ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 203 (2022) 105473

JOURNAL OF
Economic
Theory

www.elsevier.com/locate/jet

Learning with misattribution of reference dependence

Tristan Gagnon-Bartsch [a,*,1], Benjamin Bushong [b,1]

[a] *Harvard University, United States of America*
[b] *Michigan State University, United States of America*

Received 31 October 2019; final version received 26 April 2022; accepted 28 April 2022
Available online 4 May 2022

Abstract

We examine errors in learning that arise when an agent who suffers attribution bias fails to account for her reference-dependent utility. Such an agent neglects how the sensation of elation or disappointment relative to expectations contributes to her overall utility, and wrongly attributes this component of her utility to the intrinsic value of an outcome. In a sequential-learning environment, this form of misattribution generates contrast effects in evaluations and induces a recency bias: the misattributor's beliefs over-weight recent experiences and under-weight earlier ones. In the long-run, a loss-averse misattributor will grow unduly pessimistic and undervalue prospects in proportion to their variability. Both the short and long-run properties of beliefs under misattribution suggest a tendency to abandon worthwhile prospects when learning from experience. We additionally show how misattribution introduces incentives for familiar forms of expectations management.

JEL classification: D83; D84; D91

Keywords: Learning; Attribution bias; Reference dependence; Misspecified models

[*] Corresponding author.
E-mail addresses: gagnonbartsch@fas.harvard.edu (T. Gagnon-Bartsch), bbushong@msu.edu (B. Bushong).

[1] We thank Ned Augenblick, Stefano DellaVigna, Jon Eguia, Benjamin Enke, Erik Eyster, Xavier Gabaix, Benjamin Golub, David Laibson, Devin Pope, Matthew Rabin, Gautam Rao, Joshua Schwartzstein, Andrei Shleifer, Kelly Shue, Dmitry Taubinsky, Hanzhe Zhang, the editor (Pietro Ortoleva), three anonymous referees, and seminar audiences at Berkeley, Caltech, Cornell, Harvard, Harvard Business School, LSE, the Norwegian School of Economics (NHH), Princeton, SITE, Stanford, and UCSD for helpful comments.

https://doi.org/10.1016/j.jet.2022.105473
0022-0531/© 2022 Elsevier Inc. All rights reserved.

1. Introduction

When learning from personal experience, our impressions are often swayed by transitory states that affect how we feel. For instance, a person may overstate the quality of a restaurant if they try it while hungry or may overestimate the desirability of a theme park if they visit during good weather (Haggag et al., 2019). Such errors in attribution have long been explored in the psychology literature (for reviews, see Ross, 1977; Gilbert and Malone, 1995). But this research has not considered a common source of state-dependent taste: our expectations. Economists have emphasized that deviations from expectations shape our experienced utility, arguing that our utility from an outcome depends on both its "intrisinc" value and how that value compares to expectations (e.g., Kahneman and Tversky, 1979; Kőszegi and Rabin, 2006). When learning about this intrinsic value from past experiences, a person should distinguish this "reference-free" component from the sensation of elation or disappointment it generated. But the evidence on attribution errors suggests that parsing these may be difficult. In this paper, we study how an agent's learning process is distorted when she wrongly attributes reference-dependent sensations to her intrinsic taste. While learning from personal experience guides a wide range of economic decisions—shaping, e.g., our evaluations of others, preferences over consumer products, and adoption of new technologies—misattribution can hinder this common way we learn.

To illustrate, imagine a consumer learning about a new product or service. If her experience falls short of expectations, she will feel unhappy because of both the subpar quality and the negative surprise. A rational consumer will understand that part of her bad experience derived simply from her high expectations. A less introspective consumer, however, might misattribute this disappointment to the underlying quality of the product or service, and consequently underestimate how much she would enjoy it in the future. Indeed, such expectations-based disappointments have been shown to drive exit from online platforms (Backus et al., 2022) and shape employees' perceptions of their workplace (Adhvaryu et al., 2020).[2] On the other hand, an employee whose initial experience at a new workplace *beats* expectations will feel happy because of both the pleasant conditions and the positive surprise. If she misattributes this latter feeling to the quality of the workplace, she will form an overly rosy view of her new job. As these examples suggest, surprises may distort perceived outcomes: exceeding expectations inflates perceptions, and falling short deflates them.

In this paper, we model an agent who learns from repeated experience about the expected value of a prospect—e.g., the average quality of a job or a firm's service, or the typical benefit of an unfamiliar technology. How does misattribution shape this learning process? Over time, the agent's (mis)perceptions of outcomes become interdependent: a misinterpretation of today's outcome causes the agent to form biased expectations about tomorrow, which then distort her perception of that outcome. This process generates non-martingale beliefs that help explain well-documented errors such as contrast effects, overly-volatile beliefs, and a recency bias. We show how these short-run dynamics can cause a misattributor to abandon learning too quickly. Fur-

[2] Backus et al. (2022) find that new eBay users with higher expectations of winning an auction (measured by time spent in the lead) are more likely to quit using the platform if they unexpectedly lose their first auction. Relatedly, after an NGO in India improved housing conditions (Adhvaryu et al., 2020), survey responses of workers who were told of these improvements ahead of time indicated that these workers perceived their living conditions as worse than those who were not informed of the coming improvements in advance. Both of these results highlight that, *fixing the realized outcome*, falling short of expectations can lead to lower perceptions and forecasts. As discussed below, this effect on beliefs is not predicted by reference-dependence alone.

thermore, we show that even if the misattributor learns about a prospect indefinitely, then loss aversion leads her to form pessimistic beliefs in the long run. Finally, we emphasize how "expectations management"—the process of restraining an agent's initial expectations—can effectively persuade a misattributor to form inflated posterior beliefs.

