# Distrust in Experts and the Origins of Disagreement*

Ing-Haw Cheng† Alice Hsiaw‡

First draft: October 2016
This draft: June 2017

**Abstract**

Persistent disagreement about substance and expert credibility often go hand in hand. Prominent examples include disagreements in economics, climate science, and medicine. We argue that disagreement arises because individuals overinterpret how much they can learn when both substance and expert credibility are uncertain. Our proposed learning bias predicts that: 1) Disagreement about credibility drives disagreement about substance; 2) First impressions of credibility drive long-lasting disagreement; 3) Distrust is difficult to unravel; 4) Encountering experts in different order generates disagreement; and 5) Confirmation bias and/or its opposite arise endogenously. These effects provide a theory of the origins of disagreement.

*Keywords:* disagreement, polarization, learning, expectations, experts

*The authors thank Teresa Fort, Jens Großer, Adam Kleinbaum, Botond Kőszegi, Martin Oehmke, Davide Pettenuzzo, Uday Rajan, Tanya Rosenblat, Kathy Spier, Courtney Stoddard, Dustin Tingley, Wei Xiong, seminar participants at Brandeis University, Harvard University, and University of Chicago, and workshop participants at CSWEP CeMENT and the Duke Behavioral Models of Politics conference, for comments.

†Dartmouth College, Tuck School of Business. Email: ing-haw.cheng@tuck.dartmouth.edu.

‡Brandeis University, International Business School. Email: ahsiaw@brandeis.edu.

Disagreement is everywhere, over topics ranging from the causes of climate change to the consequences of immigration. A core feature of many disagreements is that individuals disagree not just about their positions ("Do humans affect climate change?"), but also about the credibility of information sources that inform those positions ("How reliable are scientists and their data?"). In debates over questions of economics ("What is the value of stimulus spending?"), medicine ("Are vaccinations safe for children?"), and politics ("Why is it hard to debunk fake news?"), one side typically expresses supreme confidence in their preferred experts while dismissing the other side's trusted sources.

Why do individuals jointly disagree about substance and expert credibility, and why does disagreement persist? The simplest model of Bayesian learning about an unknown state of the world does not answer this question, as it assumes that individuals know the expert's credibility and thus how much weight to apply to an expert's signals. Introducing unknown expert credibility requires individuals to learn about both credibility and substance. Bayesians update beliefs by letting their priors determine how much weight to give to signals, and as a consequence, two Bayesians who share common priors and signals agree.

Our theory is that disagreement about substance and credibility are inextricably linked through a learning bias we call *pre-screening*. Pre-screeners err by overinterpreting how much they can learn about both credibility and substance from just the expert's signals. They update in two steps. In the first step, they use the signals to assess credibility, before updating in a second step using this newly-assessed belief to weight the signals. This process errs by "double-dipping" the data: Bayesians take care to use only their prior beliefs to weight signals, and avoid letting signals influence the weights. This error leads to disagreement over substance that is rooted in different first impressions of expert credibility that arise when individuals experience signals—or experts—in different order. Disagreement does not require differences in factual information, and arises even when individuals share common priors, objective information, and pre-screening biases.

Previous work has had difficulty explaining why disagreement over substance and credibility are endogenously linked. Because Bayesians with common priors and information agree, one strand of work has focused on biased learning about substance as a source of disagreement. Examples include models of overconfidence (Scheinkman and Xiong, 2003;

Ortoleva and Snowberg, 2015) and confirmation bias (Rabin and Schrag, 1999). However, these theories assume that individuals hold exogenously fixed beliefs about credibility.

A second strand of work studies uncertain credibility but assumes that individuals are rational Bayesians with heterogeneous priors. Disagreement evolves from these priors (Acemoglu et al., 2016), amplified by strategic information slanting from experts or the media (Gentzkow and Shapiro, 2006). These theories have had important success in explaining polarized disagreements. However, they do not address whether rationality best describes individuals involved in high-pitched disagreements such as the debate over vaccination safety, where actress Jennifer McCarthy Wahlberg has an outsized influence, or climate change, where individuals compare each other's beliefs to a "religious mantra" (Bell, 2011) and an "intellectual stance... uncomfortably close to Hitler's" (Snyder, 2015). By assuming heterogeneous priors, these theories are also poorly positioned to make predictions about the *origins* of disagreement.

By organizing insights around biased learning and disagreement about credibility rather than substance, we make two contributions. First, we offer a parsimonious theory of the origins of joint disagreement about substance and credibility, with testable implications that capture aspects of real-world disagreements that existing theories struggle to explain. For example, in the disagreement over climate science, our theory predicts that individuals with early exposure to professional climate skeptics are more likely to become skeptics relative to those who had early exposure to the consensus first, even if both groups begin with common priors and ultimately observe the same information. Once first impressions influence beliefs about credibility, exposure to the other side's experts may harden disagreement, not resolve it, helping explain why disagreements over credibility and substance persist.

Second, our theory of biased learning about credibility links behaviors individually predicted by theories of biased learning about substance. Our theory predicts when behaviors similar to confirmation bias and overconfidence arise, while also making new predictions about when their opposites arise. Thus, pre-screening not only endogenously leads to disagreement and heterogeneous beliefs, but also heterogeneous *biases* in learning about the state. For example, climate skeptics may be under-confident in mainstream science (relative to an objective observer) if early exposure indicated low credibility, while others may be

overconfident if early exposure indicated high credibility.

Section 1 describes the information environment and learning. The key frictions we assume are that individuals lack the training to evaluate primary evidence themselves, so form opinions about an unknown state only from signals delivered by experts. Expert credibility is also unknown, and outside signals cannot fully resolve uncertainty about credibility. Although stark, these assumptions describe several learning situations in economics, climate science, medicine, and politics.

We develop a simple baseline model of an individual who learns about a binary state ($A$ or $B$) using signals from an expert who is of either high or low quality. A high-quality expert is more accurate and informative about the true state than a low-quality expert. To focus on the effect of learning, experts are simply signal sources and are not strategic. A Bayesian updates her beliefs over the joint distribution of the state and expert quality by letting her prior beliefs determine how much weight to give to signals. While signals certainly influence ex-post posterior beliefs about credibility, they do not influence the weight a Bayesian applies to signals during the updating process—only ex-ante prior beliefs determine this weight.

Pre-screening is a quasi-Bayesian process that explores a single conceptual error: the overinterpretation of information about credibility. A pre-screener updates in two steps. First, after seeing a signal, she attempts to figure out the expert's quality by forming a first-stage belief about it using Bayes' rule. For example, she forms a belief about a climate scientist's quality based on what the scientist reported ("Does her claim pass the smell test?"). Second, the pre-screener applies this first-stage updated belief about expert quality to weigh all of the expert's signals in forming her posterior belief ("Since her claim indicates she's a poor scientist, everything she has said is unreliable."). By letting signals influence the weight applied to the available data, these two steps "double-dip" the data.

Pre-screening links together two ideas in the experimental literature. First, individuals tend to overinterpret or "double-count" information, particularly in light of current beliefs (Lord, Ross and Lepper, 1979). Second, perceptions about source credibility influence message processing (Petty and Wegener, 1998). In our setting, information frictions force individuals to infer both expert quality and the state from just the experts' signals, creating a situation where the error of double-counting may escape notice. Our process also parallels

"empirical Bayes" methods in statistics, where a researcher first calibrates her prior for unknown hyperparameters using the data before performing a full analysis (Carlin and Louis, 2000). Lindley (1969) notes that "there is no one less Bayesian than an empirical Bayesian."

Section 2 shows how pre-screening generates disagreement. A pre-screener and Bayesian share the same belief about the state in expectation, but disagree about both states and quality along nearly every ex-post realized signal path. We use the following terminology: A pre-screener overtrusts (under-trusts) an expert if her posterior belief that the expert is high quality is higher (lower) than the Bayesian's. A pre-screener is optimistic (pessimistic) if she over-(under-)estimates the likelihood of the objectively more likely state as judged by a Bayesian.

The first prediction of our theory is that, starting from common priors, pre-screeners are ex-post optimistic if and only if they overtrust the expert and are pessimistic if and only if they under-trust the expert. Intuitively, over-inferring that the expert is high (low) quality means that the pre-screener overweights (underweights) the expert's signals. Disagreement between a Bayesian and pre-screener about states and quality thus go hand-in-hand. Disagreement between two pre-screeners is analogous and arises even when their signals contain the same objective information, so long as they receive signals in different orders.

Section 3 delves into the mechanism. Here, we focus on disagreement between a Bayesian and pre-screener to isolate the role of the bias. The second set of predictions shows that, for pre-screeners, first impressions about credibility have persistent influences on subsequent beliefs about the state, whereas a Bayesian's beliefs are invariant to signal order. Early sequences that have few signal reversals generate significant overtrust and optimism, or a positive first impression, while early sequences containing many signal reversals generate under-trust and pessimism, or a negative first impression. We show that these first impressions generate persistent disagreement even as the expert reports more signals.

Third, we show that pre-screening generates a tendency towards pessimism: negative first impressions are harder to unravel than positive first impressions. The overtrust and optimism from a positive first impression can be undone given enough subsequent mixed signals, which suggest low quality. In contrast, the under-trust and pessimism from a negative first impression can cast a long shadow, persisting even when subsequent identical signals indicate

high quality. This fundamental asymmetry arises because mixed signals are relatively worse news for expert quality than identical signals are good news.

Fourth, the model predicts that the order in which individuals encounter experts can drive disagreement. There is an asymmetry between "inside" experts (those who have already reported) and new unknown "outsiders" (those just reporting). If a pre-screener has a strong positive first impression of an inside expert, contrary information from the same expert will help resolve disagreement. However, the same information delivered by an outsider will lead the pre-screener to discredit the outsider and bolster her trust in the insider, paradoxically producing a "backfire effect" that strengthens disagreement (Nyhan and Reifler, 2010).

Section 4 shows how a rich set of behaviors distinguish our theory from others, including inattention, overconfidence, models of expert or media slant, and confirmation bias. Confirmation bias predicts that individuals interpret information in the direction that confirms preconceived beliefs—e.g., by interpreting contradictory signals as if they were confirmatory, or by under-reacting to them more generally (Rabin and Schrag, 1999). Pre-screening is distinct from confirmation bias about either the state or credibility because the key error is over-interpretation of information about credibility irrespective of whether signals confirm or contradict current beliefs. A fifth set of predictions shows how this generates behavior similar to confirmation bias, or opposite to it, depending on how new signals affect the pre-screener's first-stage trust in the expert. A new "undermining effect" that has received recent experimental support (De Filippis et al., 2017) can arise, where agents *over*-react to a contradictory signal because it excessively undermines the credibility of the entire history of the expert's signals ("Can I trust anything they said?").

Section 5 argues that these distinctions help explain real-world debates beyond existing theories and discusses how they are testable. For example, models of inattention study why individuals fail to pay attention to certain signals (Schwartzstein, 2014; Kominers, Mu and Peysakhovich, 2016). But when considering the debate over climate change, or differences between the public's and economists' views on many issues (Sapienza and Zingales, 2013), disagreement plausibly reflects active distrust in scientists and economists, rather than not noticing what they have said. In our framework, individuals pay a lot of attention to all signals, yet vigorously disagree about substance *because* they disagree about credibility.

Section 6 concludes with further implications of our model. For concreteness, the bulk of the paper focuses on the case where signals come from external sources. However, our theory does not require this. Alternatively, the source can be the individual's own experiences, with uncertainty arising because the individual does not know how informative her experiences are about the true state. This interpretation is broadly related to the idea that people's life experiences drive belief formation (Malmendier and Nagel, 2011, 2016).

# 1 Model

## 1.1 Information environment

An agent learns about an unknown state $\theta \in \{A, B\}$ by observing binary signals $s_t \in \{a, b\}$ in each period $t$ from an expert. The expert has quality $q \in \{L(ow), H(igh)\}$, which the agent also does not know. The high quality expert has a higher probability of correctly reporting the state than the low quality expert and is more informative: $P(s_t = a|q, A) = P(s_t = b|q, B) = p_q$, where $1 > p_H > p_L > 1 - p_H$ and $p_H > 1/2$ (the least informative expert quality is $1/2$). Experts are not strategic, and nature draws true expert quality independently from the true state. Conditional on state and expert quality, signals are independently and identically distributed. For clarity, we first assume the agent observes one signal per period from a single expert, and later generalize to multiple experts and multiple signals per period.

This simple environment captures three informational frictions: 1) the agent relies on information from the expert to form an opinion about the state of the world; 2) in addition to the state of the world, expert credibility is also unknown; and 3) outside information cannot fully resolve the uncertainty about credibility. We argue that these frictions are relevant for several real-world areas of disagreement.

**Economics.** Several topics in economics ranging from whether government stimulus promotes growth to the desirability of free trade have become the subject of recent heated discussion. Few lay individuals have the expertise or training in theory or data analysis to evaluate primary evidence on these issues, suggesting a need for economists. Yet, from the individual's perspective, the economist's ability is uncertain, and the individual must form

beliefs about it.

Recent evidence suggests that American households view economists skeptically. Sapienza and Zingales (2013) show that average American households have sharply different views than economists on questions ranging from whether it is hard to predict stock prices to whether the North American Free Trade Agreement (NAFTA) increased welfare. They find this difference tends to be large even when there is strong consensus among economists. Of course, economists may be wrong, as there is substantial uncertainty about theoretical models and evidence. Assuming that economists are type $H$, we capture this as $p_H < 1$. If there were no uncertainty about quality, individuals would also view economists as type $H$, and form beliefs about economic issues accordingly. But when told that economists agree that the stock market is unpredictable, average beliefs among households hardly moved. If anything, an even larger percentage of households thought that the market *was* predictable. This suggests the more troubling possibility that households view economists as type $L$.

Why don't independent signals about expert quality, or "credentials," resolve this gap, which persists despite doctorates, chaired professorships, and Nobel prizes? We conjecture at least three reasons. First, in some settings, credentials may not be objectively very informative about the quality of specific signals or forecasts. DellaVigna and Pope (2016) run a large experiment estimating how different incentive schemes affect effort, and ask economists to forecast the effectiveness of each treatment ex-ante. They find that the average expert forecast does well according to a wisdom-of-crowds measure, but the forecast error of experts is disperse and objective measures of expertise do not improve forecasting accuracy.

Second, the informativeness of credentials may itself be uncertain to households, even if they are informative to other experts. Programs that grant economics Ph.D.'s have varying strengths in different areas of economics, based on different publication records and practical experience. Even if an individual with no expertise knew all of these facts about, say, a particular trade economist, she would have little idea of how to evaluate them together. Another trade economist might, but that is of no help to the non-expert because that economist's quality is also unknown.

Third, perhaps the most useful tool for evaluating expert reliability—the ability to compare predictions to outcomes through repeated controlled experiments—is unavailable in

many economic settings. Economists may predict that $X$ (free trade) will cause $Y$ (growth, improved welfare) ex-ante, but evaluating whether any given instance of $X$ (NAFTA) did actually cause $Y$ ex-post is difficult, *even for economists*, because of the difficulty of empirical identification from observational data.

**Climate science, medicine, and politics.** Few individuals have the expertise to evaluate the extent to which humans affect climate change, yet many people have very strong opinions about the topic. Disagreement between climate "deniers" and supporters of the proposition is largely about the credibility of the majority the scientific community who support the proposition versus a small minority of scientists who do not. Repeated experimentation to verify expert quality is impossible: climate change is a slow-moving process we observe once, which leaves residual uncertainty about credibility.

Medicine is another area where individuals rely on experts—doctors—to provide information. For example, take the recent controversy over the safety of childhood vaccinations. The unknown true state is whether childhood vaccinations are safe, but a parent does not know how well her doctor's advice correlates with the true state. For the parent, there are limited opportunities for repeated experimentation to learn about the doctor's quality, and a lay person is unlikely to understand the specific differences between medical training and alternative medicine, let alone medical degrees from various schools in various specialties.

In politics, the proliferation of "fake news" over social media platforms, such as Facebook and Twitter, during the recent U.S. presidential election highlights the importance of uncertain expert quality (The Economist Magazine, 2016). Recent evidence suggests that fake news may be particularly hard to debunk—young individuals in particular attach credibility to content even when there are indicators that the source is not trustworthy (Stanford History Education Group, 2016; Shellenbarger, 2016).

In summary, households face substantial uncertainty about the informativeness of experts, and credentials which should help resolve this uncertainty are often of uncertain informativeness themselves. This uncertainty may remain even if one expands the set of experts an agent observes, because each source carries additional uncertainty about informativeness. To isolate how learning might occur, we focus on a stark environment with no signals that explicitly convey expert quality.

## 1.2 Learning

Suppose the agent has the prior that the state and quality are independent with marginal probabilities $(\omega_0^\theta, \omega_0^q) \in (0,1) \times (0,1)$, respectively, so that her joint prior over both, $\omega_0$, is given by Table 1. Let the agent observe a sequence of $n$ signals, denoted $\mathbf{s}^n = (s_1, s_2, \ldots s_n)$,

| | $\theta = A$ | $\theta = B$ |
|---|---|---|
| $q = H$ | $\omega_0^H \omega_0^A$ | $\omega_0^H(1-\omega_0^A)$ |
| $q = L$ | $(1-\omega_0^H)\omega_0^A$ | $(1-\omega_0^H)(1-\omega_0^A)$ |

Table 1: Joint prior beliefs $\omega_0$

where one signal is observed each period.

A Bayesian's posterior belief $P^u(q, \theta|\mathbf{s}^n)$ equals:

$$P^u(q,\theta|\mathbf{s}^n) = \frac{\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^\theta \omega_0^q}{\sum_q \sum_\theta \left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^\theta \omega_0^q}. \tag{1}$$

The Bayesian uses her prior belief $\omega_0^q$ about expert quality to weight the likelihood of signals $(\prod_{t=1}^n P(s_t|q,\theta))$, and infers her posterior belief $P^u(q,\theta|\mathbf{s}^n)$ in one step.

Our proposed bias, which we call *pre-screening*, works in two steps. First, a pre-screener applies Bayes Rule to update beliefs about the expert's quality by combining the signal's content with the joint prior on expert quality and state, denoted $\kappa_q(\mathbf{s}^n)$. Heuristically, the pre-screener checks to see if the claim passes the "smell test." Second, she weights all observed information by using her first-stage *updated* belief about the expert's quality $\kappa_q(\mathbf{s}^n)$ to form posterior beliefs on the joint distribution of state and quality, denoted $P^b(q,\theta|\mathbf{s}^n)$.

To illustrate the pre-screener's updating algorithm, suppose she observes two signals, one in each period. After observing the first signal $(s_1)$, the first-stage updated belief about the expert's quality, $\kappa_q(s_1)$, is:

$$\kappa_q(s_1) = \frac{\omega_0^q \sum_\theta P(s_1|q,\theta)\omega_0^\theta}{\sum_q \sum_\theta P(s_1|q,\theta)\omega_0^\theta \omega_0^q}.$$

Using $\kappa_q(s_1)$ to form the joint posterior belief on the state and quality, $P^b(q,\theta|s_1)$, yields the

pre-screener's posterior beliefs after the first signal:

$$P^b(q,\theta|s_1) = \frac{P(s_1|q,\theta)\kappa_q(s_1)\omega_0^\theta}{\sum_q\sum_\theta P(s_1|q,\theta)\kappa_q(s_1)\omega_0^\theta}.$$

After observing the second signal ($s_2$), the pre-screener's first-stage updated belief about the expert's quality, $\kappa_q(s_1,s_2)$, is:

$$\kappa_q(s_1,s_2) = \frac{\sum_\theta P(s_2|q,\theta)P^b(q,\theta|s_1)}{\sum_q\sum_\theta P(s_2|q,\theta)P^b(q,\theta|s_1)}.$$

The pre-screener then uses $\kappa_q(s_1,s_2)$ to form her joint posterior belief on the state and quality by re-weighting all the information from the expert. The posterior, $P^b(q,\theta|s_1,s_2)$, equals:

$$P^b(q,\theta|s_1,s_2) = \frac{P(s_2|q,\theta)P(s_1|q,\theta)\kappa_q(s_1,s_2)w_0^\theta}{\sum_q\sum_\theta P(s_2|q,\theta)P(s_1|q,\theta)\kappa_q(s_1,s_2)\omega_0^\theta}.$$

Iterating on the pre-screener's updating process allows us to characterize her posterior beliefs.

**Definition 1 (Pre-screener's beliefs)** *After observing a sequence of $n$ signals $\mathbf{s}^n$ from an expert, the **pre-screener's first-stage updated belief** about expert quality, $\kappa_q(\mathbf{s}^n)$, is given by:*

$$\kappa_q(\mathbf{s}^n) = \frac{\kappa_q(\mathbf{s}^{n-1})\sum_\theta\left(\prod_{t=1}^n P(s_t|q,\theta)\omega_0^\theta\right)}{\sum_q\kappa_q(\mathbf{s}^{n-1})\sum_\theta\left(\prod_{t=1}^n P(s_t|q,\theta)\omega_0^\theta\right)}, \tag{2}$$

*where $\kappa_q(\emptyset)=\omega_0^q$. The **pre-screener's final joint posterior** on expert quality and the state, $P^b(q,\theta|\mathbf{s}^n)$, is given by:*

$$P^b(q,\theta|\mathbf{s}^n) = \frac{\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\kappa_q(\mathbf{s}^n)\omega_0^\theta}{\sum_q\sum_\theta\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\kappa_q(\mathbf{s}^n)\omega_0^\theta}. \tag{3}$$

This definition assumes ex-ante independence of states and quality. We maintain this assumption for our analysis both for simplicity and because it isolates how pre-screening affects the evolution of correlated beliefs about states and quality without assuming any correlation ex-ante. We provide a generalized definition in Appendix A.1. If there is no uncertainty about expert quality ($p_L = p_H$), the first step becomes innocuous, so the pre-screener's

posterior beliefs are identical to the Bayesian's for any sequence of signals.

With uncertainty about quality, the pre-screener errs relative to the Bayesian in thinking that the signal provides more information about credibility than it objectively contains. She does so in two ways. First, she uses the latest signal $s_n$ to form a belief about expert quality $\kappa_q(\mathbf{s}^n)$ before updating on the joint distribution of unknowns, leading her to update twice. Second, the pre-screener also errs by re-evaluating the informativeness of all observed signals $\mathbf{s}^n$ using this updated weight. A Bayesian's updating process reduces to using the initial $\omega_0^q$ prior to weight all observed signals $\mathbf{s}^n$, whereas a pre-screener substitutes $\kappa_q(\mathbf{s}^n)$ for $\omega_0^q$ (compare Equations 1 and 3). The net result is that, by using the data to ascertain expert quality ahead of an otherwise-Bayesian update, the pre-screener over-infers expert quality by using the same signal content multiple times, "double-dipping" the data.[1]

We motivate our bias in two ways. First, our bias links together two elements with significant experimental support in economics and social psychology. The first element is that individuals over-interpret information, particularly in light of current beliefs. The information processing mechanism of erroneously using updated beliefs to form posterior beliefs was conjectured by Lord, Ross and Lepper (1979, p.2106–2107). In an experimental setting examining how subjects updated beliefs in response to information about capital punishment, they write that "Our subjects' main inferential shortcoming, in other words, did not lie in their inclination to process evidence in a biased manner...Rather, their sin lay in their readiness to use evidence already processed in a biased manner to bolster the very theory or belief that initially 'justified' the processing bias."

As Rabin and Schrag (1999, p.46–47) discuss, the "sin" is analogous to a teacher who first assigns a student a low grade because she unfavorably interprets an unclear answer

[1]One can ask what happens with one error and not the other. The first error relates to how today's information is mistakenly processed, while the second error relates how that mistake feeds back into beliefs. Having made the first error, one can change how this feeds back into beliefs. One can also make the second error without mistakenly evaluating today's information. We include both as they follow our general modeling principle that agents may over-infer expert quality: having formed an erroneous opinion using the most recent signal (the first error), one may wish to explicitly re-evaluate all previous signals in light of this belief (the second). One can also ask what happens if the agent updates first on the state, then the credibility. We focus on updating on credibility first given that our premise is that informational frictions make credibility unknown. The experimental evidence discussed in this section suggests that credibility is a key moderator in how persuasive people find messages about states. A model where the agent updates first on the state would be driven by first impressions on the state and therefore make substantially different qualitative predictions from ours and would instead coincide significantly with Rabin and Schrag (1999).

from the student as consistent with priors about low ability, but then goes on to erroneously use the low grade as *further* or *additional* evidence of low ability. They refer to this as "hypothesis-based filtering" and note that it is likely when individuals face complex information environments. In their words, after interpreting ambiguous data, "people tend to use the consequent 'filtered' evidence inappropriately as further evidence for these hypotheses."

