Taste Projection in Models of Social Learning*

Tristan Gagnon-Bartsch†

Old Version—New Version Under Revision

November 15, 2016

Abstract

This paper studies the implications of taste projection—the tendency to overestimate how similar others' preferences are to our own—within social-learning environments. Individuals sequentially choose among two options with payoffs dependent on an unknown state of the world and one's idiosyncratic, privately-observed taste. Learning about the state from others' choices requires people to assess whether uncommon actions were likely provoked by atypical tastes or private information contradicting the public belief. Taste projectors over-attribute these actions to information, which prevents some (and perhaps all) from ever learning their correct action. A player's long-run beliefs are determined by her own taste and the extent to which she and all others project. When each thinks her taste is most common, all players inevitably choose the same option and each grows certain this choice is optimal. But if some acknowledge they have an uncommon taste, social beliefs and behavior may perpetually cycle—history never provides a clear message about the optimal choice. If players additionally update their beliefs about others' tastes from their actions, an initial projection error can be intensified: agents conclude more share their preference than originally anticipated. These predictions are distinct from rational learning under uncertainty about the distribution of preferences.

JEL Classification: C72, D82, D83.
Keywords: Observational learning, herd behavior, false-consensus effect, projection bias.

*I thank Matthew Rabin for his guidance on this project. For helpful discussions and suggestions, I thank David Ahn, Ned Augenblick, Nick Barberis, Ben Bushong, Stefano DellaVigna, Erik Eyster, David Hirshleifer, Shachar Kariv, Jussi Keppo, Brian Knight, Botond Kőszegi, Kristof Madarasz, Takeshi Murooka, Omar Nayeem, Ted O'Donoghue, Antonio Rosato, Josh Schwartzstein, Adam Szeidl, Xiaoyu Xia, and various seminar audiences at Harvard and UC Berkeley. I thank the National Science Foundation Graduate Research Fellowship for generous financial support.

†Harvard University. E-mail: gagnonbartsch@fas.harvard.edu. Address: Baker Library, Bloomberg Center 433B, Boston, MA 02163, USA.

1 Introduction

We use the actions of our friends, neighbors, and peers to guide many of our own decisions. This is true of both daily choices—where to dine or which film to watch—and those with larger stakes, such as choosing a college major, selecting stocks, or deciding for whom to vote.[1] In each of these domains where social learning is likely at play, the idiosyncratic tastes of those we observe surely influence their decisions. Hence, when inferring our own optimal action from others' choices, we should account for their differing goals and motives. For instance, observing a friend dine at a particular restaurant reflects both the quality of that restaurant and her taste for the cuisine. And a stock purchase signals both a company's expected value and the investor's risk preferences. In both of these cases, two individuals with the same information may reasonably choose different actions.

How we disentangle the information driving others' choices from their tastes relies on our perceptions of how preferences are distributed within our social network. But evidence on "social projection" and the "false-consensus effect" suggests that these perceptions are often biased: we tend to overestimate how similar others' tastes are to our own.[2] For instance, people overestimate how many share their tastes for typical consumption goods (Ross, Greene, and House 1977), political candidates (Delavande and Manski 2012), and risk (Faro and Rottenstreich 2006). Van Boven and Loewenstein (2003) show that people also project transitory preferences, and hence overestimate how many share their current feelings like hunger and thirst. In social-learning settings, such misprediction of others' tastes introduces a systematic bias in what people infer from others' choices. For instance, fixing the number of people in a restaurant, those with a strong preference for its cuisine—who wrongly expect many to attend—develop a more pessimistic perception of its quality than those with a weaker preference. By incorporating taste projection into canonical models of observational learning (namely, those of Banerjee 1992; Bikhchandani et al. 1992; Smith and Sørensen 2000), this paper studies how and when these misperceptions lead society astray.

To outline the model, suppose investors with varied risk preferences wish to learn whether a new company A is riskier, but has higher potential return, than a known alternative B. A fraction λ prefers the safer investment whereas $1 - \lambda$ seeks the higher-return alternative. From experience with similar companies, investors have noisy private information about the relative risk. Thus, before performance data materializes, investors use others' choices to glean additional information. But heterogeneity in tastes complicates inference. Did a predecessor choose A because she's risk

[1] For instance, in consumption domains, Cai et al. (2009), Salganik et al. (2006), and Moretti (2010) demonstrate the impact of social learning on the demand for restaurants, music, and movies, respectively. In domains with larger stakes, social learning has been shown to influence investment in new crops (Conley and Udry 2010) and generate momentum in primary elections (Knight and Schiff 2010).

[2] While such taste-based predictions may be rational when people use their own preferences as information, Engelmann and Strobel (2012) and others show this tendency persists even when people have access to an unbiased sample of others' tastes, which is inconsistent with Bayesian updating. Section 2.2 reviews this evidence in detail.

averse with private information that A is safe? Or due to precisely the opposite preferences and 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 her same objective. The risk averse think $\hat{\lambda} > \lambda$; the return seekers think $\hat{\lambda} < \lambda$. When these two different types of traders observe somebody invest in A, they draw different conclusions about that predecessor's signal, and hence about the attributes of the asset: relative to her risk-neutral counterpart, the risk-averse investor overestimates the likelihood that A is safer than B.

How do these differing conclusions interact and shape social learning? Since each type forms different beliefs about predecessors' information, taste projectors never reach long-run agreement on the state and, thus, never mutually learn the truth. In settings where rational agents always learn, projection leads some types to continually choose incorrectly. Whether a player chooses optimally is determined by both her own taste—which dictates how she interprets past actions—and the extent of all others' biases—which determines the distribution of actions she observes. But in many of the cases I consider, projection creates a herd: all players grow confident that a particular option X is optimal for her taste, irrespective of whether this is true. These results help explain three phenomena inconsistent with rational learning. First, taste projection suggests why uniform behavior may arise despite diverse preferences. Second, it shows how society can develop and maintain confident yet false beliefs even when observing an arbitrarily large sample of informative behavior.[3] Third, naive learning can perpetuate a false-consensus bias. That is, if agents learn about others' tastes from actions but ignore differences in prior beliefs, then each type can grow *more* confident over time that her taste is most common.

Section 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, choose between two actions, A and B. An action's payoff derives from characteristics along two dimensions: (1) the "vertical" dimension measures commonly-valued quality, q, and (2) the "horizontal dimension" specifies a heterogeneously valued attribute, z.[4] For example, q may be the quality of a restaurant or film, while z represents the cuisine or genre. Or q is the expected return from a college major or stock, and z is either the major's subject or the stock's risk. Players have diverse tastes over the horizontal attribute: each prefers an attribute closest to her own taste, θ. For instance, fixing expected income, a student prefers the college major with subject z closest to her interests, θ. Agents learn about (q, z) from private signals and the history of prior choices. To crisply identify

[3]The rational-herding literature shows that, when learning from others, society may forever choose suboptimal actions. But as Eyster and Rabin (2010) note, in any setting where an incorrect herd may arise, rational agents never grow confident in the state of the world. The rational-herding literature thus does not explain how society may develop *confident* yet false beliefs.

[4]This terminology is borrowed from models of spatial differentiation, like Hotelling (1929) and Downs (1957).

the effects of projection, I focus on environments where rational agents learn the state.[5]

To model taste projection, I assume agents mispredict the distribution of tastes over horizontal attributes. Loosely, a player who prefers attribute z over z' overestimates how many share this preference, and the amount by which she overestimates is increasing in the intensity of her preference. For instance, a risk averse individual overestimates how many seek a safe investment relative to a risk neutral individual. I additionally assume that agents are naive about this bias: they neglect that those with different preferences disagree on the distribution of tastes. That is, each player wrongly assumes that others draw inference using her same model of the world, and thus thinks all players share her belief about the state after any observation.[6]

Sections 3 and 4 study the case where quality is known and the only uncertainty is over horizontal attributes—for instance, investors are uncertain about the relative risk of two startup companies. Following Smith and Sørensen (2000), I assume only two states of the world: along the horizontal dimension, A is either to the left or right of B. Section 3 first develops preliminaries on how a player updates her beliefs over these two states as a function of the actions she observes and her perceived measure of those with right-leaning preferences. This measure, 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 treats actions as overly-precise signals of the underlying private information and her beliefs overreact relative to rational beliefs. If she overestimates the variance in tastes—say, $\hat{\lambda} = 0.6$—then she treats actions as less informative and her beliefs underreact. And if she mispredicts the majority taste—say, $\hat{\lambda} = 0.4$—then her belief and the rational belief move in opposite directions after any observation.

Section 4 studies asymptotic properties of this learning process. Unlike rational beliefs, convergence to a stationary limit is not guaranteed, and convergence to fully-incorrect beliefs is possible.[7] However, limit beliefs are pinned down by the collection of all players' perceptions of the taste distribution. Characterizing limit beliefs amounts to assessing whether a system of beliefs (i.e., a belief for each type of player) is stochastically stable. I show that there exists a stable belief among taste projectors only if, at that belief, each player observes more people than expected choosing what she thought would be the majority action. It follows that taste projectors never converge on

[5] I assume private signals have unbounded informativeness. In this environment, bounded informativeness generates information cascades, whereby the information contained in the history of play eventually swamps the information contained in the most informative signal. The setting also precludes "confounded learning"—explored by Smith and Sørensen (2000)—where players may converge to an uncertain stationary belief when the quality difference between actions is not too large. Appendix A addresses when confounded learning may occur and how this possibility alters my baseline results.

[6] A variety of research demonstrates that people tend to overestimate how many share their beliefs. For instance, Egan, Merkle, and Weber (2014) survey a panel of private investors and find that people overestimate how many share their expectations about returns.

[7] Since a taste projector mispredicts the action frequency she should see conditional on her belief about the state, her beliefs do not form a martingale.

identical long-run beliefs, and thus some agents necessarily fail to learn.

When and how learning fails depends on whether people agree on the majority taste. When each thinks her taste is most common, agents inevitably herd on a single action X. All grow confident—some rightly, some wrongly—that X is optimal for their taste. This results in beliefs polarized according to preference. In the investment example, risk-averse individuals grow confident that option A is safe, while return-seeking individuals conclude it is risky.[8] Quite simply, since each thinks her taste is most common, she finds it optimal to follow the herd absent a strong contrary signal. As such, taste projection offers one reason why uniform behavior may emerge despite heterogeneity in tastes.[9] When many act per round ($N \to \infty$), the minority necessarily learns incorrectly, and all choose the action optimal for the majority taste. Hence, when people learn from others' take-up decisions, strong projection implies inadequate adoption of new technologies or welfare programs beneficial only to a minority. Observational learning is not only inefficient in this case, but potentially harmful to society.[10]

Although taste projection provides a clear logic for herding, uniform behavior is not a general consequence. When players correctly agree on the majority preference, they never settle on a fixed belief, let alone herd. Instead, society's opinion of the optimal action perpetually cycles.[11] Returning to the investment example, if society is initially confident that A is safe, then all of the risk-averse individuals, say 75% of the market, choose A. But those investors expect to see more than 75% select A. Their best explanation for such low investment is that other risk-averse traders have strong private information that B is in fact safer. As such, the risk averse switch to B, and B becomes the most prevalent choice. This sends a clear message to future investors: B is preferred by the majority, and is thus the safer asset. Once the market is confident that B is safe, the same logic repeats, which sends investors back toward A. Hence, beliefs perpetually oscillate.[12] This

[8]The beliefs of agents with opposing tastes display a strong form of polarization where they grow fully confident in alternative hypotheses. Other studies of persistent disagreement 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.

[9]Sorensen (2006) studies the selection of health-care plans and demonstrates that social learning 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 held as universally beneficial (e.g., avoiding salt), even though their efficacy likely depends on heterogeneous characteristics of patients.

[10]Although rational observational learning can cause incorrect herds (e.g., Bikchandani, et al. 1992), it is necessarily welfare improving (on average) relative to relying solely on private information. Eyster and Rabin (2010, 2014) show how a distinct form of naive learning in which people neglect redundancies in information is also socially harmful.

[11]In this case, projection generates fad-like behavior in a setting where rational behavior always converges. In other social-learning settings, however, even rational behavior may fail to converge. This may happens, for example, when players observe only their immediate predecessors (Çelen and Kariv 2004). Outside of rational models, 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; Cohen 2003).

[12]While tempting to think that this non-convergence results from the binary state space—agents expect to observe one of only two long-run action frequencies—non-convergence can arise even with a continuous state space.

cycling is most dramatic when the majority's perception of the taste distribution is less biased than those in the minority. In this case, spells where all believe A is safe are followed by even longer spells where all think B is safe, and so on. Beliefs spend roughly equal time favoring each state, causing players to make worse choices, on average, when learning from others than if they simply followed their private information.

Section 5 introduces uncertainty over quality. I assume that quality differences are potentially large enough that players may prefer the same option. For example, all diners may attend the same restaurant despite differences in taste if it has exceptional quality. I show that with two types, no matter the true quality difference between options, society necessarily concludes it is large enough that all prefer the same choice. Intuitively, projection causes agents to first herd on some action X, which consequently suggests that X has superior quality. And since people inevitably herd on the more popular action, taste-based popularity is systemically misattributed 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: people over-attribute low, selective take-up to general ineffectiveness.[13] Additionally, even when the horizontal attributes of A and B are known—for instance, diners know A serves Argentinian food while B serves Brazilian—agents still systematically mislearn quality. Fans of Brazilian food attribute moderate popularity of B to limited quality rather than accepting that fewer people enjoy such cuisine. But Argentines reach a *higher* perception of B's quality to explain higher-than-expected attendance. Those with the most positive view of B's quality are those who prefer the attributes of A.

Section 6 considers uncertainty over the distribution of tastes. In this case, agents revise their models of others' preferences as they observe actions. However, such revision does not necessarily correct errors stemming from taste projection. If agents neglect heterogeneity in priors—they fail to fully appreciate that those with different tastes have divergent beliefs about the distribution of preferences—then players may still fail to learn their optimal action. In fact, observing others can intensify the bias in perceived taste distributions, leading all agents to conclude that their taste is most common. This may happen, for example, when the history begins with many choosing A. In the investment example, a risk-averse individual infers 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 contrary information, each type best responds with A. These conflicting beliefs perpetuate the herd, leading people to conclude that *all* investors share a similar taste.

With uncertainty over the distribution of tastes, even a rational player's perceived taste distribution will depend on her own type. Throughout the paper, I contrast results stemming from such

[13] 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 idiosyncratic farm characteristics have very slow adoption rates.

rational taste-dependent perceptions with those following from naive taste projection.[14] While rational beliefs always converge and never grow fully polarized, learning may still fail. With positive probability, agents converge to an interior "confounding" belief where they cannot discern, say, if many choose A because A is safe and most are risk averse, or because A is risky and most seek a high return. Smith and Sørensen (2000) show that under perfect information about the distribution, such beliefs exist only when quality differences between options are sufficiently large. In contrast, I show that with imperfect information, they *always* exist. This extension provides a natural explanation for persistent disagreement.[15] At a confounding belief, people with different tastes continually disagree on payoffs: if most choose A, then, relative to a return-seeking agent, a risk-averse agent thinks it is more likely that A is safe.

I conclude in Section 7 by discussing related models and by considering taste projection and social "mislearning" in broader contexts. In particular, I discuss why and how taste projection can distort inference in more general environments where agents can directly communicate beliefs or payoffs. For instance, imagine learning about a restaurant's quality from online reviews. Diners with "sophisticated" tastes may report mediocre quality from a meal that typical diners find remarkable. Hence, typical diners are misled when they underestimate how often they glean advice from sophisticates. I also discuss situations where agents have biased perceptions of the type distribution distinct from projection, such as a false sense of uniqueness. Finally, I highlight some shortcomings of the model and suggest avenues for future research.

2 Model

This section describes the basic decision environment and motivates a model of social taste projection. I then define a solution concept in the presence of projection which pins down beliefs about others' perceptions and strategies. Two immediate implications follow from these assumptions: (1) players with different tastes draw a distinct inference from any history of play, and (2) each player wrongly thinks all draw the *same* inference.

2.1 Social-Learning Environment

Actions and Payoffs. There are two actions $\{A, B\} =: \mathcal{X}$. Each action $X \in \mathcal{X}$ has quality $q^X \in \mathbb{R}$, and location $z^X \in \mathbb{R}$. Like standard spatial-differentiation models (e.g., Hotelling 1929; Downs 1957), each player prefers a higher "vertical" quality, but her preference over "horizontal" location depends on her type, $\theta \in \mathbb{R}$. For simplicity, I assume players' have a utility function separable in

[14] This model, analyzed in the Appendix, is identical to Section 6 aside from the assumption of full rationality.

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

quality and location:

$$u(X, \theta) = q^X - k(z^X - \theta)^2, \tag{1}$$

where $k > 0$ is a commonly-known preference parameter.[16] Like Downs' (1957) model of political competition, q may measure the competence of a candidate, while z indicates how liberal or conservative she is. An agent's type θ indicates her most preferred attribute $z \in \mathbb{R}$.[17]

States. Agents wish to learn the collection of each options' characteristics, $((q^A, q^B), (z^A, z^B))$. To make clear how taste projection affects learning, 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, $(z^A, z^B) = (1, -1)$. Let $\zeta \in \{L, R\}$ denote the "location state", where $\zeta = L$ if and only if A is left of B.[18] Furthermore, given the utility function above, only differences in quality matter for choice, not absolute levels. Hence, let $\Delta_q := q^A - q^B$ and denote by $\mathcal{D}$ the (finite) set of possible quality differences. The payoff-relevant state is $\omega = (\zeta, \Delta_q) \in \{L, R\} \times \mathcal{D} =: \Omega$. Agents share a common prior $\pi_1 \in \Delta(\Omega)$.

