2, this is clear evidence that $\omega = A$. Since they act with confidence, all players in $t = 2$ choose A. But in autarky, the state that maximizes the share of players who choose A is state $\omega = C$. As such, Generation 3 is convinced that $\omega = C$, and all choose C. Again, under an autarkic interpretation, a herd on C indicates $\omega = B$: all players in Generation 4 choose B. Iterating forward, this cycling persists over time. Confident beliefs in any state ω leads to herd behavior that, when interpreted as autarkic, indicates some alternative state $\omega' \neq \omega$.

Proposition 4. *In the environment outlined here (Section 2.3.3.2), the set of potential limit points is empty:* $\Omega^* = \emptyset$.

Proposition 4 demonstrates how severely naivete can limit society's ability to reach a consensus over time. While previous results show that naivete reduces the set of hypotheses on which society may settle, Proposition 4 makes clear that there may not exist any such hypothesis. Each Generation t grows confident in some hypothesis distinct from that believed by the preceding generation.

2.4 Portfolio Choice

In this section, we demonstrate how naivete harms welfare within an allocation problem where investors must choose how to split wealth between a risky and safe asset. When naive investors learn about the risky asset's payoff from predecessors' allocations, two forms of inefficiency emerge: (1) when optimal to diversify, they inevitably allocate all wealth to a single asset, and (2) for some payoff realizations, they misperceive the expected payoff rankings of the two assets, whereby allocating all wealth to a dominated asset. While the logic follows along the lines of our polarization results in Section 2.3.2, this application differs in that polarized perceptions have negative welfare consequences.

2.4.1 Setting

Suppose there are two assets that pay off in terms of a consumption good at the end of each period. The "safe" asset has expected payoff equal to one unit of the consumption good. The "risky" asset has expected payoff equal to $1 + \mu$ units, but the realization of random variable μ is unknown ex ante. Investors learn about shock μ from private signals and predecessors' allocations. Additionally, both assets are subject to aggregate uncertainty—payoffs are distorted by i.i.d. mean-zero random shocks about which there is no information in the economy. We include aggregate shocks to model scenarios where rational risk-averse investors diversify even when μ is perfectly known. Summarizing, payoffs of Asset 1 (safe) and Asset 2 (risky) are

$$\begin{aligned} d_1 &= 1 + \eta_1, && (2.15) \\ d_2 &= 1 + \mu + \eta_2, && (2.16) \end{aligned}$$

where we assume μ is normal with mean μ_1 and precision τ_μ (i.e., variance $1/\tau_\mu$), and the random shocks η_m are i.i.d. normal with mean zero and precision τ_η.[27]

Investors have noisy signals about μ, but no information about η_1 or η_2. Each Investor nt receives signal $s_{nt} = \mu + \epsilon_{nt}$ where ϵ_{nt} is i.i.d. normal with mean zero and precision τ_ϵ. Investor nt's information set $I_{nt} \equiv \{s_{nt}, \bar{x}_{t-1}\}$ consists of her private signal and the aggregate share invested in the risky asset in the preceding period, denoted $\bar{x}_{t-1}$.[28]

Each Investor nt has initial wealth W_0 and allocates fraction $x_{nt} \in [0, 1]$ to the risky asset. To abstract from pricing dynamics, we assume the price of each asset is fixed at 1. Final wealth is

$$W_{nt} = W_0\big[(1 - x_{nt})\eta_1 + x_{nt}(\mu + \eta_2)\big]. \tag{2.17}$$

We assume investors have exponential utility, $u(W) = -\exp(-\alpha W)$, where α is an individual's coefficient of absolute risk aversion. With these preferences and normally distributed wealth, Investor nt's demand solves

$$x_{nt} = \arg\max_x \widehat{\mathbb{E}}[W_{nt} \mid I_{nt}] - \frac{1}{2}\alpha\widehat{\mathbb{V}}[W_{nt} \mid I_{nt}]. \tag{2.18}$$

Importantly, the expectation $\widehat{\mathbb{E}}$ and variance $\widehat{\mathbb{V}}$ are those *perceived* by naive Investor nt with information I_{nt} given her incorrect autarkic model. From budget constraint 2.17 and objective 2.18, it follows that

$$x_{nt} = \frac{1}{2 + \tau_\eta\widehat{\mathbb{V}}[\mu \mid I_{nt}]}\left\{1 + \frac{\tau_\eta}{\alpha W_0}\widehat{\mathbb{E}}[\mu \mid I_{nt}]\right\}. \tag{2.19}$$

Since we assume large markets (the number of investors $N \to \infty$), the aggregate share allocated to the risky asset in period t is $\bar{x}_t = \mathbb{E}_t[x_{nt}]$, where $\mathbb{E}_t$ is the expectation with respect to the true model as of time t.

[27] Although both assets have uncertain payoffs, we refer to Asset 2 as the "risky" asset since its payoff has an additional dimension of uncertainty, and, as typical in the literature, investors update their beliefs about this asset with the arrival of new information. There is nothing to learn about the safe asset.

[28] The game form exactly follows the general model outlined in Section 2.2.1. Specifically, we maintain the Large-Overlapping-Generations assumption. Investor nt participates in the market for a single period, t, and observes the allocation realized by the Generation immediately before her own, $t - 1$.

2.4.2 Belief and Allocation Dynamics

We now solve for belief and allocation dynamics in the naive model, and contrast them with rational dynamics. In the first period, investors act solely on private signals. Since early investors have no opportunity to mislearn from past investments, perceived expectations and variances are correct: for all n in $t = 1$, $I_{n1} = \{s_{n1}\}$,

$$
\begin{aligned}
\widehat{\mathbb{E}}[\mu \mid I_{nt}] &= \frac{\tau_\epsilon}{\tau_\epsilon + \tau_\mu} s_{nt} + \frac{\tau_\mu}{\tau_\epsilon + \tau_\mu} \mu_1, \\
\widehat{\mathbb{V}}[\mu \mid I_{nt}] &= \frac{1}{\tau_\epsilon + \tau_\mu}.
\end{aligned}
$$

Lemma 3 shows that the first-period allocation must satisfy a linear “Autarkic demand function”, $\bar{x}_1 = \beta + \gamma\mu$, where β and γ are known constants. Specifically:

Lemma 3. *In $t = 1$, the aggregate share invested in the risky asset must satisfy the autarkic equilibrium price function $\bar{x}_1 = \beta + \gamma\mu$ where*

$$
\begin{aligned}
\beta &= \frac{1}{2(\tau_\epsilon + \tau_\mu) + \tau_\eta} \left\{ \tau_\mu \left(\frac{\tau_\eta}{\alpha W_0} \mu_1 + 1 \right) + \tau_\epsilon \right\}, && (2.20) \\
\gamma &= \frac{\tau_\eta}{\alpha W_0} \left(\frac{\tau_\epsilon}{2(\tau_\epsilon + \tau_\mu) + \tau_\eta} \right). && (2.21)
\end{aligned}
$$

The first-period allocation efficiently aggregates all private signals. Since we assume the market is large, this implies that $\bar{x}_1$ perfectly reveals μ.

In periods $t \geq 2$, investors use $\bar{x}_{t-1}$ to draw inference about μ. Naive investors misinterpret $\bar{x}_{t-1}$: they think demand in $t-1$ is based solely on private information, and hence think $\bar{x}_{t-1}$ must satisfy the autarkic demand function. That is, each Generation t wrongly thinks $\bar{x}_{t-1} = \beta + \gamma\mu$. Generation t inverts this relation to arrive at their (mis)perception of predictable shock μ, $\hat{\mu}_t = (\bar{x}_{t-1} - \beta)/\gamma$.

In truth, however, investors do learn (albeit incorrectly) from past allocations. As such, for all $t \geq 2$, $\bar{x}_t$ is generically distinct from the autarkic demand. Since Generation 1 does act solely on private information, $\bar{x}_1 = \beta + \gamma\mu$. Generation 2 correctly and precisely infers μ: for all investors in $t = 2$, $\widehat{\mathbb{E}}[\mu \mid I_{n2}] = \hat{\mu}_2 = \mu$, $\widehat{\mathbb{V}}[\mu \mid I_{n2}] = 0$. Demand in $t = 2$ properly adjusts to these new perceptions. From Equation 2.19, it follows that

$$
\bar{x}_2 = x_{n2} = \frac{1}{2} \left(1 + \frac{\tau_\eta}{\alpha W_0} \mu \right). \tag{2.22}
$$

The inferential error begins among Generation 3. They neglect that investors in $t = 2$ perfectly infer μ from $\bar{x}_1$ and consequently think μ must solve $\bar{x}_2 = \beta + \gamma\mu$. It follows that for all $t \geq 2$, Generation t grows certain that μ has value

$$
\hat{\mu}_t = (\bar{x}_{t-1} - \beta)/\gamma, \tag{2.23}
$$

which leads to aggregate allocation

$$\bar{x}_t = x_{nt} = \frac{1}{2}\left(1 + \frac{\tau_\eta}{\alpha W_0}\hat{\mu}_t\right).^{29} \tag{2.24}$$

Taken together, Equations 2.23 and 2.24 recursively define the law of motion for the allocation process, $\langle \bar{x}_t \rangle$, which we characterize in Lemma 4 .

Lemma 4. *If investors are naive, then aggregate allocations $\langle \bar{x}_t \rangle$ evolve as follows: for all $t \geq 2$,*

$$\bar{x}_t = \begin{cases} 0 & \text{if} \quad \kappa_x + \kappa \bar{x}_{t-1} < 0, \\ \kappa_x + \kappa \bar{x}_{t-1} & \text{if} \quad \kappa_x + \kappa \bar{x}_{t-1} \in (0,1), \\ 1 & \text{if} \quad \kappa_x + \kappa \bar{x}_{t-1} > 1, \end{cases} \tag{2.25}$$

where

$$\kappa_x = -\frac{\tau_\mu}{2\tau_\epsilon}\left(\frac{\tau_\eta}{\alpha W_0}\mu_1 + 1\right), \tag{2.26}$$

$$\kappa = 1 + \frac{2\tau_\mu + \tau_\eta}{2\tau_\epsilon} > 1, \tag{2.27}$$

starting from initial condition $\bar{x}_1 = \beta + \gamma\mu$.

It's clear that this process implicitly defines $\langle \hat{\mu}_t \rangle$ via the linear relation in Equation 2.24.[30]

Before analyzing investment dynamics, it's worth contrasting naive and rational allocations. The rational allocation remains constant across all $t \geq 2$. Rational investors realize that predecessors efficiently use all available information, and thus perfectly infer the predictable payoff from any predecessor's split. Relative to the autarkic demand, they allocate more wealth to the risky asset if and only if they learn $\mu > 0$. Furthermore, because of aggregate uncertainty, unless μ is very large in absolute value, rational investors always diversify—that is, $0 < \bar{x}_t < 1$.[31] It's clear from Lemma 4, however, that naive allocations are not static. Investors in t form beliefs as if Generation $t-1$ used new independent information to arrive at $\bar{x}_{t-1}$. As such, naive investors always think that past demand reflects information not yet accounted for.[32]

[29] This argument assumes $\bar{x}_{t-1} \in (0,1)$. Proposition 5 deals with the case of $\bar{x}_{t-1} \in \{0,1\}$.

[30] This is true so long as the system hasn't yet reached the boundary. Proposition 5 shows that once $\bar{x}_t = 1$, then $\hat{\mu}_t = \infty$ for all $\tau > t$. Likewise, if $\bar{x}_t = 0$, then $\hat{\mu}_\tau = -\infty$ for all $\tau > t$.

[31] More specifically, a player invests all wealth to the risky asset if perceived return $\hat{\mu}$ exceeds value $\overline{\mu}$ such that Equation 2.24 equals 1. Hence, $\overline{\mu} = \alpha W_0/\tau_\eta$. Likewise, she invests all wealth to the safe asset whenever $\hat{\mu} < \underline{\mu}$, where $\underline{\mu}$ is the value of $\hat{\mu}$ such that Equation 2.24 equals 0: $\underline{\mu} = -\alpha W_0/\tau_\eta$.

[32] Naive investors *expect* allocations to be constant across all generations. They anticipate that $\bar{x}_t = \beta + \gamma\mu$ for all t. Consequently, naive investors think last-period's demand $\bar{x}_{t-1}$ is sufficient for the entire allocation path up to time t.

We now describe how allocations evolve over time. Belief dynamics reveal two ways in which naivete can lead investors astray: (1) they inevitably allocate all wealth to a single asset, which often implies under-diversification, and (2) with positive probability, they invest all wealth in an asset dominated by its alternative.

First, the perceived payoff difference between the two assets grows over time. As t grows large, naive agents form polarized perceptions about the risky asset's return: beliefs about μ diverge to positive or negative infinity. As such, the allocation converges to either 1 or 0. Formally:

Proposition 5. *For any realization of μ:*

1. *Starting in $t = 2$, $\langle \hat{\mu}_t \rangle$ and $\langle x_t \rangle$ are either both increasing or both decreasing in t.*
2. *$\langle \hat{\mu}_t \rangle$ and $\langle \bar{x}_t \rangle$ are increasing if and only if $\mu > \mu^*$ where*

$$\mu^* \equiv \frac{1}{2\tau_\mu + \tau_\eta}\big(2\tau_\mu \mu_1 - \alpha W_0\big). \tag{2.28}$$

   That is, if and only if the realized return is sufficient large relative to initial expectations.

3. *If $\langle \bar{x}_t \rangle$ is increasing, then $\lim_{t\to\infty} \bar{x}_t = 1$ and $\lim_{t\to\infty} \hat{\mu}_t = \infty$. Otherwise, $\lim_{t\to\infty} \bar{x}_t = 0$ and $\lim_{t\to\infty} \hat{\mu}_t = -\infty$.*

Part 1 of Proposition 5 shows that beliefs and allocations either increase or decrease monotonically in t. Why and when they increase (or decrease) is captured by the logic of Part 2: it increases if and only if the predictable return is sufficiently high relative to initial expectations. From Lemma 3, the autarkic demand is a linear combination of the initial expectation μ_1 and the realization of μ. When investors learn μ in $t = 2$, allocations rationally respond to this information. Roughly speaking, if μ beats expectations, allocations increase; if not, they decrease.

More precisely, the allocation to the risky asset increases between periods 1 and 2 if and only if $\mu > \mu^*$, where μ^* (Equation 2.28) is somewhat below μ_1. That is, investment may increase even when μ falls below expectations. This follows from risk aversion. Autarkic investment is cautious: since players are risk averse and uncertain of μ, they choose a conservative split. Upon learning μ in $t = 2$, the perceived variance in the risky asset's payoff decreases, increasing players' willingness to invest.

The change in allocation between periods 1 and 2 creates momentum that propagates through all future periods. Assuming that the allocation increases between periods 1 and 2, if followers treat the revised split as the autarkic demand, then they infer a higher value of μ than the previous generation. As such, today's allocation moves yet higher. This logic plays out across all periods: each new generation observes a larger "autarkic" allocation than the last, which continually leads followers to allocate more to the risky asset. Similarly, whenever the initial (rational) change in allocation is downward—which happens whenever $\mu < \mu^*$—then demand and perceived payoffs decrease over time.

The error is driven by investors continually using the past demand as if it reflects new information.[33] This is the essence of redundancy neglect. Investors neglect that observed demand already incorporates all information in the economy, and attribute any changes to private information. When the current generation incorporates this "new" information, the allocation moves yet again in the same direction as the initial (rational) adjustment. As such, naive inference provides a plausible explanation for momentum even when no new information is realized.

Part 3 of Proposition 5 establishes that aggregate allocations increase or decrease until investors either allocate all or no wealth to the risky asset. This clearly implies that naive risk-averse investors are worse off relative to rational investors whenever the true predictable return leads rational investors to diversify—whenever $\mu \in (-\alpha W_0/\tau_\eta, \alpha W_0/\tau_\eta)$. Even more damning, naive investors may fail to correctly identify which asset yields the higher payoff, whereby allocating all wealth to the dominated asset.

Corollary 1. *For any collection of parameters $(\mu_1, \tau_\mu, \tau_\epsilon, \alpha)$, there exists an open interval $\mathcal{M}$ such that whenever $\mu \in \mathcal{M}$, investors incorrectly rank the payoffs of the two assets.*

The intuition is straightforward in light of Proposition 5. First, the payoff of the risky asset is higher than the safe asset whenever $\mu > 0$. Consider the case when $\mu_1 > \mu^* > 0$: people expect the risky asset to outperform the safe asset. Whenever $\mu \in (0, \mu^*)$ this expectation is realized. But because μ falls sufficiently short of expectations, by Part 2 of Proposition 5, perceptions of μ decrease over time. Eventually, investors conclude that the safe asset yields a higher payoff and consequently allocate no wealth to the risky asset. Investors similarly come to believe μ is large even when $\mu^* < \mu < 0$.

2.5 Learning the Distribution of Information

This section explores how naive inference distorts players' perceptions of the distribution of private information in the economy. We consider settings where players not only decide which payoff structure best explains herds, but what *information* structure most likely causes autarkic players to herd. Until now, we followed the standard social-learning literature in assuming players know the distribution of private signals conditional on the payoff-relevant state. We expand on our polarization results of Section 2.3.2 by relaxing this assumption in two ways. In Section 2.5.1, the precision of private information is unknown. Since naive observers expect variation in actions proportional to the variation in private information, they conclude signals have the highest possible precision after observing a herd. In Section 2.5.2 we add aggregate uncertainty to the environment: a rational agent remains uncertain about

[33] This assumption is quite opposite that of Eyster, Rabin, and Vayanos (2013). They study asset pricing when traders fail to learn from price. Our assumptions lead investors to *over*-infer from past behavior—they draw inference even when demand provides no new information.

payoffs even if she receives an infinite number of private signals. A naive player rightfully anticipates that she'll remain uncertain about payoffs in the long run. Much to her surprise, she inevitably grows confident in some (likely false) payoff state. In attempt to rationalize herd behavior, naive observers come to believe in the state with the least aggregate uncertainty.

2.5.1 Unknown Precision of Signals

We first consider an environment where the precision of private information is unknown. Social learning causes people to herd on a single action. Naive followers, who assume those herding do so based on private information, conclude that private signals must be as precise as possible. Since the amount of variation in autarkic behavior decreases as the precision of private information increases, very precise private information is the best (naive) explanation for why all predecessors make the same choice.

To clearly make this point, we focus on the simplest variant of the independent-quality setting described in Section 2.3.2. Each of M actions has independent binary quality $q_m \in \{0,1\}$ with prior $\Pr(q_m = 1) = 1/2$. Each player receives a conditionally-independent binary signal $s^m \in \{0,1\}$ about each q_m. Assume $\Pr(s^m = q_m \mid q_m) = \rho$ for each m; a signal is accurate with probability ρ. The *precision* of private information—$\rho \in [.5,1]$—is unknown, and players share a common prior with full support over $[.5,1]$.[34] A state $\omega = (q_1, ..., q_M; \rho)$ is a vector of qualities for each m and the precision parameter: $\Omega = \{0,1\}^M \times [.5,1]$.

To determine naive long-run beliefs, we must first determine how people behave in autarky. As in Section 2.3.2, we assume $u(A_m|\omega) = q_m$. So, when acting solely on private signals, Player nt chooses at random among those options about which she receives good news (i.e, $s^m_{nt} = 1$). If $s^m_{nt} = 0$ for all m, then she chooses at random from all options. Behavior of Generation 1 follows this decision rule. Generation 2 rationally infers that the action chosen most often by Generation 1 has the highest expected quality among all options. Denote this action by $m^* \equiv \arg\max_{m=1,...,M}(a_1(1), ..., a_1(M))$. It follows that all players in Generation 2 rationally choose A_{m^*}.