Other experiments suggest that perceiving more information than objectively warranted occurs in settings when inference is difficult. In a setting where subjects receive signals from multiple sources with common underlying information, Enke and Zimmermann (2016) show that subjects fail to recognize that signals are correlated and hence double-count information, but compute beliefs correctly after this is pointed out. Eyster and Rabin (2010) study naïve herding in a similar context. Chadrasekhar, Larreguy and Xandri (2015) confirm in a field experiment that agents double-count information when learning from others. Double-counting information is the fundamental error in pre-screening.

The second element, supported by a large literature in social psychology, is that credibility influences the degree to which individuals are persuaded by a given message (Petty and Wegener, 1998; Johnson, Maio and Smith-McLallen, 2005). Subjective perceptions of credibility matter more than objective qualifications (Kruglanski and Stroebe, 2005, p.345). Consistent with this, several subjective factors intermediate the effect of objective credibility on attitude changes. Credibility matters more if, as in our information environment, direct knowledge about the topic is low (Petty and Wegener, 1998, p.344). Inference differs depending on whether experimental subjects were told objective credentials before or after a given message, consistent with our assumption that individuals process beliefs about credibility before updating further (Albarracín and Vargas, 2010, p.410).

Second, our learning process parallels certain applications of "empirical Bayes" methods in statistics, where a researcher first uses the data to calibrate a set of hyperparameters governing her prior (Carlin and Louis, 2000). Here, in the first step, a pre-screener uses the signal itself to calibrate the weight that is applied to it through Bayes' Rule in the second step. Early proponents of empirical Bayes methods were involved in a debate with proponents of "full Bayesian" methods that require a fully specified prior. Critics have noted that "double-dipping" the data using empirical Bayes methods can lead to erroneous inference; Lindley

(1969) famously noted that "there is no one less Bayesian than an empirical Bayesian."

# 2 Disagreement

We begin by focusing on disagreement between a Bayesian and pre-screener, or the bias itself. In the canonical case where the prior on the state is neutral ($\omega_0^\theta = 1/2$) and the agent observes only one signal ($n = 1$), the Bayesian and biased belief are equal: $P^b(q, \theta|\mathbf{s}^n) = P^u(q, \theta|\mathbf{s}^n)$, because $\kappa_q(\mathbf{s}^n) = w_0^q$. In this case, the first step is innocuous because the ex-ante belief is that both states are equally likely and independent of quality, so that the pre-screener finds a single signal about the state uninformative about the expert's quality. This example shows that disagreement with a Bayesian is not exogenously built in.

With more signals, the ex-ante difference between the pre-screener and Bayesian's posterior marginal beliefs about the state equals zero as long as agents begin with the prior that both states are equally likely. This is because beliefs about states are ex-ante symmetric around $A$ and $B$ and are ex-ante independent of quality, as we show in:

**Proposition 1 (No average disagreement about $\theta$)** *Let a Bayesian and pre-screener share a common prior of $(\omega_0^A, \omega_0^H) = (1/2, \hat{q})$ for any $\hat{q} \in (0, 1)$, and suppose this represents the true distribution from which nature draws $(\theta, q)$. Then $E_0[P^b(\theta = A|\mathbf{s}^n) - P^u(\theta = A|\mathbf{s}^n)] = 0$, where the expectation $E_0$ is taken over this distribution and all signal paths $\mathbf{s}^n$. However, $E_0\left[\left(P^b(\theta = A|\mathbf{s}^n) - P^u(\theta = A|\mathbf{s}^n)\right)^2\right] > 0$.*

Although there is no ex-ante disagreement about $\theta$, there is ex-post disagreement among realized paths: the expected squared (or absolute) difference in marginal posteriors about $\theta$ is strictly positive. The key reason this occurs is because the pre-screener's beliefs are path-dependent, while the Bayesian's are not path-dependent and depend only on the information content of signals, defined as:

**Definition 2 (Information content)** *The **information content** of any sequence of signals $\mathbf{s}^n$ is given by the number of "a" signals $n_a$ and the number of "b" signals $n_b$.*

To see why a pre-screener's beliefs are path-dependent, rearrange Equations 2 and 3 to

obtain:

$$P^b(q,\theta|\mathbf{s}^n) = \frac{\beta_q(\mathbf{s}^n)\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^\theta\omega_0^q}{\sum_q \beta_q(\mathbf{s}^n)\sum_\theta\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^\theta\omega_0^q}, \tag{4}$$

where $\beta_q(\mathbf{s}^n)$ is defined as:

$$\begin{aligned}\beta_q(\mathbf{s}^n) &\equiv \left(\textstyle\sum_\theta P(s_1|q,\theta)\omega_0^\theta\right) \times \left(\textstyle\sum_\theta P(s_1|q,\theta)P(s_2|q,\theta)\omega_0^\theta\right) \times \ldots \times \left(\textstyle\sum_\theta P(s_1|q,\theta)P(s_2|q,\theta)\ldots P(s_n|q,\theta)\omega_0^\theta\right) \\ &= \prod_{m=1}^n \left(\sum_\theta \left(\prod_{t=1}^m P(s_t|q,\theta)\right)\omega_0^\theta\right), \end{aligned} \tag{5}$$

and where $\beta_q(\emptyset) \equiv 1$. In these equations, $\beta_q(\mathbf{s}^n)$ reflects the cumulative effect of pre-screening and re-weighting information by updated beliefs about quality after every signal. Early signals appear more often in Equation 5, and thus carry more weight than later signals in the formation of the pre-screener's beliefs.

As an example, consider the signal sequences $\{a,a,b\}$ and $\{b,a,a\}$, which have identical information content. Let the Bayesian and pre-screener have common priors. A Bayesian's posterior beliefs depend only on the information content: $P^u(q,\theta|\{a,a,b\}) = P^u(q,\theta|\{b,a,a\})$. The pre-screener's beliefs depend both on the information content but also signal order: $P^b(q,\theta|\{a,a,b\}) \neq P^b(q,\theta|\{b,a,a\})$, because $\beta_q(\{a,a,b\}) \neq \beta_q(\{b,a,a\})$. With $\omega_0^\theta = 1/2$, under the sequence $\{a,a,b\}$, the pre-screener over-infers that the expert is type $H$ after the second signal, while under $\{b,a,a\}$, she over-infers that the expert is type $L$. The difference in this early part of the signal sequence colors how the pre-screener interprets the final signal and highlights the relevance of *first impressions about credibility.*

To characterize disagreement, we define the following terms to simplify exposition.[2]

**Definition 3 (Optimism and trust)** *Fix the information content with $n_a > n_b$ without loss of generality. Given a signal sequence $\mathbf{s}^n$,*

1. *A pre-screener is **optimistic** if $Pr^b(\theta = A|\mathbf{s}^n) > Pr^u(\theta = A|\mathbf{s}^n)$, and **pessimistic** if strictly less than ($<$).*

[2]Because there is no sense in which $A$ is a better outcome than $B$, a more precise definition would replace optimism with "over-estimates the likelihood of $A$" and pessimism with "under-estimates the likelihood of $A$." We choose "optimistic" and "pessimistic" purely for brevity.

2. *A pre-screener* ***overtrusts*** *if* $Pr^b(q = H|\mathbf{s}^n) > Pr^u(q = H|\mathbf{s}^n)$*, and* ***under-trusts*** *if strictly less than* $(<)$*.*

Proposition 2 shows that disagreement about states and expert quality between a pre-screener and Bayesian go hand-in-hand: Whether a pre-screener is ex-post optimistic or pessimistic is determined by whether she ex-post over- or under-trusts the expert.

**Proposition 2 (Correlated disagreement)** *For any* $\mathbf{s}^n$ *with* $n_a > n_b$*, and for all* $\omega_0^\theta \in (0,1)$ *and* $\omega_0^q \in (0,1)$*, the pre-screener under-trusts the expert if and only if she is pessimistic about the more likely state:* $P^b(q = H|\mathbf{s}^n) < P^u(q = H|\mathbf{s}^n)$ *if and only if* $P^b(\theta = A|\mathbf{s}^n) < P^u(\theta = A|\mathbf{s}^n)$*. The pre-screener overtrusts the expert if and only if she is optimistic in beliefs about the more likely state:* $P^b(q = H|\mathbf{s}^n) > P^u(q = H|\mathbf{s}^n)$ *if and only if* $P^b(\theta = A|\mathbf{s}^n) > P^u(\theta = A|\mathbf{s}^n)$*.*

Intuitively, if the pre-screener under-trusts the expert, she is too skeptical about the information content of the expert's signals. If the signals imply that $A$ is objectively likely, then the pre-screener will believe $A$ is less likely than it objectively is. Conversely, if the pre-screener thinks $A$ is less likely than the Bayesian, it must be because she under-trusts the expert. Analogously, overtrust is positively correlated with optimism. Note that Proposition 2 does not depend on assuming a neutral prior about the state, as it holds for any $\omega_0^\theta \in (0,1)$.

Disagreement between two pre-screeners is analogous. Proposition 3 compares the beliefs of two pre-screeners who have identical priors and observe sequences with identical information content, but who observe different signal orderings. The analog of Proposition 2 applies—a pre-screener who trusts the expert more (less) must also believe the reported state is more (less) likely, and vice versa. Disagreement arises on many paths, so that the expected squared difference in beliefs is positive. Thus, our framework generates disagreement even when agents share identical information content, learning biases, and priors, providing a foundations for the origins of disagreement.

**Proposition 3 (Origins of disagreement)** *Suppose two pre-screeners,* $J$ *and* $M$*, have identical priors* $(\omega_0^A, \omega_0^H) = (\hat{\theta}, \hat{q})$ *for any* $\hat{\theta} \in (0,1)$ *and* $\hat{q} \in (0,1)$*, and observe signal sequences* $\mathbf{s}_J^n$ *and* $\mathbf{s}_M^n$ *that have identical information content but different signal orders, where* $n_a + n_b = n$ *and* $n_a > n_b$*.*

1. *(Correlated disagreement) Agent $J$ trusts the expert more than agent $M$ does if and only if agent $J$ believes state $A$ is more likely than agent $M$ does: $P^b(q = H|\mathbf{s}_J^n) > P^b(q = H|\mathbf{s}_M^n)$ if and only if $P^b(\theta = A|\mathbf{s}_J^n) > P^b(\theta = A|\mathbf{s}_M^n)$. Likewise, Agent $J$ trusts the expert less than agent $M$ if and only if agent $J$ believes state $A$ is less likely than agent $M$ does: $P^b(q = H|\mathbf{s}_J^n) < P^b(q = H|\mathbf{s}_M^n)$ if and only if $P^b(\theta = A|\mathbf{s}_J^n) < P^b(\theta = A|\mathbf{s}_M^n)$.*

2. *(Squared disagreement about $\theta$) $E_0\left[\left(P^b(\theta = A|\mathbf{s}_J^n) - P^b(\theta = A|\mathbf{s}_M^n)\right)^2\right] > 0$, where the expectation $E_0$ is taken over the distribution of all signal paths $\mathbf{s}_i^n$ where each path $i$ has identical fixed information content.*

# 3 How Do Trust and Disagreement Evolve?

To isolate the role of the mechanism, we return to disagreement between a Bayesian and a pre-screener. Recall from the discussion following Proposition 1 that first impressions about experts matter: early signals color the interpretation of later signals through their effect on beliefs about credibility. To further characterize this, we show that there is a unique sequence of signals that generates the maximal over- and under-trust for any fixed information content:

**Lemma 1 (First impressions about experts)** *Let $(\omega_0^A, \omega_0^H) = (1/2, \hat{q})$ for any $\hat{q} \in (0, 1)$. Consider a given combination of $n_a$ $a$ signals and $n_b$ $b$ signals, where $n_a > n_b \geq 1$. The sequence in which $n_a$ consecutive $a$ signals is followed by $n_b$ consecutive $b$ signals generates the maximal degree of trust in the expert. The sequence in which $n_b$ pairs of $(a, b)$ signals is followed by $n_a - n_b$ "$a$" signals generates the minimal degree of trust in the expert.*

Lemma 1 shows that, holding information content fixed, pre-screeners erroneously believe that the timing of signal reversals is itself informative, in that a pattern of few (more) initial reversals inflates (deflates) their beliefs about expert quality and therefore the most likely state. While fewer (more) initial reversals objectively suggest that the expert is high (low) quality, the pre-screener's overweighting of this inference colors her interpretation of future signals, leading to overtrust (under-trust). Holding information content fixed, re-ordering the signals so that the longest consistent string appears first generates the most trust in the

expert, while alternating the signals first generates the least trust. In contrast, a Bayesian's final beliefs are independent of signal order given fixed information content.

## 3.1 The origins of disagreement: First impressions

First impressions generate disagreement about expert quality that persists even as the expert reports countervailing information. We characterize persistence by providing minimum bounds for how long this disagreement persists (Proposition 4) as well as conditions under which disagreement survives an arbitrary length of subsequent signals (Proposition 5). By Proposition 2, this persistence filters through to disagreement about the state.

First consider how long overtrust lasts after a positive first impression, i.e., observing an initial sequence that generates overtrust. Given that reversals are negative information about quality (Lemma 1), we ask how many pairs of $(b, a)$ signals it takes to undo the overtrust created after $n_a$ consecutive $a$ signals. Proposition 4, Part 1 shows that the answer depends on the strength of the positive impression and the informativeness of mixed signals. Weaker positive impressions (low $n_a$) unravel easily. Stronger positive impressions (high $n_a$) tend to unravel easily when $p_H$ is high, as mixed signals from the high type are unlikely and strongly indicate low quality. When $p_H$ is low, mixed signals weakly indicate low quality because neither expert type is reliable. In this case, overtrust survives at least $m' > 3$ pairs of $(b, a)$, where $m'$ increases with $n_a$.

Proposition 4, Part 2 shows that under-trust after a negative first impression, i.e. observing an initial sequence that generates under-trust, is harder to unravel than the overtrust from a positive impression. A negative first impression created by $n_b$ pairs of $(a, b)$ survives for at least $m^* > 3$ subsequent identical $a$ signals, where $m^*$ increases with $n_b$, irrespective of $p_L$ and $p_H$. The reason is that there is a fundamental asymmetry between mixed versus identical signals: mixed signals are worse news for quality than identical signals are good news. For example, in the extreme case of $p_L = 1/2$ and $p_H \approx 1$, an $(a, b)$ pair almost immediately rules out the possibility of a high type, while $(a, a)$ does not so obviously imply high quality since the low type also could have produced it by chance. Negative first impressions overweight more-informative mixed signals whereas positive first impressions overweight less-informative identical signals, making negative first impressions more robust.

**Proposition 4 (Minimum bounds for disagreement after first impressions)** *Let* $(\omega_0^A, \omega_0^H) = (1/2, \hat{q})$ *for any* $\hat{q} \in (0,1)$. *First impressions of expert quality persist in the face of contrary information about quality:*

1. *Positive first impressions: Suppose the agent observes* $n_a \geq 1$ *consecutive a signals, followed by* $m$ *pairs of* $(b,a)$ *signals:* $\mathbf{s}^n = (a, a, a, \ldots, b, a, b, a)$.

   (a) *If* $n_a \leq 2$, *the pre-screener under-trusts and is pessimistic about the most likely state for all* $m \geq 1$.

   (b) *If* $n_a \geq 3$, *then there exists some* $m' > 3$ *and* $p' \in (\frac{1}{2}, 1)$ *such that when* $m < m'$ *and* $p_L < p_H \leq p'$, *the pre-screener overtrusts and is optimistic about the most likely state, where* $m'$ *increases with* $n_a$.

2. *Negative first impressions: Suppose the agent observes* $n_b \geq 1$ *pairs of* $(a,b)$ *signals, followed by* $m \geq 1$ *consecutive a signals, where* $m \geq 1$*:* $\mathbf{s}^n = (a, b, a, b, \ldots, a, a, a)$. *Then there exists some* $m^* > 3$ *such that when* $m < m^*$, *the pre-screener under-trusts and is pessimistic about the most likely state, where* $m^*$ *increases with* $n_b$.

Proposition 5 shows that this asymmetry between mixed and identical evidence affects the degree to which first impressions persist in the limit. Enough mixed signals can always unravel a positive first impression. In contrast, arbitrarily high levels of persistence can arise for negative first impressions: fixing any $m$, there exists some combination of $(p_L, p_H)$ that makes mixed signals sufficiently worse news for quality than identical signals are good news, so that the under-trust survives $m$ subsequent identical signals of $a$.

**Proposition 5 (Long-run persistence of first impressions)** *Let* $(\omega_0^A, \omega_0^H) = (1/2, \hat{q})$ *for any* $\hat{q} \in (0,1)$. *Positive first impressions can eventually be undone, but negative first impressions may be arbitrarily persistent:*

1. *Positive first impressions: Suppose the agent observes* $n_a \geq 1$ *consecutive a signals, followed by* $m \geq 1$ *pairs of* $(b,a)$ *signals:* $\mathbf{s}^n = (a, a, a, \ldots, b, a, b, a)$. *For a given* $n_a$, *there exists* $\hat{m}$ *such that when* $m > \hat{m}$, *the pre-screener under-trusts and is pessimistic about the most likely state for any* $(p_L, p_H)$.

2. *Negative first impressions: Suppose the agent observes $n_b \geq 1$ pairs of $(a, b)$ signals, followed by $m \geq 1$ consecutive $a$ signals: $\mathbf{s}^n = (a, b, a, b, \ldots, a, a, a)$. For a given $n_b \geq 1$ and $m \geq 1$, there exists some $\check{p} > \frac{1}{2}$ and $\hat{p} < 1$ such that the pre-screener under-trusts and is pessimistic about the most likely state if $(p_L, p_H)$ satisfies one of the following sufficient conditions: (a) $\hat{p} \leq p_L < p_H$, or (b) $p_L \leq \check{p}$ and $p_H > \hat{p}$.*

## 3.2 Resolving disagreement: Insiders vs outsiders

What resolves the persistent disagreement created by first impressions? Suppose the pre-screener begins with a neutral prior on the state, and observes $k$ identical $a$ signals, resulting in a positive first impression. After an additional $k$ identical $b$ signals from the same expert, she will have the correct posterior on the state, though not necessarily on the expert's quality. Intuitively, the pre-screener understands that all signals originate from the same source, so she realizes that $n_a = n_b = k$ is equivalent to having no new information about the state, even if she is incorrect about the expert's quality.[3]

**Proposition 6 (Resolving disagreement from an overtrusted expert)** *Let $(\omega_0^A, \omega_0^H) = (\hat{\theta}, \hat{q})$ for any $\hat{\theta} \in (0, 1)$ and $\hat{q} \in (0, 1)$. After observing $n_a = k > 1$ consecutive $a$'s, a successive sequence of $n_b = k$ consecutive $b$'s returns the disagreement about the state to zero.*

A more realistic situation is one in which an agent receives additional signals from another expert, or a "second opinion." Suppose the pre-screener now receives signals from two independently drawn experts, $j = 1, 2$, with qualities $q_j$. Let $s_{tj}$ be the signal sent in period $t$ by expert $j \in \{1, 2\}$, who sends a sequence of $n_j$ signals, $\mathbf{s}^{n_j}$. Denote expert 1 as the "inside" expert who reports first, and expert 2 as the "outside" expert who reports second. Let $\mathbf{s}^{n_1, n_2}$ be the sequence of observed signals from both experts, where $\mathbf{s}^{n_1, n_2} = (\mathbf{s}^{n_1}, \mathbf{s}^{n_2})$. Let $\mathbf{s}^{n_1, 0}$ denote the sequence of signals from expert 1 when expert 2 has not yet reported.

The pre-screener now has three sources of uncertainty—the quality of each of the two experts and the state of the world. Since expert quality is independent and identically

[3]This intuition holds more generally: Given any prior on the state, $\omega_0^\theta \in (0, 1)$, observing $n_a = n_b = k$ signals in any order will return the pre-screener back to $\omega_0^\theta$, which is the correct marginal posterior.

distributed, $\omega_0^{q_j} = \omega_0^q$ for all $j$. Since the reliability of a signal $t$ from expert $j$ is independent of the other expert $k$'s quality, $P(s_{tj}|q_j, q_k, \theta) = P(s_{tj}|q_j, \theta)$ for all $t$ where $j \neq k$.

The pre-screening procedure extends naturally from one to multiple sources. First, a pre-screener updates on the joint belief about the experts' qualities, denoted $\kappa_{q_1q_2}(\mathbf{s}^{n_1,n_2})$, by combining the signal's content with the joint prior on expert qualities and state. Second, she uses the updated belief about the experts' qualities to form a joint posterior beliefs on the state and qualities. Iterating on the pre-screener's updating process allows us to characterize posterior beliefs when she receives any set of signals from both experts, $\mathbf{s}^{n_1,n_2}$. The pre-screener's beliefs after observing only expert 1 are:

$$\kappa_{q_1q_2}(\mathbf{s}^{n_1,0}) = \frac{\kappa_{q_1q_2}(\mathbf{s}^{n_1-1,0}) \left(\sum_\theta \left(\prod_{t=1}^{n_1} P(s_{t1}|q_1,\theta)\right) \omega_0^\theta\right)}{\sum_{q_1}\sum_{q_2} \kappa_{q_1q_2}(\mathbf{s}^{n_1-1,0}) \left(\sum_\theta \left(\prod_{t=1}^{n_1} P(s_{t1}|q_1,\theta)\right) \omega_0^\theta\right)}, \tag{6}$$

where $\kappa_{q_1q_2}(\emptyset) = \omega_0^{q_1}\omega_0^{q_2}$, and:

$$P^b(q_1, \theta|\mathbf{s}^{n_1,0}) = \frac{\left(\prod_{t=1}^{n_1} P(s_{t1}|q_1,\theta)\right) \kappa_{q_1q_2}(\mathbf{s}^{n_1,0})\omega_0^\theta}{\sum_q \sum_\theta \left(\prod_{t=1}^{n_1} P(s_{t1}|q_1,\theta)\right) \kappa_{q_1q_2}(\mathbf{s}^{n_1,0})\omega_0^\theta}. \tag{7}$$

After observing expert 2, her beliefs are:

$$\kappa_{q_1q_2}(\mathbf{s}^{n_1,n_2}) = \frac{\left(\sum_\theta \left(\prod_{t=n_1+1}^{n_1+n_2} P(s_{t2}|q_2,\theta)\right) \left(\prod_{t=1}^{n_1} P(s_{t1}|q_1,\theta)\right) \omega_0^\theta\right) \kappa_{q_1q_2}(\mathbf{s}^{n_1,n_2-1})}{\sum_{q_2}\sum_{q_1} \left(\sum_\theta \left(\prod_{t=n_1+1}^{n_1+n_2} P(s_{t2}|q_2,\theta)\right) \left(\prod_{t=1}^{n_1} P(s_{t1}|q_1,\theta)\right) \omega_0^\theta\right) \kappa_{q_1q_2}(\mathbf{s}^{n_1,n_2-1})}, \tag{8}$$

and:

$$P^b(q_1, q_2, \theta|\mathbf{s}^{n_1,n_2}) = \frac{\left(\prod_{t=n_1+1}^{n_1+n_2} P(s_{t2}|q_2,\theta)\right) \left(\prod_{t=1}^{n_1} P(s_{t1}|q_1,\theta)\right) \kappa_{q_1q_2}(\mathbf{s}^{n_1,n_2})\omega_0^\theta}{\sum_{q_2}\sum_{q_1}\sum_\theta \left(\prod_{t=n_1+1}^{n_1+n_2} P(s_{t2}|q_2,\theta)\right) \left(\prod_{t=1}^{n_1} P(s_{t1}|q_1,\theta)\right) \kappa_{q_1q_2}(\mathbf{s}^{n_1,n_2})\omega_0^\theta}. \tag{9}$$

As before, a Bayesian's posterior beliefs depend purely on the information content delivered by each expert, and depend neither on the order in which signals are received from a given expert, nor on the order in which experts are heard.