Preference Types. Players' preference types are i.i.d. and are distributed according to c.d.f. G with finite support $\Theta := \{\theta : \theta = \pm j\delta,\ j = 1, ..., J\}$, $\delta > 0$.[19] I call types $\theta < 0$ *left types*—they prefer the left option provided both options have identical quality; similarly, $\theta > 0$ are *right types*. As it plays a key role in inference, I denote by $\lambda := 1 - G(0)$ the measure of right types. With out loss of generality, assume right types comprise the majority: $\lambda > 1/2$.

Timing. In every period $t = 1, 2, ...$, a new set of $N \geq 1$ players is drawn, and each simultaneously takes an action $X \in \mathcal{X}$. Each player is labeled nt; t is the period in which she acts, and $n \in \{1, 2, ..., N\}$ is her label within that period. 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 the profile of actions observed in 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 (1) her preference type θ_{nt}, (2) a private i.i.d. signal $s_{nt} \in \mathbb{R}$ correlated with the state, and (3) the complete ordered history of actions, h_t. Her choice, which is based on a combination of this information, partially reveals her private signal to followers.

[16] 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 $d(z^X, \theta)$.

[17] For ease of exposition, I often refer to preference types simply as "types" despite the fact that a complete description of a type also includes a player's private information. To avoid confusion, I will be explicit whenever I refer to this complete notion of type.

[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, similarly focus on two feasible location profiles and note that additional states come at "significant algebraic cost."

[19] Aside from Section 6, I assume beliefs over the distribution of tastes, G, are degenerate. In Section 6, I introduce uncertainty over G, which implies that an agent's perceived distribution of tastes *rationally* depends on her own type θ. This generalization allows us to contrast learning in the presence of rational taste-dependent distributional beliefs with learning under taste projection.

For each $\omega \in \Omega$, let $\pi_t^\theta(\omega)$ denote a type-θ player's belief that the state is ω conditional solely on h_t and the prior; I call this the *public belief* in t.[20] Much of this paper analyzes the properties of each type's public-belief process, $\langle \pi_t^\theta \rangle$.

Finally, denote by Γ the game described above, and let $\Gamma(G)$ denote the game explicitly as a function of the taste distribution (keeping all other aspects fixed). Taste projection, which I introduce in the next section, will assume that one misperceives the taste distribution as $\widehat{G} \neq G$, but has an otherwise correct model of the game: her *perceived game* is $\Gamma(\widehat{G})$.

2.2 Taste Projection: Evidence and Model

This section reviews the literature motivating my main assumption of taste projection, and it provides a simple formulation of this bias consisting 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.

2.2.1 Motivating 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. 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.[21] Many similar studies followed that document this correlation across a wide range of domains, like preferences over political ideology and candidates, risk preferences, and preferences for income redistribution.[22]

[20] While I use the term "public belief" to match the literature, "public" in this context does not mean the belief is common across players. 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, as it is derived from publicly observable behavior.

[21] 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".

[22] 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. Across domains, 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.

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, Krueger and Clement (1994) and others provide evidence that this "bias" remains even when subjects have information about others' 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.[23] 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—that is, where the "Dawes critique" should have little bearing.[24]

Relatedly, economists have argued that *intra*-personal projection bias—exaggerating the degree to which future preferences resemble current preferences—influences behavior.[25] 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 po-

[23] 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.

[24] 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.

[25] 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.

tential buyers, causing sellers to set inefficiently high prices.

2.2.2 Perceived Distributions: Biased First-Order Beliefs

I model taste projection by assuming an individual's preference type θ influences her perceived distribution of types. In truth $\theta \sim G$. Denote type θ's perception of G by $\widehat{G}(\cdot|\theta)$. 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.

Assumption 1. (Stochastically Dominating Perceptions.) *$\widehat{G}(\theta|\theta')$ weakly first-order stochastically dominates $\widehat{G}(\theta|\theta'')$ if and only if $\theta' > \theta''$. That is, whenever $\theta' > \theta''$, $\widehat{G}(\theta|\theta') \leq \widehat{G}(\theta|\theta'')$ for all $\theta \in \Theta$.*[26]

The more right-leaning is an agent's taste, the higher is her estimate of those with right-leaning tastes. For example, people with conservative political views overestimate the share of those who prefer the conservative candidate, or those with high risk aversion overestimate the share seeking safe investment strategies. The perceived measure of right types—a key statistic in drawing inference from actions—is denoted by $\hat{\lambda}(\theta)$; dominance implies $\hat{\lambda}(\theta)$ is weakly increasing in θ.

For sake of intuition, I often consider a simple form of projection I call *choice-dependent* projection: one's perceived distribution depends only on her preferred location, not on the intensity of this preference. In this case, all left types think the distribution is $\widehat{G}_l$, whereas all right types think it's $\widehat{G}_r$, and the perceptions satisfy $\widehat{G}_l \succ G \succ \widehat{G}_r$.[27] This essentially implies 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 $\hat{\lambda}^r := 1 - \widehat{G}_l(0)$, and $\hat{\lambda}^l < \lambda < \hat{\lambda}^r$: left types underestimate the measure of right types, but right types *over*estimate it.[28] As I show in Section 4, two classes of choice-dependent projection will lead to very different learning outcomes. I define and differentiate them now.

[26] 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 $\widehat{G}(\theta|\theta')$ *strictly* first-order stochastically dominates $\widehat{G}(\theta|\theta'')$: $\widehat{G}(\theta|\theta') \leq \widehat{G}(\theta|\theta'')$ for all $\theta \in \Theta$ with strict inequality for some θ. Let $\succsim$ and $\succ$ denote weak and strict first-order stochastic dominance, respectively.

[27] 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.

[28] 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. 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 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

Definition 1. *Suppose players suffer choice-dependent projection where left and right types respectively believe the measure of right types is $\hat{\lambda}^l$ and $\hat{\lambda}^r$.*

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; each type thinks her own taste is most common. Under weak projection, all players correctly acknowledge that right types comprise the majority.

2.2.3 Naivete: Biased Second-Order Beliefs

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

Assumption 2. (Naivete.) *For all $\theta' \in \Theta$, a type-θ player believes $\widehat{G}(\cdot|\theta') = \widehat{G}(\cdot|\theta)$.*

This assumption pins down second-order beliefs—beliefs about others' perceived distributions. A player uses these second-order beliefs in thinking through how an observee learns from her predecessors, and hence in rationalizing that observee's action. Naivete implies that a player fails to engage with alternative ways of updating beliefs and instead thinks others' form beliefs as she does. For instance, a risk-averse agent who thinks 90% of investors are risk averse naively assumes that a risk-neutral agent thinks the same.

Naivete is the assumption differentiating "taste projection" from rational taste-dependent distributional beliefs, which arise whenever Bayesian players are uncertain of the distribution. In contrast to a taste projector, a rational agent knows precisely the map between an agent's type and her belief about the distribution, and hence accounts for the fact that other types interpret evidence differently than she does. More broadly, naivete departs from much of the literature on non-common priors, which assumes rational expectations about the *distribution* of priors across players (e.g., Harrison and Kreps; Morris 1996). As such, this paper explores one particular way in which people may neglect heterogeneity in beliefs.[29]

---

of $\tilde{\theta}$ such that

$$\tilde{\theta} > -\frac{\alpha}{1-\alpha}\theta. \tag{2}$$

The true measure is that of the set of types that satisfy Equation 2 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 2 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.

[29] Little work has been done in this area, however there are many domains where this form of neglect seems plausible

2.3 Naive Quasi-Bayesian Best Response

Aside from naive taste projection—incorrect first- and second-order beliefs about G—each Player nt satisfies the basic epistemic conditions governing play in a Bayesian Nash equilibrium of her perceived game, $\Gamma(\widehat{G}(\cdot|\theta_{nt}))$. First, each player is "quasi-Bayes rational" in that she maximizes expected payoffs given beliefs formed through putatively correct Bayesian updating using her (false) model of the world.[30] Second, each assumes common knowledge of Bayes rationality within her perceived game. Thus, a naive player correctly predicts others' strategies—the map $\sigma : \Theta \times \Delta(\Omega) \to \mathcal{X}$ from one's type and belief to an action.[31] But since she fails to account for others' discrepant models, she systematically mispredicts other types' *beliefs*. The model of non-rational play simply comprises a particular theory of how players form the incorrect beliefs against which they optimize.[32]

It's worth emphasizing some basic implications of these assumptions. Taste projection in social-learning environments implies that players who differ in taste draw different inference from the same history of play. It follows from projection (Assumption 1) that for any t, $\pi_t^\theta = \pi_t^{\theta'}$ if and only if $\widehat{G}(\cdot|\theta) = \widehat{G}(\cdot|\theta')$. Naivete (Assumption 2) further implies that each agent thinks her "public" belief π_t^θ is commonly shared. Simply put, agents unknowingly draw distinct beliefs from behavior. These implications suggest two ways in which a taste projector fails to understand the motives behind the actions she observes: she has incorrect theories of (1) predecessors' tastes, and (2) what predecessors have inferred from those moving before them.[33]

---

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." Although the literature on the false-consensus effect rarely elicits second-order beliefs, the few papers that do, including Egan, Merkle, and Weber (2014), find that people significantly overestimate how many share their second-order beliefs, which suggests at least some degree of naivete.

[30] 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.

[31] In this social-learning environment, this strategy is in fact the rational Bayesian-equilibrium strategy.

[32] 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.

[33] 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.

3 Learning Horizontal Attributes: Preliminaries

In this section and the next, I analyze learning about the horizontal locations of A and B when their quality difference is known. For simplicity, fix $\Delta_q = 0$. Players simply wish to learn state $\omega = \zeta \in \{L, R\}$—whether A is located to the left ($\omega = L$) or right ($\omega = R$) of B. For example, suppose the costs, q, of two risky technologies are known, but investors want to learn which is riskier. Suppose the horizontal dimension measures risk, and assets to the left are riskier, but have potential for higher return, than those to the right. Agents choose their best guess at the safer option if and only if their risk aversion is sufficiently high ($\theta > 0$).[34] The remainder of this section derives players' choice and inference rules, and discusses how projection distorts this inference rule. The implications of projection on long-run learning are analyzed in Section 4.

Private Information. Before acting, each Player nt observes a private signal $s_{nt} \in \mathbb{R}$ 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 supp(F), so that no signal perfectly reveals the state of the world.

Assumption 3. (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 4. (Unbounded Private Beliefs.) *For each ω,* co$(\text{supp}(F_\omega)) = [0, 1]$.

Assumption 3 implies private beliefs in favor of ω are relatively more likely whenever ω is true. Assumption 4 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 implies rational agents inevitably learn ω.[35]

Public Information and Individual Decision-Making. Prior to making a choice, each Player nt observes the history h_t, and computes public belief π_t, the probability of $\omega = R$ conditional on h_t. From private belief p_{nt} and π_t, she then forms posterior r that $\omega = R$ via Bayes' rule,

[34] Alternatively, the model captures agents learning about new technologies with known prices but unknown productivities that depend on θ. For example, Munshi (2003) describes hybrid seed with output dependent on soil or other input characteristics. Farmers wish to learn the seed type yielding the most productive match with their plot.

[35] 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.

$r(p,\pi) = p\pi/[p\pi+(1-p)(1-\pi)]$. Players maximize expected utility given this posterior, yielding the following decision rule: a right type chooses A iff $r(p,\pi) > 1/2$ whereas a left type chooses A iff $r(p,\pi) \leq 1/2$.

Observers draw inference about Player nt's private information p_{nt} from her action X_{nt} by inverting this decision rule to form cutoffs on p_{nt}: conditional on θ_{nt}, X_{nt} reveals if her private belief was above or below this cutoff.[36] I derive these 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 $\omega = R$. Now we can re-phrase the decision rule above as a cutoff strategy.

Lemma 1. *Let $p(\ell) := \ell/(1+\ell)$. Fixing public likelihood ratio ℓ, Player nt with private belief p has the following decision rule:*

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

The *private-belief threshold* $p(\ell)$ is the private belief that renders type θ indifferent between A and B given public likelihood ratio ℓ. Intuitively, to choose A, a "left" type must have a sufficiently strong private belief that A is located to the left—p is sufficiently low—whereas a "right" type must have a sufficiently strong private belief that A is located to the right—her p must be sufficiently high.

3.1 Belief Dynamics

This section derives equations describing the evolution of public likelihood ratios $\langle \ell_t^\theta \rangle$. As explained in Section 2.3, types with distinct perceptions of G draw different inference from history h_t, and thus 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 the case with choice-dependent projection, 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 transition equation $\ell_{t+1}^\theta = \varphi_\theta(a_t, \ell_t^\theta)$ specifying for each t how public beliefs update after observing actions a_t given public belief ℓ_t^θ. Since players update using Bayes' rule within their mispecified model, $\varphi_\theta(a_t, \ell_t^\theta) = \Psi_\theta(a_t, \ell_t^\theta)\ell_t^\theta$, where $\Psi_t(a_t, \ell_t^\theta)$ is the likelihood of

[36] While the solution concept implies that a naive projector correctly knows others' strategies, she mispredicts their private-belief thresholds since 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.

observing a_t in $\omega = L$ relative to $\omega = L$. Piecing these definitions together, beliefs move according to

$$\ell^\theta_{t+1} = \varphi_\theta(a_t, \ell^\theta_t) = \Psi_\theta(a_t, \ell)\ell^\theta_t = \frac{\psi_\theta(a_t \mid \ell^\theta_t, L)}{\psi_\theta(a_t \mid \ell^\theta_t, R)}\ell^\theta_t. \tag{3}$$

$\psi_\theta(a \mid \ell, \omega)$ in Equation 3 denotes the probability that a_t people choose A in state ω according to a θ-type player's incorrect model, which assumes that all players in t share public belief ℓ^θ_t and tastes have distribution $\widehat{G}(\cdot|\theta)$. Since behavior of each player in t is independent conditional on h_t, the perceived distribution of actions *within* a period is Binomial(N,$\alpha_\theta(\ell, \omega)$), where $\alpha_\theta(\ell, \omega) := \widehat{\Pr}_\theta(X_{nt} = A \mid \ell, \omega)$ is a type θ's perceived probability that a random player chooses A given ℓ and ω. Formally,

$$\psi_\theta(a \mid \ell, \omega) = \binom{N}{a}\alpha_\theta(\ell, \omega)^a\Big[1 - \alpha_\theta(\ell, \omega)\Big]^{N-a}, \tag{4}$$

and

$$\alpha_\theta(\ell, \omega) = \Big[1 - \hat{\lambda}(\theta)\Big]F_\omega\Big(p(\ell)\Big) + \hat{\lambda}(\theta)\Big[1 - F_\omega\Big(p(\ell)\Big)\Big]. \tag{5}$$

The first (second) term of Equation 5 is just the perceived measure of left (right) types times the perceived probability that a left (right) type takes A specified by Lemma 1. 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, $\boldsymbol{\ell}$:

$$\alpha(\boldsymbol{\ell}, \omega) = \Big[1 - \hat{\lambda}(\theta)\Big]F_\omega\Big(p(\ell^{\tilde{\theta}}_t)\Big) + \hat{\lambda}(\theta)\Big[1 - F_\omega\Big(p(\ell^{\tilde{\theta}}_t)\Big)\Big]. \tag{6}$$

Comparing Equations 5 and 6 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.

Remark on "Confounded Learning". The assumption that $\Delta_q = 0$ 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 my 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 2. *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, there exists $\tilde{\Delta}_q > 0$ such that for all $\Delta_q \in (-\tilde{\Delta}_q, \tilde{\Delta}_q)$, no confounding belief exists.

Appendix A 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 if a confounding belief exists. If one does, the logic under $\Delta_q = 0$ still holds, and results are identical aside from the possibility of convergence to the confounding belief. Consequently, results indicating the possibility, or impossibility, of society reaching some *confident* belief are unchanged by the presence of a confounding belief.

3.2 Effect of Taste Projection on Updating

This section analyzes comparative statics of $\hat{\lambda}(\theta)$ on the belief-transition equation $\ell_{t+1} = \varphi_\theta(a, \ell_t)$, making clear how taste projection distorts the interpretation of new evidence. The results established here play an important role in understanding the long-run dynamics studied in Section 4.

First, one's current belief and perception of tastes $\hat{\lambda}(\theta)$ dictates the interpretation of 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 3. *For each $\theta \in \Theta$ and perceived public likelihood ratio $\ell^\theta_t \in \mathbb{R}_+$, there exists a value $\kappa\left(\ell^\theta_t, \theta\right) \in (0,1)$ such that observation a_t is interpreted as evidence in favor of $\omega = R$ if and only if ($a_t/N > \kappa\left(\ell^\theta_t, \theta\right)$ and $\hat{\lambda}(\theta) > 1/2$) or ($a_t/N < \kappa\left(\ell^\theta_t, \theta\right)$ and $\hat{\lambda}(\theta) < 1/2$), where*

$$\kappa(\ell, \theta) = \left(1 + \log\left(\frac{\alpha_\theta(\ell, L)}{\alpha_\theta(\ell, R)}\right) \Big/ \log\left(\frac{1 - \alpha_\theta(\ell, R)}{1 - \alpha_\theta(\ell, L)}\right)\right)^{-1}. \tag{7}$$

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 return seeking must see sufficiently *few* choose A.

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, then what they subsequently observe maintains this confidence. For instance, when $\hat{\lambda}(\theta) > \frac{1}{2}$, $\lim_{\pi^\theta \to 1} \kappa(\pi^\theta, \theta) = \hat{\lambda}(\theta)$. This means that when a θ type grows nearly confident that $\omega = R$ (i.e., $\pi^\theta \approx 1$), she must observe at least $\hat{\lambda}(\theta)$ A's (on average) for her to remain confident that $\omega = R$ (i.e., for π^θ to stay near 1). But if the true fraction right types, λ, is observed and $\lambda < \hat{\lambda}(\theta)$, then π^θ instead moves downward from 1. Hence, observing exactly what a rational agent expects to see in $\omega = R$ can *reduce* the biased agent's confidence in $\omega = R$. This logic is central in understanding when some constellation of beliefs across types is stable, which is developed further in Section 4.