We introduce the model in Section 2. We consider a dynamic setting where in each period the agent realizes an i.i.d. outcome of a prospect with an unknown mean. Based on this outcome, the agent experiences utility composed of two parts: consumption utility—which depends solely on the outcome—and reference-dependent utility, which depends on the difference between her realized consumption utility and what she expected. She then updates her beliefs about the distribution of outcomes based on her experienced utility. To preview, suppose utility from outcome $x \in \mathbb{R}$ when expecting $\hat{\theta} \in \mathbb{R}$ is $u(x|\hat{\theta}) = x + \eta n(x|\hat{\theta})$, where the reference-dependent component $n(x|\hat{\theta})$ is proportional to the difference between x and $\hat{\theta}$ and parameter $\eta > 0$ measures the weight that elation and disappointment carry on total utility. We assume that a misattributor infers from her past utility *as if* she weighted these sensations by a diminished factor $\hat{\eta} < \eta$; that is, she under-appreciates the extent to which elation or disappointment contributed to her total utility. She thus infers a distorted value of each outcome. When x surpasses expectations, she infers a value $\hat{x} > x$; when x falls short, she infers $\hat{x} < x$. The agent then updates her beliefs according to Bayes' Rule as if $\hat{x}$ truly occurred.

We first show that this simple model captures several well-known ideas. First, it incorporates the basic intuition of "disconfirmation" wherein an outcome that deviates from expectations is remembered as deviating by more than it really did. Second, our model naturally links the "positive-negative asymmetry effect"—the notion that people's beliefs respond more to bad events than good ones—to reference-dependent preferences. Third, it generates sequential contrast effects: the current outcome appears better the worse was the previous one.

We then analyze learning in Section 3 by examining the short-run dynamics of a misattributor's expectations. In early rounds when beliefs are imprecise, misattribution can generate a negative correlation in mean beliefs across periods due to the logic of disconfirmation. As beliefs become more precise over time, this anomalous pattern dissipates, but a recency bias emerges—recent outcomes influence expectations more than older ones.[3] Intuitively, a high early outcome will cause later outcomes to be underestimated, reducing the overall positive impact of that high early outcome on beliefs. This countervailing force implies that the influence of an early outcome fades over time. Additionally, we show how the misattributor's overly volatile beliefs, which stem from this recency bias, can cause her to quit learning about a prospect faster than an unbiased agent.

In Section 4, we demonstrate how misattribution continues to distort beliefs in the long-run. The interplay between beliefs and perceived outcomes can prevent a misattributor from reaching correct expectations despite ample experience. Nevertheless, the agent's beliefs converge. We characterize these steady-state beliefs and show that a loss-averse misattributor underestimates a prospect's mean outcome. Furthermore, this pessimistic bias is greater for prospects that are more variable. We show that this fact, together with the preferences we examine, implies that a misattributor who learns indefinitely about two prospects can form arbitrarily pessimistic beliefs about the more variable one, even if the two in truth yield the same expected utility.

[3] While other papers incorporate an exogenous recency effect into learning models (e.g., Ehling et al., 2018; Malmendier et al., 2020), our model endogenously generates this effect.

Both the short-run belief dynamics and this long-run pessimism highlight a central implication of misattribution: a tendency to abandon worthwhile prospects when learning from experience. Our results may therefore speak to familiar patterns of "adoption and decay" in developing countries where people briefly adopt beneficial technologies but subsequently abandon them (for discussions of these patterns, see Hanna et al., 2014; Kremer et al., 2019).[4]

In Section 5, we explore how our short-run results enable the common practice of "expectations management", wherein expectations are strategically lowered to subsequently beat them. We consider a "designer" who can manipulate either prior expectations or outcomes in order to inflate the posterior beliefs of a misattributing agent. We focus only on settings where intervention would have no benefit if the agent were unbiased. When the agent's prior is sufficiently imprecise, the designer has an incentive to directly lower prior expectations. This tactic is ineffective when the prior is more precise, yet other approaches for persuasive expectations management arise. Namely, the designer has an incentive (i) to rearrange a fixed set of outcomes in low-to-high order; and (ii) to transfer utility from early to later periods in order to "walk down" expectations. We discuss how our mechanism may therefore help explain why restraining expectations (and then beating them) is persuasive.