Now consider the case where the pre-screener has a positive first impression about expert 1 because expert 1 reported $k > 1$ identical signals of $a$. Suppose expert 2 reports $k$ identical

signals of $b$. A Bayesian with $\omega_0^A = 1/2$ infers that the experts cannot both be high quality. Despite both delivering consistent messages, their signals contradict each other, and the Bayesian understands that there is insufficient evidence to deduce which expert is wrong. As a result, she concludes that neither state is more likely than the other. However, the pre-screener with $\omega_0^A = 1/2$ incorrectly trusts the first expert more than the outsider, and therefore incorrectly believes $A$ is more likely:

**Proposition 7 (Outsider rejection)** *Let* $(\omega_0^A, \omega_0^{H_1}, \omega_0^{H_2}) = (1/2, \hat{q}, \hat{q})$ *for any* $\hat{q} \in (0,1)$ *where experts 1 and 2 are independent. Let the agent observe* $k$ *a signals from expert 1, followed by* $k$ *b signals from expert 2:* $\mathbf{s}^{n_1} = (a, \ldots, a)$ *and* $\mathbf{s}^{n_2} = (b, \ldots, b)$ *where* $n_1 = n_2 = k$ *and* $k > 1$.

1. *The pre-screener believes that state A is more likely than B, and that the first expert is more likely to be high quality than the second expert:* $P^b(\theta = A|\mathbf{s}^{n_1,n_2}) > 1/2$ *and* $P^b(H_1|\mathbf{s}^{n_1,n_2}) > P^b(H_2|\mathbf{s}^{n_1,n_2})$.

2. *Persistence in the limit:* $\lim_{k\to\infty} P^b(\theta = A|\mathbf{s}^{n_1,n_2}) = 1$ *and* $\lim_{k\to\infty} P^b(H_1, L_2|\mathbf{s}^{n_1,n_2}) = 1$.

The positive impression from the first expert's consistency inflates the pre-screener's trust in the first expert and deflates trust in the outsider relative to the Bayesian inference. Like the Bayesian, the pre-screener concludes that the experts cannot both be high quality, but incorrectly concludes that the first expert is more credible than the second and therefore differentially weights information in favor of the first expert. The reason for this "backfire effect" is that her interpretation of the outsider's consistent signals are biased by the fact that the signals contradict the overtrusted insider. This asymmetry means that the outsider cannot completely unravel the insider's signals.

Furthermore, this asymmetry persists in the limit: Information that should lead to more uncertainty about qualities and no change in beliefs about the state instead leads the pre-screener to be *more sure of and more wrong in* her beliefs along both dimensions when observing opposing information from different experts sequentially.

How can signals from outsiders resolve disagreement? The fundamental problem is that the pre-screener over-infers expert quality at each step, suggesting that an outsider can better

resolve disagreement by delivering all $k$ opposing signals in one "blast." We show this in Proposition 8 by extending the model to allow multiple signals per period.

**Proposition 8 (Overcoming outsider rejection)** *Let* $(\omega_0^A, \omega_0^{H_1}, \omega_0^{H_2}) = (1/2, \hat{q}, \hat{q})$ *for any* $\hat{q} \in (0,1)$ *where experts 1 and 2 are independent. Consider a sequence of* $2k$ *observed signals such that expert 1 sends the first* $k$ *a signals, then expert 2 sends* $k$ *b signals, where* $k > 1$*.*

1. *The pre-screener overtrusts expert 1 even more when expert 1's signals are sent sequentially rather than simultaneously.*

2. *Expert 2's credibility is higher when sending her signals simultaneously rather than sequentially, but the pre-screener still believes that state* $A$ *is more likely than* $B$*.*

Whether or not expert 1's identical $a$ signals are sent simultaneously or sequentially, the pre-screener overinfers the good news about the expert's quality. But sequential signals imply overinference that compounds upon each signal, leading to more overtrust than observing simultaneous signals, where overinference occurs one time.

If expert 2 sends her signals sequentially, after each signal the pre-screener overweights the inference that expert 1 is likely to be high quality and expert 2 to be low quality because she has observed a longer sequence of $a$'s from expert 1 than $b$'s from expert 2. This overinference compounds, leading to more under-trust of expert 2 than if expert 2 sent signals simultaneously. When expert 2 delivers all countervailing signals simultaneously, the pre-screener believes it is relatively less likely that expert 2 is low quality because she compares all $k$ $b$ signals against her beliefs based on expert 1's $k$ signals. Nonetheless, the initial overtrust of expert 1, whether generated by simultaneous or sequential $a$ signals, means that expert 2 still cannot fully countervail the influence of expert 1.

To summarize, Propositions 7 and 8 suggest that the order in which experts present themselves, not just information, is highly relevant for persuasion. There is a clear first mover advantage for the first expert when sources consistently disagree. However, while overtrust in an expert is difficult to undo by an outsider, Proposition 6 shows that it is more fragile to internal contradictions by the insider.

# 4 The Central Role of Unknown Credibility

## 4.1 Confirmation bias and pre-screening

Our framework assigns a central role to expert quality for biased learning. This distinguishes it from several well-known biases, the closest of which is confirmation bias (Lord, Ross and Lepper, 1979; Griffin and Tversky, 1992; Rabin and Schrag, 1999), which is the tendency for individuals to interpret new information as confirming existing beliefs. In Rabin and Schrag (1999), agents probabilistically flip signals that oppose current beliefs. Because early signals over-influence the interpretation of subsequent signals, "first impressions matter."

Pre-screening is distinct from confirmation bias either about the state or the expert in that pre-screeners over-interpret information about credibility irrespective of whether signals confirm or contradict beliefs. This distinction matters for testable predictions in four ways.

First, our theory makes testable predictions about when confirmation bias, and behavior which is akin to its opposite, arises, summarized in Propositions 9 and 10. Proposition 9 considers when confirmation bias arises in response to the marginal signal. We conduct the following thought experiment: Suppose a pre-screener begins with prior $\omega_0^A = 1/2$, observes signals $\mathbf{s}^n$ from one expert, and has posterior $\omega_n^b$. For the sake of contrast with confirmation bias, assume that the existing evidence $\mathbf{s}^n$ objectively strictly suggests $A$ (hence $\{\mathbf{s}^n, s_{n+1}\}$ objectively weakly suggests $A$). Does the pre-screener's beliefs about the state over- or under-react in response to $s_{n+1}$, compared to a Bayesian endowed with prior $\omega_n^b$?

If the signal confirms (contradicts) beliefs ($s_{n+1} = a$ and $b$, respectively), we say the agent has over-reacted (under-reacted) if $P^b[\theta = A|\{\mathbf{s}^n, s_{n+1}\}] > P^u[\theta = A|\text{prior} = \omega_n^b]$, and under-reacted (over-reacted) if $P^b[\theta = A|\{\mathbf{s}^n, s_{n+1}\}] < P^u[\theta = A|\text{prior} = \omega_n^b]$. Relative to an "endowed Bayesian," an agent in Rabin and Schrag (1999) under-reacts on the marginal signal if it contradicts current beliefs, and correctly updates if it confirms current beliefs.

A pre-screener may over- or under-react to both confirmatory and contradictory news relative to the endowed Bayesian, depending on how the signal affects the first-stage trust ($\kappa_H(\{\mathbf{s}^n, s_{n+1}\})$) in the combined evidence from the expert ($\{\mathbf{s}^n, s_{n+1}\}$). For signals that confirm beliefs ($s_{n+1} = a$), agents will over-react if first-stage trust is high (Part 1a) and under-react if first-stage trust is low (Part 1b), relative to the endowed Bayesian. The

over-reaction is broadly consistent with confirmation bias, while the under-reaction is the opposite. The intuition for the under-reaction is that, even though the signal is confirmatory and the pre-screener and endowed Bayesian begin from the same beliefs, $\kappa_H(\{\mathbf{s}^n, s_{n+1}\})$ may be too low relative to $\kappa_H(\{\mathbf{s}^n\})$ (the effective prior over the informativeness of $\{\mathbf{s}^n, s_{n+1}\}$ for the endowed Bayesian), weighing down the pre-screener's perceived informativeness of the total evidence $\{\mathbf{s}^n, s_{n+1}\}$—including the history of signals which on balance supported $A$.

For signals that contradict beliefs ($s_{n+1} = b$), pre-screeners will under-react when first-stage trust is high (Part 1a), and over-react when first-stage trust is low (Part 1b). The under-reaction here is consistent with how a Rabin and Schrag (1999) behaves, while the over-reaction is the opposite. We label this over-reaction the *undermining effect*. In this case, contradictory information undercuts the first-stage belief $\kappa_H(\{\mathbf{s}^n, s_{n+1}\})$ and excessively undermines the history of evidence $\{\mathbf{s}^n, s_{n+1}\}$ (which supports $A$) in the second step, again relative to the endowed Bayesian. This prediction is both supported by recent experimental evidence and can be important in the real world. De Filippis et al. (2017) find evidence that individuals over-react to contradictory signals, citing a mechanism similar to the undermining effect: contradictory signals lead individuals to first revise their beliefs about the credibility of previous information, before updating on the signal. In the real world, "flip-flopping" can be very damaging for individuals seeking to establish credibility.

Overall, while our theory supports the predictions of confirmation bias, it also predicts that the opposite behavior, including the undermining effect, arises when first-stage trust in the expert is low. The crucial driver of this distinction is the fact that pre-screeners over-interpret information about credibility as reflected in $\kappa_H(\{\mathbf{s}^n, s_{n+1}\})$ irrespective of whether signals confirm or contradict current beliefs.[4]

[4]In more detail, consider first the endowed Bayesian's (EB's) beliefs. EB begins with $\omega_b^n$ and applies it to evaluate $s_{n+1}$. But because $\omega_b^n$ is generated by applying a prior of $\kappa_q(\mathbf{s}^n)$ to signals $\mathbf{s}^n$, and a Bayesian's beliefs are invariant to signal order, this is equivalent to applying a prior of $\kappa_q(\mathbf{s}^n)$ to $\{\mathbf{s}^n, s_{n+1}\}$. EB's resulting marginal posterior belief about credibility then equals $\kappa_q(\{\mathbf{s}^n, s_{n+1}\})$, the pre-screener's first-stage updated belief of credibility. However, in applying the second step, the pre-screener makes the mistake of using $\kappa_q(\{\mathbf{s}^n, s_{n+1}\})$ as a prior to to evaluate $\{\mathbf{s}^n, s_{n+1}\}$, whereas EB used $\kappa_q(\{\mathbf{s}^n\})$.

As a result, the pre-screener will think $A$ is more likely than the endowed Bayesian if and only if the *new* signal $s_{n+1}$ makes her strictly more confident that the expert is high quality relative to EB's prior over *all* signals $\{\mathbf{s}^n, s_{n+1}\}$ and the combined evidence $\{\mathbf{s}^n, s_{n+1}\}$ objectively strictly favors $A$. This is the condition that $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) > \kappa_H(\{\mathbf{s}^n\})$ in Part 1(a). Similarly, Part 1(b) describes how the pre-screener will think $A$ is less likely than the endowed Bayesian if and only if the new signal makes her strictly less confident that the expert is high quality relative to EB's prior over all signals $\{\mathbf{s}^n, s_{n+1}\}$, which is the condition that

Part 2 of Proposition 9 shows that the effect of a new signal $s_{n+1}$ on a pre-screener's beliefs cannot be summarized simply by its effect on the prior. This is because the pre-screener re-evaluates all the evidence $\{\mathbf{s}^n, s_{n+1}\}$ in light of the new first-stage belief $\kappa(\{\mathbf{s}^n, s_{n+1}\})$. In contrast, a Bayesian updates identically irrespective of whether she is endowed with a belief or observes a history of signals consistent with that belief: $P^u[\theta = A|\{\mathbf{s}^n, s_{n+1}\}] = P^u[\theta = A|prior = \omega_n^u, \{s_{n+1}\}]$, where $\omega_n^u$ equals the Bayesian posterior generated by $\mathbf{s}^n$.

**Proposition 9 (Reaction to subsequent signals)** *Let $(\omega_0^A, \omega_0^H) = (1/2, \hat{q})$ for any $\hat{q} \in (0,1)$. Let $\mathbf{s}^n$ be a sequence of $n$ observed signals (with an arbitrary number of a's and b's), let $s_{n+1}$ be the $(n+1)$th observed signal, and let $\omega_n^b$ equal the pre-screener's joint posterior after the sequence $\mathbf{s}^n$. WLOG, let the number of a's be greater than or equal to the number of b's in $\{\mathbf{s}^n, s_{n+1}\}$.*

1. *Relative to Bayesian:*

   (a) $P^b[\theta = A|\{\mathbf{s}^n, s_{n+1}\}] > P^u[\theta = A|prior = \omega_n^b, \{s_{n+1}\}]$ *if* $\{\mathbf{s}^n, s_{n+1}\}$ *has strictly more a's than b's and* $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) > \kappa_H(\mathbf{s}^n)$,

   (b) $P^b[\theta = A|\{\mathbf{s}^n, s_{n+1}\}] < P^u[\theta = A|prior = \omega_n^b, \{s_{n+1}\}]$ *if* $\{\mathbf{s}^n, s_{n+1}\}$ *has strictly more a's than b's and* $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) < \kappa_H(\mathbf{s}^n)$,

$\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) < \kappa_H(\{\mathbf{s}^n\})$. Part 1(c) reflects the edge case.

Importantly, Parts 1(a) and 1(b) can arise when new signals $s_{n+1}$ are either confirmatory or contradictory. In Part 1(b), the opposite of confirmation bias arises: there is under-reaction to a confirmatory signal and over-reaction to a contradictory signal. Under-reaction arises because we can have $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) < \kappa_H(\{\mathbf{s}^n\})$ even though the signal is confirmatory and thus $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) > \omega_b^{n,H}$—that is, even though the first-stage belief that the expert is high quality increases over the time-$n$ posterior. Intuitively, $\omega_b^{n,H}$ arises from applying $\kappa_H(\{\mathbf{s}^n\})$ to $\mathbf{s}^n$, while $\kappa_H(\{\mathbf{s}^n, s_{n+1}\})$ arises from applying $\kappa_H(\{\mathbf{s}^n\})$ to $\{\mathbf{s}^n, s_{n+1}\}$. With confirmatory news, $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) > \omega_b^{n,H}$ but may fall below $\kappa_H(\{\mathbf{s}^n\})$ if the history of evidence contains some bad news about credibility. To be precise, the pre-screener's over-inference at each step implies $\kappa_H(\{\mathbf{s}^n, s_{n+1}\})$ contains $\kappa_H(\{\mathbf{s}^n\})$ (see Equation 2). Factoring out $\kappa_H(\{\mathbf{s}^n\})$ followed by algebraic manipulation shows that the effective difference between the two equals the difference between an *objective* Bayesian's belief that the expert is high quality, $P^u(q = H|\{\mathbf{s}^n, s_{n+1}\})$ and the time-0 prior common to the pre-screener and objective Bayesian, $\omega_0^H$. Thus, even if $s_{n+1}$ is confirmatory, the pre-screener under-reacts if the *combined evidence* $\{\mathbf{s}^n, s_{n+1}\}$ is objectively bad news for credibility.

Over-reaction to the contradictory signal is similar. The intuition is more straightforward in this case because $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) < \kappa_H(\{\mathbf{s}^n\})$ and $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) < \omega_b^{n,H}$—the new signal $s_{n+1}$ is deleterious for both. Re-arranging the discussion above yields similar intuitions for why Part 1(a) and $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) > \kappa_H(\{\mathbf{s}^n\})$ can arise when new signals are either confirmatory or contradictory. Even though the resulting observed behavior is akin to confirmation bias, the mechanism is fundamentally different and arises due to the over-interpretation of signals embedded in $\kappa_H(\{\mathbf{s}^n, s_{n+1}\})$.

*(c)* $P^b[\theta = A|\{\mathbf{s}^n, s_{n+1}\}] = P^u[\theta = A|prior = \omega_n^b, \{s_{n+1}\}]$ *if* $\{\mathbf{s}^n, s_{n+1}\}$ *has an equal number of* $a$*'s and* $b$*'s or* $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) = \kappa_H(\mathbf{s}^n)$.

*Furthermore,* $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) \geq \kappa_H(\mathbf{s}^n)$ *if and only if* $P^u(q = H|\{\mathbf{s}^n, s_{n+1}\}) \geq \omega_0^H$*, with equality holding if and only if* $P^u(q = H|\{\mathbf{s}^n, s_{n+1}\}) = \omega_0^H$.

2. *History-dependence:* $P^b[q, \theta|\{\mathbf{s}^n, s_{n+1}\}] = P^b[q, \theta|prior = \omega_n^b, \{s_{n+1}\}]$ *if and only if* $P^b[q|\mathbf{s}^n] = \omega_0^q$.

Proposition 10 characterizes sufficient conditions under which pre-screeners exhibit confirmation bias or its opposite, regardless of the order in which they observed signals. Without loss of generality, assume that the expert has sent more $a$ signals than $b$ signals ($n_a > n_b$).

**Proposition 10 (Over- and under-trust without knowing signal order)** *Let* $(\omega_0^A, \omega_0^H) = (1/2, \hat{q})$ *for any* $\hat{q} \in (0, 1)$*, and let* $n_a > n_b$*. Whether pre-screeners exhibit confirmation bias or the opposite depends on the relative proportion of* $a$*'s and* $b$*'s and the distribution of beliefs about quality:*

1. *There exists some* $n_a^*$ *and* $\check{p} > \frac{1}{2}$ *such that the agent overtrusts the expert and is optimistic that the state is* $A$ *for* any *sequence with fixed* $n_a, n_b$ *when* $n_b < n_a^* < n_a$ *and* $p_L < p_H \leq \check{p}$.

2. *There exists some* $\hat{n}_b$*,* $\underline{p} > \frac{1}{2}$ *and* $\overline{p} < 1$ *such that the agent under-trusts the expert and is pessimistic that the state is* $A$ *for* any *sequence with fixed* $n_a, n_b$ *when* $0 \leq \hat{n}_b < n_b < n_a$ *and one of the following sufficient conditions is met: (a)* $\overline{p} \leq p_L < p_H$*, or (b)* $p_L \leq \underline{p}$ *and* $p_H > \overline{p}$.

When the proportion of $a$'s is much greater than the proportion of $b$'s (Part 1), a pre-screener exhibits confirmation bias in that she may always be optimistic about $A$ relative to the Bayesian. This is true even for the sequence that generates the lowest possible trust in the expert (fixing information content), so long as mixed signals do not sufficiently distinguish between high and low quality ($p_L$ and $p_H$ sufficiently low). Here, the initial negative impression from mixed signals is relatively weak, so the ensuing consistency of many $a$'s countervails the initial under-trust, creating optimism and overtrust.

When the proportion of $a$'s is sufficiently similar to the proportion of $b$'s (Part 2), this information content strongly suggests that the expert is low quality, as long as $p_H$ is sufficiently high. In this case, the pre-screener may be under-trusting and pessimistic given *any* observed order, including the sequence that generates the highest degree of trust. This is because ensuing contrary signals are extremely informative of low quality and therefore overcome even the most positive first impression, resulting in under-trust and pessimism. In contrast, on average an agent in Rabin and Schrag (1999) would exhibit optimism toward state $A$ in both of the cases described in Proposition 10.

A second reason why our framework is distinct from models of confirmation bias is that the strength of endogenous trust explains the strength of behavior similar to confirmation bias and its opposite, as shown in Proposition 9. Rabin and Schrag (1999) assume that the severity of bias is exogenous. Third, our model distinguishes how the source of information affects confirmation bias, a distinction not addressed in Rabin and Schrag (1999). Propositions 7 and 8 predict that confirmation bias arises when multiple sources contradict one another, but not when a single source is self-contradictory. Finally, while pre-screening and confirmation bias are both more likely to arise in ambiguous settings (Lord, Ross and Lepper, 1979), pre-screening occurs when the ambiguity is due to source quality. Together, these observations emphasize the central role of unknown expert credibility.

Similar to Rabin and Schrag (1999), Fryer, Harms and Jackson (2016) consider a setting where signals sometimes deliver an ambiguous signal of $ab$, which agents interpret in favor of their prior beliefs. Thus, disagreement can arise when there are many ambiguous $ab$ signals observed in different order. Because this bias is also based on beliefs about the state alone, it does not predict qualitatively different effects that vary with source and source quality. It is also less suited to describe situations in which signals are unambiguous but disagreement is still strong. Several of the examples in Section 5 fit this description.

## 4.2 Comparison with other frameworks

We now compare how our predictions distinguish our framework from theories of biased learning and disagreement other than confirmation bias. The unifying themes are that, by focusing on endogenous trust in experts, our framework 1) generates joint disagreement

about substance and credibility, even when agents share common biases, information, and priors, 2) can link several different behaviors predicted by other theories individually, 3) predicts new order effects often not predicted by other theories, and 4) operates in a context where agents are paying attention to all signals.

### 4.2.1 Overconfidence

A large strand of literature studies how agents may misinterpret signals because they misperceive their accuracy, often due to overconfidence. Scheinkman and Xiong (2003) provide a review and argue that disagreement arises because agents "agree to disagree" about how correlated signals are with the unknown state. Heidhues, Kőszegi and Strack (2017) study how overconfidence (more akin to overoptimism in their framework) can be self-defeating by leading to distorted beliefs about other decision-relevant variables; under-confidence is less deleterious. Despite the importance of overconfidence, its source is often less clear. In our framework, the correlation of signals with the underlying state is precisely what agents are trying to learn. The behavior following positive and negative first impressions endogenously generates behavior resembling over- and under-confidence while also endogenizing its strength. The most related paper that endogenizes overconfidence is Gervais and Odean (2001), where successful traders are overconfident in their financial trading skills due to a form of self-attribution bias. No actions are required by the pre-screener in our framework.

Ortoleva and Snowberg (2015) and Enke and Zimmermann (2016) consider correlation neglect, where agents under-estimate correlation among signals. This is a form of overconfidence in that agents over-estimate the amount of information in a given set of signals. Experts and signals are independent draws in our setting, so there is no correlation neglect. Disagreement in their setting requires heterogeneity in the information agents observe or in the degree of correlation neglect, whereas pre-screeners can disagree despite sharing the same information content and bias. More broadly, models of overconfidence in which the agent has mistaken yet exogenously fixed beliefs about source quality, but is otherwise Bayesian, predict beliefs that are biased but invariant to signal and source order.

### 4.2.2 Inattention

A growing literature considers boundedly rational individuals with limited ability to process information (Sims, 2003, 2006; Gabaix, Laibson, Moloche and Weinberg, 2006). Schwartzstein (2014) considers a setting where agents learn to selectively pay attention to variables after evaluating their predictive ability. However, evaluating predictive ability is difficult in several areas of major disagreement we consider—e.g., testing whether humans affect climate change. Wilson (2014) shows that confirmation bias can arise when agents have bounded memory and can forget the realized history of signals. Kominers, Mu and Peysakhovich (2016) assume that agents trade off attention costs and belief accuracy, so they screen out uninformative signals with low decision value.[5] Since contrary signals have particularly high value, such agents do not exhibit confirmation bias. While some form of inattention can potentially predict either insider unraveling (Proposition 6) or outsider rejection (Proposition 7), inattention is less able to reconcile both predictions.

Fundamentally, inattention biases revolve around agents not paying enough attention to certain signals. The central feature of our framework is that agents disagree about the credibility of signals *to which they all pay attention*. As we argue in Section 5, this feature captures the essence of several important real-world disagreements.

### 4.2.3 Media and Persuasion

The literature has also focused on the role of information supply in disagreement by showing that the media will slant news to build a reputation (Gentzkow and Shapiro, 2006) or to cater to consumers' preferences for beliefs (Mullainathan and Shleifer, 2005). Glaeser and Sunstein (2014) consider how polarization from common information can arise when consumers have sufficiently different prior beliefs about senders' motives. Because such models assume Bayesian consumers, they imply that signal and source order are irrelevant to final beliefs when given fixed information content.

In these models, exogenous heterogeneous priors drive why the media endogenously respond with biased information even when covering the same consumers. Instead, we ask how

[5]Kominers et al. (2016) briefly mention the term "pre-screen." The overlap in terminology is accidental.

heterogeneous priors arise endogenously even if sources are unbiased. This also differentiates us from the large literature on strategic experts (e.g., Hong, Scheinkman and Xiong, 2008) and persuasion (Gentzkow and Kamenica, 2011, 2016).

Nevertheless, our model of biased learning complements the predictions of these literatures by suggesting additional ways strategic manipulation of signals can amplify or mitigate disagreement. For example, Propositions 7 and 8 have speculative implications for how strategic experts can persuade pre-screeners. If the expert is first to disclose signals on a given topic, she should release consistent signals slowly in order to "build trust" among the audience. In contrast, if the expert is a second mover and knows that *the same information* contradicts a first mover, she should instead release all signals simultaneously because she has to "disprove incompetence." Alternatively, if the second mover has preliminary evidence against the prevailing theory, she should wait to amass more countervailing evidence before disclosing all of it together.