To explain the herd, Generations $t \geq 3$ come to believe in whichever state maximizes the likelihood that A_{m^*} is chosen in autarky. That is, the state that maximizes $\mathbb{P}_{(\mathbf{q},\rho)}(m^*)$. From Proposition 2, we know that for a fixed ρ, $(\omega^P_{m^*}, \rho)$ maximizes $\mathbb{P}_{(\mathbf{q},\rho)}$ if and only if $\omega^P_{m^*}$ is the polar quality vector from Definition 7—$q_{m^*} = 1$ and $q_m = 0$ for all $m \neq m^*$. So, all Generations $t \geq 3$ must believe $\mathbf{q} = \omega^P_{m^*}$. What do they think about the precision of signals? Fixing this belief $\mathbf{q} = \omega^P_{m^*}$, $\mathbb{P}_{(\omega^P_{m^*},\rho)}(m^*)$ as a function of ρ is

$$\mathbb{P}_{(\mathbf{q},\rho)}(m^*) = \sum_{k=0}^{M-1} \frac{1}{k+1} \binom{M-1}{k} \rho^{M-k}(1-\rho)^k + \frac{1}{M}(1-\rho)\rho^{M-1}. \tag{2.29}$$

.

[34] We restrict attention to $\rho \in [.5,1]$ so that signals satisfy MLRP (Assumption 11) and uniquely reveal the payoff state when aggregated: $s^m = 1$ occurs more often than $s^m = 0$ if and only if $q^m = 1$. If $\rho \in [0,1]$, then this is no longer true. Acemoglu, Chernozhukov and Yildiz (2009) study learning with this type of ambiguity about signal interpretation.

Maximizing $\mathbb{P}_{(\mathbf{q},\rho)}(m^*)$ with respect to ρ yields $\rho = 1$. This is straightforward: for any value of $\rho < 1$, autarkic players occasionally receive misleading signals and thus take actions other than A_{m^*}. The only case in which *all* players behave identically in atuarky is when there is a uniquely optimal action and misleading signals are impossible; that is, $\mathbf{q} = \omega^P_{m^*}$ and $\rho = 1$. Formally:

Proposition 6. *Let* $m^* \equiv \arg\max_{m=1,...,M}(a_1(1), ..., a_1(M))$. *Naive public beliefs converge on state* $(\omega^P_{m^*}; 1)$*: for all* $t > 2$, $\boldsymbol{\pi}_t \to \delta(\omega^P_{m^*}, 1)$ *as* $N \to \infty$. *Hence, perceived precision* $\hat{\rho} \to 1$ *in* N.

Since naive observers expect that actions are based solely on private signals, they anticipate that the dispersion in behavior reflects the dispersion in private information. And since they observe a herd, they conclude this dispersion is minimal.

An important feature of this environment is that herds are consistent with a naive player's model of the world. That is, there exist states in which all players choose identically in autarky. The environments explored in Section 2.3 don't have this feature. For instance, consider the environment here when $\rho < 1$ is known. From Section 2.3.2, we know that players herd on a single action A_{m^*} and conclude that $q_{m^*} = 1$ and $q_m = 0$ for all $m \neq m^*$. While this state best explains the herd—it minimizes the cross-entropy distance between the predicted and observed play—the long-run distribution of actions doesn't converge to the anticipated value specified by Equation 2.29. Although players expect that a strictly interior fraction of the population takes A_{m^*}, they see the full population do so. Allowing agents to simultaneously draw inference about the signal distribution and payoff structure alleviates this issue of "inexplicable" behavior, permitting naive agents to make sense of what they observe.[35]

2.5.2 Aggregate Uncertainty

We now consider an information structure which incorporates aggregate uncertainty about the payoff state: an agent remains uncertain about payoffs even when she receives an infinitely collection of signals. In this environment—where rational agents always remain uncertain about payoffs—naive individuals inevitably become confident (and, in general, wrong) about payoffs despite fully anticipating to never grow certain. In the previous case with unknown precision of signals, an arbitrarily large collection of signals identifies the payoff state. As such, naive agents expect (and do) grow fully confident about payoffs after observing a large number of predecessors. Here, naive players rationally believe ex-ante that aggregate uncertainty remains in the long-run, but their naive interpretation of a herd leads them to believe in the unlikely event that uncertainty vanishes.

To show this, we consider a slightly modified version of the setting in Section 2.5.1. For sake of exposition, we take the M options to be assets. Unlike Section 2.5.1, the "quality"

[35] Claiming that a positive-probability event is "inexplicable" with respect to an agent's model is more nuanced than than it may seem. Gagnon-Bartsch and Rabin (2014) discuss in detail the concept of "explicablity" in economic models.

of an asset $q_m \in [0,1]$ is not its payoff, but its *expected* payoff. Each asset m pays off $z_m \in \{0,1\}$ independent of one another. Payoff z_m is unknown, but has prior $\Pr(z_m = 1) = q_m$. Importantly, q_m is also unknown; each q_m is drawn independently from common prior π_1 with full support on $[0,1]$. Each player receives a conditionally-independent signal s^m about each q_m such that $\Pr(s^m = 1) = q_m$. That is, the probability of getting a good signal matches the probability the asset pays off. There are two unknown vectors: the vector of expected payoffs, $\mathbf{q} = (q_1, ..., q_M) \in [0,1]^M$, and the vector of realized payoffs, $\boldsymbol{z} = (z_1, ..., z_M) \in \{0,1\}^M$. The payoff relevant state is $\omega = \boldsymbol{z}$, while $\mathbf{q}$ dictates the distribution of information in the economy. We analyze beliefs about joint state $\omega = (\mathbf{q}, \boldsymbol{z}) \in [0,1]^M \times \{0,1\}^M$.

In autarky, players follow the same decision rule as 2.5.1. Player nt chooses at random among those options about which she receives good news (i.e, $s^m_{nt} = 1$), and if $s^m_{nt} = 0$ for all m, then she chooses at random from all options. Observing autarkic play reveals information about the signal distribution, and hence about $\mathbf{q}$. While this provides information about the likelihood that various assets payoff, it reveals nothing about *realized* payoffs. Generation 2 (correctly) infers that the action most often chosen in $t = 1$ has the highest expected payoff; without loss of generality, denote this asset by $m = 1$. In period 2, all people (rationally) herd on A_1.

How do naive players interpret this herd? There is a unique $\hat{\mathbf{q}}$ that precisely predicts an autarkic herd on A_1. Namely, the $\hat{\mathbf{q}}$ such that $\hat{q}_1 = 1$—all players receive a positive signal about Asset 1—and $\hat{q}_m = 0$ for all $m > 1$—all players receive negative signals about all other assets. Conditional on $\hat{\mathbf{q}}$, each player deterministically chooses A_1. Further, $\hat{\mathbf{q}}$ implies that all assets have deterministic payoffs: each pays off with probability 1 or 0. As such, when naive agents grow confident that $\mathbf{q} = \hat{\mathbf{q}}$, they in turn grow confident about the payoff *realization* of each asset, $\boldsymbol{z}$. Formally:

Proposition 7. *If $q_m = \max(q_1, ..., q_M)$, then naive public beliefs converge on state $(\hat{\mathbf{q}}^m, \hat{\boldsymbol{z}}^m)$ where $\hat{q}^m_m = \hat{z}^m_m = 1$ and $\hat{q}^m_j = \hat{z}^m_j = 0$ for all $j \neq m$. That is, for all $t > 2$,$\boldsymbol{\pi}_t \to \delta(\hat{\mathbf{q}}^m, \hat{\boldsymbol{z}}^m)$ as $N \to \infty$. Hence, observers grow certain of the payoff state.*

An important implication of Proposition 7 is that a naive agent grows confident about the payoff state even when she expects to remain uncertain. Naive agents rationally predict that aggregated signals, and hence social behavior, won't perfectly reveal payoffs. But behavior does reveal *some* information about payoffs. Followers use this information and herd on the asset most likely to payoff. Naive agents mistake this as evidence that all predecessors are *privately* informed about the optimal action

This form of naive overconfidence is conceptually different from that shown by Eyster and Rabin (2010). In ER, naive players may grow overconfident relative to rational inference, but not relative to their own expectations. For instance, with discrete signals and actions it is well known that rational inference leads to an information cascade in which rational players remain uncertain about payoffs. Naive players, on the other hand, continually treat actions as if they reveal new independent information. They anticipate that, in the long-run, they'll observe an infinite sequence of independent signals. Since ER assumes no aggregate uncertainty, this implies long-run confidence. With aggregate uncertainty, however, agents grow overconfident relative to their own expectations. While they anticipate observing an infinite collection of independent signals, they don't expect such information to deliver confidence.

2.6 Conclusion and Extensions

2.6.1 Robustness

This section discusses how our results extend to various observation structures. First, we believe that Lemma 1, which characterizes the set of states on which naive public beliefs can converge, holds irrespective of the observation structure so long as the number of predecessors each agent observes grows large in t. That is, Assumptions 8 and 9 are not necessary if, as $t \to \infty$, the number of actions observed by a player in t also goes to infinity. A primary example of such a structure is the canoncial "herding" model where one agent acts per round and each agent observes the full history of play (e.g., Bikchandani, Hirshleifer, and Welch, 1992, or Smith and Sørensen, 2000).

We provide a heuristic sketch demonstrating why the conclusion of Lemma 1 must still hold. Formally, suppose our only assumption on the observation structure is that for all $n = 1, ..., N$, $\lim_t |O_{nt}| = \infty$. Now suppose that $\pi_t(k)$ converges almost surely to 1: society grows confident in state ω_k. We show that ω_k must be a fixed point of ϕ. Since a naive agent treats each observation as reflecting independent private information, the observed *order* of actions does not influence her inference. Naive inference depends only on the *aggregate* distribution of behavior across all players she observes. Let $\mathbf{a}(O_{nt})$ be the distribution of actions in Player nt's observation set.

If beliefs converge, then, as t grows large, players act with near confidence. Granted that actions are continuous in beliefs in some neighborhood about $\delta(\pi_{\omega_k})$, this implies that the distribution of play converges to $\mathbb{T}_{\omega_k}$. It follows that, eventually, observed distributions resemble $\mathbb{T}_{\omega_k}$. That is, $\lim_{t\to\infty} \mathbf{a}(O_{nt}) = \mathbb{T}_{\omega_k}$. In turn, for large enough t Player nt observes an arbitrarily large population taking actions distributed according to $\mathbb{T}_{\omega_k}$. The naive probability of this observation conditional on ω is given by Equation 2.6 replacing $\mathbb{T}^*_\omega$ with $\mathbb{T}_{\omega_k}$. From Section 2.2.2, the state that maximizes this likelihood is $\hat{\omega} = \arg\min_{\omega\in\Omega} H(\mathbb{T}_{\omega_k}, \mathbb{P}_\omega) = \phi(\omega_k)$, where the final equality follows directly from the definition of map ϕ (Equation 2.10). In order for agent nt to remain confident in ω_k, it must be that $\omega_k = \phi(\omega_k)$. If not, we contradict our assumption that $\boldsymbol{\pi}_t$ converge to $\delta(\omega_k)$.

This argument and that of Lemma 1 differ only in how quickly observed behavior converges to $\mathbb{T}_{\omega_k}$. Our assumption of large generations guaranteed that behavior after a single round of observation converges to $\mathbb{T}_{\omega_k}$. Here, behavior may converge to $\mathbb{T}_{\omega_k}$ only in the limit. In either case, so long as observation sets grow large, a necessary condition for beliefs to converge to certainty in ω_k is that ω_k best explains $\mathbb{T}_{\omega_k}$ under the assumption of autarkic play. That is, we require $\omega_k = \phi(\omega_k)$.

A consequence of our large-generation assumption is that beliefs follow a deterministic sequence. An important implication of this assumption is that whenever the true state lies in Ω^*, society necessarily learns correctly. With small N, limit beliefs depends on the sample path of signals and beliefs may converge to *any* $\omega \in \Omega^*$. As such, society mislearns with positive probability even when the truth is in Ω^*. Following Eyster and Rabin (2010) Proposition 4, so long as $\omega \in \Omega^*$, there exists a sample path of signals realized with positive probability such that beliefs converge to $\delta(\omega)$ in t.

Finally, it's worth noting that we assume an extreme form of naivety in this paper: each player thinks predecessors entirely neglect the informational content of others' actions. Eyster and Rabin (2014) show that a weaker form of this bias generates the same conclusions as Eyster and Rabin (2010). They generalize this extreme form of the bias by assuming players think each predecessor's action reveals some arbitrarily small amount of her private information.[36] We conjecture that the our main results also hold under weaker generalizations of naivete. For instance, we believe that no matter the extent of redundancy neglect, naive beliefs converge over time toward polarized perceptions of payoffs.

2.6.2 Discussion

This paper explores new predictions of Eyster and Rabin's (2010) model of naive inference that emerge in an array of environments richer than those previously studied. With a range of possible payoff-relevant states, naive inference restricts the set of states upon on which society may converge. In many natural environments, there exist states that naive agents always disbelieve in the long-run no matter if they are true. Eyster and Rabin's (2010) obscured this result by focusing on a binary-state setting in which players always grow confident in one of the two states. In particular, we show that when agents care not only about ranking actions but wish to learn the size of payoff differences, naive inference leads to polarization in perceived payoffs. In settings where agents care about diversification, this distortion in beliefs generates inefficiencies through under-diversification. The force driving polarization in perceived payoffs also extends to beliefs about the distribution of information in the economy. In some settings, agents wrongly come to believe that private information is as precise as possible or that aggregate uncertainty is resolved.

This is not the first paper to study "redundancy neglect" in settings with rich state spaces. However, it does provide predictions distinct from earlier work. For instance, DeMarzo, Vayanos, and Zweibel (2003) consider a variant of the DeGroot (1974) model in which players share their signals about a normally-distributed state ω with their neighbors in a network. Each round, players observe the posterior mean belief of their neighbors and accordingly update their own beliefs. Player i treats her neighbors' reports as raw signals; she neglects that neighbors share posteriors which already incorporate information from initial signals that were previously shared. Since agents over-count signals, they grow confident in some false state whenever initial signals are misleading. The implications of this error differ from ours in two ways. First, agents in DeMarzo, Vayanos, and Zweibel's model don't

[36] In the extreme form, Eyster and Rabin (2010) assume that players think each predecessors' action fully reveals her private signal.

adopt fully polarized perceptions over time. Since players use a naive averaging rule, beliefs converge on a weighted average of initial signals rather than tending to extreme values. Second, when the number of agents observed grows large, players in DeMarzo, Vayanos, and Zweibel's model are correct in the long run. By the Law of Large Numbers, the first round of communication sends players directly to a confident and correct posterior. In our setting, even if players correctly learn the state after one round of observation, later generations mislearn by reinterpreting confident behavior as if it were autarkic. Relative to existing models, polarization and unlearning are distinct predictions Eyster and Rabin's model of naive inference.

Appendix

2.A Proofs

Proof of Lemma 1.

Proof. For a contradiction, suppose $\langle \boldsymbol{\pi}_t \rangle$ converges in t to some $\omega \notin \Omega^*$. Formally, suppose there exists τ such that in all $t > \tau$, $\lim_{N\to\infty} \boldsymbol{\pi}_t = \delta(\omega)$. Fix $t > \tau$. Since $\boldsymbol{\pi}_t = \delta(\omega)$, $\mathbf{a}_t = \mathbb{T}_\omega$. Generation $t+1$ infers $\boldsymbol{\pi}_{t+1} = \delta(\hat{\omega}_{t+1})$ where $\hat{\omega}_{t+1} = \arg\min_{\hat{\omega}\in\Omega} H(\mathbb{P}_{\hat{\omega}}, \mathbb{T}_\omega) = \phi(\omega)$. Since $\omega \notin \Omega^*$, $\phi(\omega) \neq \omega$. Thus $\hat{\omega}_{t+1} \neq \omega$. This contradicts the assumption that $\langle \boldsymbol{\pi}_t \rangle$ converges in t to ω. □

Proof of Proposition 1.

Proof. By Observation 1, each $t \geq 2$ acts with confidence. By the assumption of common tastes, in each t $\mathbf{a}_t$ represents a herd: for some $m = 1, ..., M$, $a_t(m) = 1$ and $a_t(j) = 0$ for all $j \neq m$. Denote this distribution by $\mathbf{a}^m$. For each adjacent generation, $\hat{\omega}_{t+1} = \phi(\hat{\omega}_t) = \arg\min_{\omega\in\Omega} H(\mathbb{P}_\omega, \mathbf{a}^m)$ for some m. Since H is concave, by Assumption 10, the solution is unique. Thus, there are at most M states society can conclude are true after a herd. Let $\widetilde{\Omega} \equiv \{\omega \in \Omega \mid \omega = \arg\min_{\omega'\in\Omega} H(\mathbb{P}_{\omega'}, \mathbf{a}^m),\ m = 1, ..., M\}$ be the set of states in which society grows confident after observing any herd. Since each $\mathbf{a}_t$ for $t \geq 2$ represents a herd, ϕ must map any point in $\widetilde{\Omega}$ back to $\widetilde{\Omega}$, and thus any fixed point of ϕ must lie in $\widetilde{\Omega}$. Hence, $\Omega^* \subseteq \widetilde{\Omega} \subset \Omega$. □

Proof of Lemma 2.

Proof. Without loss of generality, we prove the result for the autarkic frequency of action A_1. We make use of two well-known implications of the MLRP assumption (e.g., Milgrom, 1981):

Observation 2. *Suppose Assumption 11 holds.*

1. *For each m, $F^m(s|q_m)$ satisfies first-order stochastic dominance in s: if $q_m > q'_m$, then $F^m(s|q_m) \leq F^m(s|q'_m)$ for all $s \in S^m$, and the inequality is strict for some $s \in S^m$.*

2. *For each m, $\mathbb{E}[q_m \mid s^m]$ is increasing s^m.*

In autarky, a player with signal realization $\mathbf{s}$ chooses A_1 if $1 = \arg\max_{m\in\{1,...,M\}} \mathbb{E}[q_m \mid \mathbf{s}]$. Since we assume signals are independent across options, $\mathbb{E}[q_m \mid \mathbf{s}] = \mathbb{E}[q_m \mid s^m]$ for each m. Let $\tilde{\mathbf{s}} = (s_2, ..., s_M)$ denote the signal vector excluding the first entry. Let $v(\tilde{\mathbf{s}}) = \max_{m\geq 2} \mathbb{E}[q_m \mid s^m]$; $v(\tilde{\mathbf{s}})$ is the expected value of the best option among $m = 2, ..., M$. Fixing priors on Ω and realization $\tilde{\mathbf{s}}$, let $\sigma(s^2, ..., s^M) = \sigma(\tilde{\mathbf{s}})$ be the realization of s^1 necessary to be indifferent between A_1 and the best option among $A_2, ..., A_M$. That is, $\sigma(\tilde{\mathbf{s}})$ is defined implicitly by $\mathbb{E}[q_1 \mid \sigma(\tilde{\mathbf{s}})] = v(\tilde{\mathbf{s}})$.[37] It follows that an autarkic agent with signal realization $\mathbf{s}$ chooses $A_1 \Leftrightarrow s^1 > \sigma(\tilde{\mathbf{s}})$. Let $\mathbf{q} = (q_1, ...q_M)$ be the realized vector of qualities, and let $\tilde{\mathbf{q}} = (q_2, ..., q_M)$. Then, the autarkic frequency of action A_1 is given by

$$\mathbb{P}_\omega(1) = \Pr\left(s^1 > \sigma(\tilde{\mathbf{s}}) \;\middle|\; \mathbf{q}\right) = \mathbb{E}_{\tilde{\mathbf{s}}}\left[1 - F^1\big(\sigma(\tilde{\mathbf{s}})\big|q_1\big) \;\middle|\; \tilde{\mathbf{q}}\right] = 1 - \mathbb{E}_{\tilde{\mathbf{s}}}\left[F^1\big(\sigma(\tilde{\mathbf{s}})\big|q_1\big) \;\middle|\; \tilde{\mathbf{q}}\right], \quad \text{(A.1)}$$

where $\mathbb{E}_{\tilde{\mathbf{s}}}[\cdot \mid \tilde{\mathbf{q}}]$ indicates the expectation over random variable $\tilde{\mathbf{s}}$ given $\tilde{\mathbf{q}}$.