Figure 1 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 $\omega = L$.

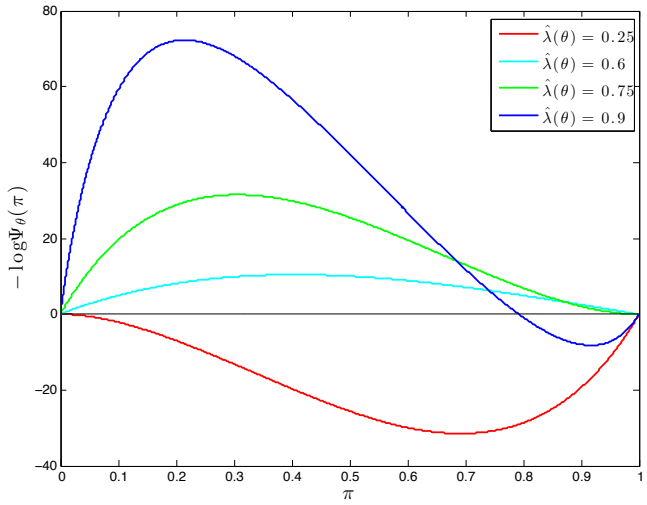

Figure 1: *Negative change the in public log-likelihood ratio,* $-\log \Psi_\theta(\pi)$*, as a function of the current belief,* π*, after observing action* $a_t/N = .75$ *for various values of* $\hat{\lambda}(\theta)$*. A type-*θ *player interprets* a_t *as evidence for* $\omega = R$ *if and only if* $-\log \Psi_\theta(\pi) > 0$*.*

Another implication of Lemma 3, evident from Figure 1, is what I call the *perceived-majority effect*: two agents who disagree on the majority preference may draw precisely opposite interpretations from the same observation.

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. *Suppose* $N = 1$*. If* $a_t = 1$ *then* $\ell_{t+1}^{\theta} < \ell_t^{\theta}$ *if and only if* $\hat{\lambda}(\theta) > \frac{1}{2}$*. Similarly, if* $a_t = 0$ *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, then one who thinks 60% are risk averse concludes A is likely safe, but another who believes only 40% are risk averse thinks A is risky. The corollary, which assumes agents act in single file ($N = 1$), is even more straightforward: an individual always interprets action 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 1); left and right types disagree on which hypotheses action A supports if and only if they suffer strong projection.

The next result, which I call the *variance effect*, describes how $\hat{\lambda}(\theta)$ affects the magnitude of changes in beliefs.

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

To interpret this result, note that $\hat{\lambda}(\theta)\left[1 - \hat{\lambda}(\theta)\right]$ is type θ's perception of the variance in tastes. Hence, Proposition 2 implies that as one's perceived variance in types decreases, her beliefs change by a greater amount after any observation. As perceived variance decreases, a player becomes more confident about the tastes of those whom she observes, so their choices are seemingly more precise signals of the their 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 under weak projection. In this case, the right-type belief changes by more than the rational belief after any action—beliefs are volatile, and over-responsive. The left-type belief, however, changes by too little relative to rational updating—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 since 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 it's roughly equally likely that her predecessor was risk averse or risk neutral, then his choice tells her relatively little about her own optimal strategy.

Figure 2 demonstrates both the "single-file" majority and variance effects. The plot shows the effect of observing choice A today on tomorrow's belief, π_{t+1}^θ, as a function of today's belief, π_t^θ. Each curves assumes a different values of $\hat{\lambda}(\theta)$. Comparing $\hat{\lambda}(\theta) = 0.25$ to the other cases highlights the perceived-majority effect: $\hat{\lambda}(\theta) < 1/2$ implies tomorrow's belief is lower than today's. We see the variance effect among the curves with $\hat{\lambda}(\theta) > 1/2$: the magnitude of changes in beliefs increases with $\hat{\lambda}(\theta)$.

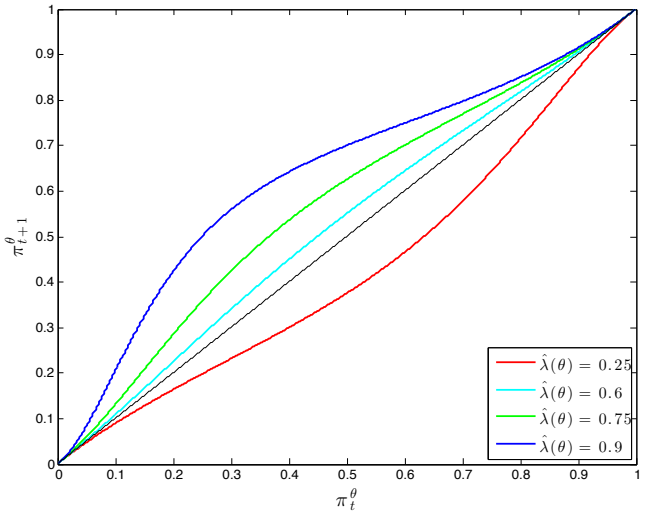

Figure 2: *Next-period's public belief π^θ_{t+1} as a function of the current public belief, π^θ_t, assuming choice A is observed in t. The $45°$-line is plotted for reference.*

4 Learning Horizontal Attributes: Long-Run Beliefs

Building on the setup of Section 3, this section investigates the effect of taste projection on long-run beliefs and behavior when players learn about horizontal locations. 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.

To arrive at these conclusions, Section 4.1 introduces the possible learning outcomes under projection, and 4.2 derives conditions on players' perceptions of λ that determine which equilibrium beliefs are stochastically stable. These conditions hold for a general model of perceptions, where each of an arbitrary finite number of types may hold a distinct perception $\hat{\lambda}(\theta)$. But to build intuition for the particular way in which learning fails as a function of the extent of projection, Sections 4.3 and 4.4 assume a simple two-type setting. There, a left and right type have distinct perceptions $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}^l, \hat{\lambda}^r)$ and beliefs, $\boldsymbol{\ell}_t = (\ell^l_t, \ell^r_t)$. Section 4.5 discusses how these "two-type" results generalize to cases with many perceptions.

4.1 Potential Learning Outcomes

I first introduce terminology for the various learning outcomes that can occur. Learning among type-θ players is (1) *complete* if π^θ_t converge almost surely to the truth; (2) *incorrect* if π^θ_t converges to certainty in a false state; (3) *incomplete* if π^θ_t 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, $\ell_t^\theta \xrightarrow{\text{a.s.}} 0$.[37]

As a benchmark, if people are fully-rational ($\hat{\lambda}(\theta) = \lambda$ for all $\theta \in \Theta$), then they necessarily learn the true state in the long run.

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

This result—first derived in Smith and Sørensen (2000)—follows from the martingale feature 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 $\omega = R$. By the Martingale Convergence Theorem, $\langle \ell_t^\theta \rangle$ converges almost surely to some random variable $\ell_\infty^\theta := \lim_{t\to\infty} \ell_t^\theta$.

With projection, however, public beliefs do not form a martingale:

Lemma 4. *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, $\psi_\theta\Big(a_t \mid \ell_t^\theta, R\Big)$, is generically not equal to the true probability. The true probability depends on *all* types' beliefs, $\boldsymbol{\ell}_t$, but a niave type θ thinks it depends solely on ℓ_t^θ.

Lemma 4 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 "candidate equilibria" such that *if* biased beliefs converge to a point belief, then these limit points must lie in $\mathscr{L}$.[38] Second, I evaluate whether these candidate equilibria are stochastically stable.

It turns out that $\mathscr{L}$ is the set of *confident beliefs*: $\boldsymbol{\ell}$ such that for each θ, $\ell^\theta \in \{0, \infty\}$. This means that if beliefs never converge to a fixed interior belief where people remain uncertain. To see this, first note that 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 $\{0, ..., N\} \times \mathbb{R}_+^{|\Theta|}$. Transitions along the θ-dimension given by

$$\ell_{t+1}^\theta = \varphi_\theta\Big(a, \ell_t^\theta\Big) \text{ with probability } \psi(a, \boldsymbol{\ell}_t) \quad (a = 0, ..., N), \tag{8}$$

where $\varphi_\theta(a, \ell)$ is the belief-transition function introduced in Section 3.1 (Equation 3) 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 $\hat{\ell}^\theta \in \text{supp}\Big(\ell_\infty^\theta\Big)$ is a fixed point

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

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

of the Markov process. Hence, for each component $\hat{\ell}^\theta$ of $\hat{\boldsymbol{\ell}} \in \text{supp}(\boldsymbol{\ell}_\infty)$ and all $a \in \{0, ..., N\}$,

$$\hat{\ell}^\theta = \varphi_\theta\Big(a, \hat{\ell}^\theta\Big). \tag{9}$$

Given the assumption of unbounded private beliefs, the only fixed points of process 8, and thus the only possible stationary beliefs, are confident beliefs.

Lemma 5. *Suppose that there exists a real, nonnegative random variable ℓ^θ_∞ such that $\ell^\theta_t \xrightarrow{\text{a.s.}} \ell^\theta_\infty$. Then* $\text{supp}\Big(\ell^\theta_\infty\Big) \subseteq \{0, \infty\}$.

From Lemma 5, $\mathscr{L} = \{0, \infty\}^{|\Theta|}$. 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. Lemma 5 implies 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 section (4.2) shows that agents' perceptions of population preferences, dictate which, if any, of these beliefs are asymptotically stable.

4.2 Stability of Confident Beliefs

This section derives, as a function of mispredictions, a condition specifying when a candidate equilibrium belief is locally stable. Section 4.2.1 derives sufficient conditions on the Markov process (8) for stability, and Section 4.2.2 establishes from these conditions a stability criterion based directly on the primitives of the model: each agent's perception of others' tastes, $\hat{\lambda}(\theta)$.

4.2.1 Local Stability of the Belief Process

Let $\hat{\boldsymbol{\ell}} \in \mathscr{L}$ denote a fixed point of process 8 with generic element $\hat{\ell}^\theta$.

Definition 2. *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 $\boldsymbol{\ell}_1 \in \mathscr{N}(\hat{\boldsymbol{\ell}})$.*

Stability means that if the belief process enters a neighborhood of the fixed point, then it will remain in that neighborhood with positive probability. Stability conditions follow from the logic of stability within linear systems. Although the belief process is nonlinear, near the fixed point we can approximate the process by its first-order Taylor series expansion; stability is assessed locally by applying standard linear-system criteria to this "linearized" approximation.

Formally, near fixed point $\hat{\boldsymbol{\ell}}$, type θ's belief process $\langle \ell^\theta_t \rangle$ is well approximated by the following stochastic difference equation: starting at $(a_t, \ell^\theta_t - \hat{\ell}^\theta)$, the continuation is $\Big(a_{t+1}, \frac{\partial}{\partial \ell}\varphi_\theta(a_t, \hat{\ell})(\ell^\theta_t - \hat{\ell}^\theta)\Big)$ with chance $\psi(a_t, \hat{\boldsymbol{\ell}})$. That is, the continuation is approximately the first-order Taylor expansion of $\varphi_\theta\Big(a_t, \ell^\theta_t\Big)$ 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 $y_t = b_0^{I_0(t)} \times ... \times b_N^{I_N(t)} y_1$ 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 fixed point $y = 0$:

$$\lim_{t\to\infty} y_t = \lim_{t\to\infty} \left(b_0^{p_0} \times ... \times b_N^{p_N}\right)^t y_1 = \lim_{t\to\infty} \chi^t y_1. \tag{10}$$

Clearly from Equation 10, 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

$$\chi_\theta(\hat{\boldsymbol{\ell}}) := \prod_{a=0}^{N} \left(\frac{\partial}{\partial \ell} \varphi_\theta\big(a, \hat{\ell}^\theta\big) \right)^{\psi(a,\boldsymbol{\ell})}. \tag{11}$$

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 (8) near $\hat{\ell}$.

Lemma 6. *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 6 is simply an extension of Smith and Sørensen's (2000) Theorem 4, which establishes this stability criterion for an arbitrary Markov process like (8) 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.

4.2.2 Characterization of Confident Equilibria

I now derive from Lemma 6 a stability criterion based directly on the primitives of the model—people's perceptions of others' tastes. This 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.[39]

This requires some final pieces of notation. Let $\widehat{\mathscr{F}_\theta} : \mathcal{X} \times \mathbb{R}_+ \to [0, 1]$ be θ'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

[39] 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 naivete" 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).

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

Proposition 4. *Let $\hat{\boldsymbol{\ell}} \in \mathscr{L}$ be a fixed point of the joint-belief process (8).*

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

$\mathscr{F}(A, \hat{\boldsymbol{\ell}})$ is the long-run frequency of action A if all players beliefs are fixed at $\hat{\boldsymbol{\ell}}$. 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 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. To remain confident, the fraction of times others' choose A must exceed $\hat{\lambda}^r$; if not, she necessarily becomes less confident over time. Essentially, observing a larger majority share than expected only reinforces a player's hypothesis about the action 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 supports the false hypothesis. Even if an 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 not all types can reach identical beliefs in the long run.

Proposition 5. *If people project tastes, i.e., there exist θ, θ' such that $\theta > \theta'$ and $\hat{\lambda}(\theta) > \lambda > \hat{\lambda}(\theta')$, then for each $\hat{\pi} \in \{0, 1\}$, $\Pr(\lim_{t\to\infty} \pi_t^\theta = \hat{\pi} \ \forall \ \theta) = 0$. That is, there is no long-run agreement across types.*

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

Corollary 2. *If people project tastes, i.e., there exist θ, θ' such that $\theta > \theta'$ and $\hat{\lambda}(\theta) > \lambda > \hat{\lambda}(\theta')$, then 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—then society observes roughly λ choose A. But people with the majority taste, who overestimate how many share this taste, 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 depend on whether projection is strong or weak (Definition 1). Section C of the Appendix also uses Proposition 4 to show how learning may fail when agents suffer alternative distributional errors distinct from projection, such as a false sense of uniqueness.

4.3 Strong Projection

This section (4.3) and the next (4.4) assume agents suffer "choice-dependent" projection. Hence, there are just two distinct perceptions of λ, $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}^l, \hat{\lambda}^r)$, and two distinct belief sequences, $\boldsymbol{\ell}_t = (\ell_t^l, \ell^\theta)$. This section studies learning under strong projection and the next analyzes weak (Definition 1). In each case, I identify $\Pi^*(\hat{\lambda}^l, \hat{\lambda}^r)$—the set of stable equilibrium beliefs given $\hat{\lambda}^l, \hat{\lambda}^r$—and show specifically how and why learning fails.

The proposition below shows that if each type thinks her taste is most common (strong projection), then people all choose the same action in the long run. Consequently, each grows confident that this action is optimal for her taste, resulting in polarized beliefs across types. To see this, consider two types of investors—risk averse ("right types") and risk neutral ("left types")—who 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: when A is safe, the risk averse expect around 80% to buy A, whereas the risk seeking expect about 40% to do so. If in fact A is safe, they observe 60%. Like in Proposition 1, the risk averse perceive this as evidence that A is safe, but the return seeking think it means A is risky. With these opposing beliefs, nearly *all* investors best respond in the next round by buying A, which only further polarizes the investors' beliefs. Eventually, all return seekers grow confident in the incorrect state.

Proposition 6. *Under strong taste projection, the following are true:*

1. $\Pi^*(\hat{\lambda}^l, \hat{\lambda}^r) = \{(0,1),(1,0)\}$.

2. *When N is finite, π_t^r converges almost surely to either 0 or 1, and each outcome arises with positive probability. If π_t^r converges to 0 (1), then π_t^l almost surely converges to 1 (0), and all players take action A (B) in the long run.*

3. *As $N \to \infty$, (π_t^l, π_t^r) converges almost surely to $(0,1)$; the majority type learns correctly, but the minority type learns incorrectly. All players take action A in the long run.*

The intuition is simple, and follows along the lines of the example above. Eventually some action, say A, earns a majority following. Projection implies that, absent strong contrary signals, each player believes it's the majority action that best suits her taste. As such, the frequency at

which A is chosen grows, reinforcing an observer's belief that A suits the more common taste—and hence *her* taste. It's worth noting that because players are naive, they don't understand that the herd results from opposing beliefs. While they think this "anomalous" herd on A is a highly unlikely event, A is any player's clear best response. The equilibrium is essentially self confirming: behavior following from polarized beliefs reinforces, and never contradicts, false beliefs.

With a finite number of players moving each round (N), either action may grow most popular in early periods. Hence, the action on which players inevitably herd is random.[40] Society suffers a form of "social" confirmation bias, where initial evidence has a lasting influence on long-run beliefs. Since people never expect a herd, the surprisingly uniform behavior moves them too quickly toward confident beliefs.[41] While either herd is possible, the majority type learns correctly more often when λ or N increases—these variables increase the likelihood that the action optimal for the majority taste, A, is most popular among early periods. As $N \to \infty$, agents almost surely herd on A.

Strong projection leads to an extremely robust 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, eliminating heterogeneity in long-run behavior. In such a "uniform herd", agents inefficiently over adopt the more popular action. Sorensen (2006) finds an example of this among workers within an academic department who observe others' choice of health-care plans before selecting their own. While employees differ significantly in their preferred plan characteristics, they tend to herd on a single plan.[42] Many employees later switch, reflecting the heterogeneity in the optimal match.

Uniform herding implies that observing others can be socially harmful. For sufficiently precise private signals, people are necessarily worse off by observing others than if they simply followed private information. Depending on which action people herd, when observing others a share $\nu \in \{1-\lambda, \lambda\}$ correctly learns, while fraction $1-\nu$ chooses the inferior option. Instead, an agent choosing solely on private information does so correctly with probability $\rho := 1 - F_R(1/2)$. So

[40] 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.