### 4.2.4 Bounded rationality

A distinguished set of models seeks to explain behavior using bounded rationality, the idea that agents follow simple heuristics due to cognitive limitations (Simon, 1957; Gigerenzer and Selten, 2002; Selten, 2002). In contrast, on a literal level, pre-screeners employ more computational horsepower than a standard Bayesian, and ours is not a model of bounded rationality in this sense. Our approach in the spirit of work that seeks to model systematic conceptual errors but nonetheless can lead to more literal computations on the part of agents in the model. For example, in Rabin and Schrag (1999), agents are Bayesian but also randomly mis-perceive contradictory signals as confirmatory. In Brunnermeier and Parker (2005), agents calculate optimal beliefs which then distort actions.[6] The error itself, discussed in Section 1.2, is rooted in the experimentally-supported idea that people double-count information. Our modeling approach starts from Bayes' Rule and asks how far this one mistake can go in explaining disagreement.

[6]Further examples include Rabin (2002), where agents mis-perceive draws from an urn that occur with replacement as occurring without replacement, requiring agents to additionally keep track of what has been drawn. Agents in Bénabou and Tirole (2002) forget bad news, but also think about whether any information lost affects the value of new information. Aforementioned models of bounded memory and inattention also involve more computations as part of the model, as do models of dual-process heuristics (Cerigioni, 2017).

# 5 Discussion

## 5.1 Real-world examples of disagreement

While no single explanation likely explains the disagreements described in the Introduction and Section 1, we argue that our mechanism, by understanding these disagreements as about credibility rather than substance, provides new insight distinct from other theories.

**Economics.** The topic of whether government stimulus promotes growth generates disagreement among both economists (e.g., Krugman, 2009; Cochrane, 2009; The Economist Magazine, 2013) and the public (Sapienza and Zingales, 2013). Regardless of who is right, disagreement among the public is likely correlated with which economists they believe are credible. For example, opponents and proponents of stimulus on the editorial pages of *The Wall Street Journal* and *The New York Times* routinely disagree about the credibility of the opposing side's economists, despite obviously paying attention to both sides (e.g., Moore, 2011; New York Times, 2014). Inattention is unlikely to be a good description of disagreement as opponents and proponents are all intensely focused on the issue.

Likewise, the opposition to free trade is plausibly related to the belief that economists are incompetent, rather than simply not noticing what they have said. Sapienza and Zingales (2013) report that providing economists' opinions that NAFTA increased welfare to the surveyed households changed their opinions about its merits very little. The widespread support among economists for free trade suggests these disagreements are also not well-described by the selective interpretation of ambiguous signals as in Fryer, Harms and Jackson (2016). The finding that survey respondents thought stock prices were easier to predict after being told the consensus opinion that they are hard to predict echoes Proposition 7.

Our interpretation is simply that trust in economists is low. In areas where economists strongly disagree among themselves (stimulus spending), legitimate scientific debate may lead to polarized opinions among different factions of the public due to differences in the order in which they were exposed to these experts. Learning about a topic early from one expert (e.g., a teacher or mentor) tends to overly influence opinions about the credibility of other experts encountered down the line. In areas with more expert consensus like free trade, households may "be their own expert" and trust their own experiences over that of

experts, an interpretation we discuss in the conclusion. Individuals may be more open to the anti-trade views of non-economists exactly because they distrust professional economists.

Notably, economists may vary widely in quality, and discerning quality may be hard for the public. Macroeconomics is a technical subject where repeated experimentation to determine expert credibility is difficult, as history does not repeat in a controlled environment.

**Climate science.** The heated disagreement between climate deniers and supporters of the proposition that humans cause climate change is also unlikely explained by inattention. Inattention suggests that climate deniers are unaware of the consensus. More plausibly, climate deniers have paid attention to what mainstream scientists say, yet actively deny their credibility, because they believe an alternative set of experts, or their own intuition, as our model suggests. The fact that the consensus is so strong may itself contribute to the disagreement (Proposition 7).

**Medical advice.** In 1998, Andrew Wakefield and co-authors held a press conference describing the results of a study they would soon publish that linked the measles vaccination with autism (Wakefield et al., 1998). Although subsequent research discredited this link, leading the journal to retract the article, the idea has lingered, potentially because Dr. Wakefield has never retracted the claim himself (Proposition 6). The role of uncertain expert quality in this controversy is laid bare by the observation that the aforementioned actress Jennifer McCarthy Wahlberg was an influential figure in the anti-vaccination movement.

**Financial advice.** In finance, investors often rely on advice from financial analysts whose expertise can be difficult to ascertain. Agnew et al. (2016) conduct an experiment and find that first impressions matter for the choice of financial advisors: after a positive first impression, an advisor can go on to "give bad advice and maintain a client's trust," consistent with Propositions 4 and 9. Jia, Wang and Xiong (2016) find that local investors react more to recommendations of local analysts, and foreign investors react more to those of foreign analysts, consistent with Propositions 2 and 3.

**Fake news and fact-checking.** The rise of fake news has coincided with a decline in trust in traditional news sources. Propositions 4 and 5 suggest that any under-trust in traditional news may be particularly difficult to unwind. Demand for the extreme news often reported by alternative sources may rise if pre-screeners over-trust these sources, amplifying

motives to slant information (Gentzkow and Shapiro, 2006, 2010).

Our theory is also consistent with the evidence that rumors and misinformation are stubbornly resistant to fact-checking or debunking by outsiders (Berinsky, 2012). Attempts at correcting false beliefs can "backfire" and harden beliefs (Nyhan and Reifler, 2010), consistent with Propositions 6–8. Corrections are more successful when they include retractions from the original source (Simonsohn, 2011; Levine and Valle, 1975) or from sources whose credibility is likely highly correlated with the original source (Berinsky, 2012).

## 5.2 Testable implications

A strength of our parsimonious modeling approach is that it generates several predictions that are testable, given appropriately rich data on individuals' opinions about a topic and where and when they obtained their information.

Our first set of predictions—Propositions 2 and 3—suggest that, in a cross-section of individuals, beliefs about source quality should predict beliefs about a given claim. These predictions also help explain why individuals who share common priors may disagree even when Bayesians should agree (Andreoni and Mylovanov, 2012).

Our second set of predictions—Propositions 4 and 5—are testable given a long panel of data, as they speak to variation within rather than across individuals. Individuals' first impressions about expert credibility should predict beliefs about credibility and the state later on. Experimental evidence from social psychology supports the idea that these "primacy effects" can arise in certain situations (Petty and Wegener, 1998, p.356).

Propositions 4 and 5 also suggest a third set of predictions—that negative first impressions are more persistent than positive first impressions. Consistent with this, a long-standing finding in social psychology is that unfavorable first impressions are more difficult to change than favorable ones, a form of "negativity bias" (Richey, McClelland and Shimkunas, 1967; Richey, Koenigs, Richey and Fortin, 1975; Weinberger, Allen and Dillon, 1981).

A fourth set of predictions—Propositions 6 through 8—suggest that varying the order in which experts present themselves should influence beliefs. If new information from outside sources conflicts with information from trusted inside sources, there may be a "backfire effect" that hardens current beliefs. If the same conflicting information came from an insider,

opinions would be more likely to change in the direction of the Bayesian. This is the most data-demanding prediction to test as it requires variation within individuals across experts through time, though experimental testing is perhaps more straightforward.

A fifth set of predictions—Propositions 9 and 10—suggest that individuals will over-react towards confirming signals and under-react towards contradictory signals when first-stage trust is high, consistent with confirmation bias. However, when first-stage trust is low, behavior akin to the opposite arises: individuals under-react towards confirmatory signals, and over-react towards contradictory signals (the "undermining effect"). These predictions are perhaps best tested experimentally, and evidence supports behavior akin to the undermining effect (De Filippis et al., 2017), as discussed in Section 4.1.

# 6 Conclusion

We argue that biased learning about credibility helps explain disagreement in novel ways across fields ranging from economics, climate science, to medicine. If individuals over-infer expert quality, they will disagree with each other about substance because they endogenously disagree about which signals are credible.

More broadly, the model can also consider how experiences affect beliefs. Although we interpret experts as external sources of signals, an alternative interpretation is that individuals have noisy experiences that inform them about an unknown true state. Here, the unknown expert quality is the informativeness of each individual's experience-generating process.

Malmendier and Nagel (2011) find that experiences affect whether individuals trust the stock market, potentially through a beliefs channel. For example, individuals born in the Depression tend to have lower stock market participation than younger generations who more recently experienced a boom. Koudijs and Voth (2016) find that personal experiences affect risk-taking, and Malmendier and Nagel (2016) find that differences in experienced inflation explain disagreement about inflation. Our model suggests that people may endogenously trust their own experiences more or less based on the consistency of those experiences, and that this may affect their trust in external sources of information. We leave a deeper exploration of this link for future research.

# References


**Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz**, "Fragility of Asymptotic Agreement under Bayesian Learning," *Theoretical Economics*, 2016, *11*, 187–227.

**Agnew, Julie R., Hazel Bateman, Christine Eckert, Fedor Iskhakov, Jordan Louviere, and Susan Thorp**, "First impressions matter: An experimental investigation of online financial advice," *Management Science*, 2016, *Forthcoming*.

**Albarracín, Dolores and Patrick Vargas**, "Attitudes and Persuasion: From Biology to Social Responses to Persuasive Intent," in Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey, eds., *Handbook of Social Psychology*, 5 ed., John Wiley & Sons, 2010, pp. 394–427.

**Andreoni, James and Tymofiy Mylovanov**, "Diverging Opinions," *American Economic Journal: Microeconomics*, 2012, *4* (1), 209–232.

**Bell, Larry**, "Climate Change As Religion: The Gospel According To Gore," *Forbes Magazine*, April 26 2011. Available online: https://www.forbes.com/sites/larrybell/2011/04/26/climate-change-as-religion-the-gospel-according-to-gore [Last accessed: May 2017].

**Bénabou, Roland and Jean Tirole**, "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics*, 2002, *117* (3), 871–915.

**Berinsky, Adam**, "Rumors, Truths, and Reality: A Study of Political Misinformation," 2012. Massachusetts Institute of Technology.

**Brunnermeier, Markus K. and Jonathan A. Parker**, "Optimal expectations," *American Economic Review*, 2005, *95* (4), 1092–1118.

**Carlin, Bradley P. and Thomas A. Louis**, "Empirical Bayes: Past, Present and Future," *Journal of the American Statistical Association*, 2000, *95* (452), 1286–1289.

**Cerigioni, Francesco**, "Dual decision processes: Retrieving preferences when some choices are automatic," 2017.

**Chadrasekhar, Arun G., Horacio Larreguy, and Juan Pablo Xandri**, "Testing Models of Social Learning on Networks: Evidence from a Lab Experiment in the Field," 2015. NBER Working Paper No. 21468.

**Cochrane, John H.**, "How did Paul Krugman get it so wrong?," September 16 2009. Available online: https://faculty.chicagobooth.edu/john.cochrane/research/papers/krugman_response.htm [Last accessed: September 2016].

**DellaVigna, Stefano and Devin Pope**, "Predicting Experimental Results: Who Knows What?," 2016.

**Enke, Benjamin and Florian Zimmermann**, "Correlation Neglect in Belief Formation," 2016.

**Eyster, Erik and Matthew Rabin**, "Naïve Herding in Rich-Information Settings," *American Economic Journal: Microeconomics*, 2010, *2*, 221–243.

**Filippis, Roberta De, Antonio Guarino, Philippe Jehiel, and Toru Kitagawa**, "Updating Ambiguous Beliefs in a Social Learning Experiment," 2017.

**Fryer, Roland G., Philipp Harms, and Matthew O. Jackson**, "Updating Beliefs when Evidence is Open to Interpretation: Implications for Bias and Polarization," 2016.

**Gabaix, Xavier, David Laibson, Guillermo Moloche, and Stephen Weinberg**, "Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model," *American Economic Review*, 2006, *96* (4), 1043–1068.

**Gentzkow, Matthew and Emir Kamenica**, "Bayesian Persuasion," *American Economic Review*, 2011, *101* (6), 2590–2615.

__ **and** __ , "Bayesian Persuasion with Multiple Senders and Rich Signal Spaces," 2016.

__ **and Jesse M. Shapiro**, "Media Bias and Reputation," *Journal of Political Economy*, 2006, *114* (2), 280–316.

__ **and** __ , "What Drives Media Slant? Evidence from U.S. Daily Newspapers," *Econometrica*, 2010, *78* (1), 35–71.

**Gervais, Simon and Terrance Odean**, "Learning to be Overconfident," *Review of Financial Studies*, 2001, *14*, 1–27.

**Gigerenzer, Gerd and Reinhard Selten**, "Rethinking Rationality," in Gerd Gigerenzer and Reinhard Selten, eds., *Bounded Rationality: The Adaptive Toolbox*, MIT Press, 2002, pp. 1–12.

**Glaeser, Edward L. and Cass R. Sunstein**, "Does More Speech Correct Falsehoods?," *Journal of Legal Studies*, 2014, *43* (1), 65–93.

**Griffin, Dale and Amos Tversky**, "The weighing of evidence and the determinants of confidence," *Cognitive Psychology*, July 1992, *24* (3), 411–435.

**Heidhues, Paul, Botond Kőszegi, and Philipp Strack**, "Unrealistic Expectations and Misguided Learning," 2017.

**Hong, Harrison, José A. Scheinkman, and Wei Xiong**, "Advisors and asset prices: a model of the origins of bubbles," *Journal of Financial Economics*, 2008, *89* (2), 268–287.

**Jia, Chunxin, Yaping Wang, and Wei Xiong**, "Market segmentation and differential reactions of local and foreign investors to analyst recommendatons," *Review of Financial Studies*, 2016, *Forthcoming.*

**Johnson, Blair T., Gregory R. Maio, and Aaron Smith-McLallen**, "Communication and Attitude Change: Causes, Processes, and Effects," in Dolores Albarracín, Blair T. Johnson, and Mark P. Zanna, eds., *Handbook of Attitudes*, Lawrence Erlbaum Associates, 2005, pp. 617–670.

**Kominers, Scott Duke, Xiaosheng Mu, and Alexander Peysakhovich**, "Paying (for) Attention: The Impact of Information Processing Costs on Bayesian Inference," 2016.

**Koudijs, Peter and Hans-Joachim Voth**, "Leverage and beliefs: personal experience and risk-taking in margin lending," *American Economic Review*, 2016, *106* (11), 3367–3400.

**Kruglanski, Arie W. and Wolfgang Stroebe**, "The Influence of Beliefs and Goals on Attitudes: Issues of Structure, Function, and Dynamics," in Dolores Albarracín, Blair T. Johnson, and Mark P. Zanna, eds., *Handbook of Attitudes*, Lawrence Erlbaum Associates, 2005, pp. 323–368.

**Krugman, Paul**, "How Did Economists Get It So Wrong?," *New York Times Magazine*, September 2 2009. Available online: http://www.nytimes.com/2009/09/06/magazine/06Economic-t.html [Last accessed: September 2016].

**Levine, John M. and Ronald S. Valle**, "The Convert as a Credible Communicator," *Social Behavior and Personality*, 1975, *3* (1), 81–90.

**Lindley, Dennis Victor**, "Compound Decisions and Empirical Bayes: Discussion," *Journal of the Royal Statistical Society*, 1969, *31* (3), 397–425.

**Lord, Charles G., Lee Ross, and Mark R. Lepper**, "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence," *Journal of Personality and Social Psychology*, 1979, *37* (11), 2098–2109.

**Malmendier, Ulrike and Stefan Nagel**, "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?," *Quarterly Journal of Economics*, 2011, *126* (1), 373–416.

**__ and __**, "Learning from Inflation Experiences," *Quarterly Journal of Economics*, 2016, *131* (1), 53–87.

**Moore, Stephen**, "Why Americans Hate Economists," August 19 2011. http://www.wsj.com/articles/SB10001424053111903596904576514552877388610 [Last accessed: September 2016].

**Mullainathan, Sendhil and Andrei Shleifer**, "The Market for News," *The American Economic Review*, 2005, *95* (4), 1031–1053.

**New York Times**, "What the stimulus accomplished," February 22 2014. http://www.nytimes.com/2014/02/23/opinion/sunday/what-the-stimulus-accomplished.html [Last accessed: September 2016].

**Nyhan, Brendan and Jason Reifler**, "When Corrections Fail: The Persistence of Political Misperceptions," *Political Behavior*, 2010, *32* (2), 303–330.

**Ortoleva, Pietro and Erik Snowberg**, "Overconfidence in Political Behavior," *American Economic Review*, 2015, *105* (2), 504–535.

**Petty, Richard E. and Duane T. Wegener**, "Attitude Change: Multiple Roles for Persuasion Variables," in Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey, eds., *Handbook of Social Psychology*, 4 ed., McGraw-Hill, 1998, pp. 323–390.

**Rabin, Matthew**, "Inference by Believers in the Law of Small Numbers," *Quarterly Journal of Economics*, 2002, *117* (3), 775–816.

__ **and Joel L. Schrag**, "First Impressions Matter: A Model of Confirmatory Bias," *Quarterly Journal of Economics*, 1999, *114* (1), 37–82.

**Richey, Marjorie H., Lucille McClelland, and Algimantas Shimkunas**, "Relative Influence of Positive and Negative Information in Impression Formation and Persistence," *Journal of Personality and Social Psychology*, 1967, *6* (3), 322–327.

__ **, Robert J. Koenigs, Harold W. Richey, and Richard Fortin**, "Negative Salience in Impressions of Character: Effects of Unequal Proportions of Positive and Negative Information," *The Journal of Social Psychology*, 1975, *97* (2), 233–241.

**Sapienza, Paola and Luigi Zingales**, "Economic Experts versus Average Americans," *American Economic Review Papers and Proceedings*, 2013, *103* (3), 636–642.

**Scheinkman, José A. and Wei Xiong**, "Overconfidence and Speculative Bubbles," *Journal of Political Economy*, 2003, *111* (6), 1183–1220.

**Schwartzstein, Joshua**, "Selective Attention and Learning," *Journal of the European Economic Association*, 2014, *12* (6), 1423–1452.

**Selten, Reinhard**, "What is bounded rationality?," in Gerd Gigerenzer and Reinhard Selten, eds., *Bounded Rationality: The Adaptive Toolbox*, MIT Press, 2002, pp. 13–36.

**Shellenbarger, Sue**, "Most Students Don't Know When News is Fake, Stanford Study Finds," *The Wall Street Journal*, November 21 2016. Available online: http://www.wsj.com/articles/most-students-dont-know-when-news-is-fake-stanford-study-finds-1479752576 [Last accessed: November 2016].

**Simon, Herbert A.**, *Models of man*, New York, New York: Wiley, 1957.

**Simonsohn, Uri**, "Lessons from an "Oops" at *Consumer Reports*: Consumers Follow Experts and Ignore Invalid Information," *Journal of Marketing Research*, February 2011, *48*, 1–12.

**Sims, Christopher A.**, "Implications of Rational Inattention," *Journal of Monetary Economics*, 2003, *50*, 665–690.

__ , "Rational Inattention: Beyond the Linear-Quadratic Case," *American Economic Review*, 2006, *96* (2), 158–163.

**Snyder, Timothy**, "The Next Genocide," *New York Times*, September 13 2015. Available online: https://www.nytimes.com/2015/09/13/opinion/sunday/the-next-genocide.html [Last accessed: May 2017].

**Stanford History Education Group**, "Evaluating Information: The Cornerstone of Civic Online Reasoning," 2016. Available online: https://sheg.stanford.edu/upload/V3LessonPlans/Executive%20Summary%2011.21.16.pdf [Last accessed: November 2016].

**The Economist Magazine**, "Sovereign doubts," September 28 2013. Available online: http://www.economist.com/news/schools-brief/21586802-fourth-our-series-articles-financial-crisis-looks-surge-public [Last accessed: September 2016].

__ , "The role of technology in the presidential election," November 20 2016. Available online: http://www.economist.com/news/united-states/21710614-fake-news-big-data-post-mortem-under-way-role-technology [Last accessed: November 2016].

**Wakefield, AJ, SH Murch, A Anthony, J Linnell, DM Casson, M Malik, M Berelowitz, AP Dhillon, MA Thomson, P Harvey, A Valentine, SE Davies, and JA Walker-Smith**, "Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children," *The Lancet*, 1998, *351*, 637–641.

**Weinberger, Marc G., Chris T. Allen, and William R. Dillon**, "Negative Information: Perspectives and Research Directions," *Advances in Consumer Research*, 1981, *8*, 398–404.

**Wilson, Andrea**, "Bounded Memory and Biases in Information Processing," *Econometrica*, 2014, *82* (6), 2257–2294.

# A Appendix (For Online Publication)

## A.1 Generalized Pre-Screening

Let $\omega_0^{q\theta}$ be the prior belief on quality $q$ and state $\theta$, where $\sum_q \sum_\theta \omega_0^{q\theta} = 1$.

When the prior beliefs about the quality and state can potentially be correlated, we cannot apply the first-stage updated belief $\kappa_q(\mathbf{s}^n)$, which is a marginal belief on quality, directly to the second stage in place of a prior belief on quality because the joint priors on quality and state are not independent. Therefore, the generalized pre-screening algorithm requires the second stage to apply Bayes' Rule to a belief whose marginal prior about quality sums to $\kappa_q(\mathbf{s}^n)$. Thus, in the second stage, we assume that the agent applies the *weighted* first-stage updated belief $\kappa_q(\mathbf{s}^n)\left(\frac{\omega_0^{q\theta}}{\sum_\theta \omega_0^{q\theta}}\right)$ in place of the existing joint prior. If the prior beliefs about quality and state are independent ($\omega_0^{q\theta} = \omega_0^q \omega_0^\theta$ for all $q$ and $\theta$), then Equations (11), (12), and (13) reduce to Equations (2), (4), and (5), respectively.

To illustrate the pre-screener's updating algorithm, suppose she observes two signals, one in each period. After observing the first signal ($s_1$), the biased agent's updated belief about the expert's quality, $\kappa_q(s_1)$, is:

$$\kappa_q(s_1) = \frac{\sum_\theta P(s_1|q,\theta)\omega_0^{q\theta}}{\sum_q \sum_\theta P(s_1|q,\theta)\omega_0^{q\theta}}.$$

Using the weighted first-stage updated belief $\kappa_q(s_1)\left(\frac{\omega_0^{q\theta}}{\sum_\theta \omega_0^{q\theta}}\right)$ to form her joint posterior belief on the state and quality, $P^b(q,\theta|s_1)$, yields her posterior beliefs after the first signal:

$$P^b(q,\theta|s_1) = \frac{P(s_1|q,\theta)\kappa_q(s_1)\left(\frac{\omega_0^{q\theta}}{\sum_\theta \omega_0^{q\theta}}\right)}{\sum_q \sum_\theta P(s_1|q,\theta)\kappa_q(s_1)\left(\frac{\omega_0^{q\theta}}{\sum_\theta \omega_0^{q\theta}}\right)}.$$

After observing the second signal ($s_2$), the biased agent's updated belief about the expert's quality, $\kappa_q(s_1,s_2)$ is

$$\kappa_q(s_1,s_2) = \frac{\sum_\theta P(s_2|q,\theta)P^b(q,\theta|s_1)}{\sum_q \sum_\theta P(s_2|q,\theta)P^b(q,\theta|s_1)}.$$

Using the weighted first-stage updated belief $\kappa_q(s_1,s_2)\left(\frac{\omega_0^{q\theta}}{\sum_\theta \omega_0^{q\theta}}\right)$ to form her joint posterior belief on the state and quality, $P^b(q,\theta|s_1,s_2)$, yields:

$$P^b(q,\theta|s_1,s_2) = \frac{P(s_2|q,\theta)P(s_1|q,\theta)\kappa_q(s_1,s_2)\left(\frac{\omega_0^{q\theta}}{\sum_\theta \omega_0^{q\theta}}\right)}{\sum_q \sum_\theta P(s_2|q,\theta)P(s_1|q,\theta)\kappa_q(s_1,s_2)\left(\frac{\omega_0^{q\theta}}{\sum_\theta \omega_0^{q\theta}}\right)}.$$

Iterating on the biased agent's updating process allows us to characterize her posterior beliefs:

Applying the generalized pre-screening procedure described above to prior beliefs $\omega_0^{q\theta}$ yields:

$$\kappa_q(\mathbf{s}^n) = \frac{\left(\frac{\kappa_q(\mathbf{s}^{n-1})}{\sum_\theta \omega_0^{q\theta}}\right)\sum_\theta\left(\prod_{t=1}^n P(s_t|q,\theta)\omega_0^{q\theta}\right)}{\sum_q\left(\frac{\kappa_q(\mathbf{s}^{n-1})}{\sum_\theta \omega_0^{q\theta}}\right)\sum_\theta\left(\prod_{t=1}^n P(s_t|q,\theta)\omega_0^{q\theta}\right)}, \tag{10}$$

where $\kappa_q(\emptyset) = \sum_\theta \omega_0^{q\theta}$.

$$P^b(q,\theta|\mathbf{s}^n) = \frac{\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\left(\frac{\kappa_q(\mathbf{s}^n)}{\sum_\theta \omega_0^{q\theta}}\right)\omega_0^{q\theta}}{\sum_q\sum_\theta\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\left(\frac{\kappa_q(\mathbf{s}^n)}{\sum_\theta \omega_0^{q\theta}}\right)\omega_0^{q\theta}} \tag{11}$$

$$= \frac{\beta_{q\theta}(\mathbf{s}^n)\left(\frac{1}{\sum_\theta \omega_0^{q\theta}}\right)^n\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^{q\theta}}{\sum_q\left(\frac{1}{\sum_\theta \omega_0^{q\theta}}\right)^n\sum_\theta \beta_{q\theta}(\mathbf{s}^n)\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^{q\theta}}. \tag{12}$$

where $\beta_{q\theta}(\mathbf{s}^n)$ is given by:

$$\begin{aligned}\beta_{q\theta}(\mathbf{s}^n) &= \left(\textstyle\sum_\theta P(s_1|q,\theta)\omega_0^{q\theta}\right)\times\left(\textstyle\sum_\theta P(s_1|q,\theta)P(s_2|q,\theta)\omega_0^{q\theta}\right)\times\ldots\times\left(\textstyle\sum_\theta P(s_1|q,\theta)P(s_2|q,\theta)\ldots P(s_n|q,\theta)\omega_0^{q\theta}\right)\\ &= \prod_{m=1}^n\left(\sum_\theta\left(\prod_{t=1}^m P(s_t|q,\theta)\right)\omega_0^{q\theta}\right),\end{aligned} \tag{13}$$

## A.2 Proof of Proposition 1

Define $D(\mathbf{s}^n) = P^b(\theta = A|\mathbf{s}^n) - P^u(\theta = A|\mathbf{s}^n)$ as the ex-post realized disagreement after any signal path. The proposition is that $E_0[D(\mathbf{s}^n)] = 0$, where the expectation $E_0$ is taken by the econometrician over the common prior of states and quality, which we assume reflects the true ex-ante distribution of $(\theta, q)$. Note that the common prior on states and quality generate a common distribution on the probability of any given signal path.