First we show that $\mathbb{P}_\omega(1)$ is increasing in q_1. It follows from Part 1 of Observation 2 that $\mathbb{E}_{\tilde{\mathbf{s}}}\big[F^1\big(\sigma(\tilde{\mathbf{s}})\big|q_1\big) \mid \tilde{\mathbf{q}}\big] < \mathbb{E}_{\tilde{\mathbf{s}}}\big[F^1\big(\sigma(\tilde{\mathbf{s}})\big|q_1'\big) \mid \tilde{\mathbf{q}}\big]$ whenever $q_1 > q_1'$. Hence, from Expression A.1, $\mathbb{P}_\omega(1)$ is increasing in q_1.

Next, we show $\mathbb{P}_\omega(1)$ is decreasing in q_m for all $m \geq 2$. Note that from Part 2 of Observation 2, $v(\tilde{\mathbf{s}})$ is weakly increasing in each s^m, $m \geq 2$; thus, by that same observation and the definition of σ, $\sigma(\tilde{\mathbf{s}})$ must also weakly increase in each s^m, $m \geq 2$. It follows from Part 2 of Observation 2 that σ is weakly increasing in s_m for each $m \geq 2$: increasing the signal of any $m \geq 2$ weakly increases one's estimate of the second-best quality. Consider any $m \geq 2$. By the law of iterated expectations and independence,

$$\mathbb{E}_{\tilde{\mathbf{s}}}\left[F^1\big(\sigma(\tilde{\mathbf{s}})\big|q_1\big) \;\middle|\; \tilde{\mathbf{q}}\right] = \mathbb{E}_{s_m}\left[\mathbb{E}_{\tilde{\mathbf{s}}}\big[F^1\big(\sigma(\tilde{\mathbf{s}})\big|q_1\big) \mid \tilde{\mathbf{q}}, s_m\big] \;\middle|\; q_m\right], \quad \text{(A.2)}$$

and since σ is increasing, $\mathbb{E}_{\tilde{\mathbf{s}}}\big[F^1\big(\sigma(\tilde{\mathbf{s}})\big|q_1\big) \mid \tilde{\mathbf{q}}, s_m\big]$ is an increasing function of s_m. So, since the distribution of s_m exhibits first-order stochastic dominance in q_m, $\mathbb{E}_{s_m}\left[\mathbb{E}_{\tilde{\mathbf{s}}}\big[F^1\big(\sigma(\tilde{\mathbf{s}})\big|q_1\big) \mid \tilde{\mathbf{q}}, s_m\big] \;\middle|\; q_m\right]$ is increasing q_m. It follows from A.1 that $\mathbb{P}_\omega(1)$ is decreasing in q_m.

Since $\mathbb{P}_\omega(1)$ is strictly increasing in q_1 and strictly decreasing in q_j for all $j \geq 2$, it follows that ω_1^P uniquely maximizes $\mathbb{P}_\omega(1)$.

□

Proof of Proposition 2.

Proof. Without loss of generality, index options such that $q_1 = \arg\max_m\{q_m\}$. By Assumption 10, $\mathbf{a}_1 = (a_1(1), ..., a_1(M))$ reveals the utility ranking of options $\{A_1, .., A_M\}$ to

[37] This definition does not restrict $\sigma(\tilde{\mathbf{s}})$ to lie in S^1—the set of signals over which s^1 has positive density. $\sigma(\tilde{\mathbf{s}})$ is the necessary signal value, whether feasible or not, such that a player is indifferent between A_1 and the best option among $A_2, ..., A_M$.

Generation 2. Rationally, all N in $t = 2$ choose A_1: $a_2(1) = N$ and $a_2(m) = 0$ for $m \geq 2$. Among Generation 3, the likelihood ratio of observing $\mathbf{a}_2$ in ω_1^P relative to any $\omega \neq \omega_1^P$ is

$$\frac{\pi_3(\omega_1^P)}{\pi_3(\omega)} = \frac{\prod_{m=1}^M \mathbb{P}_{\omega_1^P}(m)^{a_2(m)}}{\prod_{m=1}^M \mathbb{P}_{\omega}(m)^{a_2(m)}} = \left(\frac{\mathbb{P}_{\omega_1^P}(1)}{\mathbb{P}_{\omega}(1)}\right)^N .$$

Hence, naive beliefs in $t = 3$ converge to $\delta(\omega_1^P)$ in N if and only if $\omega_1^P = \arg\max_{\omega \in \Omega} \mathbb{P}(A_1 \mid \omega)$, which is true by Lemma 2.

□

Proof of Proposition 3.

Proof. Part 1. Suppose $\omega = H$. As $N \to \infty$, $\mathbf{a}_2 \to \mathbb{T}_H$ with $\mathbb{T}_H(A) = \lambda$. From Equation 2.7, the likelihood ratio $\pi_3(H)/\pi_3(L)$ entering $t = 3$ converges to

$$\lim_{N \to \infty} \left[\frac{\mathcal{L}(H \mid \mathbf{a}_2)}{\mathcal{L}(L \mid \mathbf{a}_2)}\right]^N = \lim_{N \to \infty} \left[\left(\frac{\mathbb{P}_H(A)}{\mathbb{P}_L(B)}\right)^{\lambda} \left(\frac{\mathbb{P}_H(B)}{\mathbb{P}_L(B)}\right)^{1-\lambda}\right]^N . \tag{A.3}$$

Since $\mathbb{P}_H(A) = \rho\lambda$ and $\mathbb{P}_L(A) = \rho(1-\lambda)$, Equation A.3

$$\frac{\pi_3(\omega = H)}{\pi_3(\omega = L)} \to \lim_{N \to \infty} \left[\left(\frac{\lambda}{1-\lambda}\right)^{\lambda} \left(\frac{1-\rho\lambda}{1+\rho\lambda-\rho}\right)^{1-\lambda}\right]^N ,$$

which diverges to ∞ iff $\xi(\lambda, \rho) \equiv \left(\frac{\lambda}{1-\lambda}\right)^{\lambda} \left(\frac{1-\rho\lambda}{1+\rho\lambda-\rho}\right)^{1-\lambda} > 1$. Hence to show $\pi_3(H) \to 1$, it suffices to show $\xi(\lambda, \rho) > 1$. Since $\xi(\lambda, \rho)$ is decreasing in ρ, $\xi(\lambda, \rho) > 1$ for all ρ if it's true at $\rho = 1$. Note $\xi(\lambda, 1) = \left(\frac{\lambda}{1-\lambda}\right)^{\lambda} \left(\frac{1-\lambda}{\lambda}\right)^{1-\lambda} = \left(\frac{\lambda}{1-\lambda}\right)^{2\lambda-1} > 1$ since, by assumption, $\lambda > 1/2$. Since $\pi_3(H) \to 1$, $\mathbf{a}_3 \to \mathbb{T}_3$. By induction, $\pi_t(H) \to 1$ for all $t \geq 3$.

Part 2. Suppose $\omega = L$. As $N \to \infty$, $\mathbf{a}_2 \to \mathbb{T}_L$ with $\mathbb{T}_L(A) = 1 - \lambda$. Like Equation A.3, the likelihood ratio $\pi_3(H)/\pi_3(L)$ entering $t = 3$ converges to

$$\frac{\pi_3(\omega = H)}{\pi_3(\omega = L)} \to \lim_{N \to \infty} \left[\left(\frac{\lambda}{1-\lambda}\right)^{1-\lambda} \left(\frac{1-\rho\lambda}{1+\rho\lambda-\rho}\right)^{\lambda}\right]^N ,$$

which converges to ∞ iff $\tilde{\xi}(\lambda, \rho) \equiv \left(\frac{\lambda}{1-\lambda}\right)^{1-\lambda} \left(\frac{1-\rho\lambda}{1+\rho\lambda-\rho}\right)^{\lambda} > 1$. Fixing $\lambda > 1/2$, $\tilde{\xi}(\lambda, \rho) > 1 \Leftrightarrow \left(\frac{1-\rho\lambda}{1+\rho\lambda-\rho}\right) > \Lambda$ where $\Lambda \equiv \left(\frac{1-\lambda}{\lambda}\right)^{\frac{1-\lambda}{\lambda}} < 1$. This holds so long as $\rho < \frac{1-\Lambda}{\lambda - \Lambda(1-\lambda)} \equiv \bar{\rho}(\lambda)$. Note that $\bar{\rho}(\lambda)$ is decreasing in λ: $\frac{\partial}{\partial\lambda}\bar{\rho}(\lambda) < 0 \Leftrightarrow$

$$\left(\frac{2\lambda - 1}{\lambda^2}\right) \left(\log\left(\frac{1-\lambda}{\lambda}\right) + 1\right) \Lambda < 1. \tag{A.4}$$

But for any $\lambda \in (1/2, 1)$: (i) $\frac{2\lambda-1}{\lambda^2} < 1$, (ii) $\log\left(\frac{1-\lambda}{\lambda}\right) + 1 < 1$, and (iii) $\Lambda < 1$. Thus Condition A.4 must hold on the interior of $[.5, 1]$. On the boundaries, it's straightforward to verify $\bar{\rho}(.5) = 1$ and $\bar{\rho}(1) = 0$. Thus for any $\lambda \in (1/2, 1)$, $\bar{\rho}(\lambda) \in (0, 1)$. And $\rho < \bar{\rho}(\lambda)$ implies $\pi_3(H) \to 1$ as $N \to \infty$. Thus $\mathbf{a}_3 = \mathbb{T}_H$ with $\mathbb{T}_H(A) = \lambda$. This is the initial condition of the game analyzed in Part 1, thus it must be that $\pi_t(H) \to 1$ for all $t \geq 3$. □

Proof of Lemma 3.

Proof. Since priors and signals about μ are normal, it follows (e.g., DeGroot, 1970) that an investor with signal s_{nt} has a normal posterior about μ with mean and variance

$$\begin{aligned} \widehat{\mathbb{E}}[\mu \mid I_{nt}] &= \frac{\tau_\epsilon}{\tau_\epsilon + \tau_\mu} s_{nt} + \frac{\tau_\mu}{\tau_\epsilon + \tau_\mu}\mu_1, \\ \widehat{\mathbb{V}}[\mu \mid I_{nt}] &= \frac{1}{\tau_\epsilon + \tau_\mu}. \end{aligned}$$

Given these perceptions, individual demand (Equation 2.19) is

$$x_{n1} = \frac{\tau_\epsilon + \tau_\mu}{2(\tau_\epsilon + \tau_\mu) + \tau_\eta}\left\{1 + \frac{\tau_\eta}{\alpha W_0}\left(\frac{\tau_\epsilon}{\tau_\epsilon + \tau_\mu}s_{n1} + \frac{\tau_\mu}{\tau_\epsilon + \tau_\mu}\mu_1\right)\right\}.$$

Aggregate demand is $\bar{x}_1 = \mathbb{E}_1[x_{n1}]$. Since s_{n1} is the only random variable in x_{n1} and $\mathbb{E}_1[s_{n1}] = \mu$,

$$\bar{x}_1 = \frac{\tau_\epsilon + \tau_\mu}{2(\tau_\epsilon + \tau_\mu) + \tau_\eta}\left\{1 + \frac{\tau_\eta}{\alpha W_0}\left(\frac{\tau_\epsilon}{\tau_\epsilon + \tau_\mu}\mu + \frac{\tau_\mu}{\tau_\epsilon + \tau_\mu}\mu_1\right)\right\}. \tag{A.5}$$

Defining β and γ accordingly yields the result. □

Proof of Lemma 4.

Proof. Plugging Equation 2.23, which defines $\hat{\mu}_t$, into the equation for demand (Equation 2.24) and using the definition of γ (Lemma 3) yields

$$\bar{x}_t = \frac{1}{2}\left(1 + \frac{\tau_\eta}{\alpha W_0}(\bar{x}_{t-1} - \beta)/\gamma\right) = \frac{1}{2} - \kappa\beta + \kappa\bar{x}_{t-1},$$

where $\kappa = 1 + (2\tau_\mu + \tau_\eta)/2\tau_\epsilon > 1$. It follows from the definition of β (Lemma 3) that

$$\kappa_x \equiv \frac{1}{2} - \kappa\beta = -\frac{\tau_\mu}{2\tau_\epsilon}\left(\frac{\tau_\eta}{\alpha W_0}\mu_1 + 1\right).$$

□

Proof of Proposition 5.

Proof. Part 1. We show that $\langle \bar{x}_t \rangle$ is monotonic. It's clear from Equation 2.24 that $\hat{\mu}_t$ is monotonic if and only if $\bar{x}_t$ is. From Lemma 4, $\bar{x}_t > \bar{x}_{t-1} \Leftrightarrow \kappa_x + \kappa \bar{x}_{t-1} > \bar{x}_{t-1} \Leftrightarrow \bar{x}_{t-1} > -\kappa_x/(\kappa - 1) \equiv \bar{x}^*$. (Recall from Lemma 4 that $\kappa - 1 > 0$.) Thus we need only check whether the initial value $\bar{x}_1 > \bar{x}^*$. If so, then $\bar{x}_2 > \bar{x}_1 > \bar{x}^*$, and by induction, all $\bar{x}_t > \bar{x}^*$. Thus, $\langle \bar{x}_t \rangle$ is increasing. Similarly, if $\bar{x}_1 < \bar{x}^*$, then $\langle \bar{x}_t \rangle$ is decreasing.

Part 2. From Part 1, $\langle \bar{x}_t \rangle$ is increasing if and only if $\bar{x}_1 > \bar{x}^*$. Using Lemma 4,

$$\bar{x}^* = \frac{-\kappa_p}{1-\kappa} = \frac{\tau_\mu}{2\tau_\mu + \tau_\eta}\left(\frac{\tau_\eta}{\alpha W_0}\mu_1 + 1\right). \tag{A.6}$$

It follows that $\langle \bar{x}_t \rangle$ is increasing iff $\bar{x}_1 > \bar{x}^* \Leftrightarrow \beta + \gamma\mu > \bar{x}^*$. After substituting Equations A.5 and A.6, this condition reduces to

$$\mu > \frac{1}{2\tau_\mu + \tau_\eta}\big(2\tau_\mu \mu_1 - \alpha W_0\big).$$

We define the right-hand side of the above inequality as μ^*.

Part 3. We show that allocation differences $\Delta_x(t) \equiv \bar{x}_t - \bar{x}_{t-1}$ are increasing in magnitude in t. Note $\Delta_x(t) = \bar{x}_t - \bar{x}_{t-1} = \kappa_x + (\kappa - 1)\bar{x}_{t-1}$. But $\Delta_x(t+1) = \bar{x}_{t+1} - \bar{x}_t = \kappa_x + (\kappa - 1)(\kappa_x + \kappa \bar{x}_{t-1}) = \kappa[\kappa_x - (\kappa - 1)\bar{x}_{t-1}] = \kappa \Delta_x(t)$. Thus, since $\kappa > 1$, $|\Delta_x(t+1)| > |\Delta_x(t)|$. Since t is arbitrary, $|\Delta_x(t)|$ is increasing in t. As such, $\langle \bar{x}_t \rangle$ must converge to 0 or 1. If $\langle \bar{x}_t \rangle$ is increasing, then $\lim_{t\to\infty} \bar{x}_t = 1$. Otherwise, $\lim_{t\to\infty} \bar{x}_t = 0$. Furthermore, $\bar{x}_t$ reaches boundary value 0 or 1 in finite time. We now show that $\hat{\mu}_{t+1} = \infty$ if $\bar{x}_t = 1$ and $\hat{\mu}_{t+1} = -\infty$ if $\bar{x}_t = 0$. $\bar{x}_t$ implies that each Investors nt in t chooses $x_{nt} = 1$. From Equation 2.19, if acting in autarky, $x_{nt} = 1 \Leftrightarrow s_{nt} > c$, where $c = (1-\beta)/\gamma$. For any two possible realizations of μ, μ' and μ'', the autarkic likelihood ratio of μ' relative to μ'' after observing all investors choose $x_{nt} = 1$ is

$$\left(\frac{1 - \Psi\left(\frac{c-\mu'}{\sqrt{\tau_\epsilon^{-1}}}\right)}{1 - \Psi\left(\frac{c-\mu''}{\sqrt{\tau_\epsilon^{-1}}}\right)}\right)^N,$$

where $\Psi(\cdot)$ is the standard-normal cdf. This likelihood ratio converges to 0 in N whenever $\mu' < \mu''$. Thus, for any arbitrarily high value of μ'', any value $\mu' > \mu''$ is infinitely more likely than μ''. Letting $\mu'' \to \infty$ establishes that $\hat{\mu}_{t+1}$ diverges to ∞. An analogous argument shows that $\hat{\mu}_{t+1}$ diverges to $-\infty$ whenever $\bar{x}_t = 0$. □

Proof of Corollary 1.

Proof. Since the aggregate shock is identically distributed for both assets, if μ is known, then Asset 2 dominates Asset 1 if and only if $\mathbb{E}[d_2 \mid \mu] > \mathbb{E}[d_1 \mid \mu] \Leftrightarrow \mu > 0$. There are two cases to consider: $\mu^* > 0$ and $\mu^* < 0$. First, suppose $\mu_1 > \alpha W_0/2\tau_\mu$, so $\mu^* > 0$. Define $\mathcal{M} = (0, \mu^*)$. From Proposition 5, for any realization of $\mu \in \mathcal{M}$, $\langle \hat{\mu}_t \rangle$ diverges to $-\infty$. Hence, whenever $\mu \in \mathcal{M}$, investors believe Asset 1 provides infinitely better payoff than Asset 2 despite Asset

1 being the dominated asset. Now consider the case where $\mu < \alpha W_0/2\tau_\mu$, so $\mu^* < 0$. Define $\mathcal{M} = (\mu^*, 0)$. For any realization of $\mu \in \mathcal{M}$, $\langle \hat{\mu}_t \rangle$ diverges to $+\infty$. Hence investors believe it provides infinitely higher payoff than Asset 1, despite the fact it's dominated by Asset 1. □

Proof of Proposition 6.

Proof. Fix $\rho \in [.5, 1]$ and let $m^* \equiv \arg\max_{m=1,...,M}(a_1(1), ..., a_1(M))$. Given the decision rule, it follows that Generation 2 rationally infers $m^* = \arg\max_m \mathbb{E}[q_m \mid \mathbf{a}_1]$. Since players are risk neutral, $a_2(m^*) = 1$. Generation 3 grows certain of whichever state maximizes $\mathbb{P}_{(\mathbf{q},\rho)}(m^*)$. Fixing ρ, it follows from Proposition 2 that the $\mathbf{q}$ maximizing $\mathbb{P}_{(\mathbf{q},\rho)}(m^*)$ is $\mathbf{q} = \omega^P_{m^*}$ which assigns $q_m = 0$ for all $m \neq m^*$ and $q_{m^*} = 1$. Since the maximizing $\mathbf{q}$ is independent of ρ, we simply maximize $\mathbb{P}_{(\omega^P_{m^*},\rho)}$ with respect to ρ:

$$\mathbb{P}_{(\omega^P_{m^*},\rho)}(m^*) = \sum_{k=0}^{M-1} \frac{1}{k+1}\binom{M-1}{k}\rho^{M-k}(1-\rho)^k + \frac{1}{M}(1-\rho)\rho^{M-1}.$$

Clearly as $\rho \to 1$, all terms in the expression above go to zero aside from the first term of the sum, which converges to 1. Hence, Generation 3 grows confident that $\mathbf{q} = \omega^P_{m^*}$ and that $\rho \to 1$. Given these beliefs, it's optimal for each player in $t = 3$ to choose A_{m^*}. By induction, all future Generations $t > 3$ observe the same action distribution as Generation 3, and thus draw the same inference as Generation 3. That is, for all $t > 2$, $\boldsymbol{\pi}_t \to \delta(\omega^P_{m^*}, 1)$ as $N \to \infty$.

□

Proof of Proposition 7.