[41] 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.

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

long as

$$\rho > \frac{\lambda \mathbb{E}\Big[u(A,\theta) - u(B,\theta) \mid \theta > 0\Big]}{\lambda \mathbb{E}\Big[u(A,\theta) - u(B,\theta) \mid \theta > 0\Big] + (1-\lambda)\mathbb{E}\Big[u(B,\theta) - u(A,\theta) \mid \theta < 0\Big]} = \frac{\lambda \mathbb{E}\Big[\theta \mid \theta > 0\Big]}{\lambda \mathbb{E}\Big[\theta \mid \theta > 0\Big] - (1-\lambda)\mathbb{E}\Big[\theta \mid \theta < 0\Big]}, \quad (12)$$

observing others reduces social welfare. With only two types, $\theta \in \{-1, 1\}$, condition 12 reduces to $\rho > \lambda$: social learning is harmful whenever the probability that an agent has a correct signal exceeds the chance a random other shares her taste. Finally, the welfare loss from social learning is asymmetric, as it falls entirely on agents with a particular taste. For large N, the burden falls entirely on those with the minority taste.

It's worth noting why strong projection precludes agents from converging to identical beliefs.[43] If society is nearly certain that $\omega = R$, then fraction λ chooses A. Left types, who think they're most common, think this suggests $\omega = L$. This reduces their confidence that $\omega = R$. More generally, since biased beliefs do not form a martingale, they display predictable drift. Near the truth $\hat{\boldsymbol{\ell}} = (0, 0)$, both ℓ_t^l and ℓ_t^r are strict submartingales: they increase in expectation over time, and hence drift away from the truth.

Lemma 7. *Assume strong taste projection. There exists a neighborhood $\mathscr{N}$ about the truth $\hat{\boldsymbol{\ell}} = (0, 0)$ such that for all $(\ell_t^l, \ell_t^r) \in \mathscr{N}$, $\mathbb{E}[\ell_{t+1}^r \mid \ell_t^l, \ell_t^r] > \ell_t^r$ and $\mathbb{E}[\ell_{t+1}^l \mid \ell_t^l, \ell_t^r] > \ell_t^l$.*

In terms of Proposition 4, near $\hat{\boldsymbol{\ell}} = (0, 0)$ each player sees fewer than expected choose the action she thought would be most popular. Figure 3 shows the drift in beliefs for all regions of the joint-belief space.[44] Beliefs drift *away* from each fixed point where opposing types agree, $\hat{\boldsymbol{\pi}} = (0, 0)$ and $\hat{\boldsymbol{\pi}} = (1, 1)$, but drift toward confident disagreement.

[43] This is a direct consequence of Proposition 5. This discuss provides intuition for that proposition and why different types fail to agree in the long run.

[44] As shown in Figure 3, there are four regions of "belief space", $[0, 1]^2$, 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 3 (and also Figure 4) show L_θ in units of probabilities rather than likelihood ratios. That is, the figures plot $P_\theta(\pi) := L_\theta\big(\pi/(1-\pi), \hat{\lambda}(\theta)\big)/\big[1 + L_\theta\big(\pi/(1-\pi), \hat{\lambda}(\theta)\big)\big]$.

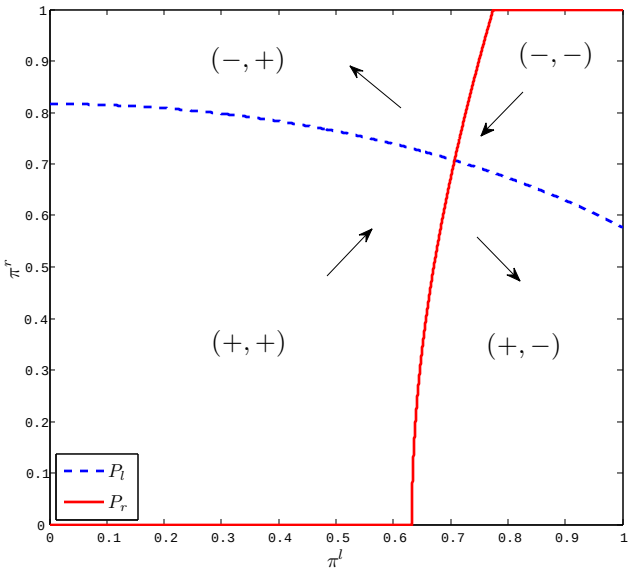

Figure 3: *Direction of drift at each point in the joint-belief space assuming "strong" taste projection.*

4.4 Weak Projection

Although "strong" projection suggests a clear logic for herding, uniform behavior is not a general consequence. When projection is weak enough so that people agree on the majority preference, players never settle on a fixed belief nor 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.

Proposition 7 implies that beliefs of each type almost surely fail to converge to a fixed value. This results from agents never observing a pattern of behavior consistent with either state given their mispecified model. As such, beliefs perpetually oscillate from favoring one state to the other. Since an agent's belief forms a martingale with respect to her *own* model, she wrongly anticipates that her opinion will eventually settle down. Instead, whenever it begins to settle down, she observes new, "shocking" evidence (with respect to her model) that pushes her back toward uncertainty.

The rationale stems from the "variance effect" discussed in Section 3.2. Since right types overestimate their frequency, $\hat{\lambda}^r > \lambda$, and consequently underestimate variance in tastes, they think actions reveal more private information than they do. 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.[45] Importantly, contrarian actions are overattributed to private information rather than taste.

[45]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_t^θ.

To see this in terms of the investment example, suppose traders initially believe A is safe: $\ell_t^\theta \approx 0$ for each θ. At first, they observe a market share for A equal to λ, the share of risk-averse traders. A risk-averse investor, however, expects a share near $\hat{\lambda}^r > \lambda$ and thus observes roughly $\hat{\lambda}^r - \lambda$ more contrarian choices (in frequency) than anticipated. She must form a theory explaining this excess demand for B. Within her model, the most likely scenario is that a significant share of fellow risk-averse investors received strong private information that B is in fact safer. From this, the risk-averse trader concludes that society was likely misled when it previously decided that A is safer than B.

This logic implies that when both types have beliefs near the truth, the belief of the majority type ℓ_t^r evolves as a submartingale. That is, ℓ_t^r drifts away from zero toward less confident beliefs. On the other hand, a left type observes more A's, and thus *fewer* contrarian actions, than anticipated. This reinforces her belief in $\omega = R$. Locally, ℓ_t^l is a supermartingale—left-type beliefs move toward zero in expectation.

Lemma 8. *Assume weak taste projection. There exists a neighborhood $\mathcal{N}$ about the truth $\hat{\boldsymbol{\ell}} = (0,0)$ such that for all $(\ell_t^l, \ell_t^r) \in \mathcal{N}$, $\mathbb{E}[\ell_{t+1}^r \mid \ell_t^l, \ell_t^r] > \ell_t^r$ and $\mathbb{E}[\ell_{t+1}^l \mid \ell_t^l, \ell_t^r] < \ell_t^l$.*

Lemma 8 confirms that the two types' beliefs move in opposite directions when both are initially quite certain of the state, but how do these dynamics play out in the long run? Like ℓ_t^r, ℓ_t^l must eventually move away from 0. Suppose instead that ℓ_t^l remained near 0 for all t. Since (1) ℓ_t^r is a submartingale conditional on $\ell_t^l = 0$ and (2) the only fixed points of $\langle \ell_t^r \rangle$ are 0 and ∞, it must be that ℓ_t^r must diverge to infinity. Hence, the frequency of action B converges to 1. But a left type is aware she is in the minority; an arbitrarily long herd on B must eventually cause her to think $\omega = L$. This logic makes clear that while right-type beliefs move from favoring $\omega = R$ to $\omega = L$, their resulting behavior compels left types to similarly revise their beliefs. But once all agree that $\omega = L$, the above logic repeats: right-type beliefs drift back toward uncertainty. No matter which state society agrees on, no action ever gains as much support as the majority anticipates. As a result, the majority never grows confident of their optimal action.

Figure 4 depicts this logic by plotting the expected drift in biased beliefs for all regions of the joint-belief space. Beliefs drift away from each fixed point and do so in a particular way: behavior near each potential equilibrium reinforces the beliefs of some types, but deteriorates the confidence of others.

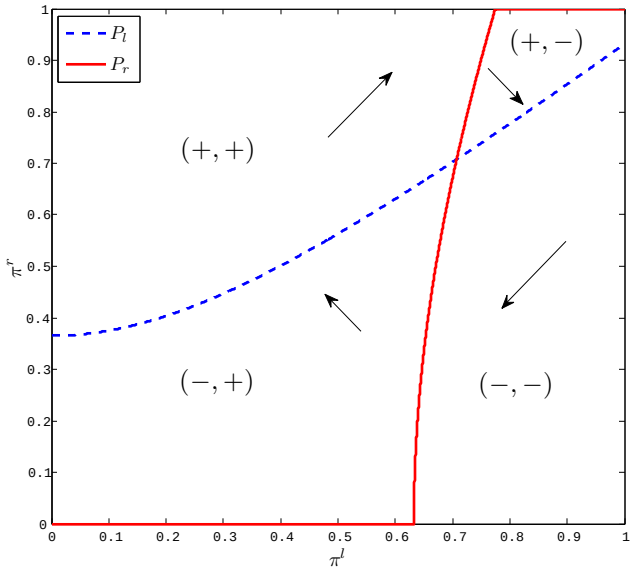

Figure 4: *Direction of drift at each point in the joint-belief space assuming "weak" taste projection.*

Weak projection generates persistent opinion fluctuations where society alternates between supporting $\omega = R$—where most people choose A—and supporting $\omega = L$—where most choose B. As such, behavior resembles "fads". Although common, such behavior is not well explained by rational learning models in settings with strongly connected networks or "unbounded" private information. For example, fad-like behavior arises in Çelen and Kariv (2004) only if rational players both observe a subset of predecessors and receive boundedly-informative private signals. Acemoglu, Como, Fagnani and Ozdaglar (2012) suggest a naive model of learning in a network where some agents are "stubborn" and never update their beliefs. Acemoglu, et al. (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 mistake the "surprising" amount of heterogeneity in choice for revelation of new private information.

Weak projection harms social welfare whenever beliefs spend a significant proportion of time below $1/2$. To determine when this occurs, we must study the long-run distribution of beliefs, which 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. A particularly interesting case arises when $\hat{\lambda}^r$ is sufficiently close to λ. Specifically, when $\hat{\lambda}^r \in (\lambda, \bar{\lambda}^r)$ where $\bar{\lambda}^r := 1 - \left(\frac{1-\lambda}{\lambda}\right)\hat{\lambda}^l$, simulations confirm that each belief process oscillates between increasingly confident beliefs in the two states. Figure 5 depicts a simulated sample path of $\log \ell_t^\theta$ in this case. When society is confident that $\omega = R$, all choose optimally—fraction λ, all

right types, take A. But when confident that $\omega = L$, all choose incorrectly—fraction $1 - \lambda$, all *left* types, take A. Figure 6 displays these swings in behavior: the frequency of choice A oscillates between λ and $1 - \lambda$.

When society has such "cyclical beliefs", expected welfare is lower than if people simply ignored others' actions. Roughly 50% of the time, a player forms nearly confident, but false, beliefs. 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.

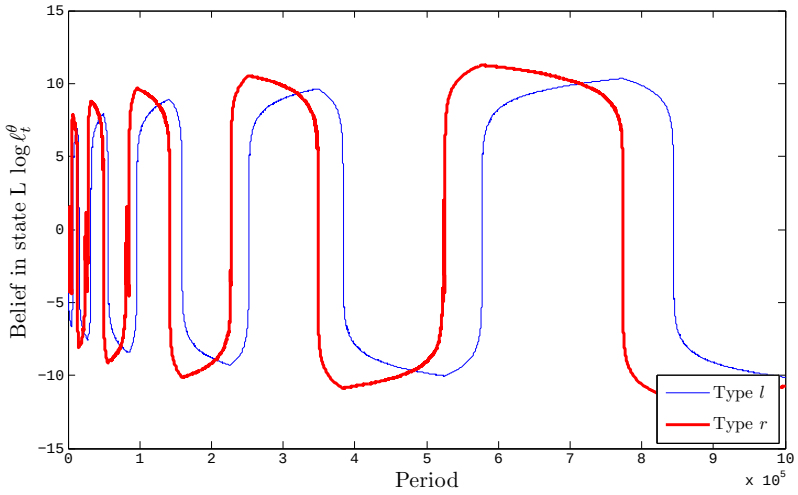

Figure 5: *Sample path of log-likelihood ratios for* $\lambda = 0.75$, $\hat{\lambda}^l = 0.55$, *and* $\hat{\lambda}^r = 0.8$.

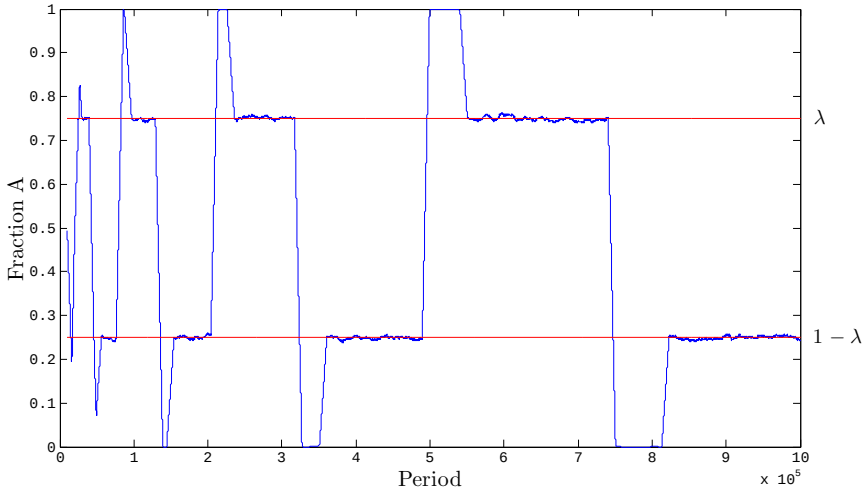

Figure 6: *Sample path of log-likelihood ratios for* $\lambda = 0.75$, $\hat{\lambda}^l = 0.55$, *and* $\hat{\lambda}^r = 0.9$.

4.5 Biased Learning Under General Taste Projection

This section characterizes learning outcomes when misperceptions may differ across an arbitrary finite number of types. The previous two sections assumed "choice-dependent" projection, which imposed that all players on a particular side of the taste spectrum share a common perception of λ. The only assumption on perceptions I make here is Assumption 1, so $\hat{\lambda}(\theta)$ is monotonically increasing in θ.

Proposition 8 provides necessary and sufficient conditions on perceptions for the existence of stable equilibria. Let W denote the share of types who wrongly think left types comprise the majority. Let $\tilde{\theta} := \arg\max_\theta \hat{\lambda}(\theta)$ subject to $\hat{\lambda}(\theta) < 1/2$ denote 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*

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 in each inequality of 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 size of the majority. So long as *all* agents observe more people than they anticipated choosing a single action, then the equilibrium is stable. In any 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. Interestingly, it's those with the most opposed tastes who choose identically. In contrast, it's 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 a general implication of a stable projection equilibrium, and strong projection demonstrates a particular 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 in Sections 4.3 and 4.4 accurately capture the essence of learning with projection, albeit in extreme fashion.

5 Learning About Quality

In many settings, beliefs about commonly-valued quality may be a greater determinant of choice than perceptions of horizontal location. For instance, imagine learning about candidates in a primary election. Early poll results reveal information about both a candidate's ideology and her competence. If a voter infers a large difference in competence between two candidates, she may choose to support the more competent one even if he is farther from the voter's preferred ideology. In this section, I consider such settings where quality differences may be large enough so that all players prefer the same option despite heterogeneity in tastes. How does taste projection distort inference about quality?

I explore two ways in which taste projectors misjudge relative quality. With two types, if people display strong projection, then society necessarily comes to believe that the quality difference between A and B is as large as possible. Such mislearning both arises from and perpetuates a herd in which all players choose the option optimal for the majority taste. With a continuum of types, stable long-run behavior implies a negative relationship between tastes and perceived quality: people who most prefer the horizontal attributes of A underestimate the quality of A relative to those who prefer the horizontal attributes of B.

5.1 Preliminaries

States. Players wish to learn both (1) the location $\zeta \in \{L, R\}$, and (2) the quality difference $\Delta_q \in \mathcal{D}$ of A relative to B. $(\zeta, \Delta_q) \in \{L, R\} \times \mathcal{D}$. Let $\underline{\Delta} := \min \mathcal{D}$ 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 $\overline{\theta} = \max \Theta$; all prefer B if $\Delta_q < -\widehat{\Delta}$. I call state $\omega = (\zeta, \Delta_q)$ *universal* whenever $|\Delta_q| > |\widehat{\Delta}|$ so all prefer the same action. Assume universal states are possible: $\overline{\Delta} > \widehat{\Delta}$ and $\underline{\Delta} < -\widehat{\Delta}$. Players expect to observe long-run uniform behavior if and only if ω is universal.

Private Information. For simplicity, I assume a unidimensional signal structure in which each player receives a signal informing them her action is optimal for her own taste. For each θ, let $\Omega_\theta \subset \Omega$ denote the set of states in which type θ weakly prefers action A. Each type-θ player receives an i.i.d. private belief that $\omega \in \Omega_\theta$ drawn according to c.d.f. F_A if $\omega \in \Omega_\theta$ or c.d.f. F_B otherwise. F_A and F_B meet the same assumptions as F_H and F_L, respectively (Assumptions 3 and 4). While private information alone leads to coarse inference over Ω, the signal structure allows agents to discern which action is optimal for each type when observing others. I assume this signal structure, which implies that agents follow decision rules analogous to those derived in Section 3, only for ease of exposition. The structure still allows rational agents to learn the optimal action, and I emphasize below that it does not drive any of the incorrect-learning results.