Divide the set of all possible signal paths $\{\mathbf{s}^n\}$ into two groups: one group $\{\mathbf{g}^n\}$ where the first signal is $a$ and another group $\{\mathbf{h}^n\}$ where the first signal is $b$. Because there are two states, there are the same number of signal paths in each group, and the union of these two groups equals $\{\mathbf{s}^n\}$.

It is clear that taking any signal path $\mathbf{g}^n$ and flipping all the $a$'s to $b$ and $b$'s to $a$ defines a one-to-one and onto mapping $F$ of $\{\mathbf{g}^n\}$ into $\{\mathbf{h}^n\}$. This mapping has two properties:

1. $P(\mathbf{g}^n|q,\theta) = P(F(\mathbf{g}^n)|q,-\theta)\ \forall(q,\theta)$, and
2. $P^b(\theta = A|F(\mathbf{g}^n)) - P(\theta = A|F(\mathbf{g}^n)) = -\left(P^b(\theta = A|\mathbf{g}^n) - P(\theta = A|\mathbf{g}^n)\right)\ \forall\mathbf{g^n}$,

where $-\theta$ is the opposite state as $\theta$. The first property says that the probability of the flipped signal sequence is the same as the original signal sequence, once the true state is flipped. The second property can be re-written as $D(F(\mathbf{g}^n)) = -D(\mathbf{g}^n)$ and says that disagreement under the flipped signal path equals the opposite disagreement under the original signal path. Intuitively, these properties follow because, starting from a neutral prior about the state which is independent from quality, the model is symmetric in $A$ and $B$ irrespective of the true expert type.

More precisely, the first property follows because:

$$P(\mathbf{g}^n|q,A) = p_q^{n_a^g}(1-p_q)^{n_b^g} = p_q^{n_b^h}(1-p_q)^{n_a^h} = P(F(\mathbf{g}^n))|q,B),$$

where $n_\theta^g, n_\theta^h$ represent the number of times a signal indicating state $\theta$ appears in signal sequence $\mathbf{g}^n$ and $F(\mathbf{g}^n)$, respectively, and $n_a^g = n_b^h, n_b^g = n_a^h$ by construction. Similarly, $P(\mathbf{g}^n|q,B) = P(F(\mathbf{g}^n)|q,A)$.

To prove the second property, note that, for the Bayesian, $P(\theta = A|\mathbf{g}^n) = P(\theta = B|F(\mathbf{g}^n)) = 1 - P(\theta = A|F(\mathbf{g}^n))$. The first equality follows from applying the first property and $\omega_0^A = \omega_0^B = 0.5$ to Equation 4, noting that a Bayesian has constant $\beta_q(F(\mathbf{g}^n))$.

Now consider the pre-screener. Given any sequence $\mathbf{g}^n$, let $g_i$ and $h_i$ be the $i$-th elements of $\mathbf{g}^n$ and $F(\mathbf{g}^n)$, respectively. Clearly, $h_i$ is the flip of $g_i$, and $P(g_i|q,\theta) = P(h_i|q,-\theta)$, as both equal $p_q$ if $g_i = \theta$ and $1-p_q$ if $g_i = -\theta$. Therefore, $\sum_\theta \left(\prod_{i=1}^m P(g_i|q,\theta)\right)\omega_0^\theta = \sum_\theta \left(\prod_{i=1}^m P(h_i|q,\theta)\right)\omega_0^\theta$ for any $m$ due to the summation over both values of $\theta$. Applying this to Equation 5, $\beta_q(\mathbf{g}^n) = \beta_q(F(\mathbf{g}^n))$. From Equation 4, $P^b(\theta = A|\mathbf{g}^n) = P^b(\theta = B|F(\mathbf{g}^n)) = 1 - P^b(\theta = A|F(\mathbf{g}^n))$, and the second property follows.

We claim that $E_0[D(\mathbf{s}^n)|q = \overline{q}] = 0$ for any $\overline{q} \in \{L, H\}$. To be clear, this conditional expectation is taken over the econometrician's information set, but the true $q$ remains unknown to the Bayesian and pre-screener. The proposition then follows due to the tower property of conditional expectations.

Let $\overline{q}$ be given. Observe that:

$$E[D(\mathbf{s}^n)|q=\overline{q}] = \omega_0^A\left(\sum_{\{\mathbf{g}^n\}} P(\mathbf{g}^n|\overline{q},A)D(\mathbf{g}^n) + \sum_{\{\mathbf{h}^n\}} P(\mathbf{h}^n|\overline{q},A)D(\mathbf{h}^n)\right) + \omega_0^B\left(\sum_{\{\mathbf{g}^n\}} P(\mathbf{g}^n|\overline{q},B)D(\mathbf{g}^n) + \sum_{\{\mathbf{h}^n\}} P(\mathbf{h}^n|\overline{q},B)D(\mathbf{h}^n)\right).$$

The two properties, along the fact that $F$ is one-to-one and onto, imply:

$$\sum_{\{\mathbf{h}^n\}} P(\mathbf{h}^n|\overline{q},A)D(\mathbf{h}^n) = -\sum_{\{\mathbf{g}^n\}} P(\mathbf{g}^n|\overline{q},B)D(\mathbf{g}^n)$$
$$\sum_{\{\mathbf{h}^n\}} P(\mathbf{h}^n|\overline{q},B)D(\mathbf{h}^n) = -\sum_{\{\mathbf{g}^n\}} P(\mathbf{g}^n|\overline{q},A)D(\mathbf{g}^n).$$

With $\omega_0^A = \omega_0^B$, the claim follows. The corollary $Var_0[D(\mathbf{s}^n)] > 0$ follows because $D(F(\mathbf{g}^n))^2 = D(\mathbf{g}^n)^2$.

## A.3 Proof of Proposition 2

**Lemma 2** *For all $\omega_0^\theta \in (0,1)$ and $\omega_0^q \in (0,1)$, $\kappa_H(\mathbf{s}^n) < w_0^H$ if and only if $\beta_H(\mathbf{s}^n) < \beta_L(\mathbf{s}^n)$. Likewise, $\kappa_H(\mathbf{s}^n) > w_0^H$ if and only if $\beta_H(\mathbf{s}^n) > \beta_L(\mathbf{s}^n)$. $\kappa_H(\mathbf{s}^n) = w_0^H$ if and only if $\beta_H(\mathbf{s}^n) = \beta_L(\mathbf{s}^n)$.*

**Proof.** For any given sequence of signals $\mathbf{s}^n = (s_1, s_2, \ldots, s_n)$, $\kappa_q(\mathbf{s}^n)$ can be re-written as

$$\begin{aligned}
\kappa_q(\mathbf{s}^n) &= \frac{\beta_q(\mathbf{s}^{n-1})\omega_0^q \sum_\theta \left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^\theta}{\sum_q \beta_q(\mathbf{s}^{n-1}) \sum_\theta \left(\prod_{t=1}^n P(s_n|q,\theta)\right)\omega_0^\theta\omega_0^q} \\
&= \frac{\left(\prod_{m=1}^{n-1}\left(\sum_\theta \left(\prod_{t=1}^m P(s_t|q,\theta)\right)\omega_0^\theta\right)\right)\omega_0^q \sum_\theta \left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^\theta}{\sum_q \left(\prod_{m=1}^{n-1}\left(\sum_\theta \left(\prod_{t=1}^m P(s_t|q,\theta)\right)\omega_0^\theta\right)\right) \sum_\theta \left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^\theta\omega_0^q} \\
&= \frac{\beta_q(\mathbf{s}^n)\omega_0^q}{\sum_q \beta_q(\mathbf{s}^n)\omega_0^q}.
\end{aligned}$$

Thus, the statement is shown for $\mathbf{s}^n = (s_1, s_2, \ldots, s_t)$.[7] ■

From Equation (4), the biased agent's posterior that the expert is high quality is lower than the Bayesian's if and only if $\kappa_H(\mathbf{s}^n) < w_0^H$. Lemma 2 shows that this is only the case if and only if $\beta_H(\mathbf{s}^n) < \beta_L(\mathbf{s}^n)$. Thus, $\beta_H(\mathbf{s}^n) < \beta_L(\mathbf{s}^n)$ if and only if $P^b(q = H|\mathbf{s}^n) < P^u(q = H|\mathbf{s}^n)$.

Consider $P^b(\theta = A|\mathbf{s}^n) < P^u(\theta = A|\mathbf{s}^n)$:

$$P^b(\theta = A|\mathbf{s}^n) < P^u(\theta = A|\mathbf{s}^n)$$
$$\frac{\omega_0^A \sum_q \beta_q(\mathbf{s}^n)\left(\prod_{t=1}^n P(s_t|q,A)\right)\omega_0^q}{\sum_q \beta_q(\mathbf{s}^n)\sum_\theta \left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^\theta\omega_0^q} < \frac{\omega_0^A \sum_q \left(\prod_{t=1}^n P(s_t|q,A)\right)\omega_0^q}{\sum_q \sum_\theta \left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^\theta\omega_0^q},$$

[7] If the signals are observed simultaneously (e.g., in period 1), then the above argument applies analogously, where $\beta_q(\mathbf{s}^{t-1}) = \beta_q(\emptyset) = 1$ instead. Thus, $\kappa_H(\mathbf{s}^n) < w_0^H$ implies $\beta_H(\mathbf{s}^n) < \beta_L(\mathbf{s}^n)$ and vice versa. Likewise when the inequality reverses or when the equality holds.

which is true if and only if

$$0 < \omega_0^A(1-\omega_0^A)\omega_0^H(1-\omega_0^H)(\beta_L(\mathbf{s}^n)-\beta_H(\mathbf{s}^n))\left(\left(\prod_{t=1}^n P(s_t|H,A)\right)\left(\prod_{t=1}^n P(s_t|L,B)\right)-\left(\prod_{t=1}^n P(s_t|H,B)\right)\left(\prod_{t=1}^n P(s_t|L,A)\right)\right)$$

$$0 < \omega_0^A(1-\omega_0^A)\omega_0^H(1-\omega_0^H)(\beta_L(\mathbf{s}^n)-\beta_H(\mathbf{s}^n))\left(p_H^{n_a}(1-p_H)^{n_b}p_L^{n_b}(1-p_L)^{n_a}-p_H^{n_b}(1-p_H)^{n_a}p_L^{n_a}(1-p_L)_b^n\right)$$

$$0 < \omega_0^A(1-\omega_0^A)\omega_0^H(1-\omega_0^H)(\beta_L(\mathbf{s}^n)-\beta_H(\mathbf{s}^n))\left(p_H(1-p_H)p_L(1-p_L)\right)^{n_b}\left((p_H(1-p_L))^{n_a-n_b}-((1-p_H)p_L)^{n_a-n_b}\right),$$

which is true when $n_a > n_b$ since $p_H > p_L$. Clearly, $P^u(A|\mathbf{s}^n) > \frac{1}{2}$ only if $n_a > n_b$, so A is the (objectively) more likely state. Note that if $n_a = n_b$, then $P^b(\theta = A|\mathbf{s}^n) = Pr^u(\theta = A|\mathbf{s}^n)$ regardless of the biased agent's beliefs on the expert's quality. Thus, for any $n_a > n_b$ set of signals and for all $\omega_0^\theta \in (0,1)$, under-trust in expert quality implies pessimism in beliefs about the more likely state: If $P^b(q = H|\mathbf{s}^n) < P^u(q = H|\mathbf{s}^n)$, then $P^b(\theta = A|\mathbf{s}^n) < P^u(\theta = A|\mathbf{s}^n)$. Likewise, $P^b(\theta = A|\mathbf{s}^n) < P^u(\theta = A|\mathbf{s}^n)$ if and only if $\beta_H(\mathbf{s}^n) < \beta_L(\mathbf{s}^n)$ when $n_a > n_b$, which implies that $P^b(q = H|\mathbf{s}^n) < P^u(q = H|\mathbf{s}^n)$. Reversing the inequalities yields that overtrust in expert quality implies optimism in beliefs about the more likely state, and vice versa.

## A.4 Proof of Proposition 3

Let $n_a^s$ be the number of $a$ signals and $n_b^s$ be the number of $b$ signals in sequence $s$. Consider any two sequences $\mathbf{x}_n$ and $\mathbf{y}_n$ with identical information content ($n_a^x = n_a^y$ and $n_b^x = n_b^y$). Let $\beta_q^s$ correspond to sequence $s \in \{\mathbf{x}_n, \mathbf{y}_n\}$ and $q \in \{L, H\}$. Let $\mathbf{s}_J^n = \mathbf{x}_n$ and $\mathbf{s}_M^n = \mathbf{y}_n$. Without loss of generality, let $n_a > n_b$.

1. Correlated disagreement

   By direct comparison of the posteriors on expert quality, a necessary and sufficient condition for $P^b(q = H|\mathbf{x}^n) > P^b(q = H|\mathbf{y}^n)$ is $\beta_H^x\beta_L^y - \beta_L^x\beta_H^y > 0$. By direct comparison of the posteriors on the most likely state (which is $A$ because $n_a > n_b$), a necessary and sufficient condition for $P^b(\theta = A|\mathbf{x}^n) > P^b(\theta = A|\mathbf{y}^n)$ is $\beta_H^x\beta_L^y - \beta_L^x\beta_H^y > 0$. Since the same condition $\beta_H^x\beta_L^y - \beta_L^x\beta_H^y > 0$ is required for both $P^b(q = H|\mathbf{x}^n) > P^b(q = H|\mathbf{y}^n)$ and $P^b(\theta = A|\mathbf{x}^n) > P^b(\theta = A|\mathbf{y}^n)$, then disagreement between biased agents is correlated. That is, $P^b(q = H|\mathbf{x}^n) > P^b(q = H|\mathbf{y}^n)$ if and only if $P^b(\theta = A|\mathbf{x}^n) > P^b(\theta = A|\mathbf{y}^n)$. Clearly reversing all the inequalities applies as well.

2. Expected disagreement

   It is sufficient to show that $P^b(\theta = A|\mathbf{x}^n) \neq P^b(\theta = A|\mathbf{y}^n)$ for at least two sequences $\mathbf{x}_n$ and $\mathbf{y}_n$ with identical information content ($n_a^x = n_a^y = n_a$ and $n_b^x = n_b^y = n_b$). Consider two sequences such that $n_a \geq n_b$, where the first $j = n_a - n_b$ signals are the same and $n-2 \geq j \geq 1$, the two sequences differ in the $j+1$ and $j+2$ signals, and then all subsequent signals are identical (i.e., terms $j+3$ through $n$). Let $x_{j+1} = a$, $x_{j+2} = b$, $y_{j+1} = b$, and $y_{j+2} = a$. Suppose the first $j$ terms contain $k$ $a$'s and $j-k$ $b$'s, where $k \geq j-k$. As shown in the proof of Lemma 1, $P^b(q = H|\mathbf{x}^n) > P^b(q = H|\mathbf{y}^n)$ when $k > j-k$. As shown in

the preceding proof of correlated disagreement between two prescreeners, this implies that $P^b(\theta = A|\mathbf{x}^n) > P^b(\theta = A|\mathbf{y}^n)$. Thus, $P^b(\theta = A|\mathbf{x}^n) \neq P^b(\theta = A|\mathbf{y}^n)$ for at least two sequences $\mathbf{x}_n$ and $\mathbf{y}_n$ with identical information content ($n_a^x = n_a^y = n_a$ and $n_b^x = n_b^y = n_b$).

## A.5 Proof of Lemma 1

Let $n_a^s$ be the number of $a$ signals and $n_b^s$ be the number of $b$ signals in sequence $s$. Consider any two sequences $\mathbf{x}_n$ and $\mathbf{y}_n$ with identical information content ($n_a^x = n_a^y = n_a$) and $n_b^x = n_b^y = n_b$). Let $\beta_q^s$ correspond to sequence $s \in \{\mathbf{x}_n, \mathbf{y}_n\}$ and $q \in \{L, H\}$. Without loss of generality, let $n_a > n_b$.

By direct comparison of the posteriors on expert quality, a necessary and sufficient condition for sequence $x$ to generate more trust than sequence $y$ (i.e., $P^b(q = H|\mathbf{x}^n) > P^b(q = H|\mathbf{y}^n)$) is $\beta_H^x\beta_L^y - \beta_L^x\beta_H^y > 0$, where $\beta_q^s$ corresponds to sequence $s \in \{\mathbf{x}_n, \mathbf{y}_n\}$ and $q \in \{L, H\}$. Consider two sequences such that $n_a \geq n_b$, where the first $j = n_a - n_b$ signals are the same and $n - 2 \geq j \geq 1$, the two sequences differ in the $j+1$ and $j+2$ signals, and then all subsequent signals are identical (i.e., terms $j+3$ through $n$). Let $x_{j+1} = a$, $x_{j+2} = b$, $y_{j+1} = b$, and $y_{j+2} = a$. (For example, sequence 1 could be *aababaa* and sequence 2 could be *aabbaaa* - here $j = 3$, $n_a = 2$, $n_b = 1$.) Then $\beta_H^x\beta_L^y - \beta_L^x\beta_H^y > 0$ whenever $n_a > n_b$ and $\beta_H^x\beta_L^y - \beta_L^x\beta_H^y = 0$ whenever $n_a = n_b$. To see this, note that, given the general expression for $\beta_q^s$, all of the terms are identical for $\beta_q^x$ and $\beta_q^y$ except term $j+1$. Suppose the first $j$ terms contain $k$ $a$'s and $j-k$ $b$'s, where $k \geq j-k$. This implies that when $\omega_0^\theta = 1/2$, then $\beta_H^x\beta_L^y - \beta_L^x\beta_H^y \geq 0$ if

$$\left(p_H^{k+1}(1-p_H)^{j-k} + (1-p_H)^{k+1}p_H^{j-k}\right)\left(p_L^k(1-p_L)^{j-k+1} + (1-p_L)^k p_L^{j-k+1}\right) -$$
$$\left(p_L^{k+1}(1-p_L)^{j-k} + (1-p_L)^{k+1}p_L^{j-k}\right)\left(p_H^k(1-p_H)^{j-k+1} + (1-p_H)^k p_H^{j-k+1}\right) \geq 0$$
$$(p_H - p_L)\left((p_H p_L)^{2k-j} - ((1-p_H)(1-p_L))^{2k-j}\right) + (p_H + p_L - 1)\left((p_H(1-p_L))^{2k-j} - (p_L(1-p_H))^{2k-j}\right) \geq 0.$$

We can verify that both terms are positive when $k > j-k$ and zero when $k = j-k$. Thus, $P^b(H|\mathbf{x}^n) > P^b(H|\mathbf{y}^n)$ when $k > j-k$. Using this result, we can iteratively apply it to order sequences of fixed composition in decreasing trust by starting with the sequence with the least reversals (all $a$'s followed by all b's), and iteratively switching the first $b$ and last $a$ to generate sequences where the first $b$ moves forward. E.g., *aaaabb* generates more trust than *aaabab*, which generates more trust than *aabaab* which generates more trust than *abaaab*. Then, *aaabba* generates more trust than *aababa* than *abaaba*, where *aaabab* generates more trust than *aaabba* and *abaaab* generates more trust than *abaaba*. We can keep doing this (and applying the result that $P^b(H|\mathbf{x}^n) > P^b(H|\mathbf{y}^n)$ when $k > j-k$) to establish that *aaaabb* generates the most trust and *ababaa* generates the least trust.

## A.6 Proof of Proposition 4

1. Positive first impressions

Suppose the agent observes $n_a \geq 1$ consecutive $a$ signals, followed by $m$ pairs of $(b,a)$ signals: $\mathbf{s}^n = (a, a, a, \ldots, b, a, b, a)$. This sequence generates[8]:

$$\beta_q(\mathbf{s}^n) = \left(\tfrac{1}{2}\right)^{n_a+m} [p_q(1-p_q)]^{m(m+1)} \left([p_q^{n_a-1} + (1-p_q)^{n_a-1}][p_q^{n_a} + (1-p_q)^{n_a}]\right)^m \left(\prod_{i=1}^{n_a}(p_q^i + (1-p_q)^i)\right) \tag{14}$$

$$\begin{aligned}\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q} = {} & \left(\tfrac{1}{2}\right)^{n_a+m} [p_q(1-p_q)]^{m(m+1)} \left([p_q^{n_a-1} + (1-p_q)^{n_a-1}][p_q^{n_a} + (1-p_q)^{n_a}]\right)^{m-1} \left(\prod_{i=1}^{n_a}(p_q^i(1-p_q)^i)\right) \\ & \Big(mp_q(1-p_q)\left((n_a-1)(p_q^{n_a-2} - (1-p_q)^{n_a-2})(p_q^{n_a} + (1-p_q)^{n_a}) + n_a(p_q^{n_a-1} + (1-p_q)^{n_a-1})(p_q^{n_a-1} - (1-p_q)^{n_a-1})\right) \\ & + (p_q^{n_a-1} + (1-p_q)^{n_a-1})(p_q^{n_a} + (1-p_q)^{n_a})\left(m(m+1)(1-2p_q) + p_q(1-p_q)\sum_{i=1}^{n_a} \tfrac{i(p_q^{i-1} - (1-p_q)^{i-1})}{p_q^i + (1-p_q)^i}\right)\Big).\end{aligned} \tag{15}$$

Since $\beta_q$ is symmetric about $p_q = 1/2$, we focus on the case of $p_q \geq 1/2$ but the conclusion extends to $1/2 > p_L > 1 - p_H$. From Equations (14) and (15), we can see that $\frac{\partial}{\partial p_q}(\beta_q(\mathbf{s}^n)) = 0$ when $p_q \in \{\frac{1}{2}, 1\}$, $\beta_q(\mathbf{s}^n) > 0$ when $p_q = \frac{1}{2}$, and $\beta_q(\mathbf{s}^n) = 0$ when $p_q = 1$. Moreover, using the fact that $\beta_q(\mathbf{s}^n) = 0$ when $p_q = 1/2$, then

$$\begin{aligned}\frac{\partial^2 \beta_q(\mathbf{s}^n)}{\partial p_q^2}\Big|_{p_q=\frac{1}{2}} = {} & \beta_q(\mathbf{s}^n)[p_q(1-p_q)(p_q^{n_a} + (1-p_q)^{n_a})(p_q^{n_a-1} + (1-p_q)^{n_a-1})]^{-1} \\ & \left(\frac{1}{2}\right)^{2n_a-3} \left(2m[(n_a-1)^2 - (m+1)] + \frac{1}{3}n_a(n_a-1)(n_a+1)\right)\end{aligned} \tag{16}$$

Note that the last term of Equation (16) increases in $n_a$ for all $n_a > 1$, and that it is positive for all $m > m'$ where

$$2m'[(n_a-1)^2 - (m'+1)] + \frac{1}{3}n_a(n_a-1)(n_a+1) = 0.$$

By direct computation, we can see that $\frac{\partial \beta_q}{\partial p_q} < 0$ for all $m \geq 1$ and $p_q \in (1/2, 1)$ for $n_a \in \{1, 2\}$. This implies that the agent under-trusts and is pessimistic about the most likely state for all $m \geq 1$ when $n_a \leq 2$. Consider the case of $n_a \geq 3$. Since $\beta_q(\mathbf{s}^n) > 0$ when $p_q = 1/2$, $\beta_q(\mathbf{s}^n) = 0$ when $p_q = 1$, $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q}\Big|_{p_q=1/2} = 0$, and $\beta_q(\mathbf{s}^n) \geq 0$ for any $p_q \in [0,1]$, then there exists some threshold $\frac{1}{2} < p' < 1$ such that $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q} > 0$ for all $p_q < p'$ when $\frac{\partial^2 \beta_q(\mathbf{s}^n)}{\partial p_q^2}\Big|_{p_q=\frac{1}{2}} > 0$, which holds when $m < m'$. This implies that the agent overtrusts and is optimistic about the most likely state when $m < m'$ and $p_L < p_H \leq p'$. Since the last term of Equation (16) increases in $n_a$ for all $n_a > 1$ and

[8] We use the property that a product of multiple factors is given by $\frac{d}{dx}\left(\prod_{i=1}^k f_i(x)\right) = \left(\prod_{i=1}^k f_i(x)\right)\left(\sum_{i=1}^k \frac{f_i'(x)}{f_i(x)}\right)$.