Proof. This result follows trivially from Proposition 2. Suppose for the moment that $\mathbf{q}$ represent payoffs rather than expected payoffs. Since this environment meets the assumptions of Proposition 2, society grows confident that $\mathbf{q} = \omega^P_{m^*}$ if $m^* = \arg\max_m q_m$. Since players are risk neutral, the proof of Proposition 2 holds no matter if q_m represents the expected payoff from A_m or the realized payoff from A_m. As such, agents grow certain that m^* pays off for sure (i.e., $z_{m^*} = 1$) and all other assets payoff with probability zero (i.e., $z_m = 0$ for all $m \neq m^*$).

□

Chapter 3

Projection of Private Values in Auctions

3.1 Introduction

Evidence from social psychology and economics suggests that people mispredict others' preferences in a systematic way: people perceive their own taste as more common than it is. For instance, those with a particular taste for art, sports, or wine tend to overestimate how many share that taste (Ross, Greene, and House, 1977). Such misperceptions also pervade domains with greater social importance, like preferences for income redistribution (Cruces, Perez-Truglia, and Tetaz, 2013) and political candidates (Delavande and Manski, 2012). While a large literature provides empirical evidence demonstrating this bias—commonly known as the "false-consensus effect" or "taste projection"—very little research studies its implications.[1] This paper explores how taste projection affects bidding strategies, efficiency, and revenue across a variety of auction environments and formats. We find that projection induces overbidding in private-value auctions and reduces efficiency in auctions with both private and common value—relative to rational bidding, players with optimistic signals about the common value are more likely to win than those with the highest private value.

Why taste projection might affect behavior in auctions is straightforward. Many auction mechanisms induce bidding strategies dependent on a bidder's perceived distribution of others' valuations. Projection implies that players with different private values perceive this distribution differently. Those with high valuations overestimate the likelihood that others have high valuations—they overestimate the extent of competition in private-value auctions—while those with low valuations overestimate the prevalence of low valuations—they underestimate competition. Projection introduces an additional error when the good up for auction has some common-value element. Since bids depend on both private tastes and private signals about the common value, players draw inference about the common value from others' bids. But since a taste projector has wrong beliefs about others' tastes, she systematically draws biased estimates of others' signals when conditioning on their bids. Specifically, we show that a player's biased estimate of the common value of the good is

[1] Section 3.2.2.1 reviews some of this evidence.

inversely related to her private value.

To see this, consider a real-estate auction for a house with a modern architectural design. Bidders may differ in their taste for modern versus traditional architecture—how much one likes the property's design determines her private value. But bidders also care about the future resale value of the home—which is commonly valued—and each has a private signal of this value. To see how projection distorts inference, consider an English auction with three bidders, Frank, Ludwig, and Andrea. Ludwig is fanatical about modern architecture, but Andrea is a traditionalist. Suppose Frank withdraws first, and does so at price p. Ludwig and Andrea glean information about Frank's common-value signal from p. But projection implies they draw different—and incorrect—conclusions from p. Ludwig assumes that Frank most likely has a high private value for the house, and thus attributes his withdrawal to a pessimistic common-value signal. Andrea, however, assumes Frank likely has a low private value, which he thinks explains Frank's withdrawal. Consequently, Andrea develops a more optimistic belief about Frank's common-value signal. Importantly, it is Ludwig—the bidder with the high private value for the house—who develops the most pessimistic belief about the property's common value. As such, Ludwig bids less aggressively than if he were rational. This is the source of the additional inefficiency which projection adds to auctions with private and common value.

Section 2 presents our basic model, which incorporates Gagnon-Bartsch's (2014) model of taste projection into Goeree and Offerman's (2003) analysis of auctions with private and common value. Players have a private value for the good and a noisy signal of the common value. In truth, private values t are independently drawn from distribution G.[2] To model projection, we assume that a player with private valuation t_i wrongly thinks $t \sim \widehat{G}(\cdot \mid t_i)$. Specifically, players with above-average private values perceive a distribution of valuations that's shifted to the right of the true distribution whereas those with below-average private values perceive a distribution shifted to the left.[3] The amount by which Player i's perceived distribution is shifted is increasing in a parameter ρ and in the distance between her value and the mean value. The parameter $\rho \in [0, 1]$ provides a natural measure of the extent of the bias: $\rho = 0$ corresponds with rational exceptions whereas $\rho = 1$ implies each player believes she has the mean valuation.

We close the model by assuming each player is naive about the heterogeneity in perceptions. That is, each Player i believes she is playing a Bayesian game in which all players agree that $t \sim \widehat{G}(\cdot \mid t_i)$. With this assumption, solving the model is relatively straightforward. Player i plays her Bayesian-Nash-equilibrium strategy of the auction where $\widehat{G}(\cdot \mid t_i)$ is commonly known. For instance, if $\beta(t_i; G)$ is the Bayesian Nash bidding strategy in the rational common-prior auction with distribution G, then Player i follows strategy $\beta\big(t_i; \widehat{G}(\cdot \mid t_i)\big)$ in an auction with projection.

[2] For simplicity, we focus on uniformly distributed private values. We discuss throughout how the intuitions behind our results are independent of this uniform assumption.

[3] This implies that a player's perceived distribution first-order stochastically dominates the perceptions of those with lower valuations. This structure allows for a tractable ordering of perceptions.

As a benchmark, Section 3.3 explores the effect of projection when the good has only private value. In both a sealed-bid second-price or English auction, projection has no effect on the symmetric equilibrium in which players simply bid their private value. This follows naturally from the fact that the bidding strategy doesn't depend on beliefs about the taste distribution. Strategies in a first-price auction, however, call for each Player i to bid her estimate of the second-highest valuation conditional on herself having the highest. A player with above-average private value overestimates the share with a valuation above her own. As such, she overestimates the second-highest valuation. She perceives competition as more fierce than it is, and consequently overbids. Conversely, a player with below-average private value underestimates the extent of competition, and thus underbids. Since those who overbid are most likely to win, projection increases expected revenue. Revenue equivalence does not hold: first-price auctions revenue dominate second-price auctions.

Section 3.4 adds a common-value component to the model. In addition to misperceiving the extent of competition, projection distorts bidders' equilibrium inference. As such, we show that projection affects outcomes even in second-price or English auctions where perceptions of the competition only influence bidding strategies through inference—unlike first-price auctions, there is no direct incentive to bid higher when one thinks others likely have high private valuations. Our main result is that, in both second-price and English auctions, efficiency—the probability that the player with the highest value is allocated the good—is decreasing in the extent of projection, ρ.

First, Section 3.4.1 derives biased bidding strategies for second-price auctions with private and common values. Projection biases bids in exactly the opposite way that it does in first-price auctions with private values: those with above-average taste draw overly-pessimistic inference about the common value and underbid whereas those with below-average taste draw overly-optimistic inference and overbid. The rationale is similar to that in the real-estate-auction example above. In equilibrium, Player i bids her expected value of the object conditional on tying with the highest bidder, Player j. Supposing i has high private value, what does she infer about j's signal conditional on a tie? Holding j's bid fixed, since i exaggerates the chance that j has high private value, she overestimates the chance that j has a bad private-value signal. As such, relative to the rational inference, she forms a pessimistic estimate of the common value. The logic is reversed for those with low private values. Since they overestimate how many have low taste for the good, they overestimate the likelihood that the highest bidder has *optimistic* information about the common value. The overall effect is a compression of expected valuations across players—those with low taste think too highly of the common value, while those with high taste think too poorly of the common value. Consequently, relative to rational bidding, it's more likely that somebody with low private value wins the auction. This probability is naturally increasing in ρ, since the larger is ρ, the more distorted are common-value inferences.

Second, Section 3.4.2 derives the biased bidding strategies for an English auction. Projection has very much the same effect here as it does in second-price auctions. However, players additionally draw biased inference from prices at which early bidders withdraw. Again, this inference is increasingly pessimistic the higher is one's private value. For instance, if Player

i has high private value, she thinks most others similarly have high private value. Thus, when Player i observes Player k withdraw at a low price, she overestimates the likelihood that k does so on the basis of pessimistic information about the common value. Like the second-price auction, biased inference decreases allocational efficiency.

This paper adds to a growing literature that incorporates behavioral biases into auction theory. Many of these papers offer explanations for the widely-documented phenomenon that people overbid in a variety of auctions in ways inconsistent with classical theory. Overbidding is most famously associated with common-value auctions, where bidders must avoid falling victim of the "winner's curse". Explanations include cursed thinking a la Eyster and Rabin (2005) and level-k thinking a la Crawford and Iriberri (2007). Eyster and Rabin (2005) assume people neglect the informational content of others' behavior and consequently fail to understand that bids reveal private information in equilibrium.[4]

Although cursed thinking predicts rational bidding in private-value auctions, overbidding is also observed in this environment. There is an experimental literature on independent-private-value auctions that documents a widespread (though not universal) tendency for subjects to bid higher than the risk-neutral Bayesian Nash equilibrium (RNNE) benchmark. Many researchers attribute this finding to cognitive errors. For instance, Goeree, Holt, and Palrefy (2002) provide lab evidence of overbidding and attempt to explain it with Quantal-Response equilibrium. Level-k thinking also predicts overbidding in first-price auctions so long as the distribution of values is not uniform.[5] Compte (2002, 2004) suggests a model in which bidders overestimate the precision of their private-value signals, generating a "winner's curse" within private-value auctions. Finally, some researchers attribute this phenomenon to non-standard preferences. Cox, Smith, and Walker (1992), for instance, argue that bidders display "joy of winning". A weakness of such preference-based theories, however, is that behavior consistent with such preferences is observed in some auction formats, but not all.

We conclude in Section 3.5 by discussing the limitations of our model and how our notion of taste projection relates to other approaches in the literature. We also review avenues for drawing further implications from our model.

3.2 Model

This section presents our model. Section 3.2.1 introduces bidders' preferences and the various auction formats we consider. Section 3.2.2 describes our model of projection: we specify how players form taste-dependent perceptions of the distribution of valuations. We also provide some motivating evidence for taste projection.

[4] Our model proposes an alternative form of "belief neglect". Unlike Eyster and Rabin (2005), players in our model understand that bids depend on private information. But, because they neglect heterogeneity in beliefs about the distribution of values, they fail to understand the map from others' signals to their bids. As a consequence, they draw incorrect inference when conditioning on others' behavior.

[5] In contrast, we show that projection predicts overbidding despite uniformly distributed private values.

3.2.1 Auction Environment

We consider auctions for a good with both private and common value. Our basic setup follows Goeree and Offerman (2003) who characterize the rational equilibria of auctions with private and common values.[6] Let $N = \{1, \ldots, N\}$ denote the set as well as number of players. Each Player i has valuation (net of price) $u(t_i, v) = t_i + v$, where $t_i \in T \equiv [\underline{t}, \overline{t}]$ denotes i's private taste and $v \in V \equiv [\underline{v}, \overline{v}]$ denotes the commonly-valued quality of the good. Private tastes are independently and identically distributed across players according to c.d.f. G. Common value v is unknown; each Player i receives an i.i.d. private signal $\theta_i \in \Theta$ about v's value. Let $\theta_i \sim F$, and take F as common knowledge. The total common value is defined as the sum of all signals: $v \equiv \sum_{i=1}^{N} \theta_i$.[7] For tractability, we assume both F and G are uniform on $[0, 1]$.[8] However, we emphasize throughout how the intuition of our results extends to general distributions. Finally, we denote by p the price dictated by the auction mechanism.

We consider three auction formats in this paper: (1) first-price sealed bid, (2) second-price sealed bid, and (3) ascending English auction. In the sealed-bid formats, each player i simultaneously submits bid $b_i \in \mathbb{R}$, and Player $i^* = \arg\max_{i \in N} b_i$ is allocated the good. Under the first-price mechanism, $p = b_{i^*}$; under the second-price mechanism the price equals the second highest bid, $p = \arg\max_{i \in N \setminus \{i^*\}} b_i$. Denote by $\beta : T \times \Theta \to \mathbb{R}$ a type's bidding strategy. Our model of the English auction follows Milgrom and Weber (1982): the auctioneer continuously raises the price and bidders publicly reveal when they withdraw from the auction; exit decisions are irreversible. We describe the formal strategies when we analyze English auctions in Section 3.4. Fixing the auction format, let $\Gamma(G)$ denote a generic auction in which G is the commonly known distribution of private values. The purpose of this notation is made clear when we introduce taste projection: each Player i misperceives the game as $\Gamma(\widehat{G}(\cdot \mid t_i))$, and, hence, plays her part in a Bayesian Nash equilibrium of $\Gamma(\widehat{G}(\cdot \mid t_i))$.

To motivate our private-and-common-value setting, we provide a few examples. First, suppose firms are competing for a license to operate in a market. The cost structure of a firm—dictated by the firm's private technology—constitutes a private value element. The demand in the market constitutes the common-value component. In the canonical mineral-rights example, a firm's profit depends on the common value of the well and its private idiosyncratic drilling costs. Projection will imply firms overestimate how similar others' costs are to one's own. Alternatively, consider auctions for consumer goods with resale markets, like real estate, art, or wine. Bidders competing for a painting are motivated

[6] For other articles analyzing auctions with private and common signals see Compte and Jehiel (2002), Jehiel and Moldovanu (2001), Pagnozzi (2007) and Pesendorfer and Swinkels (2000).

[7] We model the common-value component of utility with an additive value function as in the "Wallet Game" introduced in Klemperer (1998) and Bulow and Klemperer (2002). Goeree and Offerman (2003) assume that the common value component of a bidder's utility is equal to the average of the bidder's common value signals: $v \equiv \frac{1}{n} \sum_{i=1}^{N} \theta_i$. Our formulation is qualitatively equivalent to theirs but easier to work with under taste projection.

[8] In Section 3.3, we consider an environment with purely private values. That is, $v = 0$, and F is degenerate on zero. When we add common value to the model in Section 3.4, we maintain that $\theta_i \overset{\text{iid}}{\sim} \mathrm{U}[0, 1]$.

by two factors: (1) their idiosyncratic taste for the painting—which determines immediate consumption utility when they hang the painting on their wall, and (2) the common resale value of the painting—which is realized when they ultimately sell the painting. Projection implies that those who get the most pleasure from consuming the product overestimate it's resale value.

3.2.2 Projection: Motivation and Model

This section reviews the literature motivating our main assumption of taste projection and, following Gagnon-Bartsch (2014), provides a simple formulation of this bias. The model consists of two key assumptions: (1) an agent's perceived preference distribution depends on her own taste, and (2) she neglects that others' perceptions depend on their tastes.

3.2.2.1 Evidence of Taste Projection

The notion that people systematically misptredict others' tastes is supported by several strands of research. A large literature in social psychology studies inter-personal projection—the idea that people's own habits, values, and behavioral responses bias their estimates of how common are such habits, values, and actions in the general population. Early work, including Ross, Greene, and House (1977)—who coin the term "false-consensus effect"—find positive correlation between subjects' own preference responses and their estimates of others' responses. Subjects in Ross, Greene, and House (1977) gave their own (binary) response to a question, and predicted the fraction of subjects who answered similarly. (E.g., "Are you politically left of center?"; "Do you prefer basketball over football?"; "Will there be women in the supreme court in the next decade?"; "Do you prefer Italian movies over French?") Out of 34 questions, 32 were consistent with taste projection: those who answer "yes" to a question overestimate how many others will answer "yes" relative to those who answer "no". Many similar studies followed, documenting this correlation across a wide range of domains, including preferences over political candidates and ideology, perceptions of the income distribution and preferences for redistribution, and risk preferences.[9]

Each of these studies, however, simply document correlation between a subject's own taste and her prediction. Is such correlation necessarily indicative of an error? If there is uncertainty about others' tastes, the answer is no. As first noted by Dawes (1989), with uncertainty, a Bayesian should use her own taste as information, resulting in *rational* type-dependent estimates that appear consistent with a "false-consensus" bias.

[9] Marks and Miller (1987) review 45 different studies documenting the false-consensus effect published over the decade following Ross, Greene, and House (1977). Mullen, Atkins, Champion, Edwards, Hardy, Story, and Vanderlok (1985) find robust evidence of this correlation in a meta-study of 115 tests. Evidence of type-dependent misprediction has been found in a variety of domains. For instance, Brown (1982) and Rouhana, O'Dwyer and Vaso (1997) find type-dependent perceptions of political preference. Cruces, et al. (2013) find type-dependent misprediction of the income distribution in Argentina, and demonstrate that this leads to misprediction of population preferences for income redistribution. Faro and Rottenstreich (2006) find correlation between subjects' own risk preference and their perception of others' risk preferences.

Motivated by this critique to demonstrate a systematic error, Krueger and Clement (1994) and others provide evidence that this "bias" remains even when subjects have information about other' preferences. They find that subjects use their own preference information more so than that of anonymous others when making population predictions, inconsistent with Bayesian rationality.[10] In incentivized settings, Engelmann and Strobel (2012) verify that a truly-false-consensus bias remains so long as subjects must exert a small amount of effort to get information on others' choices; when this information is not freely available or made salient, people rely too heavily on their own choice when predicting the choices of others. So long as attending to others' tastes comes at some cost, this result suggests that people can hold incorrect type-dependent beliefs about population preferences even in settings with ample opportunity to observe others—where the "Dawes critique" should have little bearing.[11]

Relatedly, economists have argued *intra*-personal projection bias—exaggerating the degree to which future preferences resemble current preferences—influences behavior.[12] To the extent that preferences of contemporaneous others are similarly difficult to predict, we should expect the logic of intrapersonal projection bias to suggest *inter*personal-projection. An intuition for intrapersonal projection is that we "mentally trade places" with our future selves, and in doing so, project our current preference states. But this exact logic applies when empathizing with another. Indeed, Van Boven and Loewenstein (2003) show that the same transient preference states shown to warp subjects' perceptions of own future preferences also distort predictions of *others'* preferences. Subjects' predictions of whether thirst or hunger would be more bothersome to hypothetical hikers lost without food or water were biased in the direction of subjects' own exercise-induced thirst. More economically relevant, Van Boven, Dunning and Loewenstein (2000, 2003) show that sellers who experience an endowment effect project their high valuation of a good onto the valuations of potential buyers, causing sellers to set inefficiently high prices.

3.2.2.2 General Properties of Projection

Our model of taste projection, which follows Gagnon-Bartsch (2014), assumes a player's private value t influences her perceived distribution of types. In truth, tastes t are i.i.d. across players with according to c.d.f. G. Denote a type-t's perception of $G(\cdot)$ by $\widehat{G}(\cdot \mid t)$.

[10] Krueger and Clement (1994) deduce that when estimating the percent of subjects that endorse some action or preference, subjects use their own response nearly twice as much as the response of an anonymous other. A rational Bayesian should, of course, use these two responses equally.

[11] Using data from the American Life Panel, Delavande and Manski (2012) show that perceptions of others' candidate preferences in the 2008 U.S. presidential election and 2010 congressional election were consistent with the false-consensus effect even after the release of poll results. While this finding may indicate additional statistical biases (e.g., failure to appreciate the Law of Large Numbers—see Benjamin, Raymond, and Rabin, 2013), it shows that taste-dependent perceptions can persist despite opportunity to learn about others' tastes.

[12] For empirical studies see Busse, Pope, Pope, and Silva-Risso (2012), Simonsohn, (2010), and Conlin, O'Donoghue, and Vogelsang (2007). For example, Busse, et al. shows that projection bias affects demand and prices in large, high-stakes markets for cars and houses. Loewenstein, O'Donoghue, and Rabin (2003) provide a general overview of the evidence and draw out implications of a formal theoretical model.

Before proposing a simple specification for $\widehat{G}$, we make 2 general assumptions on players perceptions.

First, consistent with the false-consensus effect, we assume high types think high private valuations, or "tastes", are relatively more common, while low types think the opposite. We capture this intuition by assuming $\widehat{G}(\cdot \mid t)$ dominates $\widehat{G}(\cdot \mid t')$ in the sense of first-order stochastic dominance whenever $t > t'$:

Assumption 12. (Stochastically Dominating Perceptions.) *$\widehat{G}(\cdot \mid t_i)$ first-order stochastically dominates $\widehat{G}(\cdot \mid t_j)$ if and only if $t_i > t_j$. That is, whenever $t_i > t_j$, $\widehat{G}(t \mid t_i) \leq \widehat{G}(t \mid t_j)$ for all $t \in T$, and the inequality is strict for some $t \in T$.*

Fixing any threshold $\bar{t}$, the higher is a player's private value, the higher is her estimate of the share of competitors with value exceeding $\bar{t}$. For example, lovers of Italian wine think 90% of bidders in a wine auction have a private valuation for Italian wine that exceeds \$300. But those who prefer French wine think only 50% of bidders have a private valuation for Italian wine that exceeds \$300.