5.2 Quality Distortions with Two Types

In this section, I show that if two types suffer strong projection, then society necessarily comes to believe in a universal state. Thus, differences in vertical quality are always weakly exaggerated.

Suppose there are two types—a left type ($\theta = -1$) and a right type ($\theta = 1$). $\lambda := \Pr(\theta_{nt} = 1)$ denotes the fraction of players with right-leaning tastes.[46] With two types, observing a history generates public beliefs over a partition of Ω comprised of four elemnts. Essentially, players can only determine which action is optimal for each type. Hence, each partition element—denoted $\Omega^{XX'}$ for $X, X' \in \{A, B\}$—contains all states in which it is optimal for left types to chose X and right types to choose X'. Let $\pi_t^\theta(XX')$ denote type θ's belief that the state is in $\Omega^{XX'}$ 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. *Under strong taste projection, the following are true:*

1. *Any joint belief with $\pi^r(AB) = \pi^l(AB) = 1$ or $\pi^r(BA) = \pi^l(BA) = 1$ is unstable. Hence, 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 conclude that the quality difference is large enough so that all players prefer the same action.*

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 4.3). That is, 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, and a uniform herd on A results.

Part 2 of Proposition 9 is a consequence of how agents explain this uniform herd. Since quality differences might be large, agents have a perfectly good explanation: if all choose A, it must simply be that A has high 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, implies that belief dynamics are deterministic, and is assumed merely to make precise claims about limit beliefs.

Under strong projection, payoff differences along the quality dimension are always perceived as greater than those along the idiosyncratic "horizontal" dimension. This misperception has important consequences in markets for niche goods that are appealing only to a minority of consumers. In

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

such cases, low demand over attributed to poor quality, rather than limited appeal. Thus, even those who would enjoy the good assume it's not worthwhile.[47] Further, this is independent of priors: even if a universal state is very unlikely, people still conclude one option has such a large quality advantage that all should choose it irrespective of taste.

This result adds a new perspective to puzzles surrounding the slow adoption of useful agricultural practices. Consider a setting where farmers learn whether to adopt a new strain of hybrid rice (A) or use the status-quo crop (B). Hybrid rice grows well only in specific types of soil; for instance, some strains require either high or low salinity (Munshi, 2003). Suppose in truth this seed is only worthwhile for low-salinty farms, which comprise 40% of the region. But farmers don't know this: they learn about the optimal soil by observing how many others have adopted. Nor do they know the potential yield of the new seed. It's conceivable that even when sowed in suboptimal soil, the hybrid may trump the alternative. Before investing in the new crop, farmers cultivate a small test plot—they have noisy signals about the match between the seed and their farm. Initial adoption is based on this private information. In $t = 2$, additionally use 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. They've concluded that the new technology is globally, rather than selectively, ineffective.

5.3 Quality Distortions with Many Types

With many types, there may exist long-run stable equilibria in which those with different tastes choose different options. In such cases, agents hold correct beliefs along the horizontal dimension, but display an interesting form of mislearning along the quality dimension. Instead of universally concluding one action has superior quality, perceptions of quality are negatively correlated with tastes. Specifically, if people 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. This section, like those before, reiterates an implication of projection: people 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.

There is a simple logic for why projection induces negative correlation between quality perception and taste. Suppose agents wish to learn the future health benefits of exercise. People vary in how pleasurable—or painful—they find exercise, but each person knows their own idiosyncratic taste. Upon observing how many others regularly attend the gym, exercise fans—who overestimate

[47] 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).

the share with similar taste—find attendance lower than expected. They attribute this in part to its health benefit, and conclude these benefits must be limited. Those who find the gym particularly unpleasant draw precisely the opposite conclusion. They see more than expected attending the gym—they think most find exercising a painful endeavor—and thus infer that the health benefits must be high.

To show this result formally in the domain of choice-dependent projection with a continuum of types, $\Theta = [\underline{\theta}, \overline{\theta}]$. Suppose $\theta < 0$ think $\theta \sim \widehat{G}_l$ and $\theta > 0$ think $\theta \sim \widehat{G}_r$; $\widehat{G}_r$ dominates $\widehat{G}_l$ in the sense of FOSD. 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. While I assume just two distinct perceptions of G for simplicity, it will be clear how the logic of the equilibria discussed here extends to the case where each type may hold a distinct perception.

Suppose that in the long-run, a fraction α of players choose A. When agents face no uncertainty over location—they know $\zeta = R$—how does each type rationalize α? Given their differing beliefs about how many have right-leaning tastes, different types must form conflicting theories of Δ_q. Denote by Δ_q^l and Δ_q^r the perceived quality differences of left and right types, respectively. All players correctly understand that the marginal type—the type indifferent between A and B—is $\hat{\theta} = -\Delta_q/4k$ and that those who choose A have $\theta \geq \hat{\theta}$. People simply use the wrong model when deciding how many players have taste $\theta > \hat{\theta}$. Letting $\hat{\theta}^l$ and $\hat{\theta}^r$ denote each type's perception of $\hat{\theta}$, equilibrium requires $\alpha = 1 - \widehat{G}_l(\hat{\theta}^l)$ and $\alpha = 1 - \widehat{G}_r(\hat{\theta}^r)$. Thus left and right types respectively conclude $\Delta_q^l = -4k\widehat{G}_l^{-1}(1-\alpha)$ and $\Delta_q^r = -4k\widehat{G}_r^{-1}(1-\alpha)$. Since $\widehat{G}_r(x) \leq \widehat{G}_l(x)$, it follows that $\Delta_q^r \leq \Delta_q^l$.

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 $\alpha = 1 - \widehat{G}(\hat{\theta}(\theta)|\theta)$ for all θ, where $\widehat{G}(\cdot|\theta)$ is a type θ's perceived distribution and $\hat{\theta}(\theta)$ is her perception of the marginal agent. This condition does not necessarily hold—existence requires the speed at which $\widehat{G}(\cdot|\theta)$ varies across θ to be small.[48] However, in any such equilibrium, it's clear that the perceived quality advantage of A is decreasing in type. To see this, let $\Delta_q(\theta)$ denote type θ's perception of Δ_q. $\alpha = 1 - \widehat{G}(\hat{\theta}(\theta)|\theta)$ implies $\hat{\theta} = \widehat{G}^{-1}(1-\alpha|\theta)$, and using $\hat{\theta} = -\Delta_q(\theta)/4k$ yields

$$\Delta_q(\theta) = -4k\widehat{G}^{-1}(1-\alpha|\theta). \tag{13}$$

By first-order stochastic dominance (Assumption 1), $\widehat{G}^{-1}(1-\alpha|\theta)$ is increasing in θ, and thus $\Delta_q(\theta)$ is decreasing in θ.

[48] Specifically, a sufficient condition is that $-\widehat{G}^{-1}(1-\alpha|\theta) + \theta$ is increasing, hence $\frac{\partial}{\partial\theta}\widehat{G}^{-1}(1-\alpha|\theta) < 1$, on Θ.

6 Learning About Preferences

This section explores learning about horizontal differentiation, as in Sections 3 and 4, among agents who revise their models of others' preferences after observing actions. Until now, I assumed agents have fixed perceptions: they believe the distribution of tastes (which they mispredict) is perfectly known by all agents.[49] This section considers a more realistic model where all agents perceive some uncertainty over the distribution and learn about others' tastes through their actions. If the true taste distribution lies in the support, will 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, I assume agents with different tastes *rationally* form divergent priors over the distribution. But a naive agent errs by assuming all share her prior. She thus develops incorrect beliefs about what other types infer. This demonstrates that it's not heterogeneous priors, per se, that lead agents astray, but rather their neglect of others' discrepant beliefs. I show that a particular class of priors can cause agents to become fully biased in their perceptions of others tastes. That is, each wrongly concludes that all players share her preference.

Subsection 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 6.2 explores properties of biased long-run learning.

6.1 Uncertainty Over the Taste Distribution

Consider the model of Section 3 and 4 with no uncertainty over quality. Suppose there are two types—a left type ($\theta = -1$) and a right type ($\theta = 1$).[50] $\lambda := \Pr(\theta_{nt} = 1)$ denotes the fraction of players with right-leaning tastes. Learning the taste distribution entails estimating then 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 $\{L, R\} \times \Lambda$, consisting of payoff states, $\omega \in \{L, R\}$, and distribution states, λ_k. Denote by $\pi_t^\theta(\omega, \lambda_k)$ a θ type's public belief that the state is (ω, λ_k) upon observing h_t. Without loss of generality, suppose the state is (R, λ^*) for some $\lambda^* \in \Lambda$. Finally, let the conditional distributions of private beliefs F_ω meet Assumptions 3 and 4 from Section 3.

Priors. Assume $\Pr(\omega = R) = 1/2$, and let $\mu_0 \in \Delta(\Lambda)$ denote the prior over λ. Since one's

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

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

taste is information about λ, each type updates μ_0 using θ according to Bayes' rule. Let μ^θ denote each types' revised "type-dependent" prior.[51] I model taste projection as a biased perception of revised priors over λ. As before, I assume agents are naive in that they neglect that players with differing taste have different priors. A θ type thinks *all* players share her prior μ^θ regardless of their tastes.[52] 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.

Assuming naivete—that each thinks others' priors exactly match her own—is admittedly strong.[53] It is, however, stronger than necessary, and I assume this only because the error is particularly simple. The results below essentially rely on players (unknowingly) inferring too much, relative to a Bayesian, from their own tastes. Hence, they underappreciate the extent to which priors differ across types. I examine the extreme case in which players think priors don't differ at all.

Decision Making and Updating. While beliefs about λ dictate the interpretation of actions, an individual's decision relies solely on her belief about ω.[54] Denote this belief, the marginal probability of $\omega = R$, by $\pi_t^\theta := \sum_k \pi_t^\theta(R, \lambda_k)$. Since a naive agent thinks all share her prior, she assumes all types share her public belief $\pi_t^\theta(\omega, \lambda_k)$ in each state, for all h_t. The key difference between rational and naive updating is that a rational player has correct second-order beliefs. Hence, she knows that left and right types have different public beliefs.

6.2 Biased Long-Run Learning

This section shows that naivete—incorrect second-order beliefs about λ—can generate polarized beliefs about ω and λ. 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, all players take the same action. In explaining this herd, agents' perceptions of others' tastes also polarize: each thinks the herd indicates that her taste is most common. I provide sufficient conditions on priors guaranteeing that such outcomes occur with positive probability.

Before turning to formal results, I first provide some intuition. Uniform herding can occur whenever a herd on an action, say A, is forever "polarizing" This means that left and right types always (unknowingly) disagree on the interpretation of A no matter how often it is played. Hence, the herd on A leads left and right types to believe $\omega = L$ and $\omega = R$, respectively. To check

[51] Specifically, for θ and λ_k, $\mu^\theta(\lambda_k) = \Pr(\lambda_k \mid \theta)$. So, $\mu^r(\lambda_k) = \lambda_k\mu_0(\lambda_k)/\sum_i \lambda_i\mu_0(\lambda_i) = \lambda_k\mu_0(\lambda_k)/\mathbb{E}[\lambda]$ and $\mu^l(\lambda_k) = (1-\lambda_k)\mu_0(\lambda_k)/\sum_i(1-\lambda_i)\mu_0(\lambda_i) = (1-\lambda_k)\mu_0(\lambda_k)/(1-\mathbb{E}[\lambda])$.

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

[53] This notion of naivete is consistent with the earlier definition (Assumption 2). 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 2 follows from this definition in settings with no uncertainty.

[54] Agents follow the same decision rule as in Section 3 (Lemma 1).

whether this is possible within a given environment, first suppose people act in single file and let $\pi_t^\theta(h_t^A)$ be type θ's belief that $\omega = R$ entering period t following history h_t^A, where h_t^A is a history of length $t-1$ consisting of all A's. Then fully-polarized beliefs may occur if for all $t \in \mathbb{N}$, $\pi_{t+1}^r(A, h_t^A) > \pi_t^r(h_t^A)$ and $\pi_{t+1}^l(A, h_t^A) < \pi_t^l(h_t^A)$. That is, each types' beliefs are monotonic in h_t^A.

Rational beliefs, of course, never satisfy this condition. However, they may be polarized by *finite* sequences of A's. To see this, consider an investment setting where the fraction of risk-averse agents is $\lambda \in \Lambda = \{\frac{1}{4}, \frac{3}{4}\}$ with prior $\mu_0(3/4) = 0.6$. After using their own risk preference as information, risk-averse and risk neutral agents respectively think $\mu^r(3/4) \approx 0.82$ and $\mu^l(3/4) = 1/3$. Hence, initially, each type *rationally* believes her taste is most common. If the first investor chooses A, each agent reasons that the investor likely shared her taste. And so a rational risk-neutral agent infers A is likely risky, whereas the risk-averse infer A is likely safe. Action A temporarily polarizes beliefs.

But so long as agents are rational, A cannot forever polarize beliefs. Since agents have correct second-order beliefs, they know exactly what people of opposite tastes infer actions. A rational player cannot grow confident of some hypothesis while fully aware that another rational agent is confident of an alternative hypothesis.[55] In the example above, observing a second A reveals little information—all know that each type likely chooses A irrespective of their private signal. As such, after a long enough sequence of A's, people eventually rely on the original prior μ_0 to draw conclusions instead of their taste dependent prior. All people eventually agree that a long sequence of A's is strong evidence for $(R, 3/4)$.

Naive agents aren't so clever. In the example above, naive players neglect that the first A sends beliefs in opposite directions. Hence, upon observing a second A, observers fail to limit their inference. Instead of understanding that *each* type is inclined to pick A in $t = 2$, a naive risk-averse agent overestimates the likelihood that the second A results from a fellow risk-averse agent with private information that A is safe. This over-inference from relatively uninformative behavior sends naive beliefs of each types toward opposite extremes.

In general, if actions can have a lasting polarizing effect, then with positive probability agents with different tastes converge to confident beliefs in opposite payoff states and a herd results. This happens on sample paths that begin with a long sequence of A's. From this "initial condition" in which people unknowingly disagree on the state, most continue to choose A—risk neutral grow confident A is risky and most are risk neutral all thew while the risk averse are confident it's safe and that they comprise the majority preference. I show that these polar-opposite beliefs are stable: they lead all agents to play A with high probability, which only strengthens players' beliefs. Thus,

[55] Although Acemoglu, Chernozhukov, and Yildiz (2007, 2009) show that rational agents may "agree to disagree" on the interpretation of an infinite sequence of evidence, players in their model never fully disagree on the state.

so long as people can reach a neighborhood of the polar-opposite beliefs, then they may forever remain at there.

6.2.1 Two-Point Taste 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 naive player comes to believe after observing an arbitrarily long herd on A as a function of her prior. In this setting with $|\Lambda| = 2$, let μ^θ denote type θ's perceived probability that $\lambda = \overline{\lambda}$.

Lemma 9. *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 9 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 $\underline{\lambda} > 1/2$ or $\underline{\lambda} < 1/2$, so that both $\underline{\lambda}$ and $\overline{\lambda}$ lie on the same side of $1/2$, then the two types always agree on the interpretation of a herd.

Proposition 11. *Suppose $\underline{\lambda} < \frac{1}{2} < \overline{\lambda}$ and $\mu^l < \hat{\mu}(\underline{\lambda}, \overline{\lambda}) < \mu^r$. 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 most share her taste. In the next subsection, I explain how this logic extends to more general distributions of λ, and discuss the intuition and significance of these results.

6.2.2 General Taste Distributions

I now discuss informally how this logic should extend to settings with $\Lambda = [0,1]$. Suppose type-dependent priors μ^l and μ^r are respectively strictly decreasing and increasing on Λ.[56] 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.

To provide intuition, 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

[56] This is true, for instance, when λ is drawn from a uniform prior on $[0,1]$.

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 $(\omega, \lambda) = (L, 0)$.

With uncertainty over tastes, players explain a herd by assuming common preferences. We saw a similar logic in Section 5, where players explained an otherwise anomalous herd by inferring that one option had high relative quality. Essentially, people use alternative dimensions of uncertainty to explain the seemingly unusual behavior that results from projection. So long as players' models are able to explain herds—whether it's a theory of common tastes or large quality differences—their erroneous beliefs are essentially self confirming: agent's incorrect theories perpetuate the herd, and thus never generate evidence inconsistent with their false beliefs.

It's worth emphasizing that naive learning can exacerbate biased perceptions of others' tastes. In both cases studied above, beliefs about the average taste grow polarized across types. Agents move from a seemingly mild error—they assume others share their uncertain beliefs about λ—to a growing *confident* that most share their taste. In this sense, naive learning can generate a strong taste projection, where each type thinks their own preference is most common. Even though agents in this model have precisely correct theories of the world aside from mispredicting others' priors, the fact that they ignore heterogeneity in beliefs when learning can potentially lead them far from the truth.

7 Conclusion

7.1 Relation to Previous Research

This paper contributes to a growing literature studying how informational biases can lead to the persistence of false or divergent beliefs.[57] Ellison and Fudenberg (1993), who were among the first to study biased social learning among agents with heterogeneous tastes, explore the efficiency of "rule-of-thumb" learning in a setting with observable payoffs. In their model, agents with heterogeneous tastes simply choose whichever action performed best among those observed. This naive

[57] One strand of this literature studies the consequences of probabilistic errors—such as over-inferring from small samples (Rabin, 2002; Rabin and Vayanos 2010) or under-appreciating properties of statistical processes (Barberis, Shleifer, and Vishny, 1998). A distinct strand studies agents who neglect the informational content of others' behavior, providing explanations for the winner's curse and excessive trading in asset markets (e.g., Eyster and Rabin, 2005; Eyster, Rabin and Vayanos, 2013). Taste projection at its root is a probabilistic error, but since it leads to inaccurate perceptions of others' information, taste projectors additionally misinfer from others' behavior.