$\frac{\partial^2 \beta_q(\mathbf{s}^n)}{\partial p_q^2}\Big|_{p_q=\frac{1}{2}} > 0$ for $m \leq 3$, then $m' > 3$ for all $n_a \geq 3$.

2. Negative first impressions

Suppose the agent observes $n_b \geq 1$ pairs of $(a, b)$ signals, followed by $m \geq 1$ consecutive $a$ signals, where $m \geq 1$: $\mathbf{s}^n = (a, b, a, b, \ldots, a, a, a)$. This sequence generates:

$$\beta_q(\mathbf{s}^n) = \left(\frac{1}{2}\right)^{n_b} (p_q(1-p_q))^{n_b^2} \left(\prod_{i=1}^{m} \frac{1}{2}\left(p_q^{i+n_b}(1-p_q)^{n_b} + p_q^{n_b}(1-p_q)^{i+n_b}\right)\right) \quad (17)$$

$$= \left(\frac{1}{2}\right)^{n_b+m} (p_q(1-p_q))^{n_b(n_b+m)} \left(\prod_{i=1}^{m}(p_q^i + (1-p_q)^i)\right). \quad (18)$$

$$\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q} = \left(\frac{1}{2}\right)^{n_b+m} (p_q(1-p_q))^{n_b(n_b+m)-1} \left(\prod_{i=1}^{m}(p_q^i + (1-p_q)^i)\right)$$
$$\left(n_b(n_b+m)(1-2p_q) + p_q(1-p_q)\sum_{i=1}^{m}\left(\frac{i(p_q^{i-1} - (1-p_q)^{i-1}}{p_q^i + (1-p_q)^i}\right)\right) \quad (19)$$

Since $\beta_q$ is symmetric about $p_q = 1/2$, we focus on the case of $p_q \geq 1/2$ but the conclusion extends to $1/2 > p_L > 1 - p_H$. Evaluating Equation (19) when $m = 1$ (i.e., $n_a = n_b + 1$ where $n_b \geq 1$), $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q}|_{m=1} < 0$. Thus, by Proposition 2, the pre-screener under-trusts and is pessimistic about the mostly likely state, A, when she observes a sequence $\mathbf{s}^n = (a, b, a, b, \ldots, a, b, a)$ where $n_a = n_b + 1$. Further, evaluating Equation (19) when $m = 2$ (i.e., $n_a = n_b + 2$ where $n_b \geq 1$), $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q}|_{m=1} < 0$. Thus, the pre-screener still under-trusts and is pessimistic about the most likely state, A, when $\mathbf{s}^n = (a, b, a, b, \ldots, a, a)$ where $n_a = n_b + 2$ for all $n_b \geq 1$. Further, evaluating Equation 19 when $m = 3$ (i.e., $n_a = n_b + 3$ where $n_b \geq 1$), $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q}|_{m=3} \leq 0$ with equality at $p_q = \frac{1}{2}$ only if $n_b = 1$. Since the third term of Equation (19) is decreasing in $n_b$ for all $p_q \in (\frac{1}{2}, 1]$, then $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q}|_{m=3} < 0$ for all $n_b > 1$. Thus, the pre-screener still under-trusts and is pessimistic about the most likely state, A, when $\mathbf{s}^n = (a, b, a, b, \ldots, a, a)$ where $n_a = n_b + 3$ for all $n_b \geq 1$. Therefore, there exists some $m^* > 3$ such that $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q} < 0$ for all $m < m^*$, which implies that the pre-screener will under-trust for $m < m^*$. Moreover, since the third term of Equation (19) is decreasing in $n_b$ for all $p_q \in (\frac{1}{2}, 1]$, then $m^*$ is increasing in $n_b$.

## A.7 Proof of Proposition 5

1. Since $\beta_q$ is symmetric about $p_q = 1/2$, we focus on the case of $p_q \geq 1/2$ but the conclusion extends to $1/2 > p_L > 1 - p_H$. Note that Equation (15) can be re-written as

$$\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q} = \left(\frac{1}{2}\right)^{n_a+m} m[p_q(1-p_q)]^{m(m+1)} \left([p_q^{n_a-1} + (1-p_q)^{n_a-1}][p_q^{n_a} + (1-p_q)^{n_a}]\right)^{m-1} \left(\prod_{i=1}^{n_a}(p_q^i(1-p_q)^i)\right)(Z),$$

where $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q}$ is negative whenever $Z$ is negative and $p_q \in (\frac{1}{2}, 1)$, and

$$Z = p_q(1-p_q)\left((n_a-1)(p_q^{n_a-2} - (1-p_q)^{n_a-2})(p_q^{n_a} + (1-p_q)^{n_a}) + n_a(p_q^{n_a-1} + (1-p_q)^{n_a-1})(p_q^{n_a-1} - (1-p_q)^{n_a-1})\right)$$
$$+ (p_q^{n_a-1} + (1-p_q)^{n_a-1})(p_q^{n_a} + (1-p_q)^{n_a})\left((m+1)(1-2p_q) + \left(\frac{1}{m}\right)p_q(1-p_q)\sum_{i=1}^{n_a}\frac{i(p_q^{i-1} - (1-p_q)^{i-1})}{p_q^i + (1-p_q)^i}\right).$$

For given $n_a$, $Z$ is more than linearly decreasing in $m$. Thus, there exists $\hat{m}$, defined by $Z(\hat{m}) = 0$, such that $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q} < 0$ for all $p_q \in (\frac{1}{2}, 1)$ when $m > \hat{m}$. Thus for any given $n_a$, there exists $\hat{m}$ such that when $m > \hat{m}$, the pre-screener under-trusts and is pessimistic about the most likely state for any $(p_L, p_H)$.

2. Since $\beta_q$ is symmetric about $p_q = 1/2$, we focus on the case of $p_q \geq 1/2$ but the conclusion extends to $1/2 > p_L > 1 - p_H$. From Equations (18) and (19), we can see that $\frac{\partial}{\partial p_q}(\beta_q(\mathbf{s}^n)) = 0$ when $p_q \in \{\frac{1}{2}, 1\}$, $\beta_q(\mathbf{s}^n) > 0$ when $p_q = \frac{1}{2}$, and $\beta_q(\mathbf{s}^n) = 0$ when $p_q = 1$. Since $\beta_q(\mathbf{s}^n) = 0$ when $p_q = 1$, $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q}\big|_{p_q=1} = 0$, and $\beta_q(\mathbf{s}^n) \geq 0$ for any $p_q \in [0, 1]$, then there exists some threshold $\hat{p} < 1$ such that $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q} < 0$ and $\beta_q(\mathbf{s}^n) < \beta_q(\mathbf{s}^n)\Big|_{p_q=\frac{1}{2}}$ for all $p_q > \hat{p}$. Therefore, $\beta_L(\mathbf{s}^n) > \beta_H(\mathbf{s}^n)$ so the pre-screener under-trusts and is pessimistic about the most likely state if $\hat{p} \leq p_L < p_H$.

Moreover, using the fact that $\beta_q(\mathbf{s}^n) = 0$ when $p_q = 1/2$, then

$$\frac{\partial^2 \beta_q(\mathbf{s}^n)}{\partial p_q^2}\Big|_{p_q=\frac{1}{2}} = \beta_q(\mathbf{s}^n)\left(-8n_b(n_b+m) + \sum_{i=1}^{m} 4i(i-1)\right)$$
$$= \beta_q(\mathbf{s}^n)\left(-8n_b(n_b+m) + \frac{4}{3}m(m-1)(m+1)\right). \tag{20}$$

Thus there exists some threshold $\frac{1}{2} < \check{p} < 1$ such that $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q} > 0$ for all $p_q < \check{p}$ when $\frac{\partial^2 \beta_q(\mathbf{s}^n)}{\partial p_q^2}\Big|_{p_q=\frac{1}{2}} > 0$. Note that $\check{p} > \frac{1}{2}$ for any given $n_b$ if $m$ is sufficiently large that $\frac{\partial^2 \beta_q(\mathbf{s}^n)}{\partial p_q^2}\Big|_{p_q=\frac{1}{2}} > 0$. Thus the pre-screener also under-trusts and is pessimistic about the most likely state if $p_L \leq \check{p}$ and $p_H > \hat{p}$ where $\check{p} \geq \frac{1}{2}$. Note that we have already shown

directly (in the preceding proof) that the pre-screener under-trusts and is pessimistic for $m = 1, 2, 3$ regardless of $p_L$, $p_H$, and $n_b$.

## A.8 Proof of Proposition 6

Shown in Proof of Proposition 2.

## A.9 Proof of Proposition 7

1. To show the results when agents receive signals from multiple experts, note that Equation (9) can also be re-written as

$$P^b(q_1, q_2, \theta|\mathbf{s}^{n_1,n_2}) = \frac{\left(\prod_{t=n_1+1}^{n_1+n_2} P(s_{t2}|q_2,\theta)\right)\left(\prod_{t=1}^{n_1} P(s_{t1}|q_1,\theta)\right)\omega_0^{q_1}\omega_0^{q_2}\omega_0^{\theta}\beta_{q_1}(\mathbf{s}^{n_1})\beta_{q_2q_1}(\mathbf{s}^{n_1,n_2})}{\sum_{q_2}\sum_{q_1}\sum_{\theta}\left(\prod_{t=n_1+1}^{n_1+n_2} P(s_{t2}|q_2,\theta)\right)\left(\prod_{t=1}^{n_1} P(s_{t1}|q_1,\theta)\right)\omega_0^{q_1}\omega_0^{q_2}\omega_0^{\theta}\beta_{q_1}(\mathbf{s}^{n_1})\beta_{q_2q_1}(\mathbf{s}^{n_1,n_2})}, \tag{21}$$

where the functions $\beta_{q_1}(\mathbf{s}^{n_1})$ and $\beta_{q_2q_1}(\mathbf{s}^{n_1,n_2})$ reflect the path dependency of the biased agent's beliefs and $\beta_{q_1}(\emptyset) = 1$:

$$\begin{aligned}
\beta_{q_1}(\mathbf{s}^{n_1}) &= \prod_{m=1}^{n_1}\left(\sum_{\theta}\left(\prod_{t=1}^{m} P(s_{t1}|q_1,\theta)\right)\omega_0^{\theta}\right) \\
\beta_{q_2q_1}(\mathbf{s}^{n_1,n_2}) &= \prod_{m=n_1+1}^{n_1+n_2}\left(\sum_{\theta}\left(\prod_{t=n_1+1}^{m} P(s_{t2}|q_2,\theta)\right)\left(\prod_{t=1}^{n_1} P(s_{t1}|q_1,\theta)\right)\omega_0^{\theta}\right).
\end{aligned}$$

Consider a sequence of signals such that the agent observes $k$ $a$ signals from expert 1, followed by $k$ $b$ signals from expert 2: $\mathbf{s}^{n_1} = (a, \ldots, a)$ and $\mathbf{s}^{n_2} = (b, \ldots, b)$ where $n_1 = n_2 = k$.

To show this, note that the following properties hold when $\omega_0^{\theta} = 1/2$ and the two experts send either (1) an equal number $k$ of opposing signals, or (2) an equal number of completely mixed signals: $\prod_{i=k+1}^{2k} P(s_{i2}|H_2, A) = \prod_{i=1}^{k} P(s_{i1}|H_1, B)$, $\prod_{i=k+1}^{2k} P(s_{i2}|L_2, A) = \prod_{i=1}^{k} P(s_{i1}|L_1, B)$, $\prod_{i=k+1}^{2k} P(s_{i2}|H_2, B) = \prod_{i=1}^{k} P(s_{i1}|H_1, A)$, and $\prod_{i=k+1}^{2k} P(s_{i2}|L_2, B) = \prod_{i=1}^{k} P(s_{i1}|L_1, A)$.

For all $\omega_0^{\theta} \in (0, 1)$, then $P^b(\theta|\mathbf{s}^{n_1,n_2}) > 1/2$ only if

$$\begin{aligned}
&\omega_0^H(1-\omega_0^H)\left(\left(\prod_{i=k+1}^{2k} P(s_{i2}|H_2, A)\right)\left(\prod_{i=k+1}^{2k} P(s_{m2}|L_2, B)\right) - \left(\prod_{i=k+1}^{2k} P(s_{i2}|L_2, A)\right)\left(\prod_{i=2}^{2k} P(s_{i2}|H_2, B)\right)\right) \\
&(\beta_{L_1}(\mathbf{s}^n)\beta_{H_2L_1}(\mathbf{s}^n) - \beta_{H_1}(\mathbf{s}^n)\beta_{L_2H_1}(\mathbf{s}^n)) > 0.
\end{aligned} \tag{22}$$

When the two experts send an equal number of opposing signals in sequence (and

suppressing the arguments of $\beta_{q_1}(\mathbf{s}^n)$ and $\beta_{q_1q_2}(\mathbf{s}^n)$ for brevity of exposition), we also know

$\prod_{i=k+1}^{2k} P(s_{i2}|H_2, A) = \prod_{i=1}^{k} P(s_{i1}|H_1, B) = (1-p_H)^k$
$\prod_{i=k+1}^{2k} P(s_{i2}|L_2, A) = \prod_{i=1}^{k} P(s_{i1}|L_1, B) = (1-p_L)^k$
$\prod_{i=k+1}^{2k} P(s_{i2}|H_2, B) = \prod_{i=1}^{k} P(s_{i1}|H_1, A) = p_H^k$
$\prod_{i=k+1}^{2k} P(s_{i2}|L_2, B) = \prod_{i=1}^{k} P(s_{i1}|L_1, A) = p_L^k$
$\beta_{q_1} = \prod_{i=1}^{k}(\frac{1}{2})(p_{q_1}^i + (1-p_{q_1})^i)$, where we have previously shown that $\beta_{H_1} > \beta_{L_1}$
$\beta_{q_2q_1} = \prod_{i=1}^{k}(\frac{1}{2})\left((1-p_{q_2})^i p_{q_1}^k + p_{q_2}^i(1-p_{q_1})^k\right)$

Substituting all of these into the biased agent's posterior on the state, $P^b(\theta|\mathbf{s}^{n_1,n_2}) > 1/2$ only if

$$\omega_0^H(1-\omega_0^H)\left((1-p_H)^k p_L^k - (1-p_L)^k p_H^k\right)(\beta_{L_1}\beta_{H_2L_1} - \beta_{H_1}\beta_{L_2H_1}) > 0. \qquad (23)$$

The first term of Equation (23) is positive and the second term is clearly negative, since $p_H > p_L$. Note that $\beta_L < \beta_H$ for $n_a > 1$ and $\beta_L = \beta_H$ for $n_a = 1$. Comparing a given $m$th term of $\beta_{L_2H_1} - \beta_{H_2L_1}$ yields

$$\begin{aligned}&(\frac{1}{2})\left((1-p_L)^m p_H^k + p_L^m(1-p_H)^k - (1-p_H)^m p_L^k - p_H^m(1-p_L)^k\right)\\ &\quad = (\frac{1}{2})\left(p_H^m(1-p_L)^m(p_H^{k-m} - (1-p_L)^{k-m}) + p_L^m(1-p_H)^m((1-p_H)^{k-m} - p_L^{k-m})\right),\end{aligned}$$

which is zero if $k = m$ and positive if $m < k$. Thus, each $m$th term of $\beta_{L_2H_1}$ is strictly greater than the $m$th term of $\beta_{H_2L_1}$ for $m < k$ and is equal when $m = k$, implying that $\beta_{L_2H_1} > \beta_{H_2L_1}$ if $k > 1$ (and $\beta_{L_2H_1} = \beta_{H_2L_1}$ if $k = 1$) . This implies that the third term of Equation (23) is strictly negative when $k > 1$, so Equation (23) is satisfied. Thus, $P^b(\theta = A|\mathbf{s}^{n_1,n_2}) > 1/2$ when $\omega_0^A = 1/2$ and $k > 1$, and $P^b(\theta = A|\mathbf{s}^{n_1,n_2}) = 1/2$ when $\omega_0^A = 1/2$ and $k = 1$.

Substituting all of these into the biased agent's posteriors on expert qualities, we have that $P^b(q_1|\mathbf{s}^{n_1,n_2}) > P^b(q_2|\mathbf{s}^{n_1,n_2})$ only if $\omega_0^H(1-\omega_0^H)\left((1-p_H)^k p_L^k - (1-p_L)^k p_H^k\right)(\beta_{L_1}\beta_{H_2L_1} - \beta_{H_1}\beta_{L_2H_1}) > 0$, which is exactly Equation (23) again. Thus, the biased agent believes that the first expert is more likely to be high quality than the second expert: $P^b(q_1|\mathbf{s}^{n_1,n_2}) > P^b(q_2|\mathbf{s}^{n_1,n_2})$.

2. The biased agent's joint posteriors on experts' qualities are of the form $P^b(H_1, H_2|s_{1j}, s_{2j}) = \frac{w}{w+x+y+z}$, $P^b(L_1, H_2|s_{1j}, s_{2j}) = \frac{x}{w+x+y+z}$, $P^b(H_1, L_2|s_{1j}, s_{2j}) = \frac{y}{w+x+y+z}$, and $P^b(H_1, H_2|s_{1j}, s_{2j}) =$

$\frac{z}{w+x+y+z}$, where

$$w = 2(\omega_0^H)^2(1-p_H)^k p_H^{2k(k+1)} \left( \prod_{m=1}^{k} (1 + (\frac{1-p_H}{p_H})^m)((\frac{1-p_H}{p_H})^m + (\frac{1-p_H}{p_H})^k) \right)$$

$$x = \omega_0^H(1-\omega_0^H)\left((1-p_H)^k p_L^k + p_H^k(1-p_L)^k\right) p_L^{\frac{k(k+1)}{2}} p_H^{\frac{k(k+1)}{2}} p_L^{k^2} \left( \prod_{m=1}^{k} (1 + (\tfrac{1-p_L}{p_L})^m)((\tfrac{1-p_H}{p_H})^m + (\tfrac{1-p_L}{p_L})^k) \right)$$

$$y = \omega_0^H(1-\omega_0^H)\left((1-p_L)^k p_H^k + p_L^k(1-p_H)^k\right) p_H^{\frac{k(k+1)}{2}} p_L^{\frac{k(k+1)}{2}} p_H^{k^2} \left( \prod_{m=1}^{k} (1 + (\tfrac{1-p_H}{p_H})^m)((\tfrac{1-p_L}{p_L})^m + (\tfrac{1-p_H}{p_H})^k) \right)$$

$$z = 2(1-\omega_0^H)^2(1-p_L)^k p_L^{2k(k+1)} \left( \prod_{m=1}^{k} (1 + (\frac{1-p_L}{p_L})^m)((\frac{1-p_L}{p_L})^m + (\frac{1-p_L}{p_L})^k) \right).$$

Letting $k \to \infty$ and factoring, we can re-write the terms $w$, $x$, $y$, and $z$ as

$$\begin{aligned} w &= 2(\omega_0^H)^2(1-p_H)^k p_H^{2k(k+1)} (\frac{1-p_H}{p_H})^{\frac{k(k+1)}{2}} \left( \prod_{m=1}^{\infty} (1 + (\frac{1-p_H}{p_H})^m) \right) \\ &= p_H^{\frac{3k(k+1)}{2}} (1-p_L)^{k+\frac{k(k+1)}{2}} \left( 2(\omega_0^H)^2 (\frac{1-p_H}{1-p_L})^{k+\frac{k(k+1)}{2}} \left( \prod_{m=1}^{\infty} (1 + (\frac{1-p_H}{p_H})^m) \right) \right) \\ x &= \omega_0^H(1-\omega_0^H)\left((1-p_H)^k p_L^k + p_H^k(1-p_L)^k\right) p_L^{\frac{k(k+1)}{2}} p_H^{\frac{k(k+1)}{2}} p_L^{k^2} (\frac{1-p_H}{p_H})^{\frac{k(k+1)}{2}} \left( \prod_{m=1}^{\infty} (1 + (\frac{1-p_L}{p_L})^m) \right) \\ &= p_H^{\frac{3k(k+1)}{2}} (1-p_L)^{k+\frac{k(k+1)}{2}} \left( \omega_0^H(1-\omega_0^H) \left( 1 + (\tfrac{p_L(1-p_H)}{p_H(1-p_L)})^k \right) \left( \tfrac{1-p_H}{1-p_L} \right)^{\frac{k(k+1)}{2}} (\tfrac{p_L}{p_H})^{k^2+\frac{k(k+1)}{2}} \left( \prod_{m=1}^{\infty} (1 + (\tfrac{1-p_L}{p_L})^m) \right) \right) \\ y &= \omega_0^H(1-\omega_0^H)\left((1-p_L)^k p_H^k + p_L^k(1-p_H)^k\right) p_H^{\frac{k(k+1)}{2}} p_L^{\frac{k(k+1)}{2}} p_H^{k^2} (\frac{1-p_L}{p_L})^{\frac{k(k+1)}{2}} \left( \prod_{m=1}^{\infty} (1 + (\frac{1-p_H}{p_H})^m) \right) \\ &= p_H^{\frac{3k(k+1)}{2}} (1-p_L)^{k+\frac{k(k+1)}{2}} \left( \omega_0^H(1-\omega_0^H) \left( 1 + \left( \frac{p_L(1-p_H)}{p_H(1-p_L)} \right)^k \right) \left( \prod_{m=1}^{\infty} (1 + (\tfrac{1-p_H}{p_H})^m) \right) \right) \\ z &= 2(1-\omega_0^H)^2(1-p_L)^k p_L^{2k(k+1)} (\frac{1-p_L}{p_L})^{\frac{k(k+1)}{2}} \left( \prod_{m=1}^{\infty} (1 + (\frac{1-p_L}{p_L})^m) \right) \\ &= p_H^{\frac{3k(k+1)}{2}} (1-p_L)^{k+\frac{k(k+1)}{2}} \left( 2(1-\omega_0^H)^2 \left( \frac{p_L}{p_H} \right)^{\frac{3k(k+1)}{2}} \left( \prod_{m=1}^{\infty} (1 + (\frac{1-p_L}{p_L})^m) \right) \right) \end{aligned}$$

Note that the term $p_H^{\frac{3k(k+1)}{2}} (1-p_L)^{k+\frac{k(k+1)}{2}}$ drops out since it is in every term when calculating the joint posteriors. Also, note that a necessary and sufficient condition for $\prod_{m=1}^{\infty}(1+(\frac{1-p_q}{p_q})^m)$ to converge is that $\sum_{m=1}^{\infty}(\frac{1-p_q}{p_q})^m$ is absolutely convergent, which is clearly satisfied when $p_q > \frac{1}{2}$. Thus, when $1 > p_H > p_L > \frac{1}{2}$, $\lim_{k\to\infty} w = 0, \lim_{k\to\infty} x = 0, \lim_{k\to\infty} y = \omega_0^H(1-\omega_0^H)\left(\prod_{m=1}^{\infty}(1+(\frac{1-p_H}{p_H})^m)\right), \lim_{k\to\infty} z = 0$. This implies that when $1 > p_H > p_L > \frac{1}{2}$, $\lim_{k\to\infty} P^b(H_1, H_2|\mathbf{s}^{n_1,n_2}) = 0$, $\lim_{k\to\infty} P^b(L_1, H_2|\mathbf{s}^{n_1,n_2}) = 0$, $\lim_{k\to\infty} P^b(H_1, L_2|\mathbf{s}^{n_1,n_2}) = 1$, $\lim_{k\to\infty} P^b(L_1, L_2|\mathbf{s}^{n_1,n_2}) = 0$.