Second, we assume that a projector is *naive* about her bias: she neglects that those with different tastes have alternative perceptions of the distribution. She thinks *all* agents share a common perception.

Assumption 13. (Naivete.) *For any player j with private valuation t_j, a player with valuation t_i believes $\widehat{G}(\cdot \mid t_j) = \widehat{G}(\cdot \mid t_i)$. That is, player i thinks that all players $\neg i$ have identical perceptions equal to $\widehat{G}(\cdot \mid t_i)$.*

Player i with taste t_i thinks type distribution $\widehat{G}(\cdot \mid t_i)$ is common knowledge to all players. In essence, agents imagine they are playing a game with common priors, when in fact priors are heterogeneous. With this assumption, solving the model is relatively straightforward. Player i plays her Bayesian-Nash-equilibrium strategy of the auction where $\widehat{G}(\cdot \mid t_i)$ is the commonly-known taste distribution. For instance, if $\beta(t_i; G)$ is the Bayesian Nash bidding strategy in the rational common-knowledge sealed-bid auction with distribution G, then Player i follows strategy $\beta\big(t_i; \widehat{G}(\cdot \mid t_i)\big)$ in an auction with projection.[13] It follows that each Player i's strategy is part of a Bayesian Nash equilibrium of game $\Gamma\big(\widehat{G}(\cdot \mid t_i)\big)$. We call the resulting profile of of strategies a "naive equilibrium" of Γ.[14]

[13] Henceforth, we simply write $\beta(t_i)$ to denote t_i's bidding strategy, which implicitly depends on her perception $\widehat{G}(\cdot \mid t_i)$.

[14] Note that a "naive equilibrium" is *not* a true equilibrium since players' beliefs are systematically biased away from observed outcomes. Instead, each player's strategy is a best response to her incorrect belief about others' actions. "Naive equilibrium" is a very weak concept as it only imposes that each player plays a Bayesian Nash equilibrium strategy of her perceived game. In this setting, however, we impose further structure by focusing on symmetric equilibria. We solve for the symmetric monotone Bayesian Nash equilibrium as if all players were fully rational, then consider the naive equilibrium in which all players think they play their part in this particular symmetric monotone equilibrium. Essentially, players agree on which equilibrium they are in, but miscalculate their optimal strategy within this equilibrium.

Aside from these errors in perceptions about G, players are Bayes rational. They draw inference using Bayes' Rule given their incorrect model of the game, and maximize expected payoffs given these inferred beliefs.[15]

3.2.2.3 Parametric Model

For the purpose of this paper, we propose a simple parametric specification for the family of perceived distributions $\{\widehat{G}(\cdot \mid t)\}_{t\in T}$. We assume that each type perceives a distribution with the same shape as G, but shifted to the right or left by an extent proportional to her private value. Those with low private value perceive a distribution shifted to the left; those with high private value perceive one shifted to the right. Our specification captures the notion that people overestimate how "representative" their valuation is. The perceived support is shifted such that a player's valuation is closer to the "center" of the support. To proceed, we first define the shifted support and then define the perceived distribution as a "re-normalization" of the true distribution over this new support.

To derive the perceived support, it is useful to think of the true support $T = [\underline{t}, \bar{t}]$ as an interval of about some central statistic μ:

$$T = \left[\mu - \underline{\Delta}, \mu + \overline{\Delta}\right] \tag{3.1}$$

where $\underline{\Delta} \equiv \mu - \underline{t}$ and $\overline{\Delta} = \bar{t} - \mu$. Bidders project with respect to this central statistic μ: each player wrongly thinks her private value is closer to statistic μ of G than it really is. Our model is agnostic about μ—it could be the mean, mode, or some other salient statistic. Player i with taste t_i thinks μ has value $\hat{\mu}(t_i)$, which we define as the convex combination of her taste and the true value of μ:

$$\hat{\mu}(t_i) = \rho t_i + (1-\rho)\mu. \tag{3.2}$$

We call parameter $\rho \in [0,1]$ the *extent* of projection.[16] $\rho = 0$ implies correct perceptions, while $\rho = 1$ implies each type thinks her valuation lies exactly at the μ statistic. Intermediate ρ implies for all $t_i \neq \mu$, $|\hat{\mu}(t_i) - t_i| < |\mu - t_i|$: people perceive their value as more central than it is. Only the type with valuation exactly equal to μ has a correct perception of μ.

[15] Naivete is the key assumption that differentiates "taste projection" from a model with rational taste-dependent distributional beliefs. Rational agents know precisely the map between an agent's type and her belief about the distribution. Further, naivete departs from much of the literature on non-common priors, which assumes individuals have rational expectations about the *distribution* of heterogeneous beliefs across players. See, for example, Harrison and Kreps (1979) or Morris (1996). Here, however, players assume the distribution of beliefs (about G) is degenerate on their own perception. As such, within the particular domain of auctions, this paper provides a first step in analyzing the implications of neglecting heterogeneity in beliefs. The importance of this line of research has been previously emphasized by Nisbett and Ross (1980): "The real source of difficulty does not lie in the fact that human beings subjectively define the situations they face, nor even in the fact that they do so in variable and unpredictable ways. Rather, the problem lies in their failure to recognize and make adequate inferential allowance for this variability and unpredictability."

[16] For simplicity, we assume each player has the same extent of projection.

A player with taste t_i thinks the support is $\widehat{T}(t_i)$. The misperceived support is

$$\widehat{T}(t_i) = \big[\hat{\mu}(t_i) - \underline{\Delta}, \hat{\mu}(t_i) + \overline{\Delta}\big]. \tag{3.3}$$

Player i wrongly thinks the support is "centered" about the taste-dependent estimate of $\hat{\mu}(t_i)$.[17]

We now describe a player's perceived distribution. We assume each perceived distribution $\widehat{G}(\cdot \mid t_i)$ has precisely the same shape as the true distribution G, but is shifted to fit over $\widehat{T}(t_i)$. If t denotes the true random variable from which private values are drawn, a player with valuation t_i thinks types are drawn from random variable $\hat{t}(t \mid t_i) = t + [\hat{\mu}(t_i) - \mu]$; $\hat{t}(t_i)$ is a simple linear shift of the true random variable. It follows that the c.d.f. of $\hat{t}(t_i)$ is

$$\widehat{G}(\tau \mid t_i) = G\Big(\hat{t}^{-1}(\tau \mid t_i)\Big) = G\Big(\tau - \big[\hat{\mu}(t_i) - \mu\big]\Big).^{18} \tag{3.4}$$

Equation 3.4 pins down our family of perceived distributions in terms of the true distribution and projection parameter ρ. Let $\widehat{\mathbb{E}}_i[\cdot]$ denote the expectations operator for agent i with valuation t_i, which is with respect to $\widehat{G}(\cdot \mid t_i)$.

How should we interpret $\widehat{G}(\cdot \mid t_i)$? While it has precisely the same shape as G, it's worth clarifying what this means. First, players correctly perceive the percentile of $\hat{\mu}(t_i)$. It follows directly from Equation 3.4 that for any t_i, $\widehat{G}(\hat{\mu}(t_i) \mid t_i) = G(\mu)$. For instance, if μ is the median so $G(\mu) = 0.5$, then $\hat{\mu}(t_i)$ is the median of $\widehat{G}$: $\widehat{G}(\hat{\mu}(t_i) \mid t_i) = 0.5$. Players systematically mispredict the percentile of other types in a similar way. Consider Player i's perception of opponent τ's percentile. Player i thinks that τ's position in her perceived support is such that fraction y of the perceived support falls below τ:

$$y = \frac{\tau - \underline{t}(t_i)}{\overline{t} - \underline{t}}$$

Our model says that i's perception of τ's percentile equals the *true percentile* of the type τ who is such that faction y of the *true support* falls below τ'—τ' solves

$$y = \frac{\tau' - \underline{t}}{\overline{t} - \underline{t}}.$$

[17] To be clear, our model of projection relies on 2 parameters, μ and ρ. Projection is with respect to some statistic of the private-value distribution, denoted μ. ρ measures the extent to which players underestimate the distance between their own value and statistic μ. μ is specified by the game, while ρ is an exogenous "behavioral" parameter.

[18] Denote by $\hat{g}(\cdot \mid t_i)$ Player i's perceived density, which we obtain by differentiating 3.4:

$$\hat{g}(t \mid t_i) = \frac{1}{\sigma} g\left(\frac{1}{\sigma}(t - \hat{\mu}(t_i)) + \mu\right).$$

In other words, if $\tau \in \widehat{T}(t_i)$ and $\tau' \in T$ have identical proportional positions within T and $\widehat{T}$, $\tau - \underline{t}(t_i) = \tau' - \underline{t}$, then i perceives τ at the same percentile as τ''s true percentile.[19]

More generally, the model implies that players misperceive their own percentile in the taste distribution. All players underestimate the distance (in terms of percentile) between their own valuation and $\hat{\mu}(t_i)$.

Lemma 1. *Let $\hat{\mu}(t_i) = \rho t_i + (1-\rho)\mu$. For all t_i, $\left|\widehat{G}\big(t_i \mid t_i\big) - \widehat{G}\big(\mu(t_i) \mid t_i\big)\right| < |G(t_i) - G(\mu)|$.*

Lemma 1 implies that those below the $G(\mu)$-percentile overestimate their percentile—those with low private value overestimate how many have valuations lower than themselves. And those above the $G(\mu)$-percentile underestimate their percentile—those with high private value overestimate how many have valuations higher than themselves. Players with extreme values fail to appreciate how much their taste differs from the general population. For instance, suppose μ is the median of G so $G(\mu) = 0.5$. An agent at the 70^{th} percentile estimates that she's at the π^{th} percentile for some $\pi \in (50, 70)$. Likewise, an agent at the 10^{th} percentile estimates that she's at the $\tilde{\pi}^{th}$ percentile for some $\tilde{\pi} \in (10, 50)$.

For a graphical example of our model, suppose in truth $T = [0, 1]$, $t \sim \text{Beta}(3, 2)$, and μ is the mean of t ($\mu = 0.6$). The plot below shows the true density (thick black curve) and the perceived densities for types $t_i \in \{0, .2, .4, .6, .8, 1\}$ for $\rho = 0.5$.

While we provide a general model of projection, in our auction environments we focus on the case where $t \sim \text{U}[0, 1]$. With uniformly distributed values, we believe that $\mu = 1/2$ is the most the natural candidate for the statistic about which players project. As such, we assume $\mu = 1/2$ unless explicitly assume otherwise. In this case, our model implies that Player i thinks $t \sim \text{U}[\underline{t}(t_i), \bar{t}(t_i)]$ where

$$\begin{aligned} \underline{t}(t_i) &= \rho t_i + (1-\rho)/2 - 1/2 \\ \bar{t}(t_i) &= \rho t_i + (1-\rho)/2 + 1/2. \end{aligned} \tag{3.5}$$

Player i thinks the mean valuation is $\widehat{\mathbb{E}}_i[t] = \rho t_i + (1-\rho)/2$.

[19] Since our model assumes that agents misperceive the support of valuations, it is conceivable that agents observe behavior they thought was impossible. To rule this out, we can slightly augment our definition of the perceived distribution so that the true support T is a subset of the perceived support for all t. Let $\widehat{G}(\cdot \mid t_i)$ and $\widehat{T}(t_i)$ be defined as above, and let the perceived support be $\widetilde{T} \equiv \widehat{T}(t_i) \cup T$, and define the perceived cdf as $\widetilde{G}(t \mid t_i) = (1-\epsilon)\widehat{G}(t \mid t_i) + \epsilon\frac{t-\underline{t}}{\bar{t}-\underline{t}}$ for some $\epsilon > 0$. That is, the agent puts weight $1-\epsilon$ on the perceived distribution defined in Equation 3.4, and weight ϵ on a uniform distribution over the true support. In the limit as $\epsilon \to 0$, all of our analysis using $\widehat{G}$ rather than $\widetilde{G}$ consists of nearly-exact approximations, and agents never observe behavior they thought was impossible. This is, of course, not a realistic model of perceptions, but it eliminates any possibility of observing supposedly-impossible behavior. Essentially, anything that can happen in reality can happen in the false model, but things that the agent thinks can happen in the false model never happen in reality. An agent never observes anything that contradicts the model.

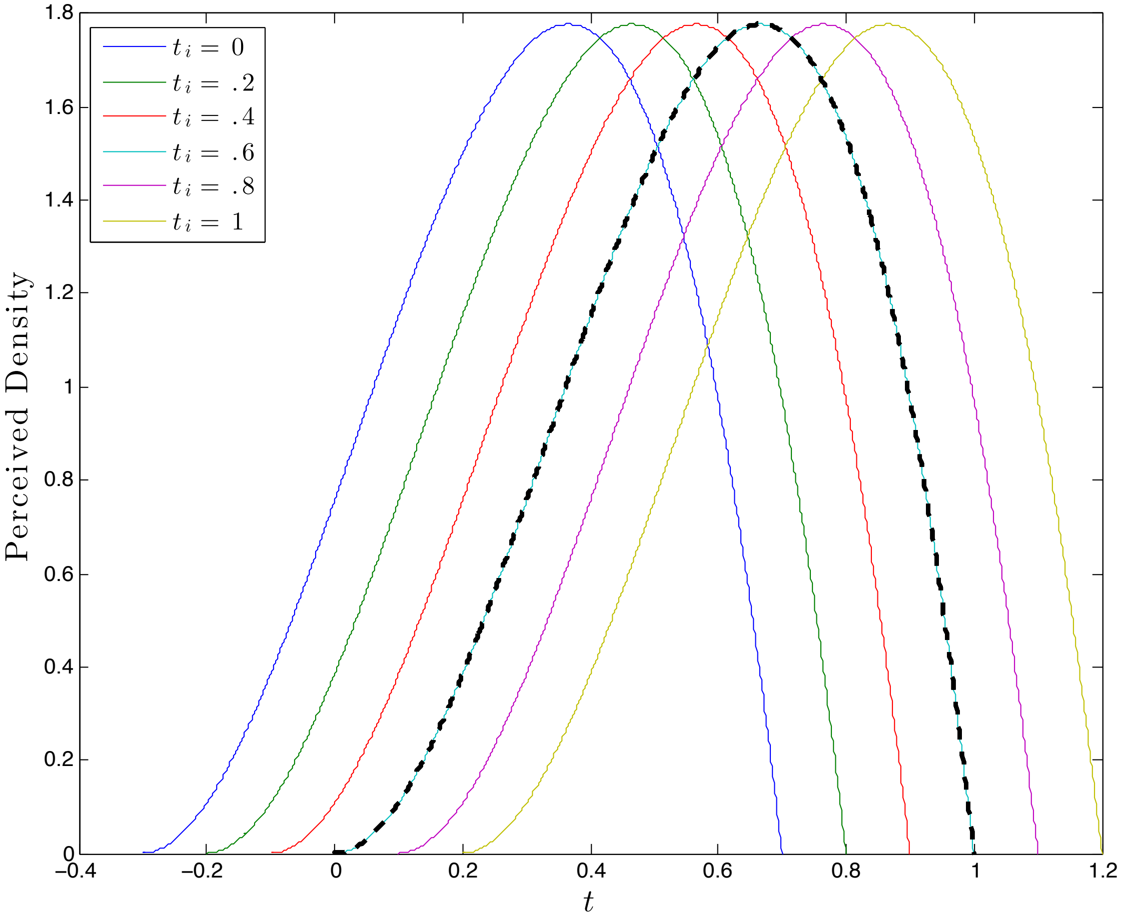

Figure 3.1: *The true density (dashed black curve) and perceived densities for* $t_i = \{0, .2, .4, .6, .8, 1\}$. *when* $t \sim \text{Beta}(3, 2)$*,* $\rho = .5$*.*

3.3 Private Values: Overbidding

In this section, we study how taste projection affects bidding and revenue in auctions with independent private values. Since there is no common-value component, players infer nothing about the value of the good from others' bids. Despite this, projection distorts a players' perception of the competition. We first consider second-price and English auctions, then turn to first-price auctions.

3.3.1 Second-Price and English Auctions

We begin with the most basic scenario: a second-price sealed-bid auction with independent private values. With rational bidders, the symmetric Bayesian equilibrium bidding strategy is simply $\beta(t_i) = t_i$—a player bids her valuation of the good. No matter the extent of the bias, this is still an equilibrium under projection. The English auction has an equivalent equilibrium: each player remains in the auction until the price exceeds her private value.

Hence, if we restrict attention to these symmetric strategies, we conclude that projection has no effect on bidding:

Proposition 1. *In both a second-price or English auction with independent private values, the strategy profile where all players bid their private value is a naive equilibrium no matter the extent of projection. That is, $\beta(t_i)$ is constant in ρ.*

It follows that expected revenue is identical with biased or rational bidders. Further, the auction is efficient—the player with the highest private value is always allocated the good. Note that these results are entirely independent of any assumptions on the parametric form of projection or the distribution of private values. It follows entirely from the fact that a player's strategy does not depend on her beliefs about others' private values. As we show next, this is not the case in a first-price auction.

3.3.2 First-Price Auction

Suppose players engage in a first-price sealed-bid auction with independent private values. The symmetric Bayesian Nash equilibrium calls for Player i to bid her estimate of the second-highest valuation conditional on her having the highest valuation: $\beta(t_i) = \widehat{\mathbb{E}}_i[t_k = \max_{j \in N \setminus \{i\}} t_j \mid t_i = \max_{j \in N} t_j]$. Clearly, this bid depends on Player i's perception of others' valuations, and thus projection distorts her bid.

To solve for biased bidding strategies, some notation is in order. Let $\tau_1(t_i)$ be the first-order statistic of tastes among all bidders other than i. That is, $\tau_1(t_i) = \max_{j \in N \setminus \{i\}} t_j$. Let $\widehat{G}_1(\cdot \mid t_i)$ and $\hat{g}_1(\cdot \mid t_i)$ denote the c.d.f. and p.d.f. of $\tau_1(t_i)$, respectively.[20] Bidder i solves

$$\max_{b_i} \widehat{G}_1\left(\beta^{-1}(b_i) \mid t_i\right) \times (t_i - b_i), \tag{3.6}$$

yielding first-order condition

$$\frac{\hat{g}_1\left(\beta^{-1}(b_i) \mid t_i\right)}{\beta'\left(\beta^{-1}(b_i)\right)}(t_i - b_i) - \widehat{G}_1\left(\beta^{-1}(b_i) \mid t_i\right) = 0.$$

In a symmetric equilibrium $b_i = \beta(t_i)$, yielding the following differential equation:

$$\frac{d}{dt_i}\left(\widehat{G}_1(t_i \mid t_i)\,\beta(t_i)\right) = t_i \hat{g}_1(t_i \mid t_i).$$

To solve the above differential equation, we need an initial condition. In the rational model, this condition is $\beta(\underline{t}) = \underline{t}$, since the lowest type knows she will never win. With projection,

[20] Given our assumption of independent private values, $\widehat{G}_1(t \mid t_1) = \widehat{G}(t \mid t_i)^{N-1}$.

type t_i assumes an analogous condition: she thinks that type $\underline{t}(t_i) = \rho t_i + (1-\rho)/2 - \sigma/2$ bids $\beta\left(\underline{t}(t_i)\right) = \underline{t}(t_i)$. Hence:

$$\begin{aligned} \beta\left(t_i\right) &= \frac{1}{\widehat{G}_1\left(t_i \mid t_i\right)} \int_{\underline{t}(t_i)}^{t_i} y \hat{g}_1\left(y \mid t_i\right) dy \\ &= \widehat{\mathbb{E}}_i\left[\tau_1(t_i) \mid \tau_1\left(t_i\right) < t_i\right]. \end{aligned}$$

Since Bidder i has perception that t is uniform on $[\underline{t}(t_i), \bar{t}(t_i)]$, $\widehat{\mathbb{E}}_i\left[\tau_1(t_i) \mid \tau_1\left(t_i\right) < t_i\right] = \underline{t}(t_i) + \frac{N-1}{N}\left(t_i - \underline{t}(t_i)\right)$. From the definition of $\underline{t}(t_i)$ in Equation 3.5, the naive bidding strategy is

$$\beta(t_i) = \left(\frac{N-1}{N}\right) t_i + \frac{\rho}{N}\left(t_i - \frac{1}{2}\right). \tag{3.7}$$

In contrast, the rational bidding strategy $\beta^R(t)$ conditions on the fact that $t \sim [0,1]$. Setting $\rho = 0$ in Equation 3.7 yields

$$\beta^R(t_i) = \left(\frac{N-1}{N}\right) t_i. \tag{3.8}$$

The next proposition shows that the naive bidding strategy leads players with high private valuations to overbid while others underbid.