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 (2014) studies a variant of the canonical model from Bikhchandani et al. (1992) where only a fraction of players observe the history and players mispredict this fraction. As with taste projection, various degrees of misprediction can lead to both stable, incorrect herds or persistent fluctuations in beliefs. Bohren's focus, however, is on a commonly-held misprediction, while I emphasize the interaction of misperceptions that differ across types of agents. Furthermore, the inferential error studied by Bohren (2014) 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). These papers also show how biased observational learning generates confident, yet false, beliefs.

Finally, taste projection is closely related to information projection, explored in Madarasz (2012). That model assumes agents overestimate the likelihood that others have the same private information as themselves. Madarasz explores 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 nearly 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.[58] Further, the literature on non-common priors typically assumes that agents 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 who thus wrongly believe in a commonly-shared interpretation of public information.

7.2 Discussion

Throughout this paper, I highlight how one's interpretation of others' behavior depends on the lens through which it is viewed—those with differing perceptions of tastes develop inconsistent beliefs about the state of the world. In many cases, this discrepancy in beliefs can lead behavior far from

[58]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.

the optimum. Furthermore, these results 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 despite 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, taste projection has important consequences in other natural social-learning environments as well. For instance, consider agents who directly share their experiences. In word-of-mouth learning (e.g., Banerjee and Fudenberg, 2004) or learning from online reviews—where players observe the actions and *payoffs* of predecessors—projection still leads learning astray. To see this, suppose restaurant Y generates stochastic outcomes y, which provide type θ with utility $u(y, \theta)$, and an observer sees a large collection of payoffs from a random sample of the population. With correct knowledge of the distribution of θ, a rational observer infers 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. For instance, if some unsophisticated diners earn high payoffs from average-quality meals, then those who enjoy only exceptional meals will be mislead by the shining reviews of those with limited taste, and vice versa.

More broadly, a novel feature of this paper is the assumption that agents within non-common-prior environments 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 an investor may under-appreciate the influence of her own experience on perceptions, and thus conclude that investors from different generations share her perceptions. How do conflicting expectations interact in the market and shape the perceptions of the current young generation? How will this naive learning process play out in the long run? These questions are left open for future research.

References

[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] 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.

[11] BOHREN, A. (2014): "Informational Herding with Model Mispecification," Working Paper, University of Pennsylvania.

[12] BROWN, C. (1982): "A False Consensus Bias in 1980 Presidential Preferences," *Journal of Social Psychology*, 118, 137–138.

[13] 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*.

[14] CAI, H., Y. CHEN, AND H. FANG (2009): "Observational Learning: Evidence from a Randomized Field Experiment," *American Economic Review*, 99(3), 864-882.

[15] ÇELEN, B. AND S. KARIV (2004): "Observational Learning Under Incomplete Information," *Games and Economic Behavior*, 47(1), 72–86.

[16] 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.

[17] CONLEY, T. AND C. UDRY (2010): "Learning About a New Technology: Pineapples in Ghana," *American Economic Review*, 100(1), 35–69.

[18] CONLIN, M., T. O'DONOGHUE, AND T. VOGELSANG (2007): "Projection Bias in Catalog Orders," *American Economic Review*, 97(4): 1217–1249.

[19] 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.

[20] DAWES, R. (1989): "Statistical criteria for establishing a truly false consensus effect," *Journal of Experimental Social Psychology*, 25(1), 1–17.

[21] 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.

[22] 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.

[23] DEMARZO, D. VAYANOS, AND J. ZWIEBEL (2003): "Persuasion Bias, Social Influence, and Uni- Dimensional Opinions," *Quarterly Journal of Economics*, 118: 909–968.

[24] DOWNS, A. (1957): "An Economic Theory of Political Action in a Democracy," *Journal of Political Economy*, 65(2): 135-150.

[25] EGAN, D., C. MERKLE AND M. WEBER (2014): "Second-Order Beliefs and the Individual Investor," *Journal of Economic Behavior & Organization*, forthcoming.

[26] ELLISON, G., AND D. FUDENBERG (1993): "Rules of thumb for social learning," *Journal of Political Economy* 101(4), 612–643.

[27] ENGELMANN, D. AND M. STROBEL (2001): "The False Consensus Effect Disappears if Representative Information and Monetary Incentives Are Given," *Experimental Economics*, 3, 241–643.

[28] ENGELMANN, D. AND M. STROBEL (2012): "Deconstruction and Reconstruction of an Anomaly," *Games and Economic Behavior*, 76, 678–689.

[29] EYSTER, E. AND M. RABIN (2005): "Cursed Equilibrium," *Econometrica*, 73(5), 1623-1672.

[30] EYSTER, E. AND M. RABIN (2010): "Naïve Herding in Rich-Information Settings," *American Economic Journal: Microeconomics*, 2(4), 221–243.

[31] EYSTER, E., M. RABIN, AND D. VAYANOS (2013): "Financial Markets where Traders Neglect the Informational Content of Asset Prices," Mimeo.

[32] FARO, D., AND Y. ROTTENSTREICH (2006): "Affect, Empathy, and Regressive Mispredictions of Others' Preferences Under Risk," *Management Science*, 52(4), 529–541.

[33] FUDENBERG, D. AND D. LEVINE (1993): "Self-Confirming Equilibrium," Econometrica, 61(3), 523–545.

[34] GAGNON-BARTSCH, T., AND M. RABIN (2014): "Naive Social Learning, Mislearning, and Unlearning" Mimeo.

[35] GOEREE, J., T. PALFREY, AND B. ROGERS (2006): "Social Learning with Private and Common Values," *Economic Theory*, 28(2), 245–264.

[36] HOTELLING, H. (1929): "Stability in Competition," *The Economic Journal*, 39: 41–57.

[37] 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.

[38] JACKSON, M. AND E. KALAI (1997): "Social Learning in Recurring Games," *Games and Economic Behavior*, 21, 102–134.

[39] KNIGHT, B. AND N. SCHIFF (2010): "Momentum and Social Learning in Presidential Primaries," *Journal of Political Economy*, 118(6), 1110–1150.

[40] KRAMER, G. H. (1971): "Short-Term Fluctuations in U.S. Voting Behavior: 1896-1964," *American Political Science Review*, 65(1): 131–143.

[41] 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.

[42] LOEWENSTEIN, G., T. O'DONOGHUE, AND M. RABIN (2003): "Projection Bias in Predicting Future Utility," *Quarterly Journal of Economics*, 118(4): 1209–1248.

[43] MADARASZ, K. (2012): "Information Projection: Model and Applications," *Review of Economic Studies*, 79, 961–985.

[44] MALMENDIER, U. AND S. NAGEL (2011): "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?" *Quarterly Journal of Economics*, 126, 373–416.

[45] 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.

[46] MORETTI, E. (2011): "Social Learning and Peer Effects in Consumption: Evidence from Movie Sales," *Review of Economic Studies*, 78(1): 356–393.

[47] MORRIS, S. (1995): "The Common Prior Assumption in Economic Theory," *Economics and Philosophy*, 11: 227–253.

[48] MORRIS, S. (1996): "Speculative Investor Behavior and Learning", *Quarterly Journal of Economics*, 111: 1111–1133.

[49] MOSSEL, E., A. SLY, AND O. TAMUZ (2012): "From Agreement to Asymptotic Learning," Mimeo.

[50] 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.

[51] MUNSHI, K. (2003): "Social Learning in a Heterogeneous Population: Technology Diffusion in the Indian Green Revolution," *Journal of Development Economics*, 73(1), 185–213.

[52] RABIN, M. (2002): "Inference by Believers in the Law of Small Numbers," *Quarterly Journal of Economics*, 117(3): 775–816.

[53] RABIN, M. AND J. SCHRAG (1999): "Inference by Believers in the Law of Small Numbers," *Quarterly Journal of Economics*, 114(1): 37–82.

[54] RABIN, M. AND D. VAYANOS (2010): "The Gambler's and Hot-Hand Fallacies: Theory and Applications," *Review of Economic Studies*, 77: 730–778.

[55] 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.

[56] 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.

[57] SALGANIK, M., P. DODDS, AND D. WATTS (2006): "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science*, 311(2), 854–856.

[58] SCHEINKMAN, J. AND W. XIONG (2003): "Overconfidence and Speculative Bubbles", *Journal of Political Economy*, 111(6): 1183–1219.

[59] SMITH, L. AND P. SØRENSEN (2000): "Pathological outcomes of observational learning," *Econometrica*. 68(2), 371–398.

[60] SMITH, L. AND P. SØRENSEN (2008): "Rational Social Learning with Random Sampling," Mimeo.

[61] SETHI, R. AND M. YILDIZ (2012): "Public Disagreement," *American Economic Journal: Microeconomics*, 4(3): 57–95.

[62] SIMONSOHN, U. (2010): "Weather to Go to College," *Economic Journal*, 120(543), 270–280.

[63] SORENSEN, A. (2006): "Social Learning and Health Plan Choice," *RAND Journal of Economics*, 37(4), 929–945.

[64] 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.

[65] VAN BOVEN, L., AND G. LOEWENSTEIN (2003): "Social Projection of Transient Drive States," *Personality and Social Psychology Bulletin*, 29(9): 1159–1168.

[66] VAN BOVEN, L., G. LOEWENSTEIN, AND D. DUNNING (2003): "Mispredicting the Endowment Effect: Underestimation of Ownersąŕ Selling Prices by Buyerąŕs Agents," *Journal of Economic Behavior and Organization*, 51: 351–365.

[67] WALLACE, D. F. (1996): *Infinite Jest*. New York: Little, Brown.

A Smith and Sørensen's Confounded Learning

Consider the model of Sections 3 and 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.

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

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

where

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

B Rational Learning with 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 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.

I consider a model identical to that in Section 6, but with the following exception: agents are fully-rational, so second-order beliefs are correct. Essentially, each player knows precisely the priors of all others. Despite this, 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 $\zeta, \zeta' \in \{L, R\}$ 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.*

To show learning is incomplete, 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. This is because at a confounding belief, new observations reveal no new information. As such, long-run beliefs across types are interior and thus 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—does occur under taste projection.

C 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 in Sections 4.1 and 4.2 was independent of assumptions placed on $\hat{\boldsymbol{\lambda}}$. Hence, we can directly apply those results to $\hat{\boldsymbol{\lambda}}$ exhibiting any particular pattern of error.

Proposition 4, which tells us when a confident equilibrium belief is stable, yields the following general result for any form of misprediction of type proportion λ:

Proposition A.3. *As $N \to \infty$, universal learning is complete if and only if for all $\theta \in \Theta$, $\hat{\lambda}(\theta) \in (1/2, \lambda]$.*

When all individuals mutually underestimate the share of people with the majority preference, then the truth is asymptotically stable. Near an equilibrium, people observe more people taking the majority action than they anticipated, which only strengthens 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.[59] In such a case, it's intuitive that action frequencies

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

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.

D Proofs

Proof of Lemma 1.

Proof. This follows immediately by rewriting the posterior $r(p, \pi)$ as $r = p/(p + (1-p)\ell)$ and solving the decision rule in the text for a threshold on p. □

Proof of Lemma 2.

Proof. See Lemma A.1. □

Proof of Lemma 3.

Proof. Fix $\theta \in \Theta$ and $\ell_t^\theta \in \mathbb{R}_+$. Suppose $a_t/N > \hat{\lambda}(\theta)$. From Equation 3, $\ell_{t+1}^\theta < \ell_t^\theta \Leftrightarrow \psi(a_t \mid \ell_t^\theta, L) < \psi(a_t \mid \ell_t^\theta, R) \Leftrightarrow \alpha_\theta(\ell_t^\theta, \omega)^a \big[1 - \alpha_\theta(\ell_t^\theta, L)\big]^{N-a} < \alpha_\theta(\ell_t^\theta, R)^a \big[1 - \alpha_\theta(\ell_t^\theta, R)\big]^{N-a}$,

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

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.1 holds iff
$a/N < \left(1 + \log\left(\frac{\alpha_\theta(\ell_t^\theta, L)}{\alpha_\theta(\ell_t^\theta, R)}\right) / \log\left(\frac{1-\alpha_\theta(\ell_t^\theta, R)}{1-\alpha_\theta(\ell_t^\theta, L)}\right)\right)^{-1} =: \kappa(\ell_t^\theta, \theta)$. Otherwise, A.1 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 3, since MLRP implies first-order stochastic dominance.

□

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

from everyone else."

$\lim_{\ell\to 0}\kappa(\ell,\theta)$ 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:

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

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

and, since Equation 5 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

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

Thus, $\lim_{\ell\to 0}\kappa(\ell,\theta)=\hat{\lambda}(\theta)$ 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 3 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 3 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 a_t at public belief ℓ_t^θ, Equation 3 implies $\ell_{t+1}^\theta>\ell_t^\theta\Leftrightarrow\Psi_\theta(a_t,\ell_t^\theta)>1\Leftrightarrow\psi(a_t\mid\ell_t^\theta,L)>\psi(a_t\mid\ell_t^\theta,R)$. Suppose $N=1$ and $a_t=1$ and let $\bar{p}:=p(\ell_t^\theta)$ denote type θ's private-belief threshold in t. Then Equation 5 implies $\psi(a_t\mid\ell_t^\theta,L)>\psi(a_t\mid\ell_t^\theta,R)$ 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 $\big[1-2\hat{\lambda}(\theta)\big]F_L(\bar{p})>\big[1-2\hat{\lambda}(\theta)\big]F_R(\bar{p})$. By Assumption 3, $F_L(\bar{p})>F_R(\bar{p})$ for all $\ell_t^\theta\in\mathbb{R}_+$, this inequality 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 $a_t=1$, Proposition 1 $\Psi_\theta(a_t=1,\ell_t^\theta)>1\Leftrightarrow\hat{\lambda}(\theta)<\frac{1}{2}$. First consider $\hat{\lambda}(\theta)\in(\frac{1}{2},1)$ so $\Psi_\theta(a_t=1,\ell_t^\theta)<1$. We want to show $|\ell_{t+1}^\theta-\ell_t^\theta|$ is increasing in $\hat{\lambda}(\theta)$ on this domain. Note that $|\ell_{t+1}^\theta-\ell_t^\theta|=\ell_t^\theta|\Psi_\theta(A,\ell_t^\theta)-1|$, which is increasing in $\hat{\lambda}(\theta)\Leftrightarrow\Psi_\theta(A,\ell_t^\theta)$ is decreasing in $\hat{\lambda}(\theta)$. Let $\bar{p}:=p(\ell_t^\theta)$ denote θ's private-belief threshold in t. Note

$$\Psi_\theta(A,\ell_t^\theta)=\frac{\big[1-\hat{\lambda}(\theta)\big]F_L(\bar{p})+\hat{\lambda}(\theta)\big[1-F_L(\bar{p})\big]}{\big[1-\hat{\lambda}(\theta)\big]F_R(\bar{p})+\hat{\lambda}(\theta)\big[1-F_R(\bar{p})\big]}=\frac{\hat{\lambda}(\theta)\big[1-2F_R(\bar{p})\big]+F_R(\bar{p})}{\hat{\lambda}(\theta)\big[1-2F_L(\bar{p})\big]+F_R(\bar{p})}$$

so $\frac{\partial}{\partial\hat{\lambda}(\theta)}\Psi_\theta(A,\ell_t^\theta)<0$ if and only if $\hat{\lambda}(\theta)\big[1-2F_R(\bar{p})\big]\big[1-2F_L(\bar{p})\big]+F_R(\bar{p})\big[1-2F_L(\bar{p})\big]>$
$\hat{\lambda}(\theta)\big[1-2F_L(\bar{p})\big]\big[1-2F_R(\bar{p})\big]+F_L(\bar{p})\big[1-2F_R(\bar{p})\big]$, which holds if and only if $F_L(\bar{p})>F_R(\bar{p})$, which is true for all $\ell_t^\theta\in\mathbb{R}_+$. Next, suppose that $\hat{\lambda}(\theta)\in(0<\frac{1}{2})$ so $\Psi_\theta(a_t=1,\ell_t^\theta)>1$. We want to show that $|\ell_{t+1}^\theta-\ell_t^\theta|=\ell_t^\theta|\Psi_\theta(A,\ell_t^\theta)-1|$ is decreasing in $\hat{\lambda}(\theta)$ on this domain. This is true iff $\Psi_\theta(A,\ell_t^\theta)$ is

decreasing in $\hat{\lambda}(\theta)$, which was shown in the case above. The logic is identical for $a_t = 1$, but uses the fact that $\Psi_\theta(a_t, \ell_t^\theta) > 1 \Leftrightarrow \hat{\lambda}(\theta) > \frac{1}{2}$. □

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 $t \in \mathbb{N}$, $\pi_t^\theta = \pi_t^{\theta'}$ 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 4, $\langle \ell_t^\theta \rangle$ forms a conditional martingale on $\omega = R$. By the Martingale Convergence Theorem, it must converge almost surely to some stationary limit. By Lemma 5, the only stationary limit points are $\ell \in \{0, \infty\}$. But rational beliefs never converge to fully-incorrect beliefs, so it must be that $\ell_t^\theta \to 0$ a.s. □

Proof of Lemma 4.