An extremely similar proof applies to show the result when $1 > p_H > \frac{1}{2} > p_L > 1-p_H$,

where we can instead factor out $1-p_L$ instead of $p_L$, so all the $p_L$ and $1-p_L$ terms are exchanged and the proof applies.

The result also holds if $p_L = \frac{1}{2}$. Letting $k \to \infty$ and $p_L = \frac{1}{2}$ and factoring, we can re-write the terms $w$, $x$, $y$, and $z$ as

$$
\begin{aligned}
w &= 2(\omega_0^H)^2(1-p_H)^k p_H^{2k(k+1)}(\frac{1-p_H}{p_H})^{\frac{k(k+1)}{2}}\left(\prod_{m=1}^{\infty}(1+(\frac{1-p_H}{p_H})^m)\right)\\
&= 2(\omega_0^H)^2(1-p_H)^{k+\frac{k(k+1)}{2}} p_H^{\frac{3k(k+1)}{2}}\left(\prod_{m=1}^{\infty}(1+(\frac{1-p_H}{p_H})^m)\right)\\
&= p_H^{\frac{3k(k+1)}{2}}(1-p_L)^{k+\frac{k(k+1)}{2}}\left(2(\omega_0^H)^2(\frac{1-p_H}{1-p_L})^{k+\frac{k(k+1)}{2}}\left(\prod_{m=1}^{\infty}(1+(\frac{1-p_H}{p_H})^m)\right)\right)\\
x &= \omega_0^H(1-\omega_0^H)(\frac{1}{2})^k\left((1-p_H)^k+p_H^k\right)(\frac{1}{2})^{k^2+\frac{k(k+1)}{2}} p_H^{\frac{k(k+1)}{2}}(2)^k\left(\prod_{m=1}^{\infty}(1+(\frac{1-p_H}{p_H})^m)\right)\\
&= \omega_0^H(1-\omega_0^H)(\frac{1}{2})^{k^2+\frac{k(k+1)}{2}}\left((1-p_H)^k+p_H^k\right) p_H^{\frac{k(k+1)}{2}}\left(\prod_{m=1}^{\infty}(1+(\frac{1-p_H}{p_H})^m)\right)\\
&= p_H^{\frac{3k(k+1)}{2}}(1-p_L)^{k+\frac{k(k+1)}{2}}\left(\omega_0^H(1-\omega_0^H)\left(\frac{2}{(2p_H)^k}\right)^k(1+(\tfrac{1-p_H}{p_H})^k)\left(\textstyle\prod_{m=1}^{\infty}(1+(\frac{1-p_H}{p_H})^m)\right)\right)\\
y &= \omega_0^H(1-\omega_0^H)(\tfrac{1}{2})^{k+\frac{k(k+1)}{2}} p_H^{k^2+\frac{k(k+1)}{2}}(p_H^k+(1-p_H)^k)\left(1+(\tfrac{1-p_H}{p_H})^k\right)^k\left(\textstyle\prod_{m=1}^{\infty}(1+(\frac{1-p_H}{p_H})^m)\right)\\
&= p_H^{\frac{3k(k+1)}{2}}(1-p_L)^{k+\frac{k(k+1)}{2}}\left(\omega_0^H(1-\omega_0^H)(1+(\tfrac{1-p_H}{p_H})^k)(1+(\tfrac{1-p_H}{p_H})^k)^k\left(\textstyle\prod_{m=1}^{\infty}(1+(\frac{1-p_H}{p_H})^m)\right)\right)\\
z &= 2(1-\omega_0^H)^2(\frac{1}{2})^{2k^2+3k}(2)^{2k} = 2(1-\omega_0^H)^2(\frac{1}{2})^{2k^2+k}\\
&= p_H^{\frac{3k(k+1)}{2}}(1-p_L)^{k+\frac{k(k+1)}{2}}\left(2(1-\omega_0^H)^2(\frac{1}{2p_H})^{\frac{3k(k+1)}{2}}(\frac{1}{2})^{\frac{k(k-1)}{2}}\right)
\end{aligned}
$$

The term $p_H^{\frac{3k(k+1)}{2}}(1-p_L)^{k+\frac{k(k+1)}{2}}$ drops out since it is in every term when calculating the joint posteriors. A necessary and sufficient condition for $\prod_{m=1}^{\infty}(1+(\frac{1-p_q}{p_q})^m)$ to converge is that $\sum_{m=1}^{\infty}(\frac{1-p_H}{p_H})^m$ is absolutely convergent, which is clearly satisfied when $p_H > \frac{1}{2}$.

Terms $w$, $x$, and $z$ converge to 0. Term $y$ converges to $\omega_0^H(1-\omega_0^H)\left(\prod_{m=1}^{\infty}(1+(\frac{1-p_H}{p_H})^m)\right)$, which is a finite number, because $\lim_{k\to\infty}(1+(\frac{1-p_H}{p_H})^k)^k = 1$ (re-arranging and using

L'Hopital's rule several times):

$$\begin{aligned}
\lim_{k\to\infty}(1+(\frac{1-p_H}{p_H})^k)^k &= \lim_{k\to\infty}\left(\exp\left(\ln(1+(\frac{1-p_H}{p_H})^k)\right)^k\right) = \exp\lim_{k\to\infty}\left(k\ln(1+(\frac{1-p_H}{p_H})^k)\right)\\
&= \exp\lim_{k\to\infty}\frac{\ln(1+(\frac{1-p_H}{p_H})^k)}{\frac{1}{k}} = \exp\lim_{k\to\infty}\frac{\frac{(\frac{1-p_H}{p_H})^k\ln(\frac{1-p_H}{p_H})}{1+(\frac{1-p_H}{p_H})^k}}{-(\frac{1}{k})^2}\\
&= \exp\lim_{k\to\infty}(\ln(\frac{1-p_H}{p_H}))\left(\frac{-k^2}{\frac{1+(\frac{1-p_H}{p_H})^k}{(\frac{1-p_H}{p_H})^k}}\right) = \exp\lim_{k\to\infty}(\ln(\frac{1-p_H}{p_H}))\left(\frac{-2k}{-\frac{\ln(\frac{1-p_H}{p_H})}{(\frac{1-p_H}{p_H})^k}}\right)\\
&= \exp\lim_{k\to\infty}2\left(\frac{k}{\frac{1}{(\frac{1-p_H}{p_H})^k}}\right) = \exp\lim_{k\to\infty}2\left(\frac{1}{-\frac{\ln(\frac{1-p_H}{p_H})}{(\frac{1-p_H}{p_H})^k}}\right) = \exp\lim_{k\to\infty}2\left(\frac{(\frac{1-p_H}{p_H})^k}{-\ln(\frac{1-p_H}{p_H})}\right)\\
&= \exp(0) = 1.
\end{aligned}$$

This implies that when $1 > p_H > p_L = \frac{1}{2}$, $\lim_{k\to\infty} P^b(H_1, H_2|\mathbf{s}^{n_1,n_2}) = 0$, $\lim_{k\to\infty} P^b(L_1, H_2|\mathbf{s}^{n_1,n_2}) = 0$, $\lim_{k\to\infty} P^b(H_1, L_2|\mathbf{s}^{n_1,n_2}) = 1$, and $\lim_{k\to\infty} P^b(L_1, L_2|\mathbf{s}^{n_1,n_2}) = 0$. An extremely similar proof applies to show that $\lim_{k\to\infty} P^b(\theta = A|\mathbf{s}^{n_1,n_2}) = 1$ where $n_1 = n_2 = k$ when $1 > p_H > p_L = \frac{1}{2}$.

## A.10 Proof of Proposition 8

Let $s_{itj}$ be the $i$th signal, observed in period $t$, sent by expert $j \in \{1, 2\}$. As before, expert 1 reports first, and expert $j$ sends $n_j$ signals in total. For example, if the agent observes one signal per period from expert 1, followed by one signal per period from expert 2, then $\mathbf{s}^{n_1} = (s_{111}, s_{221}, \ldots s_{n_1,n_1,1})$ and $\mathbf{s}^{n_2} = (s_{n_1+1,n_1+1,2}, s_{n_1+2,n_1+2,2}, \ldots s_{n_1+n_2,n_1+n_2,2})$. If expert 1 sends all of her $n_1$ signals in period 1, then $\mathbf{s}^{n_1} = (s_{111}, s_{211}, \ldots s_{n_1,1,1})$. Since the reliability of a signal $i$ from expert $j$ is independent of the period in which it is observed and the other expert's quality, note that $P(s_{itj}|q_j, q_k, \theta) = P(s_{ij}|q_j, \theta)$ for all $t$ where $j \neq k$.

The pre-screener's updating after observing generalize to Equations (8) and (9), where the functions $\beta_{q_1}(\mathbf{s}^{n_1})$ and $\beta_{q_1q_2}(\mathbf{s}^{n_1,n_2})$ reflect the path dependency of the biased agent's beliefs and will differ depending on both timing and order of signals.

1. If the pre-screener receives $n$ signals simultaneously (say, from expert 1 in period 1), then her posterior after observing $\mathbf{s}^{n_1} = (s_{111}, s_{211}, \ldots s_{n_1,1,1})$ all together is still described by Equation 4, but her $b^{q_1}(\mathbf{s}^n)$ is instead given by

$$\beta_{q_1}(\mathbf{s}^n) = \left(\sum_\theta P(s_1|q,\theta)P(s_2|q,\theta)\ldots P(s_n|q,\theta)\omega_0^\theta\right) = \sum_\theta\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^\theta. \qquad (24)$$

Let $x$ be the event in which expert 1 sends $k$ $a$'s simultaneously: $\mathbf{s}_x^{n_1} = (s_{111}, s_{211}, \ldots s_{n_1,1,1})$. Let $y$ be the event in which expert 1 sends $k$ $a$'s sequentially $\mathbf{s}_y^{n_1} = (s_{111}, s_{221}, \ldots s_{n_1,n_1,1})$. Then $\beta_{q_1}^x = (\frac{1}{2})(p_{q_1}^k + (1-p_{q_1})^k)$ and $\beta_{q_1}^y = \prod_{i=1}^k (\frac{1}{2})(p_{q_1}^i + (1-p_{q_1})^i)$. By direct comparison of the posteriors, $P^b(q_1 = H|\mathbf{s}_x^{n_1}) < P^b(q_1 = H|\mathbf{s}_y^{n_1})$ only if $\beta_H^x \beta_L^y - \beta_H^y \beta_L^x < 0$, which is satisfied:

$$\begin{aligned}\beta_H^x \beta_L^y - \beta_H^y \beta_L^x &= (\tfrac{1}{2})(p_H^k + (1-p_H)^k)\left(\textstyle\prod_{i=1}^k (\frac{1}{2})(p_L^i + (1-p_L)^i)\right) - (\tfrac{1}{2})(p_L^k + (1-p_L)^k)\left(\textstyle\prod_{i=1}^k (\frac{1}{2})(p_H^i + (1-p_H)^i)\right) \\ &= (\tfrac{1}{2})^{k+1}(p_L^k + (1-p_L)^k)(p_H^k + (1-p_H)^k)\left(\left(\textstyle\prod_{i=1}^{k-1}(p_L^i + (1-p_L)^i)\right) - \left(\textstyle\prod_{i=1}^{k-1}(p_H^i + (1-p_H)^i)\right)\right),\end{aligned}$$

since $\frac{\partial}{\partial p_q}(p_q^k + (1-p_q)^k) < 0$ for $k > 1$ and the relevant parameter restrictions on $p_L$ and $p_H$. Thus, if expert 1 sends $k$ identical signals, the pre-screener with $\omega_0^A = 1/2$ trusts him more when they are sent sequentially than simultaneously (though there is still overtrust in both cases, which is straightforward to show given $\beta_{q_1}^x$ and $\beta_{q_1}^y$).

2. First, we show that the pre-screener still believes that state $A$ is more likely than $B$. Note that Equation (22) must be satisfied for this to be true, whether either expert sends signals simultaneously or sequentially. What differs based on simultaneous versus sequential signals is the terms $\beta_{q_1}$ and $\beta_{q_2 q_1}$. Let $W$ be the event in which expert 1 sends $k$ $a$'s simultaneously and expert 2 sends $k$ $b$'s simultaneously: $\mathbf{s}_W^{n_1} = (s_{111}, s_{211}, \ldots s_{n_1,1,1})$, $\mathbf{s}_W^{n_2} = (s_{n_1+1,2,2}, s_{n_1+2,2,2}, \ldots s_{n_1+n_2,2,2})$. Let $X$ be the event in which expert 1 sends $k$ $a$'s simultaneously and expert 2 sends $k$ $b$'s sequentially: $\mathbf{s}_X^{n_1} = (s_{111}, s_{211}, \ldots s_{n_1,1,1})$, $\mathbf{s}_X^{n_2} = (s_{n_1+1,n_1+1,2}, s_{n_1+2,n_1+2,2}, \ldots s_{n_1+n_2,n_1+n_2,2})$. Let $Y$ be the event in which expert 1 sends $k$ $a$'s sequentially and expert 2 sends $k$ $b$'s simultaneously: $\mathbf{s}_Y^{n_1} = (s_{111}, s_{221}, \ldots s_{n_1,n_1,1})$, $\mathbf{s}_Y^{n_2} = (s_{n_1+1,n_1+1,2}, s_{n_1+2,n_1+1,2}, \ldots s_{n_1+n_2,n_1+1,2})$. Let $Z$ be the event in which expert 1 sends $k$ $a$'s sequentially and expert 2 sends $k$ $b$'s sequentially: $\mathbf{s}_Z^{n_1} = (s_{111}, s_{221}, \ldots s_{n_1,n_1,1})$ and $\mathbf{s}_Z^{n_2} = (s_{n_1+1,n_1+1,2}, s_{n_2+1,n_2+1,2}, \ldots s_{n_1+n_2,n_1+n_2,2})$. Note that in each of these events, expert 1's signals are sent strictly before expert 2's signals.

   When $n_1 = n_2 = k$ where expert 1's signals are all $a$'s and expert 2's signals are all $b$'s, then $\beta_{q_1}^W = \beta_{q_1}^X = (\frac{1}{2})(p_{q_1}^k + (1-p_{q_1})^k)$, $\beta_{q_1}^Y = \beta_{q_1}^Z = \prod_{i=1}^k (\frac{1}{2})(p_{q_1}^i + (1-p_{q_1})^i)$, $\beta_{q_2 q_1}^W = \beta_{q_2 q_1}^Y = (\frac{1}{2})\left((1-p_{q_2})^k p_{q_1}^k + p_{q_2}^k (1-p_{q_1})^k\right)$, and $\beta_{q_2 q_1}^X = \beta_{q_2 q_1}^Z = \prod_{i=1}^k (\frac{1}{2})\left((1-p_{q_2})^i p_{q_1}^k + p_{q_2}^i (1-p_{q_1})^k\right)$. Also, note that $\beta_H^E > \beta_L^E$ for all $k > 1$ and $E \in \{W, X, Y, Z\}$. We have already shown previously that $\beta_{L_2 H_1}^X > \beta_{H_2 L_1}^X$ for $k > 1$, and obviously $\beta_{L_2 H_1}^W = \beta_{H_2 L_1}^W$. Using these properties in Equation (22), we can verify that $P_E^b(\theta = A|\mathbf{s}^{n_1,n_2}) > \frac{1}{2}$ when $\omega_0^\theta = \frac{1}{2}$, $k > 1$, and $E \in \{W, X, Y, Z\}$.

   To show that sending the $k$ $b$ signals simultaneously rather than sequentially gives more credibility to expert 2, it is sufficient to show that $P_X^b(\theta = A|\mathbf{s}^{n_1,n_2}) > P_Y^b(\theta = A|\mathbf{s}^{n_1,n_2})$

and $P_W^b(\theta = A|\mathbf{s}^{n_1,n_2}) > P_X^b(\theta = A|\mathbf{s}^{n_1,n_2})$.

$P_Y^b(\theta = A|\mathbf{s}^{n_1,n_2}) > P_Z^b(\theta = A|\mathbf{s}^{n_1,n_2})$ is satisfied only if

$$\begin{aligned}\omega_0^H(1-\omega_0^H)[p_H^k(1-p_L)^k - p_L^k(1-p_H)^k]\left((\omega_0^H)^2 p_H^k(1-p_H)^k(b_H^Y)^2\left(b_{H_2H_1}^Y(b_{L_2H_1}^Z - b_{H_2L_1}^Z) - b_{H_2H_1}^Z(b_{L_2H_1}^Y - b_{H_2L_1}^Y)\right)\right.\\ \left.+(1-\omega_0^H)^2 p_L^k(1-p_L)^k(b_L^Y)^2\left(b_{L_2L_1}^Y(b_{L_2H_1}^Z - b_{H_2L_1}^Z) - b_{L_2L_1}^Z(b_{L_2H_1}^Y - b_{H_2L_1}^Y)\right)\right)\\ +(\omega_0^H)^2(1-\omega_0^H)^2[p_H^{2k}(1-p_L)^{2k} - p_L^{2k}(1-p_H)^{2k}]b_H^Y b_L^Y\left(b_{L_2H_1}^Z b_{H_2L_1}^Y - b_{L_2H_1}^Y b_{H_2L_1}^Z\right) > 0.\end{aligned}$$

Note that $\beta_{L_2H_1}^Y = \beta_{H_2L_1}^Y$ and $\beta_{L_2H_1}^X > \beta_{H_2L_1}^X$ for $k > 1$, so the third term is positive. For the first and second terms, $\beta_{L_2H_1}^X > \beta_{H_2L_1}^X$, and $\beta_{L_2H_1}^Y > \beta_{H_2L_1}^Y$, so it is sufficient to show that $\beta_{qq}^Y - \beta_{qq}^X$ for them to each be positive:

$$\begin{aligned}\beta_{qq}^Y - \beta_{qq}^X &= p_q^k(1-p_q)^k - \prod_{i=1}^{k}\frac{1}{2}\left((1-p_q)^i p_q^k + p_q^i(1-p_q)^k\right)\\ &= p_q^k(1-p_q)^k\left(1 - \prod_{i=1}^{k-1}\frac{1}{2}\left((1-p_q)^i p_q^k + p_q^i(1-p_q)^k\right)\right),\end{aligned}$$

where each term of $(1-p_q)^i p_q^k + p_q^i(1-p_q)^k$ is bounded above by $\frac{1}{2}$ for $k > 1$. Thus, $\beta_{qq}^Y - \beta_{qq}^X > 0$ for $k > 1$ so the first and second terms are positive when $k > 1$.

The same argument applies for $P_W^b(\theta = A|\mathbf{s}^{n_1,n_2}) > P_X^b(\theta = A|\mathbf{s}^{n_1,n_2})$. Thus, sending the $k$ $b$ signals simultaneously rather than sequentially gives more credibility to expert 2, given expert 1's signals.

## A.11 Proof of Proposition 9

1. Let $\mathbf{s}^n$ be a sequence of $n$ observed signals with $n_a$ $a$'s and $n_b$ $b$'s, let $s_{n+1}$ be the $(n+1)$th observed signal, and let $\omega_n^b$ equal the pre-screener's joint posterior after the sequence $\mathbf{s}^n$.

   First, note that each joint belief on the state and quality for the prior $\omega_n^b$, denoted $\omega_n^{q\theta}$, is given by

$$\omega_n^{q\theta} \equiv P^b(q,\theta|\{\mathbf{s}^n\}) = \frac{\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^\theta\omega_0^q\beta_q(\mathbf{s}^n)}{\sum_\theta\sum_q\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^\theta\omega_0^q\beta_q(\mathbf{s}^n)}, \tag{25}$$

   where

$$\beta_q(\mathbf{s}^n) = \prod_{m=1}^{n}\left(\sum_\theta\left(\prod_{t=1}^{m} P(s_t|q,\theta)\right)\omega_0^\theta\right). \tag{26}$$

   Thus, the Bayesian's posterior belief given the biased prior is

$$P^u(\theta = A|\text{prior} = \omega_n^b, \{s_{n+1}\}) = \frac{\omega_0^A\sum_q\left(\prod_{t=1}^{n+1} P(s_t|q,A)\right)\omega_0^q\beta_q(\mathbf{s}^n)}{\sum_\omega\omega_0^\theta\sum_q P(s_{n+1}|q,\theta)\left(\prod_{t=1}^n P(s_t|q,\theta)\right)\omega_0^q\beta_q(\mathbf{s}^n)}.$$

In contrast, the pre-screener's posterior belief after observing $\{\mathbf{s}^n, s_{n+1}\}$ is

$$P^b(\theta = A|\{\mathbf{s}^n, s_{n+1}\}) = \frac{\omega_0^A \sum_q \left(\prod_{t=1}^{n+1} P(s_t|q, A)\right) \omega_0^q \beta_q(\{\mathbf{s}^n, s_{n+1}\})}{\sum_\theta \omega_0^\theta \sum_q \left(\prod_{t=1}^{n+1} P(s_t|q, \theta)\right) \omega_0^q \beta_q(\{\mathbf{s}^n, s_{n+1}\})},$$

where

$$\beta_q(\{\mathbf{s}^n, s_{n+1}\}) = \prod_{m=1}^{n+1} \left(\sum_\theta \left(\prod_{t=1}^{m} P(s_t|q, \theta)\right) \omega_0^\theta\right) = \beta_q(\mathbf{s}^n) \left(\sum_\theta \left(\prod_{t=1}^{n+1} P(s_t|q, \theta)\right) \omega_0^\theta\right). \quad (27)$$

Substituting all of the preceding information into $P^b(\omega = A|\{\mathbf{s}^n, s_{n+1}\}) > P^u(\omega = A|prior = \omega_n^b, \{s_{n+1}\})$, the inequality is only satisfied if

$$\omega_0^A(1-\omega_0^A)\omega_0^H(1-\omega_0^H)\beta_L(\mathbf{s}^n)\beta_H(\mathbf{s}^n) \underbrace{\left(\left(\sum_\theta \left(\prod_{t=1}^{n+1} P(s_t|H, \theta)\right) \omega_0^\theta\right) - \left(\sum_\theta \left(\prod_{t=1}^{n+1} P(s_t|L, \theta)\right) \omega_0^\theta\right)\right)}_{X}$$
$$\underbrace{\left(\left(\prod_{t=1}^{n+1} P(s_t|H, A)\right)\left(\prod_{t=1}^{n+1} P(s_t|L, B)\right) - \left(\prod_{t=1}^{n+1} P(s_t|H, B)\right)\left(\prod_{t=1}^{n+1} P(s_t|L, A)\right)\right)}_{Y} > 0, \quad (28)$$

Without loss of generality, suppose that $n_a \geq n_b$.