Proposition 2. *Consider a first-price auction with independent private values.*

1. *The naive bidding function is $\beta(t_i) = \left(\frac{N-1}{N}\right) t_i + \frac{\rho}{N}\left(t_i - \frac{1}{2}\right)$.*
2. *The naive bid is larger than the rational bid if and only if $t_i > \mu = \frac{1}{2}$. All players with above-average taste overbid. Players with below-average taste underbid.*

The intuition for Part 2 of Proposition 2 is as follows. Each bidder i attempts to guess the value of the second-highest valuation (conditional on i having the highest) and slightly outbid her. A player with high private value ($t_i > \mu$) uses a distribution shifted to the right: her estimate of the second-highest value is necessarily too high. She perceives competition as more fierce than it is, and consequently overbids. Conversely, a player with low private value ($t_i < \mu$) overestimates the share of bidders with valuation below her own. She underestimates the extent of competition, and, thus, underbids. Since the bidder with the highest valuation wins the auction, projection has no effect on efficiency. Further, since high value players set the price in the auction, projection increases expected revenue so long as types who overbid are sufficiently common. We provide this result in the following section.

3.3.3 Revenue Comparisons

We now assess how projection affects expected revenue across auction formats and how revenue changes with the extent of projection. In a first-price auction, revenue is increasing in the extent of projection, ρ. In second-price or English auctions, ρ has no effect on revenue.

Proposition 3. *Consider tn auction with independent private values, and suppose bidders play a symmetric equilibrium.*

1. *In a first-price auction, expected revenue is higher with projection than with rational bidders. Furthermore, expected revenue is increasing in ρ.*
2. *In a second-price or English auction, expected revenue with projection is identical to expected revenue with rational bidders.*

The intuition for Part 1 of Proposition 3 follows from the fact that the player with the highest value sets the price. Since the distribution of the highest value among N places most density over high values of t, the winner is typically of the type who overbids. That is, $t^* = \max_{i \in N} t_i$ exceeds $1/2$ frequently enough to off-set any revenue-reducing effects of those who underbid.[21] Part 2 of Proposition 3 follows immediately from Proposition 1: the standard second-price and English equilibria are unchanged by projection, thus revenue is constant in ρ. Taken together, Parts 1 and 2 of Proposition 3 imply that revenue equivalence doesn't hold across first-price and second-price auctions.

Corollary 2. *Revenue equivalence does not hold with projection: if $\rho > 0$, then a first-price auction revenue dominates a second-price or English auction.*

Other models of bidding behavior similarly predict a failure of revenue equivalence, with risk aversion and loser regret (see Filiz-Ozbay and Ozbay, 2007) being among the most widely-cited explanations. However, both risk aversion and loser regret predict that all bidders, irrespective of their taste, bid above the RNNE benchmark. Taste projection, instead, predicts a "bifurcation": bidders with above-average valuations overbid compared to the RNNE benchmark whereas bidders with below-average valuations underbid. Hence, analyzing the full spectrum of bidding data could differentiate these two theories.

3.4 Private and Common Values: Inefficiency

We now analyze the effect of projection in auctions when the good also has some common-value component. Intuitively, others' bids now provide an agent with additional information about her own value of the object. Hence, in equilibrium, she must condition on these

[21] In the model we analyze here, this is always the case. But it needn't be true in general. Recall that our model of projection has players misperceive the value of statistic μ. Those with $t_i > \mu$ overestimate μ, while $t_i < \mu$ underestimate μ. The value μ is the turning point at which all $t_i > \mu$ necessarily overbid. As μ increases, the fraction of over-bidders decreases, making it less likely that such a type is present in the auction. That is, the event in which the highest type is an under-bidder becomes more likely. In general, there is a cutoff value $\bar{\mu}$ such that $\mu < \bar{\mu}$ implies revenue increases with projection. For instance, when perceptions are uniformly distributed, expected revenue increases with projection so long as $\mu < 1 - \frac{1}{N}$—the expected value of the first-order statistic of t. As argued above, we believe that $\mu = \frac{1}{2}$ is most consistent with taste projection when $t \sim \mathrm{U}[0, 1]$.

bids. However, projection causes her to incorrectly assess the likely motivation for a bid. Those with high tastes over-attribute a competitor's bid to taste, while those with low tastes over-attribute a competitor's bid to the common-value signal. Now, in addition to the "competition effect" highlighted above, projection distorts bids by biasing players' estimates of the common-value component. We make clear both the intuition and effect of biased inference within the context of each particular auction format and show that it always reduces efficiency.

3.4.1 Second-Price Auctions

We first consider a second-price auction, where the role of misinfernce in equilibrium bidding strategies is most straightforward. While we solve for the bidding function for a general N number of bidders below, we begin by assuming $N = 2$ in order to build intuition.

Suppose $N = 2$. Following Georee and Offerman (2003), the symmetric Bayesian Nash equilibrium calls for Player i to bid her expected value of the object conditional on tying with her opponent. That is:

$$\beta(t_i, \theta_i) = t_i + \theta_i + \widehat{\mathbb{E}}_i[\theta_j \mid t_j + \theta_j = t_i + \theta_i]. \tag{3.9}$$

As we formally show in the next lemma, projection distorts a player's equilibrium inference about her opponent's signal in a systematic way.

Lemma 2. *The difference between Player i's inference and the rational inference,* $\widehat{\mathbb{E}}_i[\theta_j \mid t_j + \theta_j = t_i + \theta_i] - \mathbb{E}[\theta_j \mid t_j + \theta_j = t_i + \theta_i]$*, is decreasing in Player i's private value,* t_i*.*

Lemma 2 shows that, relative to rational inference, the higher is one's private value, the more pessimistic is her inference about the common value of the good. The intuition is as follows. A Player i with high private value overestimates the average taste of her opponent; thus, conditional on a tie, Player i must *under*estimate her opponent's private signal. That is, holding fixed the opponent's surplus $t_j + \theta_j$, the higher is j's expected taste, the lower must be j's expected signal.[22] Since the perceived expected value of t_j is increasing in t_i, the higher is a player's taste, the more pessimistic an inference about θ_j she draws.

Specifically, if i thinks $t \sim U[\rho t_i + (1-\rho)/2 - 1/2, \rho t_i + (1-\rho)/2 + 1/2]$, then

$$\begin{aligned}
\widehat{\mathbb{E}}_i[\theta_j \mid t_j + \theta_j = t_i + \theta_i] &= \frac{1}{2}\Big(\widehat{\mathbb{E}}_i[t_j + \theta_j \mid t_j + \theta_j = t_i + \theta_i] - \underline{t}(t_i)\Big) \\
&= \frac{1}{2}\Big(t_i + \theta_i - (1-\rho)t_i - (1-\rho)/2 + 1/2\Big) \\
&= \frac{1}{2}\Big(\theta_i + (1-\rho)t_i + \rho/2\Big).
\end{aligned} \tag{3.10}$$

[22] Following Goeree and Offerman (2003), we refer to the sum of a player's private value and her signal as her private "surplus".

Combining this expression with the bidding function in Equation 3.9 yields

$$\beta(t_i,\theta_i) = \frac{1}{2}\bigg((3-\rho)t_i + 3\theta_i + \rho/2\bigg).$$

A rational player, however, bids

$$\beta^R(t_i,\theta_i) = \frac{3}{2}(t_i+\theta_i).$$

Upon comparing the biased and rational bidding strategies, it's straightforward that projection reduces efficiency relative to rational bidding. Since $\beta^R(t_i,\theta_i) = \frac{3}{2}(t_i+\theta_i)$, even a rational auction may be inefficient—the player with the highest t_i loses—when a player with a low taste has a very high common-value signal. Note that rationality implies that a bidder equally weights her taste and signal. With projection, the winner is he who has the highest value of $(3-\rho)t_i + 3\theta_i$: players put relatively more weight on their signal than on their taste. As such, inefficient outcomes occur more frequently relative to the case where bidders are rational.

The intuition is as follows. From Lemma 2, the naive equilibrium inference biases high private-value bidders' estimates of the common value downward, while it biases low private-value bidders' estimates upward. Thus, the difference in the expected *total* value of the good between high private-value bidders and low private-value bidders is reduced. Since low and high types now have more similar valuations, it's more likely that a low type is misallocated the good.

We now show that this same intuition holds with more than two bidders. To extend the analysis to N bidders, we need a new piece of notation. Let $(t+\theta)_{(k)}$ be the k^{th} highest value of $t+\theta$ among the N bidders. Let $\theta_{(k)}$ be the value of θ in $(t+\theta)_{(k)}$ Following Goeree and Offerman (2003), Player i bids the expected value of the item given that she wins but ties with the player owning the second-highest surplus. That is:

$$\beta(t_i,\theta_i) = t_i + \theta_i + \sum_{j=2}^{N} \widehat{\mathbb{E}}_i\left[\theta_{(j)} \mid (\theta+t)_{(2)} = \theta_i + t_i\right]. \tag{3.11}$$

To evaluate this expression, first note that with perception $t \sim \mathrm{U}[\underline{t}(t_i), \overline{t}(t_i)]$,

$$\widehat{\mathbb{E}}_i\big[\theta_{(j)} \mid (\theta+t)_{(2)} = \theta_i + t_i\big] = \frac{1}{2}\bigg(\widehat{\mathbb{E}}_i\big[(\theta+t)_{(j)} \mid (\theta+t)_{(2)} = \theta_i+t_i\big] - \underline{t}(t_i)\bigg). \tag{3.12}$$

Thus,

$$\begin{aligned}\sum_{j=2}^{N} \widehat{\mathbb{E}}_i\left[\theta_{(j)} \mid (\theta+t)_{(2)} = \theta_i + t_i\right] &= \frac{1}{2}\sum_{j=2}^{N}\bigg(\widehat{\mathbb{E}}_i\big[(\theta+t)_{(j)} \mid (\theta+t)_{(2)} = \theta_i+t_i\big] - \underline{t}(t_i)\bigg) \\ &= \frac{1}{2}\bigg(\widehat{\mathbb{E}}_i\big[\theta_j + t_j \mid \theta_j + t_j = \theta_i + t_i\big] + \\ &\qquad (N-2)\widehat{\mathbb{E}}_i\big[\theta_j + t_j \mid \theta_j + t_j < \theta_i + t_i\big] - (N-1)\underline{t}(t_i)\bigg). \end{aligned} \tag{3.13}$$

The second equality above follows from the fact that the expectation of the sum of all order statistics of N i.i.d. draws is simply N times the expected value of a single draw. Finally, since $\widehat{\mathbb{E}}_i\big[\theta_j + t_j \;\big|\; \theta_j + t_j = \theta_i + t_i\big] = t_i + \theta_i$, plugging Equation 3.13 into 3.11 yields the following bidding strategy:

$$\beta(t_i, \theta_i) = \frac{3}{2}(t_i + \theta_i) + \frac{N-2}{2}\widehat{\mathbb{E}}_i\big[\theta_j + t_j \;\big|\; \theta_j + t_j < \theta_i + t_i\big] - (N-1)\rho\big(t_i - 1/2\big). \tag{3.14}$$

Proposition 4. *Consider a second-price auction with $N \geq 2$ bidders. With projection, the bidding strategy is*

$$\beta(t_i, \theta_i) = \begin{cases} \beta^2_L(t_i, \theta_i) & \text{if} \quad \rho(t_i - 1/2) < t_i + \theta_i \leq \rho(t_i - 1/2) + 1 \\ \beta^2_H(t_i, \theta_i) & \text{if} \quad \rho(t_i - 1/2) + 1 < t_i + \theta_i \leq \rho(t_i - 1/2) + 2 \end{cases} \tag{3.15}$$

where

$$\beta^2_L(t_i, \theta_i) = \frac{2N+5}{6}\left(\theta_i + t_i\right) - \frac{\rho\left(2N-1\right)}{6}\left(t_i - 1/2\right), \tag{3.16}$$

$$\beta^2_H(t_i, \theta_i) = \frac{3}{2}\left(\theta_i + t_i\right) + \frac{N-2}{2}\left(\frac{A_2(t_i, \theta_i)}{B_2(t_i, \theta_i)}\right) - \frac{\rho}{2}\left(N-1\right)\left(t_i - 1/2\right), \tag{3.17}$$

$$\begin{aligned} A_2(t_i, \theta_i) = {} & 16\left(\theta_i + t_i\right)^3 - 12\left(\theta_i + t_i\right)^2\left(-\rho + 2\rho t_i + 4\right) \\ & + \rho\left(2t_i - 1\right)\left(-12\rho + \rho^2 + 4\rho^2 t_i^2 + 24\rho t_i - 4\rho^2 t_i + 24\right) + 16, \end{aligned} \tag{3.18}$$

and

$$\begin{aligned} B_2(t_i, \theta_i) = {} & 24\left(\theta_i + t_i\right)^2 - 24\left(\theta_i + t_i\right)\left(-\rho + 2\rho t_i + 4\right) \\ & + 6\left(-8\rho + \rho^2 + 4\rho^2 t_i^2 + 16\rho t_i - 4\rho^2 t_i + 8\right). \end{aligned} \tag{3.19}$$

The logic of Lemma 2 implies that those with high tastes underbid, and those with low tastes overbid. While this is not immediately clear from the bidding function in Proposition 4, Figure 3.1 clearly reveals this bias in bidding strategies.

The following proposition formalizes our intuition about efficiency. Efficiency is always lower with projection and decreases in the extent of projection.

Proposition 5. *Consider a second-price sealed-bid auction with N bidders for a good with both common and private value. The probability that the auction is efficient is decreasing in the extent of projection, ρ.*

The logic is similar to the case discussed above with $N = 2$. Those with high private values underestimate the common value and those with low private values overestimate it. As such, relative to rational inference, the total perceived value of the good varies by less than it should across players with different tastes. As a consequence, it's more likely for Player i with a low private value to outbid Player i^* with the highest private value when i has a high signal. From Lemma 2, i^*'s perception of the common value is biased downward by more than any other player's. Increasing ρ only further decreases i^* perceptions relative to the rational inference, which further decreases efficiency. Figure 3.2 shows how efficiency changes with ρ.

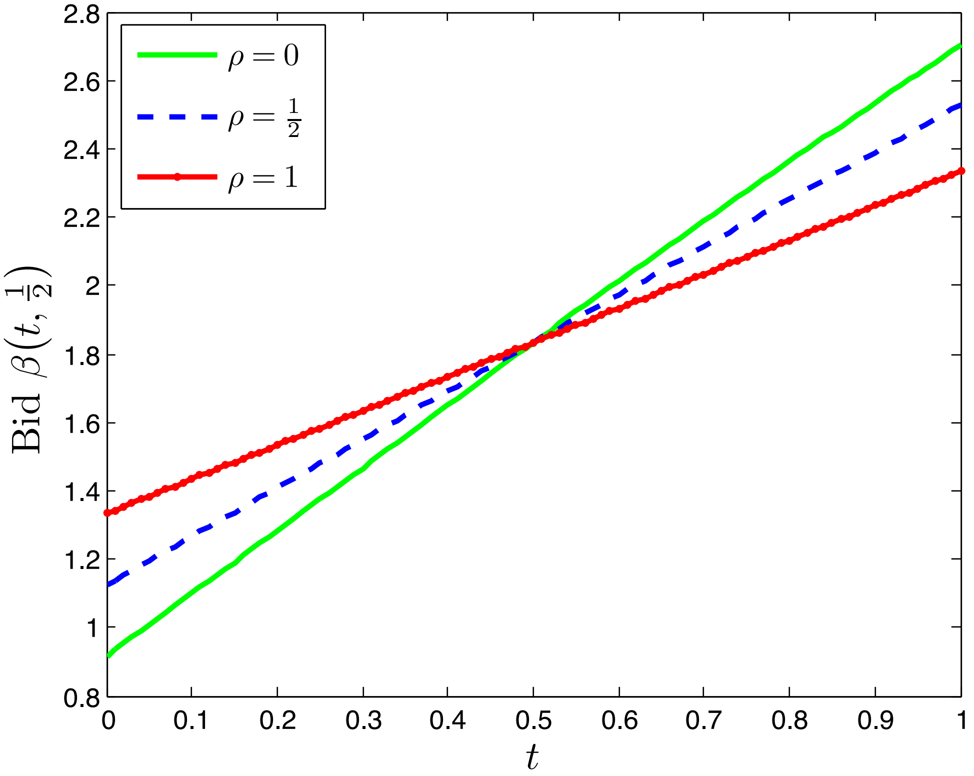

Figure 3.1: *The second-price bidding function as a function of* t *with* $\theta = 1/2$ *for various values of* ρ*.*

3.4.2 English Auctions

We now consider an ascending-price English auction. Biased bidders suffer a similar misinference in the English auction as they do in the second-price auction. When inferring from the prices at which other bidders have exited, a player still in the auction misinterprets the signals of those already withdrawn. A high-taste player forms overly pessimistic beliefs about a withdrawn bidder's signal, while a low-taste player forms overly optimistic beliefs.

For simplicity, we focus on the case with $N = 3$ bidders. Each Player i's strategy consists of two dropout prices, p_i^1 and p_i^2. First, $p_i^1 = p^1(t_i, \theta_i)$ is the price at which i exits the auction conditional on no other player dropping out prior to price p_i^1. The price at which the first player drops out of the auction is thus $\bar{p}^1 = \min_i p_i^1$. Second, $p_i^2 = p^2(t_i, \theta_i, \bar{p}^1)$ is the price at which i drops out if one player has already dropped out at $\bar{p}^1$. p_i^2 naturally depends on the previous dropout price $\bar{p}^1$, which reveals information about the private signal of the player who drops out first.