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

$$\mathbb{E}[\ell_{t+1}^\theta \mid \boldsymbol{\ell}_t] = \sum_{a_t=0}^{N} \psi(a_t \mid \boldsymbol{\ell}_t, R)\Psi_\theta(a_t, \ell_t^\theta)\ell_t^\theta \tag{A.2}$$

Thus in order for $\langle \ell_t^\theta \rangle$ to form a Martingale conditional on R, we would need $\sum_{a_t=0}^{N} \psi(a_t \mid \boldsymbol{\ell}_t, R)\Psi_\theta(a_t, \ell_t^\theta) = 1$ for all $\ell_t^\theta \in \mathbb{R}_+$. But note

$$\sum_{a_t=0}^{N} \psi(a_t \mid \boldsymbol{\ell}_t, R)\Psi_\theta(a_t, \ell_t^\theta) = \sum_{a_t=0}^{N} \psi(a_t \mid \boldsymbol{\ell}_t, R)\frac{\psi_\theta(a_t \mid \ell_t^\theta, L)}{\psi_\theta(a_t \mid \ell_t^\theta, L)} = \sum_{a_t=0}^{N} \frac{\psi(a_t \mid \boldsymbol{\ell}_t, R)}{\psi_\theta(a_t \mid \ell_t^\theta, R)}\psi_\theta(a_t \mid \ell_t^\theta, L).$$

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 the Martingale condition above 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 5.

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 $\ell^\theta \in \text{supp}(\ell_\infty^\theta)$, 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 6.

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 8. From Lemma 6, $\hat{\boldsymbol{\ell}}$ 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 $\alpha_\theta(\hat{\ell}^\theta, \omega)$; the true frequency is $\alpha(\hat{\boldsymbol{\ell}})$. To determine whether this unexpected frequency reinforces each θ's beliefs, we must calculate $\chi_\theta(\hat{\boldsymbol{\ell}}) = \prod_{a=0}^{N} \left(\frac{\partial}{\partial \ell}\varphi_\theta(a, \hat{\ell}^\theta)\right)^{\psi(a, \boldsymbol{\ell})}$ 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 4, it follows that

$$\begin{aligned}\frac{\partial}{\partial \ell}\psi_\theta(a \mid \ell,\omega) &= \binom{N}{a}\Big(a\alpha_\theta(\ell,\omega)^{a-1}\big[1-\alpha_\theta(\ell,\omega)\big]^{N-a}\frac{\partial}{\partial \ell}\alpha_\theta(\ell,\omega) \\ &\quad -(N-a)\alpha_\theta(\ell,\omega)^{a}\big[1-\alpha_\theta(\ell,\omega)\big]^{N-a-1}\frac{\partial}{\partial \ell}\alpha_\theta(\ell,\omega)\Big) \\ &= \frac{\partial}{\partial \ell}\alpha_\theta(\ell,\omega)\left(a\frac{\psi_\theta(a \mid \ell,\omega)}{\alpha_\theta(\ell,\omega)} - (N-a)\frac{\psi_\theta(a \mid \ell,\omega)}{1-\alpha_\theta(\ell,\omega)}\right). \qquad \text{(A.3)}\end{aligned}$$

From Equation 5 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.3 and using the fact $p(\ell) = \ell/(1+\ell) \Rightarrow \frac{\partial}{\partial \ell}p(\ell) = 1/(1+\ell)^2$ yields

$$\frac{\partial}{\partial \ell}\psi_\theta(a \mid \ell,\omega) = \frac{\big[1-2\hat{\lambda}(\theta)\big]}{(1+\ell)^2}\psi_\theta(a \mid \ell,\omega)f_\omega\big(p(\ell)\big)\left(\frac{a-N\alpha_\theta(\ell,\omega)}{\alpha_\theta(\ell,\omega)\big[1-\alpha_\theta(\ell,\omega)\big]}\right). \qquad \text{(A.4)}$$

Using the definition of $\Psi_\theta(a,\ell)$ and Equation A.3,

$$\begin{aligned}\frac{\partial}{\partial \ell}\Psi_\theta(a,\ell) = \Psi_\theta(a,\ell)\Bigg\{\frac{\big[1-2\hat{\lambda}(\theta)\big]}{(1+\ell)^2}\Bigg[f_L\big(p(\ell)\big)\left(\frac{a-N\alpha_\theta(\ell,L)}{\alpha_\theta(\ell,L)\big[1-\alpha_\theta(\ell,L)\big]}\right) \\ - f_R\big(p(\ell)\big)\left(\frac{a-N\alpha_\theta(\ell,R)}{\alpha_\theta(\ell,R)\big[1-\alpha_\theta(\ell,R)\big]}\right)\Bigg]\Bigg\}. \qquad \text{(A.5)}\end{aligned}$$

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.5 implies

$$\begin{aligned}\frac{\partial}{\partial \ell}\varphi_\theta(a,\ell) = \Psi_\theta(a,\ell)\Bigg\{1+\frac{\big[1-2\hat{\lambda}(\theta)\big]\ell}{(1+\ell)^2}\Bigg[f_L\big(p(\ell)\big)\left(\frac{a-N\alpha_\theta(\ell,L)}{\alpha_\theta(\ell,L)\big[1-\alpha_\theta(\ell,L)\big]}\right) \\ - f_R\big(p(\ell)\big)\left(\frac{a-N\alpha_\theta(\ell,R)}{\alpha_\theta(\ell,R)\big[1-\alpha_\theta(\ell,R)\big]}\right)\Bigg]\Bigg\}. \qquad \text{(A.6)}\end{aligned}$$

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.6 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 $p(\hat{\ell}^\theta) = 1$ and $\alpha_\theta(\infty,\omega) = 1-\hat{\lambda}(\theta)$. In either case, $\alpha_\theta(\hat{\ell}^\theta,\omega)$ is independent of ω, so it follows immediately from Equation 4 that $\psi_\theta(a \mid \hat{\ell}^\theta,\omega)$ is also independent of ω. Hence $\Psi_\theta(a,\hat{\ell}^\theta) = \psi_\theta(a \mid \hat{\ell}^\theta, L)/\psi_\theta(a \mid \hat{\ell}^\theta, R) = 1$. So for any $\hat{\boldsymbol{\pi}} \in \Pi$ and corresponding likelihood ratios $\hat{\boldsymbol{\ell}}$,

$$\frac{\partial}{\partial \ell}\varphi_\theta(a,\hat{\ell}^\theta)\Big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} = 1. \tag{A.7}$$

It follows from Equation 11 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 6: $\hat{\boldsymbol{\ell}}$ 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

$$\begin{aligned}
\frac{\partial}{\partial \ell^\theta}\log\chi_\theta(\boldsymbol{\ell})\Big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} &= \sum_{a=0}^{N}\psi(a,\hat{\boldsymbol{\ell}})\left(\frac{\partial}{\partial \ell}\varphi_\theta(a,0)\right)^{-1}\left(\frac{\partial^2}{\partial \ell^2}\varphi_\theta(a,\ell^\theta)\Big|_{\hat{\ell}^\theta=0}\right) \\
&\quad + \sum_{a=0}^{N}\left(\frac{\partial}{\partial \hat{\ell}^\theta}\psi(a,\boldsymbol{\ell})\Big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}}\right)\log\left(\frac{\partial}{\partial \ell}\varphi_\theta(a,0)\right) \\
&= \sum_{a=0}^{N}\psi(a,\hat{\boldsymbol{\ell}})\left(\frac{\partial^2}{\partial \ell^2}\varphi_\theta(a,\ell^\theta)\Big|_{\hat{\ell}^\theta=0}\right)
\end{aligned} \tag{A.8}$$

where the final equality follows from $\frac{\partial}{\partial \ell}\varphi_\theta(a,0) = 1$ (as shown above in Equation A.7). Since $\frac{\partial^2}{\partial \ell^2}\varphi_\theta(a,\ell) = \frac{\partial^2}{\partial \ell^2}\{\Psi_\theta(a,\ell)\ell\} = 2\frac{\partial}{\partial \ell}\Psi_\theta(a,\ell) + \ell\frac{\partial^2}{\partial \ell^2}\Psi_\theta(a,\ell)$, Equation A.8 reduces to

$$\frac{\partial}{\partial \ell^\theta}\log\chi_\theta(\boldsymbol{\ell})\Big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} = \sum_{a=0}^{N}2\psi(a,\hat{\boldsymbol{\ell}})\frac{\partial}{\partial \ell}\Psi_\theta(a,0) \tag{A.9}$$

From Equation A.5 and using the fact that $p(0) = 0 \Rightarrow \alpha_\theta(0,\omega) = \hat{\lambda}(\theta)$ and $\Psi_\theta(a,0) = 1$,

$$\frac{\partial}{\partial \ell}\Psi_\theta(a,0) = \big[1 - 2\hat{\lambda}(\theta)\big]\big[f_L(0) - f_R(0)\big]\left(\frac{a - N\hat{\lambda}(\theta)}{\hat{\lambda}(\theta)\big[1-\hat{\lambda}(\theta)\big]}\right), \tag{A.10}$$

so Equation A.9 implies

$$\begin{aligned}
\frac{\partial}{\partial \ell^\theta}\log\chi_\theta(\boldsymbol{\ell})\Big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} &= \frac{2\big[1-2\hat{\lambda}(\theta)\big]\big[f_L(0)-f_R(0)\big]}{\hat{\lambda}(\theta)\big[1-\hat{\lambda}(\theta)\big]}\sum_{a=0}^{N}\psi(a,\hat{\boldsymbol{\ell}})(a - N\hat{\lambda}(\theta)) \\
&= \frac{2\big[1-2\hat{\lambda}(\theta)\big]\big[f_L(0)-f_R(0)\big]}{\hat{\lambda}(\theta)\big[1-\hat{\lambda}(\theta)\big]}\Big(N\alpha(\hat{\boldsymbol{\ell}}) - N\hat{\lambda}(\theta)\Big).
\end{aligned} \tag{A.11}$$

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.11 implies the following result:

$$\frac{\partial}{\partial \ell^\theta}\chi_\theta(\boldsymbol{\ell})\Big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} < 0 \Leftrightarrow \begin{cases} \hat{\lambda}(\theta) < \alpha(\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ \hat{\lambda}(\theta) > \alpha(\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}. \end{cases} \tag{A.12}$$

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 $\imath_t^\theta := (\ell_t^\theta)^{-1}$. 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 $\imath$ at 0. The stability coefficient of interest is now that of the inverse likelihood ratio:

$$\tilde{\chi}_\theta(\hat{\imath}) = \prod_{a=0}^{N} \left(\frac{\partial}{\partial \imath} \tilde{\varphi}_\theta(a, \hat{\imath}^\theta) \right)^{\tilde{\psi}(a,\imath)} \tag{A.13}$$

where $\tilde{\varphi}_\theta(a, \imath)$ is the transition equation for the process $\langle \imath_t^\theta \rangle$: $\tilde{\varphi}_\theta(a, \imath) = \tilde{\Psi}_\theta(a, \imath)\imath$ with $\tilde{\Psi}_\theta(a, \imath) = \tilde{\psi}_\theta(a \mid \imath, R)/\tilde{\psi}_\theta(a \mid \imath, L)$. $\tilde{\psi}_\theta(a \mid \imath, \omega)$ is the direct analog of $\psi_\theta(a \mid \ell, \omega)$: it is the probability of observing a at belief $\imath$ in state ω according to type-θ's theory of tastes.

As above, $\tilde{\chi}_\theta(\hat{\imath}) = 1$ if $\hat{\imath}^\theta = 0$, so we must calculate the derivative or $\tilde{\chi}_\theta(\hat{\imath})$ with respect to $\imath^\theta$ 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

$$\frac{\partial}{\partial \imath^\theta} \log \tilde{\chi}_\theta(\imath)\Big|_{\imath=\hat{\imath}} = \sum_{a=0}^{N} 2\tilde{\psi}(a, \hat{\imath}) \frac{\partial}{\partial \imath} \tilde{\Psi}_\theta(a, \imath)\Big|_{\imath=\hat{\imath}}. \tag{A.14}$$

Note that

$$\frac{\partial}{\partial \imath} \tilde{\Psi}_\theta(a, \imath) = \tilde{\Psi}(a, \imath) \Bigg\{ \frac{[1 - 2\hat{\lambda}(\theta)]}{(1+\imath)^2} \Bigg[f_L(p(\imath)) \left(\frac{a - N\alpha_\theta(\imath, L)}{\alpha_\theta(\imath, L)[1 - \alpha_\theta(\imath, L)]} \right) - f_R(p(\imath)) \left(\frac{a - N\alpha_\theta(\imath, R)}{\alpha_\theta(\imath, R)[1 - \alpha_\theta(\imath, R)]} \right) \Bigg] \Bigg\}. \tag{A.15}$$

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

$$\frac{\partial}{\partial \imath^\theta} \tilde{\Psi}_\theta(a, \imath^\theta)\Big|_{\imath=\hat{\imath}} = [1 - 2\hat{\lambda}(\theta)]\,[f_L(1) - f_R(1)] \left(\frac{a - N(1 - \hat{\lambda}(\theta))}{\hat{\lambda}(\theta)[1 - \hat{\lambda}(\theta)]} \right). \tag{A.16}$$

Plugging into Equation A.14,

$$\frac{\partial}{\partial \imath^\theta} \log \tilde{\chi}_\theta(\imath)\Big|_{\imath=\hat{\imath}} = \frac{2[1 - 2\hat{\lambda}(\theta)]\,[f_L(1) - f_R(1)]}{\hat{\lambda}(\theta)[1 - \hat{\lambda}(\theta)]} \Big(N\alpha(\hat{\imath}) - N(1 - \hat{\lambda}(\theta)) \Big). \tag{A.17}$$

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

$$\frac{\partial}{\partial \imath^\theta} \tilde{\chi}_\theta(\imath)\Big|_{\imath=\hat{\imath}} < 0 \Leftrightarrow \begin{cases} 1 - \hat{\lambda}(\theta) > \alpha(\hat{\imath}) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ 1 - \hat{\lambda}(\theta) < \alpha(\hat{\imath}) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2} \end{cases} \tag{A.18}$$

Step 5. Linking stability to expected action frequencies.

Finally, I write the stability conditions derived in Steps 3 and 4—Results A.12 and A.18—in terms of

the expected and true action frequencies at $\hat{\boldsymbol{\ell}}$. First, note that

$$\widehat{\mathscr{F}}_\theta(M_\theta(\hat{\ell}),\hat{\ell}) = \begin{cases} \hat{\lambda}(\theta) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ 1-\hat{\lambda}(\theta) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}. \end{cases} \tag{A.19}$$

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.12 and A.18 respectively yield

$$\frac{\partial}{\partial \ell^\theta}\chi_\theta(\boldsymbol{\ell})\Big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} < 0 \Leftrightarrow \begin{cases} \widehat{\mathscr{F}}_\theta(M_\theta(0),0) < \alpha(\hat{\boldsymbol{\ell}}) = \mathscr{F}(A,\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ \widehat{\mathscr{F}}_\theta(M_\theta(0),0) < 1-\alpha(\hat{\boldsymbol{\ell}}) = \mathscr{F}(B,\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}, \end{cases} \tag{A.20}$$

and

$$\frac{\partial}{\partial \imath^\theta}\tilde{\chi}_\theta(\imath)\Big|_{\imath=\hat{\imath}} < 0 \Leftrightarrow \begin{cases} \widehat{\mathscr{F}}_\theta(M_\theta(\infty),\infty) < 1-\alpha(\hat{\boldsymbol{\ell}}) = \mathscr{F}(B,\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ \widehat{\mathscr{F}}_\theta(M_\theta(\infty),\infty) < \alpha(\hat{\boldsymbol{\ell}}) = \mathscr{F}(A,\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}. \end{cases} \tag{A.21}$$

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

$$M_\theta(0) = \begin{cases} A & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ B & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}, \end{cases} \quad \text{and} \quad M_\theta(\infty) = \begin{cases} B & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ A & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}. \end{cases} \tag{A.22}$$

Appropriately incorporating these identities into A.23 and A.24 finally yields the following stability conditions:

$$\frac{\partial}{\partial \ell^\theta}\chi_\theta(\boldsymbol{\ell})\Big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} < 0 \Leftrightarrow \begin{cases} \widehat{\mathscr{F}}_\theta(M_\theta(0),0) < \mathscr{F}(M_\theta(0),\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ \widehat{\mathscr{F}}_\theta(M_\theta(0),0) < \mathscr{F}(M_\theta(0),\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}, \end{cases} \tag{A.23}$$

and

$$\frac{\partial}{\partial \imath^\theta}\tilde{\chi}_\theta(\imath)\Big|_{\imath=\hat{\imath}} < 0 \Leftrightarrow \begin{cases} \widehat{\mathscr{F}}_\theta(M_\theta(\infty),\infty) < \mathscr{F}(M_\theta(\infty),\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) > \frac{1}{2} \\ \widehat{\mathscr{F}}_\theta(M_\theta(\infty),\infty) < \mathscr{F}(M_\theta(\infty),\hat{\boldsymbol{\ell}}) & \text{if} \quad \hat{\lambda}(\theta) < \frac{1}{2}. \end{cases} \tag{A.24}$$

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 $\widehat{\mathscr{F}}_\theta(M(\hat{\ell}(\theta)),\hat{\ell}(\theta)) < \mathscr{F}(M(\hat{\ell}(\theta)),\hat{\boldsymbol{\ell}})$, 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 belief 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 $\alpha_\theta(\ell^\theta,\omega)$. $\Delta_q \neq 0$ and $\frac{\partial}{\partial \ell}p(0,\theta) = 1/v(\theta)$ implies

$$\frac{\partial}{\partial \ell}\alpha_\theta(0,\omega) = f_\omega(0)\left[\sum_{\tilde{\theta}\in\Theta^l}\frac{\hat{g}(\tilde{\theta}|\theta)}{v(\tilde{\theta})} - \sum_{\tilde{\theta}\in\Theta^r}\frac{\hat{g}(\tilde{\theta}|\theta)}{v(\tilde{\theta})}\right]. \tag{A.25}$$