If $s_{n+1} = a$, then $\{\mathbf{s}^n, s_{n+1}\}$ has $n_a + 1$ $a$'s and $n_b$ $b$'s. Then the term $Y$ is given by

$$\begin{aligned} &p_H^{n_a+1}(1-p_H)^{n_b}(1-p_L)^{n_a+1}(p_L)^{n_b} - (1-p_H)^{n_a+1}(p_H)^{n_b}(p_L)^{n_a+1}(1-p_L)^{n_b} \\ &= [p_H p_L(1-p_H)(1-p_L)]^{n_b}[(p_H(1-p_L))^{n_a-n_b+1} - (p_L(1-p_H))^{n_a-n_b+1}] \end{aligned}$$

Thus if $s_{n+1} = a$, then

$$Y(s_{n+1} = a) \begin{cases} > 0 & \text{if } n_a + 1 > n_b \\ = 0 & \text{if } n_a + 1 = n_b \\ < 0 & \text{if } n_a + 1 < n_b. \end{cases}$$

If $s_{n+1} = b$, then $\{\mathbf{s}^n, s_{n+1}\}$ has $n_a$ $a$'s and $n_b + 1$ $b$'s. Then the term $Y$ is given by

$$\begin{aligned} &p_H^{n_a}(1-p_H)^{n_b+1}(1-p_L)^{n_a}(p_L)^{n_b+1} - (1-p_H)^{n_a}(p_H)^{n_b+1}(p_L)^{n_a}(1-p_L)^{n_b+1} \\ &= [p_H p_L(1-p_H)(1-p_L)]^{n_b}[(p_H(1-p_L))^{n_a-n_b}p_L(1-p_H) - ((1-p_H)p_L)^{n_a-n_b}p_H(1-p_L)] \end{aligned}$$

Thus if $s_{n+1} = b$, then

$$Y(s_{n+1} = b) \begin{cases} > 0 & \text{if } n_a > n_b + 1 \\ = 0 & \text{if } n_a = n_b + 1 \\ < 0 & \text{if } n_a < n_b + 1. \end{cases}$$

Thus, $Y$ is positive if $\{\mathbf{s}^n, s_{n+1}\}$ has more $a$'s than $b$'s, $Y$ is negative if $\{\mathbf{s}^n, s_{n+1}\}$ has more $b$'s than $a$'s, and $Y$ is zero if $\{\mathbf{s}^n, s_{n+1}\}$ has an equal number of $a$'s and $b$'s.

Moreover, note that $\kappa_{\mathbf{s}^n} \equiv \frac{\beta_q(\mathbf{s}^n)\omega_0^q}{\sum_q \beta_q(\mathbf{s}^n)\omega_0^q}$. Then Equation (27) implies that $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) > \kappa_H(\mathbf{s}^n)$ if and only if
$\omega_0^H(1-\omega_0^H)\beta_H(\mathbf{s}^n)\beta_L(\mathbf{s}^n)\left(\left(\sum_\theta \left(\prod_{t=1}^{n+1} P(s_t|H,\theta)\right)\omega_0^\theta\right) - \left(\sum_\theta \left(\prod_{t=1}^{n+1} P(s_t|L,\theta)\right)\omega_0^\theta\right)\right) > 0$, which is the requirement that $X > 0$.

In other words,

$$X \begin{cases} > 0 & \text{if and only if } \kappa_H(\{\mathbf{s}^n, s_{n+1}\}) > \kappa_H(\mathbf{s}^n) \\ = 0 & \text{if and only if } \kappa_H(\{\mathbf{s}^n, s_{n+1}\}) = \kappa_H(\mathbf{s}^n) \\ < 0 & \text{if and only if } \kappa_H(\{\mathbf{s}^n, s_{n+1}\}) < \kappa_H(\mathbf{s}^n). \end{cases}$$

From above, we can see that the sign of Equation (28) depends on the sign of $XY$. Putting everything together, then

- $P^b(\theta = A|\{\mathbf{s}^n, s_{n+1}\}) = P^u(\theta = A|prior = \omega_n^b, \{s_{n+1}\})$ if either (1) $\{\mathbf{s}^n, s_{n+1}\}$ has an equal number of $a$'s and $b$'s or (2) $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) = \kappa_H(\mathbf{s}^n)$
- $P^b(\theta = A|\{\mathbf{s}^n, s_{n+1}\}) > P^u(\theta = A|prior = \omega_n^b, \{s_{n+1}\})$ if (3) $\{\mathbf{s}^n, s_{n+1}\}$ has more $a$'s than $b$'s and $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) > \kappa_H(\mathbf{s}^n)$ or (4) $\{\mathbf{s}^n, s_{n+1}\}$ has more $b$'s than $a$'s and $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) < \kappa_H(\mathbf{s}^n)$
- $P^b(\theta = A|\{\mathbf{s}^n, s_{n+1}\}) < P^u(\theta = A|prior = \omega_n^b, \{s_{n+1}\})$ if (5) $\{\mathbf{s}^n, s_{n+1}\}$ has more $a$'s than $b$'s and $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) < \kappa_H(\mathbf{s}^n)$ or (6) $\{\mathbf{s}^n, s_{n+1}\}$ has more $b$'s than $a$'s and $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) > \kappa_H(\mathbf{s}^n)$

Note that the statement $P^b(\theta = A|\{\mathbf{s}^n, s_{n+1}\}) > P^u(\theta = A|prior = \omega_n^b, \{s_{n+1}\})$ if $\{\mathbf{s}^n, s_{n+1}\}$ has more $a$'s than $b$'s and $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) > \kappa_H(\mathbf{s}^n)$ is equivalent to the statement $P^b(\theta = A|\{\mathbf{s}^n, s_{n+1}\}) < P^u(\theta = A|prior = \omega_n^b, \{s_{n+1}\})$ if $\{\mathbf{s}^n, s_{n+1}\}$ has more $b$'s than $a$'s and $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) > \kappa_H(\mathbf{s}^n)$. Likewise, the statement $P^b(\theta = A|\{\mathbf{s}^n, s_{n+1}\}) > P^u(\theta = A|prior = \omega_n^b, \{s_{n+1}\})$ if $\{\mathbf{s}^n, s_{n+1}\}$ has more $b$'s than $a$'s and $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) < \kappa_H(\mathbf{s}^n)$ is equivalent to the statement $P^b(\theta = A|\{\mathbf{s}^n, s_{n+1}\}) < P^u(\theta =$

$A|prior = \omega_n^b, \{s_{n+1}\})$ if $\{\mathbf{s}^n, s_{n+1}\}$ has more $a$'s than $b$'s and $\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) < \kappa_H(\mathbf{s}^n)$. Therefore, we can state the proposition assuming that the number of $a$'s be greater than or equal to the number of $b$'s in $\{\mathbf{s}^n, s_{n+1}\}$ without loss of generality.

Moreover, note that $Pr^u(q = H|\{\mathbf{s}^n, s_{n+1}\}) = \frac{\omega_0^H \sum_\theta \left(\prod_{t=1}^{n+1} P(s_t|q,\theta)\right)\omega_0^\theta}{\sum_q \omega_0^q \sum_\theta \left(\prod_{t=1}^{n+1} P(s_t|q,\theta)\right)\omega_0^\theta}$. From this definition, we know that $Pr^u(q = H|\{\mathbf{s}^n, s_{n+1}\}) > \omega_0^H$ if and only if $\left(\left(\sum_\theta \left(\prod_{t=1}^{n+1} P(s_t|H,\theta)\right)\omega_0^\theta\right) - \left(\sum_\theta \left(\prod_{t=1}^{n+1} P(s_t|L,\theta)\right)\omega_0^\theta\right)\right) > 0$, which is the requirement that $X > 0$. Thus,

$$\begin{aligned}
&\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) > \kappa_H(\mathbf{s}^n) \text{ if and only if } Pr^u(q = H|\{\mathbf{s}^n, s_{n+1}\}) > \omega_0^H \\
&\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) = \kappa_H(\mathbf{s}^n) \text{ if and only if } Pr^u(q = H|\{\mathbf{s}^n, s_{n+1}\}) = \omega_0^H \\
&\kappa_H(\{\mathbf{s}^n, s_{n+1}\}) < \kappa_H(\mathbf{s}^n) \text{ if and only if } \ Pr^u(q = H|\{\mathbf{s}^n, s_{n+1}\}) < \omega_0^H.
\end{aligned}$$

2. First, note that $P^b[q, \theta|\{\mathbf{s}^n, s_{n+1}\}]$ is equal to

$$P^b[q, \theta|\{\mathbf{s}^n, s_{n+1}\}] = \frac{\beta_q(\{\mathbf{s}^n, s_{n+1}\}) \left(\prod_{t=1}^{n+1} P(s_t|q,\theta)\right)\omega_0^\theta \omega_0^q}{\sum_q \beta_q(\{\mathbf{s}^n, s_{n+1}\}) \sum_\theta \left(\prod_{t=1}^{n+1} P(s_t|q,\theta)\right)\omega_0^\theta \omega_0^q} \tag{29}$$

where $\beta_q(\{\mathbf{s}^n, s_{n+1}\})$ is described by Equation (27). Second, applying the generalized pre-screening described in A.1, $P^b[q, \theta|\text{prior} = \omega_n^b, \{s_{n+1}\}]$ is equal to

$$P^b[q, \theta|\text{prior} = \omega_n^b, \{s_{n+1}\}] = \frac{\beta_{q\theta}(s_{n+1}) \left(\frac{1}{\sum_\theta \omega_n^{q\theta}}\right) P(s_{t+1}|q,\theta)\omega_n^{q\theta}}{\sum_q \sum_\theta \beta_{q\theta}(s_{n+1}) \left(\frac{1}{\sum_\theta \omega_n^{q\theta}}\right) P(s_{t+1}|q,\theta)\omega_n^{q\theta}}, \tag{30}$$

where $\beta_{q\theta}(s_{n+1}) = \sum_\theta P(s_{n+1}|q,\theta)\omega_n^{q\theta}$ and $\omega_n^{q\theta}$ is described by Equation (25) and $\beta_q(\mathbf{s}^n)$ is described by Equation (26). Substituting this into $P^b[q, \theta|\text{prior} = \omega_n^b, \{s_{n+1}\}]$ yields:

$$\begin{aligned}
P^b[q, \theta|\text{prior} = \omega_n^b, \{s_{n+1}\}] &= \frac{\beta_{q\theta}(s_{n+1})\left(\frac{1}{\sum_\theta \omega_n^{q\theta}}\right) P(s_{t+1}|q,\theta)\beta_q(\mathbf{s}^n)\left(\prod_{t=1}^{n} P(s_t|q,\theta)\right)\omega_0^\theta \omega_0^q}{\sum_q \sum_\theta \beta_{q\theta}(s_{n+1})\left(\frac{1}{\sum_\theta \omega_n^{q\theta}}\right) P(s_{t+1}|q,\theta)\beta_q(\mathbf{s}^n)\left(\prod_{t=1}^{n} P(s_t|q,\theta)\right)\omega_0^\theta \omega_0^q} \\
&= \frac{\beta_{q\theta}(s_{n+1}) \left(\frac{1}{\sum_\theta \omega_n^{q\theta}}\right) \beta_q(\mathbf{s}^n) \left(\prod_{t=1}^{n+1} P(s_t|q,\theta)\right)\omega_0^\theta \omega_0^q}{\sum_q \sum_\theta \beta_{q\theta}(s_{n+1}) \left(\frac{1}{\sum_\theta \omega_n^{q\theta}}\right) \beta_q(\mathbf{s}^n) \left(\prod_{t=1}^{n+1} P(s_t|q,\theta)\right)\omega_0^\theta \omega_0^q},
\end{aligned} \tag{31}$$

where

$$\beta_{q\theta}(s_{n+1})\left(\frac{1}{\sum_\theta \omega_n^{q\theta}}\right) = \sum_\theta \left(\prod_{i=1}^{n+1} P(s_t|q,\theta)\right)\beta_q(\mathbf{s}^n)\omega_0^\theta \left(\frac{\omega_0^q}{\sum_\theta \omega_n^{q\theta}}\right)$$
$$= \beta_q(\mathbf{s}^n)\left(\sum_\theta \left(\prod_{t=1}^{n+1} P(s_t|q,\theta)\right)\omega_0^\theta\right)\left(\frac{\omega_0^q}{\sum_\theta \omega_n^{q\theta}}\right).$$

Equation (27) implies that Equation (31) equals Equation (29) if and only if $\omega_0^q = \sum_\theta \omega_n^{q\theta}$. Since $\omega_n^{q\theta} \equiv P^b(q,\theta|\{\mathbf{s}^n\})$, then $P^b[q,\theta|\{\mathbf{s}^n, s_{n+1}\}] \neq P^b[q,\theta|\text{prior} = \omega_n^b, \{s_{n+1}\}]$ if $P^b[q|\mathbf{s}^n] \neq \omega_0^q$ and $P^b[q,\theta|\{\mathbf{s}^n, s_{n+1}\}] = P^b[q,\theta|\text{prior} = \omega_n^b, \{s_{n+1}\}]$ if $P^b[q|\mathbf{s}^n] = \omega_0^q$.

## A.12 Proof of Proposition 10

1. **Lemma 3** *Suppose the agent observes $n_a = n_b$ signals of* a*'s and* b*'s in alternating order: $\mathbf{s}^n = (a, b, \ldots a, b)$ where $n_a = n_b = k$. Then the biased agent is always underconfident that the agent is high quality.*

   **Proof.** An alternating sequence of $n_a = n_b = k$ signals of $a$'s and $b$'s generates:

$$\beta_q(\mathbf{s}^n) = \left(\frac{1}{2}\right)^k \left(\prod_{i=1}^{k-1}(p_q(1-p_q))^{2i}\right)(p_q(1-p_q))^k = \left(\frac{1}{2}\right)^k (p_q(1-p_q))^{k^2}.$$

   This implies that $\frac{\partial}{\partial p_q}(\beta_q(\mathbf{s}^n)) < 0$ for all $p_q$:

$$\frac{\partial}{\partial p_q}(\beta_q(\mathbf{s}^n)) = \left(\frac{1}{2}\right)^k k^2 (p_q(1-p_q))^{k^2-1}(1-2p_q),$$

   Since $(p_H(1-p_H))^{k^2} < (p_L(1-p_L))^{k^2}$ whenever $p_H > p_L \geq \frac{1}{2}$ or $\frac{1}{2} \geq p_L > 1-p_H$, then $\beta_H(\mathbf{s}^n) < \beta_L(\mathbf{s}^n)$ for all $p_H > p_L \geq \frac{1}{2}$, which implies that the biased agent's belief that the expert is high quality is underconfident relative to the Bayesian: $Pr^b(H|\mathbf{s}^n) < Pr^u(H|\mathbf{s}^n)$. ■

   Suppose the agent observes $n_a > n_b$ signals, where $n_b$ $a$'s and $n_b$ $b$'s alternate followed by the remaining $m \equiv n_a - n_b$ $a$'s where $m \geq 1$: $\mathbf{s}^n = (a, b, a, b, \ldots, a, a, a)$. Since $\beta_q$ is symmetric about $p_q = 1/2$, we focus on the case of $p_q \geq 1/2$ but the conclusion extends to $1/2 > p_L > 1 - p_H$. From Equations (18) and (19), we can see that $\frac{\partial}{\partial p_q}(\beta_q(\mathbf{s}^n)) = 0$ when $p_q \in \{\frac{1}{2}, 1\}$, $\beta_q(\mathbf{s}^n) > 0$ when $p_q = \frac{1}{2}$, and $\beta_q(\mathbf{s}^n) = 0$ when $p_q = 1$. Moreover, using the fact that $\beta_q(\mathbf{s}^n) = 0$ when $p_q = 1/2$, then

$$\frac{\partial^2 \beta_q(\mathbf{s}^n)}{\partial p_q^2}\Big|_{p_q=\frac{1}{2}} = \beta_q(\mathbf{s}^n)\left(-8n_b(n_b+m) + \sum_{i=1}^m 4i(i-1)\right) = \beta_q(\mathbf{s}^n)\left(-8n_b(n_b+m) + \tfrac{4}{3}m(m-1)(m+1)\right).$$

Thus for any given $n_b$, there exists some threshold $\frac{1}{2} < \check{p} < 1$ whenever $m > m^*$, where $-8n_b(n_b + m^*) + \frac{4}{3}m^*(m^* - 1)(m^* + 1) = 0$. Let $n_a^* = n_b + m^*$. Then for $n_a, n_b$ where $0 \geq n_b < n_a^* < n_a$ and $p_L < p_H \leq \check{p}$, the agent overtrusts and is optimistic that the state is A. Since this is the sequence that generates the least trust by Lemma 1, then if it results in overtrust then all other sequences of such a combination must generate overtrust and optimism as well.

2. **Lemma 4** *After observing $n_a > 1$ and $n_b = 0$ signals in sequence or simultaneously, the pre-screener overtrusts and is overoptimistic about the reported state.*

**Proof.** Without loss of generality, suppose the sequence is $n_a$ $a$'s: $\mathbf{s}^n = (a, a, \ldots, a)$ where $n_a = n$ and $n_b = 0$. Then $\beta_q(\mathbf{s}^n) = \prod_{i=1}^{n_a}(\frac{1}{2})(p_q^i + (1-p_q)^i)$. Considering each $i$th component of $\beta_q(\mathbf{s}^n)$, $p_H^i + (1-p_H)^i > p_L^i + (1-p_L)^i$ is positive for $i > 0$ when $p_H > p_L \geq \frac{1}{2}$ or when $\frac{1}{2} \geq p_L > 1 - p_H$, which implies that $\beta_H(s_{11} = a, s_{22} = a, \ldots s_{n_a,n_a} = a) > \beta_L(s_{11} = a, s_{22} = a, \ldots s_{n_a,n_a} = a)$. Thus, applying Proposition 2, the pre-screener overtrusts and is overoptimistic about the reported state when she observes $n_a > 1$ and $n_b = 0$ signals in sequence. Since the simultaneous case implies $\beta_q(\mathbf{s}^n) = p_q^{n_a} + (1-p_q)^{n_a}$, then this argument also shows the result when the biased agent observes $n_a > 1$ signals simultaneously. ■

**Lemma 5** *Consider a sequence of signals such that the first $k$ observed signals are $a$, followed by $k$ $b$ signals: $\mathbf{s}^n = (a, a, \ldots, a, b, b, \ldots, b)$ where $n_a = n_b = k$. There exists some $\underline{p} > \frac{1}{2}$ and $\overline{p} < 1$ such that the pre-screener under-trusts if (1) $k \in \{1, 2, 3\}$, (2) if $\overline{p} \leq p_L < p_H$, or (3) if $p_L \leq \underline{p}$ and $p_H \geq \overline{p}$.*

**Proof.** WLOG, suppose the sequence is $n_a$ $a$'s, then $n_b$ $b$'s. Then

$$\beta_q(\mathbf{s}^n) = \left(\prod_{i=1}^{n_a}(\frac{1}{2})(p_q^i + (1-p_q)^i)\right)\left(\prod_{i=1}^{n_b}(\frac{1}{2})(p_q^{n_a}(1-p_q)^i + p_q^i(1-p_q)^{n_a})\right)$$

In particular, if the sequence is $n_a = n_b = k$, then:

$$\beta_q(\mathbf{s}^n) = (\frac{1}{2})^{2k}\prod_{i=1}^{k}\left(p_q^i + (1-p_q)^i\right)\left(p_q^k(1-p_q)^i + p_q^i(1-p_q)^k\right) \tag{32}$$

Characterizing $\frac{\partial}{\partial p_q}(\beta_q(\mathbf{s}^n))$ when $n_a = n_b = k$ yields

$$\frac{\partial}{\partial p_q}(\beta_q(\mathbf{s}^n)) = (\frac{1}{2})^{2k}\left(\prod_{i=1}^{k}\left(p_q^i + (1-p_q)^i\right)\left(p_q^k(1-p_q)^i + p_q^i(1-p_q)^k\right)\right)$$
$$\left(\sum_{i=1}^{k}\frac{i(p_q^{i-1} - (1-p_q)^{i-1})(p_q^k(1-p_q)^i + p_q^i(1-p_q)^k) + \left(p_q^i + (1-p_q)^i\right)\left(k\left(p_q^{k-1}(1-p_q)^i - p_q^i(1-p_q)^{k-1}\right) + i\left(p_q^{i-1}(1-p_q)^k - p_q^k(1-p_q)^{i-1}\right)\right)}{\left(p_q^i + (1-p_q)^i\right)\left(p_q^k(1-p_q)^i + p_q^i(1-p_q)^k\right)}\right) \tag{33}$$

Since $\beta_q$ is symmetric about $p_q = 1/2$, we focus on the case of $p_q \geq 1/2$ but the conclusion extends to $1/2 > p_L > 1 - p_H$. From Equation (32) we can see that $\frac{\partial}{\partial p_q}(\beta_q(\mathbf{s}^n)) = 0$ when $p_q \in \{\frac{1}{2}, 1\}$, $\beta_q(\mathbf{s}^n) > 0$ when $p_q = \frac{1}{2}$, and $\beta_q(\mathbf{s}^n) = 0$ when $p_q = 1$. Moreover, using the fact that $\beta_q(\mathbf{s}^n) = 0$ when $p_q = 1/2$, then

$$\frac{\partial^2 \beta_q(\mathbf{s}^n)}{\partial p_q^2}\Big|_{p_q=\frac{1}{2}} = \beta_q(\mathbf{s}^n)\left(\sum_{i=1}^{k} 4(2i(i-1) + k(k-1) - 2ki)\right) = \beta_q(\mathbf{s}^n)\left(\frac{8}{3}k(-3k + k^2 - 1)\right),$$

so $\frac{\partial^2 \beta_q(\mathbf{s}^n)}{\partial p_q^2}\Big|_{p_q=\frac{1}{2}}$ is negative when $k < \frac{3+\sqrt{13}}{2} \approx 3.3028$ and positive when $k > \frac{3+\sqrt{13}}{2}$.

Since $\beta_q(\mathbf{s}^n) = 0$ when $p_q = 1$, $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q}\big|_{p_q=1} = 0$, and $\beta_q(\mathbf{s}^n) \geq 0$ for any $p_q \in [0, 1]$, then there exists some threshold $\overline{p} < 1$ such that $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q} < 0$ and $\beta_q(\mathbf{s}^n) < \beta_q(\mathbf{s}^n)\Big|_{p_q=\frac{1}{2}}$ for all $p_q > \overline{p}$.

Since $\beta_q(\mathbf{s}^n) > 0$ when $p_q = \frac{1}{2}$, $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q}\Big|_{p_q=\frac{1}{2}} = 0$, and $\frac{\partial^2 \beta_q(\mathbf{s}^n)}{\partial p_q^2}\Big|_{p_q=\frac{1}{2}} > 0$ when $k > \frac{3+\sqrt{13}}{2}$, then there exists some threshold $\underline{p} > \frac{1}{2}$ such that $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q} > 0$ and $\beta_q(\mathbf{s}^n) > \beta_q(\mathbf{s}^n)\Big|_{p_q=\frac{1}{2}}$ for all $p_q < \underline{p}$ when $k > \frac{3+\sqrt{13}}{2}$. When $k \leq \frac{3+\sqrt{13}}{2}$, we can show by direct computation of $\beta_q(\mathbf{s}^n)$ that $\frac{\partial \beta_q(\mathbf{s}^n)}{\partial p_q} < 0$ for all $p_q \in (\frac{1}{2}, 1)$ when $k \in \{1, 2, 3\}$. This implies that the pre-screener under-trusts for all values of $p_L < p_H$ whenever $k \leq 3$, since $\beta_H(\mathbf{s}^n) < \beta_L(\mathbf{s}^n)$. When $k > 3$, there are two other sufficient conditions for the pre-screener to under-trust: (1) if $\overline{p} \leq p_L < p_H$, or (2) if $p_L \leq \underline{p}$ and $p_H > \overline{p}$ where $\underline{p} > \frac{1}{2}$ and $\overline{p} < 1$. If either of these sufficient conditions is met, then $\beta_H(\mathbf{s}^n) < \beta_L(\mathbf{s}^n)$ for $k > 3$. ■

Lemma 4 shows that the agent overtrusts and is overoptimistic about the reported state for a given $n_a > 1$ and $n_b = 0$. Clearly, the agent's degree of overtrust is monotonically decreasing as $n_b$ increases. Lemma 5 shows that there exists some $\underline{p} > \frac{1}{2}$ and $\overline{p} < 1$ such that the pre-screener under-trusts if (1) $k \in \{1, 2, 3\}$, (2) if $\overline{p} \leq p_L < p_H$, or (3) if $p_L \leq \underline{p}$ and $p_H \geq \overline{p}$. By the intermediate value theorem, there exists some $\hat{n}_b$ such that the agent under-trusts when $\mathbf{s}^n = (a, a, \ldots, a, b, b, \ldots, b)$ where $0 \leq \hat{n_b} \leq n_b < n_a$. By Lemma 1, this is the sequence most likely to generate overtrust, so *all other sequences* of such a fixed combination $(n_a, n_b)$ will also result in under-trust and pessimism about the mostly likely state. Thus, if one of the last two sufficient conditions for Lemma 5 is satisfied, then there exists some $\hat{n}_b$ such that the agent under-trusts when $\mathbf{s}^n = (a, a, \ldots, a, b, b, \ldots, b)$ where $0 \leq \hat{n_b} \leq n_b < n_a$.