Following Goeree and Offerman (2003), p_i^1 is the price at which i is indifferent between winning at p_i^1 and dropping out, conditional on no player previously dropping out. Since i wins at p_i^1 only if all players dropout at p_i^1, $p_i^1 = p^1(t_i, \theta_i) = \widehat{\mathbb{E}}_i[v + t_i \mid \theta_j + t_j = \theta_i + t_i \; \forall \; j] =$

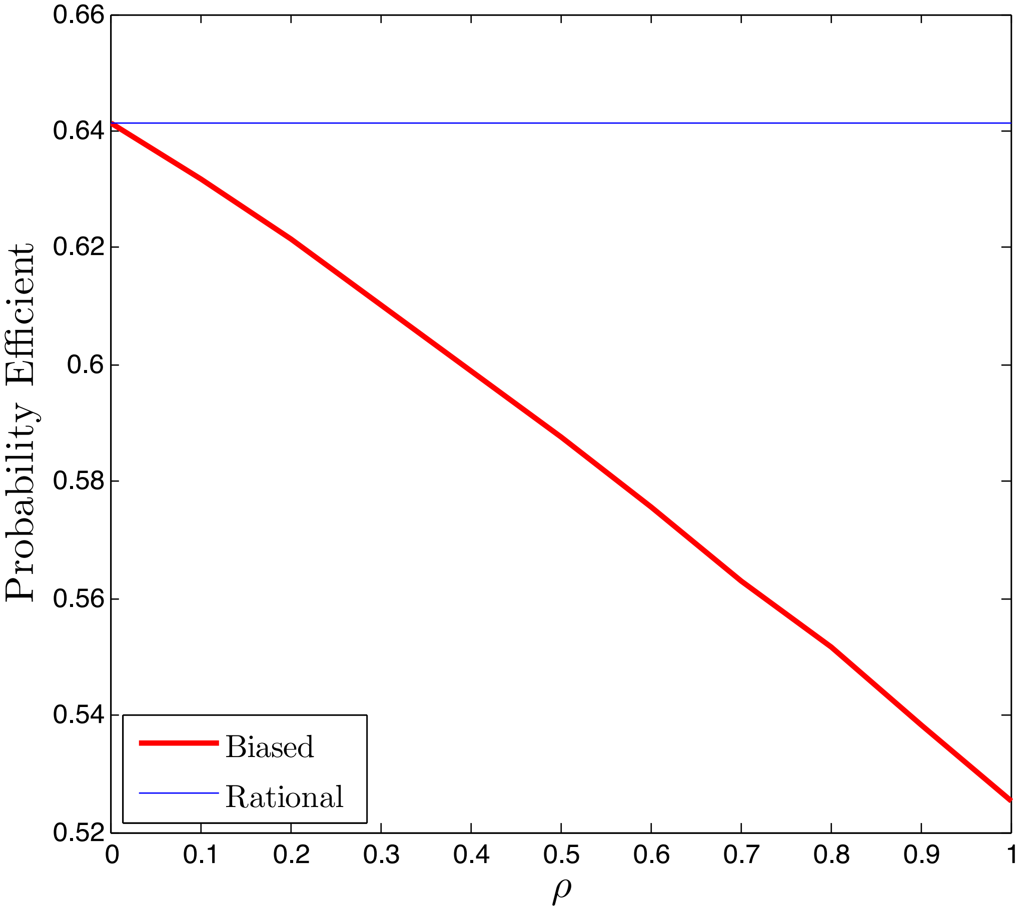

Figure 3.2: *The probability that the player with the highest taste t wins the auction as a function of ρ. (Estimated from 1,000,000 simulated auctions.)*

$\theta_i + t_i + 2\widehat{\mathbb{E}}_i[\theta_j \mid \theta_j + t_j = \theta_i + t_i]$. Since $\widehat{\mathbb{E}}_i[\theta_j \mid \theta_j + t_j = \theta_i + t_i] = \frac{1}{2}[\theta_i + (1-\rho)t_i + \rho/2]$ by Equation 3.10, it follows that $p^1(t_i, \theta_i) = \phi(\theta_i + t_i) - \gamma_i$ where ϕ is a fixed constant across all players—$\phi = \frac{3}{2}$—and γ_i is a type dependent distortion caused by misperception of the taste distribution—$\gamma_i = \rho(t_i - 1/2)$. As will be crucial below, Player i thinks γ_i depends solely on the distribution of tastes, and is thus constant across players. Simplifying, the first player to dropout does so at price

$$\bar{p}^1 = \min_i \left\{ 2\theta_i + (2-\rho)t_i + \rho/2 \right\}.$$

To derive p_i^2, we must solve for a remaining players' inference from observing the first player drop at $\bar{p}^1$. Say that k drops at $\bar{p}^1$. Then i thinks $\bar{p}^1 = \phi(\theta_j + t_j) - \gamma_i$; hence

$$\widehat{\mathbb{E}}_i[\theta_k \mid \phi(\theta_k + t_k) - \gamma_i = \bar{p}^1] \quad = \quad \frac{1}{2}\Big((\bar{p}^1 + \gamma_i)/\phi - \rho(t_i - 1/2) \Big)$$

$$= \frac{1}{4}\Big(\bar{p}^1 - \rho(t_i - 1/2)\Big). \tag{3.20}$$

The next lemma follows immediately from the previous expression:

Lemma 3. *Suppose the first player k drops out at price $\bar{p}^1$. For each $i \neq k$, $\widehat{\mathbb{E}}_i[\theta_k \mid \bar{p}^1]$ is decreasing in t_i.*

Lemma 3 is the dynamic analogue of Lemma 2. Since Player i assumes the first bidder to withdraw (k) has taste similar to her own, i's taste distorts her perception of k's private information. Rational inference about θ_k shouldn't depend on one's own taste whatsoever. Here, of course, it does in a systematic way. The higher is t_i, the lower is i's expectation of θ_k. A bidder with a strong private taste for the object thinks there is no way anybody would dropout early unless she has very negative private information about the common value.

Supposing that players i and j remain, i bids up to

$$p^2(t_i, \theta_i, \bar{p}^1) = t_i + \theta_i + \widehat{\mathbb{E}}_i[\theta_j \mid \theta_j + t_j = \theta_i + t_i] + \widehat{\mathbb{E}}_i[\theta_k \mid \bar{p}^1].$$

From Equations 3.10 and 3.20

$$\begin{aligned} p^2(t_i, \theta_i, \bar{p}^1) &= t_i + \theta_i + \frac{1}{2}\Big(\theta_i + (1-\rho)t_i + \rho/2\Big) + \frac{1}{4}\Big(\bar{p}^1 - \rho(t_i - 1/2)\Big) \\ &= \frac{3}{2}[t_i + \theta_i] + \frac{1}{4}p_1^* - \frac{3}{4}\rho(t_i - 1/2). \end{aligned} \tag{3.21}$$

Intuitively, the bid is increasing in surplus, past drop-out price, and decreasing in the extent of the rightward shift of the support. Making explicit the weight $\bar{p}_i^2$ puts on θ_i relative to t_i,

$$p^2(t_i, \theta_i, \bar{p}^1) = \frac{3}{2}\theta_i + \frac{3}{2}\left(1 - \rho/2\right)t_i + \frac{1}{4}p_1^* + \frac{3}{8}\rho.$$

That is, people put too little weight on t_i relative to θ_i. Their bid should weight these equally, but they weight t_i less by a factor of $1 - \rho/2 \in (.5, 1)$. This nicely shows how inefficiency is increasing in ρ. The weight on t_i approaches the rational benchmark only as $\rho \to 0$.

The final price of the good is the minimum of the expression among those players still in the auction: $\bar{p}^2 = \min_{l \in \{i,j\}} p^2(t_l, \theta_l)$.

Proposition 6. *Consider an English auction with $N = 3$. With projection, the bidding strategy consists of the following type- and history-dependent dropout prices:*

$$\begin{aligned} p^1(t_i, \theta_i) &= 2\theta_i + (2-\rho)t_i + \frac{1}{2}\rho \\ p^2(t_i, \theta_i, \bar{p}^1) &= \frac{3}{2}\theta_i + \frac{3}{4}\left(2 - \rho\right)t_i + \frac{1}{4}p_1^* + \frac{3}{8}\rho. \end{aligned} \tag{3.22}$$

As in the second-price auction, the fact that inference about the common value is increasingly pessimistic in a player's taste, inefficiency increases with the extent of projection.

Proposition 7. *Consider an English auction with $N = 3$. The probability that the auction is efficient is decreasing in ρ.*

The logic behind 7 is essentially a dynamic analogue of the logic as to why projection decreases efficiency in second-price auctions. Suppose i has the highest private value and k withdraws first at price $\bar{p}^1$. Lemma 3 implies that i draws a more pessimistic inference from $\bar{p}^1$ than does j. As such, if i projects, she's more likely to drop out early relative to a rational Player i. Figure 3.3 shows how efficiency changes with ρ.

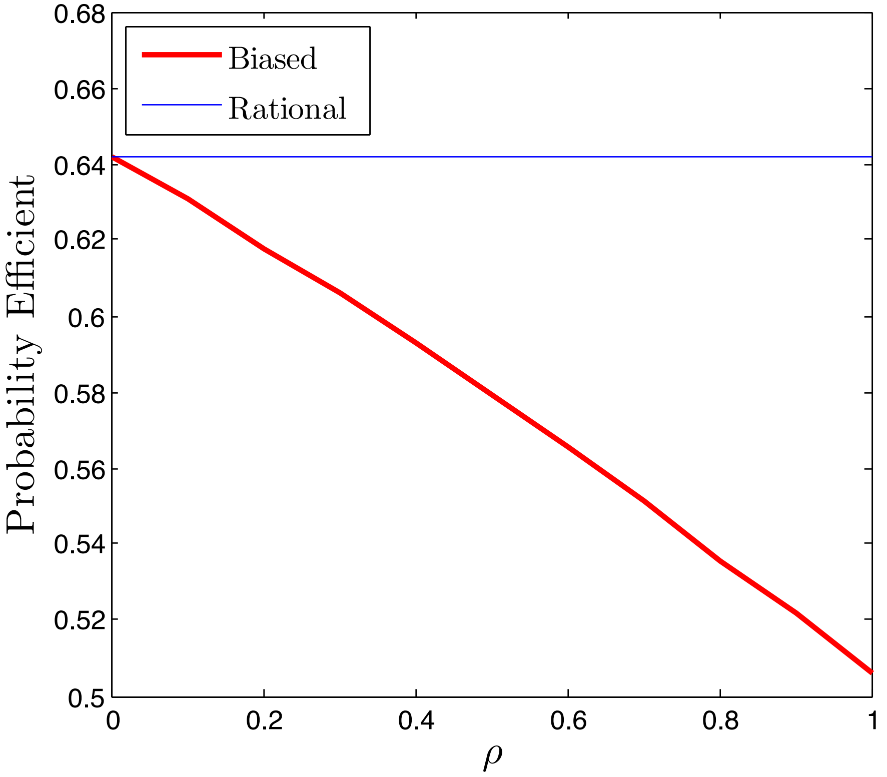

Figure 3.3: *The probability that the player with the highest taste t wins the auction as a function of ρ. (Estimated from 1,000,000 simulated auctions.)*

3.5 Conclusion

This paper explores how systematically mispredicting others' tastes affects the revenue and efficiency of several auction formats. In particular, we show that in private-value settings, projection leads to overbidding in first-price auctions. When the good also has some common-value component, projection causes misinference about the common value and consequently distorts bidding strategies regardless of the auction format. A player's biased estimate of the common value is always inversely related to her private value of the good. This misinference leads to reduced efficiency in both second-price and English auctions.

While a large literature provides empirical evidence that people project preferences, this is one of very few papers to formally draw out implications of the bias. Goeree and Grosser (2007) explore the consequences of a consensus effect in two-party voting settings—liberals overestimate the fraction of liberals, and conservatives overestimate the fraction of conservatives. Miscalculated probabilities of being pivotal can lead to inefficient election outcomes. Gagnon-Bartsch (2014) examines the impact of projection on social learning in general settings where players learn which action is best for themselves by observing predecessors' choices. Projection can cause society to believe a single practice is best for all individuals when in fact different types should optimally choose different actions.

This paper leaves open several questions that we hope to address in future research. While we characterize projection's effect on efficiency in settings with both private and common value, we do not fully explore how it affects expected revenue. This question is particularly complex in first-price auctions with private and common value. On the one hand, those with high private values overestimate the extent of competition, which pushes their bids upward. But, on the other hand, they draw the most pessimistic inference about the common value, pushing their bids downward. Understanding which of these forces dominates, and why, is still a work in progress.

Additionally, there are two related models—information projection and rational uncertainty about tastes—which may yield similar predictions to naive taste projection. First, the model of information projection by Madarasz (2012), in its most simple form, posits that player i with private signal θ_i thinks that all other players observe θ_i in addition to their own other private information. In essence, people treat their private signal as if it were public information. It is an open question how the predictions of taste projection and information projection differ in auctions with both common and private value. Second, the rational explanation for the false-consensus effect assumes the true distribution of values is unknown. A Bayesian agent uses her own valuation to update her beliefs about the distribution. Players with different private values update differently, arriving at conflicting perceptions of G. The key difference between naive projection and rational misprediction is that rational players know that people have conflicting beliefs and account for this disagreement when drawing inference. Differentiating the predictions of these various models will guide us in inferring from auction data which, if any, of these models are likely driving behavior.

Appendix

3.A Omitted Proofs

Proof of Lemma 1

Proof. Let $\hat{\mu}(t_i) = \rho t_i + (1-\rho)\mu$. From Equation 3.4,

$$\widehat{G}(t_i \mid t_i) = G\bigg(t_i - \big[\hat{\mu}(t_i)) - \mu\big]\bigg) = G\bigg((1-\rho)t_i + \rho\mu\bigg),$$

and

$$\widehat{G}(\hat{\mu}(t_i) \mid t_i) = G(\mu).$$

Thus, we want to show $\big|\widehat{G}\big(t_i \mid t_i\big) - \widehat{G}\big(\mu(t_i) \mid t_i\big)\big| = \big|G\big((1-\rho)t_i + \rho\mu\big) - G(\mu)\big| < |G(t_i) - G(\mu)|$. Since G is increasing, $G\big((1-\rho)t_i + \rho\mu\big) - G(\mu) > 0 \Leftrightarrow (1-\rho)t_i + \rho\mu \Leftrightarrow t_i > \mu \Leftrightarrow G(t_i) - G(\mu) > 0$. Thus, $G\big((1-\rho)t_i + \rho\mu\big) - G(\mu) > 0$ if and only if $G(t_i) - G(\mu) > 0$, which is true if and only if $t_i > \mu$. So, we consider two cases: (1) $t_i > \mu$, and (2) $t_i < \mu$. Suppose $t_i > \mu$. Our claim holds iff $G\big((1-\rho)t_i + \rho\mu\big) < G(t_i)$, which holds iff $(1-\rho)t_i + \rho\mu < t_i$, which is true by assumption $t_i > \mu$ if $\rho \in (0,1]$. Similarly, suppose $t_i < \mu$. Then our claim holds iff $(1-\rho)t_i + \rho\mu < t_i$, which is true by assumption $t_i < \mu$ if $\rho \in (0,1]$. □

Proof of Proposition 1

Proof. From our naivete assumption, each player i thinks she's participating in an auction where it's common knowledge that $t \sim \widehat{G}(\cdot \mid t_i)$. Let Γ denote either the second-price or English auction. As such, she plays a strategy that's part of a rational Bayesian Nash equilibrium of game $\Gamma(\widehat{G}(\cdot \mid t_i))$. Let $\beta(t_i)$ denote type t_i's bidding strategy—in the second-price auction, $\beta(t_i)$ denotes the bid, and in the English auction it denotes t_i's exit price. For any belief $\widehat{G}$, it is well known that $\beta(t_i) = t_i$ is a rational Bayesian Nash equilibrium strategy of $\Gamma(\widehat{G}(\cdot \mid t_i))$. □

Proof of Proposition 2

Proof. Part 1. In text.
Part 2. From Equations 3.7 and 3.8, $\beta(t_i) - \beta^R(t_i) = \frac{\rho}{N}t_i - \frac{\rho}{2N} > 0 \Leftrightarrow t_i > \frac{1}{2}$. □

Proof of Proposition 3

Proof. We must compare the expected winning bid under projection with the expected winning bid under rational expectations. Since both biased and rational bidding functions are increasing in t_i, we take expectations with respect to t_i conditional on $t_i = \max_{j \in N} t_j$. Note that $t_i = \max_{j \in N} t_j$ has distribution G^N, so $g(t_i \mid t_i \max_{j \in N} t_j) = NG^{N-1}(t)g(t)$. Assuming $t \sim \mathrm{U}[0,1]$, $g(t_i \mid t_i \max_{j \in N} t_j) = Nt^{N-1}$ and expected revenue is

$$\begin{aligned}
\mathbb{E}[\beta(t_i) \mid t_i = \max_{j \in N} t_j] &= N \int_0^1 \beta(t_i) t_i^{N-1} dt_i \\
&= \left(\frac{N-1+\rho}{N} \right) \frac{N}{N+1} - \frac{\rho}{2N} \\
&= \frac{N-1+\rho}{N+1} - \frac{\rho}{2N}. \qquad \text{(A.1)}
\end{aligned}$$

Letting $\rho = 0$, it follows that expected revenue under rational bidding is $\frac{N-1}{N+1}$. Expected revenue with projection is larger so long as

$$\frac{\rho}{N+1} > \frac{\rho}{2N} \Leftrightarrow N > 1$$

so long as $\rho > 0$. Furthermore, from Equation A.1 the derivative of expected revenue with respect to ρ is $(N-1)/[2N(N+1)] > 0$ for $N > 1$. Thus, expected revenue is increasing in ρ. □

Proof of Lemma 2

Proof. Player i perceives $t \sim \mathrm{U}[\underline{t}(t_i), \bar{t}(t_i)]$ where $\underline{t}(t_i)$ and $\bar{t}(t_i)$ are given in Equation 3.5. It follows that $\widehat{\mathbb{E}}_i[\theta_j \mid t_j + \theta_j = t_i + \theta_i] = \frac{1}{2}\big[\theta_i + t_i - \underline{t}(t_i)\big] = \frac{1}{2}\big[\theta_i + (1-\rho)t_i + \rho/2\big]$. A rational player forms expectations with respect to $t \sim \mathrm{U}[0,1]$, hence $E[\theta_j \mid t_j + \theta_j = t_i + \theta_i] = \frac{1}{2}\big[\theta_i + t_i\big]$. Hence,

$$\widehat{\mathbb{E}}_i[\theta_j \mid t_j + \theta_j = t_i + \theta_i] - E[\theta_j \mid t_j + \theta_j = t_i + \theta_i] = \frac{\rho}{4}\Big(1 - 2t_i\Big),$$

which is clearly decreasing in t_i. □

Proof of Proposition 4

Proof. Consider a second-price auction with $N \geq 2$ bidders. From Equation 3.14,

$$\beta(t_i, \theta_i) = \frac{3}{2}(t_i + \theta_i) + \frac{N-2}{2}\widehat{\mathbb{E}}_i\big[\theta_j + t_j \bigm| \theta_j + t_j < \theta_i + t_i\big] - (N-1)\rho\big(t_i - 1/2\big). \qquad \text{(A.2)}$$

Thus, we need only compute $\widehat{\mathbb{E}}_i\big[\theta_j + t_j \bigm| \theta_j + t_j < \theta_i + t_i\big]$. Since Player i thinks $t \sim \mathrm{U}[\underline{t}(t_i), \underline{t}(t_i) + 1]$ where $\underline{t}(t_i) = \rho\,(t_i - 1/2)$, Player i thinks $z \equiv (\theta + t)$ follows a triangular distribution on with density

$$f_Z\left(z_i\right) = \begin{cases} z - \underline{t}(t_i) & \text{if} \quad \underline{t}(t_i) \leq z \leq \underline{t}(t_i) + 1 \\[2ex] 2 + \underline{t}(t_i) - z & \text{if} \quad \underline{t}(t_i) + 1 \leq z \leq \underline{t}(t_i) + 2 \end{cases}$$

and c.d.f.

$$F_Z\left(z_i\right) = \begin{cases} \frac{\left(z - \underline{t}(t_i)\right)^2}{2} & \text{if} \quad \underline{t}(t_i) \leq z \leq \underline{t}(t_i) + 1 \\[2ex] z\left(\underline{t}(t_i) + 2\right) - \frac{z_i^2}{2} - \frac{\underline{t}(t_i)\left(\underline{t}(t_i)+4\right)}{2} - 1 & \text{if} \quad \underline{t}(t_i) + 1 \leq z \leq \underline{t}(t_i) + 2 \end{cases}.$$

Hence,

$$\widehat{\mathbb{E}}_i\left[\theta_j + t_j \;\middle|\; \theta_j + t_j < \theta_i + t_i\right] = \\ \begin{cases} \frac{2\int_{\underline{t}(t_i)}^{\theta_i+t_i} z\left(z - \underline{t}(t_i)\right) dz}{\left(\theta_i + t_i - \underline{t}(t_i)\right)^2} & \text{if} \quad \underline{t}(t_i) \leq z \leq \underline{t}(t_i) + 1 \\ \frac{2\int_{\underline{t}(t_i)}^{\underline{t}(t_i)+1} z\left(z - \underline{t}(t_i)\right) dz + \int_{\underline{t}(t_i)+1}^{\theta_i+t_i} z\left(\underline{t}(t_i) - z + 2\right) dz}{2(\theta_i+t_i)\left(\underline{t}(t_i)+2\right) - (\theta_i+t_i)^2 - \underline{t}(t_i)\left(\underline{t}(t_i)+4\right) - 4} & \text{if} \quad \underline{t}(t_i) + 1 \leq z \leq \underline{t}(t_i) + 2 \end{cases}$$