It follows from Proposition 4 that

$$\begin{aligned}\frac{\partial}{\partial \ell^\theta}\log\chi_\theta(\boldsymbol{\ell})\Big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} &= \sum_{a=0}^{N} 2\psi(a,\hat{\boldsymbol{\ell}})\frac{\partial}{\partial \ell}\Psi_\theta(a,0) \\ &= \frac{2\big[f_L(0)-f_R(0)\big]}{\alpha_\theta(0,\omega)\big[1-\alpha_\theta(0,\omega)\big]}\left[\sum_{\tilde{\theta}\in\Theta^l}\frac{\hat{g}(\tilde{\theta}|\theta)}{v(\tilde{\theta})}-\sum_{\tilde{\theta}\in\Theta^r}\frac{\hat{g}(\tilde{\theta}|\theta)}{v(\tilde{\theta})}\right]\sum_{a=0}^{N}\psi(a,\hat{\boldsymbol{\ell}})(a-N\alpha_\theta(0,\omega)) \qquad \text{(A.26)}\end{aligned}$$

The first equality follows from Equation A.9. To arrive at the second equality, first plug $\frac{\partial}{\partial \ell}\alpha_\theta(0,\omega)$ from A.25 into the expression for $\frac{\partial}{\partial \ell}\psi_\theta(a \mid \ell,\omega)$ in Equation A.3, then plug the result into Equation A.5, 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 $\alpha_\theta(0,L) = \alpha_\theta(0,R) = 1 - \sum_{\tilde{\theta}\in\Theta^l}\hat{g}(\tilde{\theta}|\theta)$—$\theta$'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 $\tilde{a} \sim$ Binomial$(N,\alpha(\hat{\boldsymbol{\ell}}))$, and since $f_L(0) > f_R(0)$, it follows from Equation A.26 that $\frac{\partial}{\partial \ell^\theta}\log\chi_\theta(\boldsymbol{\ell})\big|_{\boldsymbol{\ell}=\hat{\boldsymbol{\ell}}} > 0$ if and only if

$$\left[\sum_{\tilde{\theta}\in\Theta^l}\frac{\hat{g}(\tilde{\theta}|\theta)}{v(\tilde{\theta})}-\sum_{\tilde{\theta}\in\Theta^r}\frac{\hat{g}(\tilde{\theta}|\theta)}{v(\tilde{\theta})}\right]\big[\alpha(\hat{\boldsymbol{\ell}})-\alpha_\theta(0,\omega)\big] > 0. \qquad \text{(A.27)}$$

I now argue that, generically, Condition A.27 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.27 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 $v(\theta) = (4k\theta + \Delta_q)/(4k\theta - \Delta_q)$. By definition of Θ^r, $1/v(\theta)$ is increasing on Θ^r. Because $\widehat{G}(\tilde{\theta}|\theta)$ first-order stochastically dominates $\widehat{G}(\tilde{\theta}|\theta')$ 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.27 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. (Sketch.) 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 $\langle \ell_t^l, \ell_t^r\rangle$ reaches the north-west quadrant of belief space (see Figure 3), which we define by all points $\boldsymbol{\ell}_t$ such that $\ell_t^r > L_l(\ell_t^l)$ and $\ell_t^l < L_r(\ell_t^r)$ (see footnote 44). Call this set L_{NW}. Restricted to L_{NW}, each $\langle \ell_t^l\rangle$ and $\langle 1/\ell_t^r\rangle$ 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 $\ell_t^r < L_l(\ell_t^l)$ and $\ell_t^l > L_r(\ell_t^r)$. Call this space L_{SE}. Restricted to L_{SE}, each $\langle \ell_t^r\rangle$ and $\langle 1/\ell_t^l\rangle$ are non-negative supermartingales, and thus converge. Hence, if process $\langle \ell_t^l, \ell_t^r\rangle$ 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 7.

Proof. Since $\mathbb{E}[\ell_{t+1}^\theta \mid \boldsymbol{\ell}_t] = \sum_{a_t=0}^N \psi(a_t \mid \boldsymbol{\ell}_t, R)\Psi_\theta(a_t, \ell_t^\theta)\ell_t^\theta$, $\mathbb{E}[\ell_{t+1}^\theta \mid \boldsymbol{\ell}_t] > \ell_t^\theta \Leftrightarrow \xi_\theta(\ell_t^l, \ell_t^r) \equiv \sum_{a_t=0}^N \psi(a_t \mid \boldsymbol{\ell}_t, R)\Psi_\theta(a_t, \ell_t^\theta) > 1$. 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 $\xi_\theta(\ell_t^l, \ell_t^r)$ near $\mathbf{0}$. Note that

$$\xi_\theta(\epsilon,\epsilon) \approx \xi_\theta(0,0) + \sum_{a=0}^N \psi(a \mid \mathbf{0}, R)\frac{\partial}{\partial \ell^\theta}\Psi_\theta(a,0) + \epsilon\left(\sum_{a=0}^N \left(\frac{\partial}{\partial \ell^l}\psi(a \mid \mathbf{0}, R) + \frac{\partial}{\partial \ell^r}\psi(a \mid \mathbf{0}, R)\right)\Psi_\theta(a,0)\right). \quad \text{(A.28)}$$

From Equation A.29,

$$\frac{\partial}{\partial \ell^\theta}\psi(a \mid \mathbf{0}, R) = (1-2\lambda)\psi(a \mid \mathbf{0}, R) f_R(0)\left(\frac{a - N\lambda}{\lambda(1-\lambda)}\right), \quad \text{(A.29)}$$

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

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

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.28 yields

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

Finally, recall that $\mathbb{E}[\ell_{t+1}^\theta \mid \boldsymbol{\ell}_t = (\epsilon,\epsilon)] > \ell_t^\theta = \epsilon \Leftrightarrow \xi_\theta(\epsilon,\epsilon) > 1 \Leftrightarrow \sum_{a=0}^N \psi(a \mid \mathbf{0}, R)\frac{\partial}{\partial \ell^\theta}\Psi_\theta(a,0) > 0$. From Equation A.10,

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

so

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

which exceeds 0 if and only if $\left[1 - 2\hat{\lambda}(\theta)\right]\left[\lambda - \hat{\lambda}(\theta)\right] > 0$. With Strong projection, $\hat{\lambda}^r > \lambda > 1/2$, so $[1 - 2\hat{\lambda}^r][\lambda - \hat{\lambda}^r] > 0$. Hence, ℓ_t^r 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_t^l 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_t^r is unstable near 0. ℓ_t^l must also be unstable near 0: by Lemma 8, there exists an $\epsilon > 0$ such that ℓ_t^r is submartingale so long as $\ell_t^l < \epsilon$. If $\ell_t^l < \epsilon$ for all t, then ℓ_t^r diverges to ∞ and the frequency of A converges to 1, which necessarily implies $\ell_t^l \to \infty$, a contradiction. The analogous argument holds at *any*

potential limit point $\hat{\boldsymbol{\ell}}$: for some $\theta \in \{l, r\}$, ℓ_t^θ is immediately unstable by Proposition 4, and the martingale property of the unstable ℓ_t^θ, which moves away from $\hat{\ell}^\theta$ in expectation, implies $\ell_t^{\theta'}$ $\theta' \neq \theta$ necessarily exits a neighborhood about $\hat{\ell}^{\theta'}$, contradicting stability of $\ell_t^{\theta'}$. □

Proof of Lemma 8.

Proof. The proof of Lemma 7 shows that $\mathbb{E}[\ell_{t+1}^\theta \mid \boldsymbol{\ell}_t = (\epsilon, \epsilon)] > \ell_t^\theta = \epsilon \Leftrightarrow [1 - 2\hat{\lambda}(\theta)][\lambda - \hat{\lambda}(\theta)] > 0$. This holds for $\hat{\lambda}^r > \lambda > 1/2$, but fails for $\hat{\lambda}^l \in (1/2, \lambda)$. Hence ℓ_t^r is locally a submartingale in the neighborhood of $\boldsymbol{\ell} = (0, 0)$ whereas ℓ_t^l 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 $G(\tilde{\theta}) + \lambda > \max\{1 - \hat{\lambda}(\underline{\theta}), \hat{\lambda}(\overline{\theta})\}$. 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 $(1 - \lambda) + 1 - G(\tilde{\theta}) = 2 - (\lambda - -G(\tilde{\theta})) > \max\{1 - \hat{\lambda}(\underline{\theta}), \hat{\lambda}(\overline{\theta})\}$. □

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.[60] 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

$$\begin{aligned}
\mathcal{L}^r &= \left[\left(\frac{\widehat{\Pr}^\theta(X_{n2} = A \mid \omega \in \Omega^{BA})}{\widehat{\Pr}^\theta(X_{n2} = A \mid \omega \in \Omega^{AB})} \right)^{a_2/N} \left(\frac{1 - \widehat{\Pr}^\theta(X_{n2} = A \mid \omega \in \Omega^{BA})}{1 - \widehat{\Pr}^\theta(X_{n2} = A \mid \omega \in \Omega^{AB})} \right)^{1 - a_2/N} \right]^N \\
&= \left(\frac{\hat{\lambda}^r - \hat{\epsilon}(N)}{1 - \hat{\lambda}^r + \hat{\epsilon}(N)} \right)^{\lambda - \epsilon_N} \left(\frac{1 - \hat{\lambda}^r + \hat{\epsilon}(N)}{\hat{\lambda}^r - \hat{\epsilon}(N)} \right)^{1 - \lambda + \epsilon_N} = \left(\frac{\hat{\lambda}^r - \hat{\epsilon}(N)}{1 - \hat{\lambda}^r + \hat{\epsilon}(N)} \right)^{2\lambda - 1 - 2\epsilon_N} \qquad \text{(A.30)}
\end{aligned}$$

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,

$$\mathcal{L}^l = \left(\frac{\hat{\lambda}^l - \hat{\epsilon}^l(N)}{1 - \hat{\lambda}^l + \hat{\epsilon}^l(N)} \right)^{2\lambda - 1 - 2\epsilon(N)}, \qquad \text{(A.31)}$$

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

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 9, 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

$$\frac{\partial}{\partial \ell(\omega)} \varphi(a, \boldsymbol{\ell}, \omega') = \ell(\omega') \frac{\partial}{\partial \ell(\omega)} \left(\frac{\psi_\theta(a \mid \boldsymbol{\ell}, \omega')}{\psi_\theta(a \mid \boldsymbol{\ell}, \omega^*)} \right) + \frac{\partial \ell(\omega')}{\partial \ell(\omega)} \left(\frac{\psi_\theta(a \mid \boldsymbol{\ell}, \omega')}{\psi_\theta(a \mid \boldsymbol{\ell}, \omega^*)} \right) \tag{A.32}$$

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 11) for each ω and $a \in \{0, 1, ..., N\}$ is less than one at $\hat{\boldsymbol{\ell}}^l = \mathbf{0}$, $\hat{\boldsymbol{\ell}}^r = \mathbf{0}$:

$$\chi_\theta(\hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r, \omega)\Big|_{(\hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r) = (\mathbf{0}, \mathbf{0})} < 1, \tag{A.33}$$

where

$$\chi_\theta(\hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r, \omega) = \prod_{a=0}^{N} \left(\frac{\partial}{\partial \ell(\omega)} \varphi_\theta(a, \boldsymbol{\ell}^\theta, \omega) \right)^{\psi(a, \hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r)}, \tag{A.34}$$

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, $\chi_\theta(\mathbf{0}, \mathbf{0}, \omega) < 1 \Leftrightarrow \frac{\partial}{\partial \ell(\omega)} \varphi_\theta(N, \mathbf{0}, \omega) < 1$. From A.32, for any ω, $\frac{\partial}{\partial \ell(\omega)} \varphi_\theta(N, \boldsymbol{\ell}^\theta, \omega) = \psi_\theta(N \mid \mathbf{0}, \omega)/\psi_\theta(N \mid \mathbf{0}, \omega^*) = \alpha_\theta(\mathbf{0}, \omega)/\alpha_\theta(\mathbf{0}, \omega^*)$, 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.10,

$$\frac{\partial}{\partial \ell^\theta(\omega)} \log \chi_\theta(\hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r, \omega)\Big|_{(\hat{\boldsymbol{\ell}}^l, \hat{\boldsymbol{\ell}}^r) = (\mathbf{0}, \mathbf{0})} = 2 \sum_{z=0}^{N} \psi(a, \mathbf{0}, \mathbf{0}) \frac{\partial}{\partial \ell^\theta(\omega)} \left(\frac{\psi_\theta(a \mid \mathbf{0}, \omega)}{\psi_\theta(a \mid \mathbf{0}, \omega^*)} \right). \tag{A.35}$$

For $\omega = (\zeta, \lambda)$ and $\omega^* = (\zeta^*, \lambda^*)$, analogous to Equation A.5

$$\begin{aligned} \frac{\partial}{\partial \ell^\theta(\omega)} \left(\frac{\psi_\theta(a \mid \boldsymbol{\ell}^\theta, \omega)}{\psi_\theta(a \mid \boldsymbol{\ell}^\theta, \omega^*)} \right) = \left(\frac{\psi_\theta(a \mid \boldsymbol{\ell}^\theta, \omega)}{\psi_\theta(a \mid \boldsymbol{\ell}^\theta, \omega^*)} \right) \Bigg\{ \frac{\partial p^\theta(\boldsymbol{\ell}^\theta)}{\partial \ell^\theta(\omega)} \Bigg[[1 - 2\lambda] f_\zeta(p^\theta(\boldsymbol{\ell}^\theta)) \left(\frac{a - N\alpha_\theta(\boldsymbol{\ell}^\theta, \omega)}{\alpha_\theta(\boldsymbol{\ell}^\theta, \omega)[1 - \alpha_\theta(\boldsymbol{\ell}^\theta, \omega)]} \right) \\ - [1 - 2\lambda^*] f_{\zeta^*}(p^\theta(\boldsymbol{\ell}^\theta)) \left(\frac{a - N\alpha_\theta(\boldsymbol{\ell}^\theta, \omega^*)}{\alpha_\theta(\boldsymbol{\ell}^\theta, \omega^*)[1 - \alpha_\theta(\boldsymbol{\ell}^\theta, \omega^*)]} \right) \Bigg] \Bigg\}, \end{aligned} \tag{A.36}$$

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.35 is less than 0 so long as A.36 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

$$C^\theta(\omega) := \frac{\partial p^\theta(\mathbf{0})}{\partial \ell^\theta(\omega)} [1 - 2\lambda^*][1 - \alpha_\theta(\mathbf{0}, \omega^*)] \left[f_\zeta(p^\theta(\mathbf{0})) - f_{\zeta^*}(p^\theta(\mathbf{0})) \right] < 0. \tag{A.37}$$

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, $\theta = l \Rightarrow \omega^* = (L, \underline{\lambda}) \Rightarrow C^l(R, \underline{\lambda}) < 0 \Leftrightarrow \underline{\lambda}[1 - 2\underline{\lambda}][f_R(1) - f_L(1)] > 0$, which holds since $f_R(1) > f_L(1)$ and $\underline{\lambda} < \frac{1}{2}$. And, $\theta = r \Rightarrow \omega^* = (R, \overline{\lambda}) \Rightarrow C^r(L, \overline{\lambda}) < 0 \Leftrightarrow (1 - \overline{\lambda})[1 - 2\overline{\lambda}][f_L(0) - f_R(0)] < 0$, which holds since $f_L(0) > f_R(0)$ and $\overline{\lambda} > \frac{1}{2}$.

□

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 $\alpha(\hat{\pi}^l,\hat{\pi}^r,\omega)=\lambda[1-F_\zeta(1-\hat{\pi}^r)]+(1-\lambda)F_\zeta(1-\hat{\pi}^l)$. 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, $\lim_{p\to0}p^\theta(p)=0$ and $\lim_{p\to1}p^\theta(p)=1$. Now consider the condition for confounding beliefs: $\Pr(a_t\mid\hat{\pi}^l,\hat{\pi}^r,L,\underline{\lambda})=\Pr(a_t\mid\hat{\pi}^l,\hat{\pi}^r,R,\overline{\lambda})\Leftrightarrow\alpha(\hat{\pi}^l,\hat{\pi}^r,\underline{\omega})=\alpha(\hat{\pi}^l,\hat{\pi}^r,\overline{\omega})\Leftrightarrow$

$$\underline{\lambda}[1-F_L(p^r(p))]+(1-\underline{\lambda})F_L(p^l(p))=\overline{\lambda}[1-F_R(p^r(p))]+(1-\overline{\lambda})F_R(p^l(p)). \tag{A.38}$$

I now argue that there must exist $p\in(0,1)$ such that Equation A.38 holds. At $p=0$, the left-hand side is $\underline{\lambda}$, and the right-hand side 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.38 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.38. 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_t^n. Let $\psi(a\mid\ell_t^n,\omega)$ be the probability of observation $a\in\{0,1,...,N\}$ in state ω given neutral belief ℓ_t^n. Fix $\omega=\overline{\omega}$. Then process $\langle\ell_t^n\rangle$ evolves according to $\ell_{t+1}^n=\ell_t^n\psi(a\mid\ell_t^n,\underline{\omega})/\psi(a\mid\ell_t^n,\overline{\omega}):=\varphi(a,\ell_t^n)$ with transition probability $\psi(a\mid\ell_t^n,\overline{\omega})$. 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 6 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, $\langle\ell_t^n\rangle$ forms a martingale conditional on $\omega=\overline{\omega}$: $\mathbb{E}[\ell_{t+1}^n\mid\ell_t^n]=\sum_{a=0}^N\psi(a\mid\ell_t^n,\overline{\omega})\varphi(a,\ell_n^t)=\ell_n^t\sum_{a=0}^N\psi(a\mid\ell_t^n\underline{\omega})=\ell_t^n$. □