Using $\underline{t} = \rho(t_i - 1/2)$, it follows that

$$\widehat{\mathbb{E}}_i\left[\theta_j + t_j \;\middle|\; \theta_j + t_j < \theta_i + t_i\right] = \begin{cases} \frac{2}{3}\left(\theta_i + t_i\right) + \frac{\rho}{3}\left(t_i - \frac{1}{2}\right) & \text{if} \quad \underline{t}(t_i) \leq z \leq \underline{t}(t_i) + 1 \\[2ex] \dfrac{A_2(t_i, \theta_i)}{B_2(t_i, \theta_i)} & \text{if} \quad \underline{t}(t_i) + 1 \leq z \leq \underline{t}(t_i) + 2 \end{cases}$$

where

$$\begin{aligned} A_2(t_i, \theta_i) = 16\left(\theta_i + t_i\right)^3 - 12\left(\theta_i + t_i\right)^2\left(-\rho + 2\rho t_i + 4\right) \\ + \rho\left(2t_i - 1\right)\left(-12\rho + \rho^2 + 4\rho^2 t_i^2 + 24\rho t_i - 4\rho^2 t_i + 24\right) + 16, \end{aligned}$$

and

$$\begin{aligned} B_2(t_i, \theta_i) = 24\left(\theta_i + t_i\right)^2 - 24\left(\theta_i + t_i\right)\left(-\rho + 2\rho t_i + 4\right) \\ + 6\left(-8\rho + \rho^2 + 4\rho^2 t_i^2 + 16\rho t_i - 4\rho^2 t_i + 8\right). \end{aligned}$$

Plugging this expression for $\widehat{\mathbb{E}}_i\left[\theta_j + t_j \;\middle|\; \theta_j + t_j < \theta_i + t_i\right]$ into the bidding function (Euation A.2) yields the strategy stated in the proposition. □

Proof of Proposition 5

Proof. (Sketch.) As ρ increases, the weight that the bidding function places on t decreases relative to the weight on θ. To see this, consider low types (i.e., $t+\theta<\rho(t-1/2)+1$) who bid $\beta(t,\theta)=\beta_L(t,\theta)$ (see Equation 3.15). It's clear that $\frac{\partial}{\partial\theta}\beta(t,\theta)-\frac{\partial}{\partial t}\beta(t,\theta)=\rho(2N-1)/6$. Thus, when $\rho=0$, an incremental change in θ has the same effect on β as an incremental change in t. But when $\rho>0$, increasing θ has a larger effect on β than increasing t. Holding types fixed, increasing ρ alters the ordering of bids in a way that makes those with high signals more likely to win. More precisely, fix the vector of tastes $(t_1,...,t_N)$ and assume $t_1>t_i$ for all $i=2,...,N$. For any $i=2,...,N$ and any realization of θ_1, the measure of signals θ_i such that $\beta(t_i,\theta_i)>\beta(t_1,\theta_1)$ is larger the larger is ρ. This follows from the fact that the difference in β_i and β_1 due solely to taste is smaller the larger is ρ. As such, Player i with $t_i<t_1$ doesn't need as high a signal in order to outbid Player 1 the higher is ρ. One can arrive at similar conclusions for $\beta(t,\theta)=\beta_L(t,\theta)$, although the algebra is significantly more involved. □

Proof of Proposition 7

Proof. Let $i=\arg\max_{l\in N}t_l$. We show that the probability that Player i drops out in the first or second stage is increasing in ρ. Player i drops out in the first stage if $p^1_{i^*}=\min_{l\in N}p^1_l$. This occurs with probability $\Pr\left(p^1(t_i,\theta_i)<p^1(t_k,\theta_k)\right)\Pr\left(p^1(t_i,\theta_i)<p^1(t_j,\theta_j)\right)$. Since for any $l\in N$, $p^1(t_l,\theta_l)=\theta_i+(2-\rho)t_i+\frac{1}{2}\rho$,

$$\begin{aligned}\Pr\left(p^1(t_i,\theta_i)<p^1(t_l,\theta_l)\right) &= \Pr\left(\theta_i+(2-\rho)t_i<\theta_l+(2-\rho)t_l\right)\\ &= \Pr\left(\theta_l-\theta_i>(2-\rho)(t_i-t_l)\right).\end{aligned}$$

Since $t_i>t_l$, the right-hand side of the inequality above is decreasing in ρ. Thus, the probability is clearly increasing in ρ. As such, the probability that i drops out first is increasing in ρ. Now suppose that k drops out first at price $\bar{p}$. The probability that i loses is $\Pr(p^2(t_i,\theta_i,\bar{p}^1)<p^2(t_j,\theta_j,\bar{p}^1))$. From Proposition 6,

$$\begin{aligned}\Pr\left(p^2(t_i,\theta_i,\bar{p}^1)<p^2(t_j,\theta_j,\bar{p}^1)\right) &= \Pr\left(\frac{3}{2}\theta_i+\frac{3}{4}\left(2-\rho\right)t_i<\frac{3}{2}\theta_j+\frac{3}{4}\left(2-\rho\right)t_j\right)\\ &= \Pr\left(\theta_j-\theta_i>\frac{1}{2}(2-\rho)(t_i-t_j)\right).\end{aligned}$$

Again, since the right-hand side is decreasing in ρ, the probability is increasing. Thus, inefficiency is increasing in ρ. □

Bibliography

[1] Acemoglu, D., V. Chernozhukov, and M. Yildiz (2007): "Learning and disagreement in an uncertain world," Working Paper, MIT.

[2] Acemoglu, D., V. Chernozhukov, and M. Yildiz (2009): "Fragility of Asymptotic Agreement under Bayesian Learning," Working Paper, MIT.

[3] Acemoglu, D., G. Como, F. Fagnani, and A. Ozdaglar (2013): "Opinion Fluctuations and Disagreement In Social Networks," *Mathematics of Operations Research*, 28(1): 1–27.

[4] Acemoglu, D., M. Dahleh, I. Lobel, and A. Ozdaglar (2011): "Bayesian Learning in Social Networks," *Review of Economic Studies*, 78(4): 1201–1236.

[5] Andreoni, J. and T. Mylovanov (2012): "Diverging Opinions," *American Economic Journal: Microeconomics*, 4(1): 209–232.

[6] Banerjee, A. (1992): "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, 107, 797–817.

[7] Banerjee, A. and D. Fudenberg (2004): "Word-of-mouth Learning," *Games and Economic Behavior*, 46(1), 1–22.

[8] Barberis, N., A. Shleifer, and R. Vishny (1998): "A model of investor sentiment," *Journal of Financial Economics*, 49(3), 307–343.

[9] Benjamin, D., C. Raymond, and M. Rabin (2012): "A Model of Non-Belief in the Law of Large Numbers," Working Paper, Cornell University.

[10] Berk, R. (1966): "Limiting Behavior of Posterior Distributions When the Model Is Incorrect," *Annals of Mathematical Statistics*, 37, 51–58.

[11] Bikchandani, S., D. Hirshleifer, and I. Welch (1992): "A Theory of Fads, Fashion, Custom and Cultural Change as Informational Cascades," *Journal of Political Economy*, 100, 992–1026.

[12] Bohren, A. (2010): "Information-processing bias in social learning," Working Paper, University of California, San Diego.

[13] Brown, C. (1982): "A False Consensus Bias in 1980 Presidential Preferences," *Journal of Social Psychology*, 118, 137–138.

[14] Bulow, J. and P. Klemperer (2002): "Prices and the Winner's Curse," *RAND Journal of Economics*, 33(1), 1–21.

[15] Burnham, K., and D. Anderson (1989): *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer, New York.

[16] Busse, B., D. Pope, J. Pope, and J. Silva-Risso (Forthcoming): "The Overinfluence of Weather Fluctuations on Convertible and 4-Wheel Drive Purchases," *Quarterly Journal of Economics.*

[17] Cai, H., Y. Chen, and H. Fang (2009): "Observational Learning: Evidence from a Randomized Field Experiment," *American Economic Review*, 99(3), 864-882.

[18] Çelen, B. and S. Kariv (2004): "Observational Learning Under Incomplete Information," *Games and Economic Behavior*, 47(1), 72–86.

[19] Cohen, G. (2003): "Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs," *Journal of Personality and Social Psychology*, 85(5): 808–822.

[20] Compte, O. and P. Jehiel (2002): "On the Value of Competition in Procurement Auctions," *Econometrica*, 70(1), 343–355.

[21] Conley, T. and C. Udry (2010): "Learning About a New Technology: Pineapples in Ghana," *American Economic Review*, 100(1), 35–69.

[22] Conlin, M., T. O'Donoghue, and T. Vogelsang (2007): "Projection Bias in Catalog Orders," *American Economic Review*, 97(4): 1217–1249.

[23] Cox, J., V. Smith, and J. Walker (1992): "Theory and Misbehavior of First-Price Auctions: Comment," *American Economic Review* 82(5), 1392-1412.

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

[25] Cruces, G., R. Perez-Truglia, and M. Tetaz (2013): "Biased Perceptions of Income Distribution and Preferences for Redistribution: Evidence from a Survey Experiment," *Journal of Public Finance*, 98, 100–112.

[26] Dawes, R. (1989): "Statistical criteria for establishing a truly false consensus effect," *Journal of Experimental Social Psychology*, 25(1), 1–17.

[27] Dawes, R., and M 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.

[28] DeGroot, M. (1970): *Optimal Statistical Decisions*, McGraw-Hill, New York.

[29] DeGroot, M. (1974): "Reaching a Consensus," *Journal of the American Statistical Association*, 69(345), 118–121.

[30] Delavande, A., and C. Manski (2012): "Candidate Preferences and Expectations of Election Outcomes," *Proceedings of the National Academy of Sciences of the United States*, 109(10): 3711–3715.

[31] DeMarzo, D. Vayanos, and J. Zwiebel (2003): "Persuasion Bias, Social Influence, and Uni- Dimensional Opinions," *Quarterly Journal of Economics*, 118: 909–968.

[32] Downs, A. (1957): "An Economic Theory of Political Action in a Democracy," *Journal of Political Economy*, 65(2): 135-150.

[33] Egan, D., C. Merkle and M. Weber (2012): "The Beliefs of Others: Naive Realism and Investment Decisions," Mimeo.

[34] Ellison, G., and D. Fudenberg (1993): "Rules of thumb for social learning," *Journal of Political Economy* 101(4), 612–643.

[35] Engelmann, D. and M. Strobel (2001): "The False Consensus Effect Disappears if Representative Information and Monetary Incentives Are Given," *Experimental Economics*, 3, 241–643.

[36] Engelmann, D. and M. Strobel (2012): "Deconstruction and Reconstruction of an Anomaly," *Games and Economic Behavior*, 76, 678–689.

[37] Enke, B. and F. Zimmerman (2013): "Correlation Neglect in Belief Formation," Mimeo.

[38] Esponda, I. and D. Pouzo (2013): "An Equilibrium Framework for Modeling Bounded Rationality," Mimeo.

[39] Eyster, E. and M. Rabin (2005): "Cursed Equilibrium," *Econometrica*, 73(5), 1623-1672.

[40] Eyster, E. and M. Rabin (2010): "Naïve Herding in Rich-Information Settings," *American Economic Journal: Microeconomics*, 2(4), 221–243.

[41] Eyster, Erik and Matthew Rabin (2013): "Extensive Imitation is Irrational and Harmful," Mimeo.

[42] Eyster, E., M. Rabin, and D. Vayanos (2013): "Financial Markets where Traders Neglect the Informational Content of Asset Prices," Mimeo.

[43] Eyster, E., M. Rabin, and G. Weizsäcker (2013): "An Experiment on Social Mislearning," Mimeo.

[44] Faro, D., and Y. Rottenstreich (2006): "Affect, Empathy, and Regressive Mispredictions of Others' Preferences Under Risk," *Management Science*, 52(4), 529–541.

[45] Filiz-Ozbay E. and E. Ozbay (2007): "Auctions with Anticipated Regret: Theory and Experiment," *American Economic Review*, 97(4), 1407-1418.

[46] Flynn, F. and S. Wiltermuth (2010): "Who's With Me? False Consensus, Brokerage, and Ethical Decision Making in Organizations," *Academy of Management Journal*, 53(5): 1074–1089.

[47] Fudenberg, D. and D. Levine (1993): "Self-Confirming Equilibrium," Econometrica, 61(3), 523–545.

[48] Gagnon-Bartsch, T. (2014): "Taste Projection in a Model of Social Learning." Working paper, University of California, Berkeley.

[49] Gagnon-Bartsch, T. and M. Rabin. (2014): "Explicability," Mimeo, UC Berkeley.

[50] Gagnon-Bartsch, T., and M. Rabin (2014): "Naive Social Learning, Mislearning, and Unlearning" Mimeo.

[51] Goeree, J., T. Palfrey, and B. Rogers (2006): "Social Learning with Private and Common Values," *Economic Theory*, 28(2), 245–264.

[52] Golub, Benjamin, and Matthew O. Jackson (2010): "NaÃŕve Learning in Social Networks and the Wisdom of Crowds," *American Economic Journal: Microeconomics*, 2(1): 112–49.

[53] Grossman, S. (1976): "On the Efficiency of Competitive Stock Markets when Traders Have Diverse Information," *Journal of Finance*, 31, 573–585.

[54] Hotelling, H. (1929): "Stability in Competition," *The Economic Journal*, 39: 41–57.

[55] Harrison, J. M. and D. Kreps (1978): "Speculative Investor Behavior in a Stock Market with Heterogenous Expectations", Quarterly Journal of Economics 93(2): 323–336.

[56] Jackson, M. and E. Kalai (1997): "Social Learning in Recurring Games," *Games and Economic Behavior*, 21, 102–134.

[57] Jehiel, P. and B. Moldovanu (2001): “Efficient Design with Interdependent Valuations,” *Econometrica*, 69(5), 1237–1259

[58] Klemperer, P. (1998): “Auctions with Almost Common Value: The "Wallet Game" and its Applications,” *European Economic Review*, 42(3-5), 757-769.

[59] Knight, B. and N. Schiff (2010): “Momentum and Social Learning in Presidential Primaries,” *Journal of Political Economy*, 118(6), 1110–1150.

[60] Kramer, G. (1971): “Short-Term Fluctuations in U.S. Voting Behavior: 1896-1964,” *American Political Science Review*, 65(1): 131–143.

[61] Krueger, J., and R. Clement (1994): “The truly false consensus effect - an ineradicable and egocentric bias in social-perception,” *Journal of Personality and Social Psychology*, 67(4), 596–610.

[62] Kübler, D., and G. Weizsäcker (2004): “Limited Depth of Reasoning and Failure of Cascade Formation in the Laboratory,” *Review of Economic Studies*, 71(2): 425–441.

[63] Loewenstein, G., T. O’Donoghue, and M. Rabin (2003): “Projection Bias in Predicting Future Utility,” *Quarterly Journal of Economics*, 118(4): 1209–1248.

[64] Madarasz, K. (2012): “Information Projection: Model and Applications,” *Review of Economic Studies*, 79, 961–985.

[65] Malmendier, U. and S. Nagel (2011): “Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?” *Quarterly Journal of Economics*, 126, 373–416.

[66] Marks, G. and N. Miller (1987): “10 years of research on the false-consensus effect: An empirical and theoretical review,” *Psychological Bulletin*, 102(1), 72-90.

[67] Milgrom, P. (1981): “Good News and Bad News: Representation Theorems and Applications,” *Bell Journal of Economics*. 12(2), 380–391.

[68] Milgrom, P. and R. Weber (1982): “A Theory of Auctions and Competitive Bidding,” *Econometrica*, 50(5), 1089-1112.

[69] Moretti, E. (2011): “Social Learning and Peer Effects in Consumption: Evidence from Movie Sales,” *Review of Economic Studies*, 78(1): 356–393.

[70] Morris, S. (1995): “The Common Prior Assumption in Economic Theory,” *Economics and Philosophy*, 11: 227–253.

[71] Morris, S. (1996): “Speculative Investor Behavior and Learning”, *Quarterly Journal of Economics*, 111: 1111–1133.

[72] Mossel, E., A. Sly, and O. Tamuz (2012): “From Agreement to Asymptotic Learning,” Mimeo.

[73] Mullen, B., J. Atkins, D. Champion, C. Edwards, D. Hardy, J. Story, and M. Vanderlok (1985): “The false consensus effect: A meta-analysis of 115 hypothesis tests,” *Journal of Experimental Social Psychology*, 21(3), 262–283.

[74] Munshi, K. (2003): “Social Learning in a Heterogeneous Population: Technology Diffusion in the Indian Green Revolution,” *Journal of Development Economics*, 73(1), 185–213.

[75] Pagnozzi, M. (2007): “Sorry Winners,” *Review of Industrial Organization*, 30(3), 203–225.

[76] Pesendorfer, W., and J. Swinkels (2000): “Efficiency and Information Aggregation in Auctions,” *American Economic Review*, 90(3), 499-525.

[77] Rabin, M. (2002): “Inference by Believers in the Law of Small Numbers”, *Quarterly Journal of Economics*, 117(3): 775–816.

[78] Rabin, M. and J. Schrag (1999): “Inference by Believers in the Law of Small Numbers”, *Quarterly Journal of Economics*, 114(1): 37–82.

[79] Rabin, M. and D. Vayanos (2010): “The Gambler’s and Hot-Hand Fallacies: Theory and Applications,” *Review of Economic Studies*, 77: 730–778.

[80] Ross, L., D. Greene, and P. House (1977): “The False Consensus Effect: An Egocentric Bias in Social Perception and Attribution Processes,” *Journal of Experimental Social Psychology*, 13, 279–301.

[81] Rouhana, N., A. O’Dwyer, and S. Vaso (1997): “Cognitive biases and political party affiliation in intergroup conflict,” *Journal of Applied Social Psychology*, 27(1), 37–57.

[82] Salganik, M., P. Dodds, and D. Watts (2006): “Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market,” *Science*, 311(2), 854–856.

[83] Scheinkman, J. and W. Xiong (2003): “Overconfidence and Speculative Bubbles”, *Journal of Political Economy*, 111(6): 1183–1219.

[84] Schwartzstein, J. (Forthcoming): “Selective Attention and Learning,” *Journal of the European Economic Association.*

[85] Sethi, R. and M. Yildiz (2012): “Public Disagreement,” *American Economic Journal: Microeconomics*, 4(3): 57–95.

[86] Simonsohn, U. (2010): "Weather to Go to College," *Economic Journal*, 120(543), 270–280.

[87] Smith, L. and P. Sørensen (2000): "Pathological outcomes of observational learning," *Econometrica*. 68(2), 371–398.

[88] Smith, L. and P. Sørensen (2008): "Rational Social Learning with Random Sampling," Mimeo.

[89] Sorensen, A. (2006): "Social Learning and Health Plan Choice," *RAND Journal of Economics*, 37(4), 929–945.

[90] Van Boven, L., D. Dunning, and G. Loewenstein (2000): "Egocentric Empathy Gaps Between Owners and Buyers: Misperceptions of the Endowment Effect," *Journal of Personality and Social Psychology*, 79(1), 66–76.

[91] Van Boven, L., and G. Loewenstein (2003): "Social Projection of Transient Drive States," *Personality and Social Psychology Bulletin*, 29(9): 1159–1168.

[92] Van Boven, L., G. Loewenstein, and D. Dunning (2003): "Mispredicting the Endowment Effect: Underestimation of Owners' Selling Prices by Buyer's Agents," *Journal of Economic Behavior and Organization*, 51: 351–365.

[93] Wallace, D. F. (1996): *Infinite Jest*. New York: Little, Brown.