We conclude in Section 6 by further discussing our assumptions—including their motivation, interpretation, and generalizability—and ways that future research could distinguish misattribution from similar models. We also note how our model naturally extends to a bias in social learning where observers under-appreciate how expectations shape others' experiences. In this sense, we caution policy makers and researchers to account for expectations when interpreting satisfaction surveys. Our model also suggests a more basic point: what organizations or policy makers promise about new technologies or reforms may ultimately shape their perceived value. Thus, hype may backfire.

Broadly, our paper extends the literature on attribution bias by characterizing the dynamics of beliefs and their long-run convergence. Moreover, we are the first paper to explore attribution bias with respect to reference-dependent utility. In doing so, we highlight a simple avenue through which reference-dependent preferences directly alter *beliefs*, which would not arise absent misattribution. In this way, we capture the intuition that the emotional affect of an event may alter its perceived informational content.

Related literature

A sizeable literature in psychology explores "attribution biases" wherein extraneous situational factors are incorrectly attributed to a stable characteristic of a person or good (see Ross, 1977 for an early review of this literature or Gilbert and Malone, 1995 for a more recent one).

Recent empirical work in economics has explored such attribution biases, though the focus in these papers is distinct from ours. For example, Haggag et al. (2019) show that experimental participants have a higher willingness to pay for an unfamiliar drink when they first experienced it while thirsty. Additionally, frequent patrons of an amusement park are more likely to return when their most recent visit was during good weather. These findings suggest that state-dependent utilities—stemming from thirst and weather—are wrongly attributed to stable characteristics—

[4] These articles discuss behavioral explanations for a number of empirical patterns in development contexts. For a specific example of adoption and decay, see Hanna et al. (2016), who find that in-person demonstrations encourage people to adopt improved cooking stoves, but inconsistent personal experience thereafter leads them to stop using them.

the quality of the drink or the desirability of the theme park.[5] Relatedly, Haggag et al. (2021) show that West Point cadets are less likely to study a given major if their initial coursework in that area was scheduled very early in the morning. Those authors suggest that the sensation of "tiredness" colors these students' opinions of a given area of study. Our model builds from a shared intuition: transient sensations—in our case, elation and disappointment—are misattributed to stable characteristics. However, we extend this literature by analyzing the dynamic learning process under attribution bias.

Bushong and Gagnon-Bartsch (2022) utilize a series of experiments which carefully control expectations to provide evidence of our *specific* notion of attribution bias with reference-dependent utility. In that paper's main experiment, participants completed one of two unfamiliar tasks: a neutral task or that same task with an unpleasant noise played in the background. At the start of the experiment, we manipulated participants' expectations about which task they would face. Participants in one treatment group determined their task by flipping a coin just before working, while participants in a second group faced only a little uncertainty over their task. Hence, the task assignment for some came as a larger positive surprise (or a larger disappointment). Hours after participants first worked on their assigned task, we elicited their willingness to continue working (WTW) on that task for additional pay. Relative to the group with little ex-ante uncertainty, those assigned by coin flip exhibited an inflated WTW on the neutral task and a depressed WTW on the unpleasant task. These results, which we show are inconsistent with reference-dependence absent misattribution, suggest that the magnitude of the sensations of elation or disappointment influenced participants' ex-post evaluations of those outcomes, in line with the theory we present here.[6]

Other social sciences have noted that prior expectations can alter a person's evaluation and memory of an experience (see, e.g., Wilson and Gilbert, 2003 for a review).[7] This research highlights that when an outcome deviates from expectations, a person might either assimilate it—interpreting it as favoring their current beliefs—or contrast it—interpreting it against their current beliefs. Our model, capturing the latter notion, thus stands at odds with "confirmation bias", wherein new evidence is wrongly interpreted as conforming to one's expectations (e.g., Rabin and Schrag, 1999; Fryer et al., 2019). However, these two mechanisms are not mutually exclusive. Empirical tests of order effects in belief updating find support for both confirmation and disconfirmation effects (e.g., Hogarth and Einhorn, 1992; Geers and Lassiter, 1999).[8] And there are many settings where the logic of disconfirmation—where perceptions are inversely re-

[5] We differ from Haggag et al. (2019) by examining (i) a different domain of attribution bias, and (ii) learning over time, with a formal analysis of belief dynamics and long-run perceptions. In contrast to our results, biased forecasts in Haggag et al.'s framework vanish with experience. This distinction arises because Haggag et al. rule out complementarities where past experiences influence current consumption utility. This is central to our model, as past experiences form the reference point against which current consumption is evaluated.

[6] This evidence is consistent with the model we propose below and specifically provides strong support for Observation 1 and suggestive support for Observation 3. However, Bushong and Gagnon-Bartsch (2022) cannot test the dynamic implications of our model; in that way, it primarily serves as a foundation for the extensions we explore here.

[7] For instance, political scientists have argued that discrepancies between a politician's performance and citizens' expectations shape how citizens perceive that politician (Kimball and Patterson, 1997; Waterman et al., 1999). Marketing has also emphasized the role of expectations on the perceived quality of service (Oliver, 1977, 1980; Boulding et al., 1993).

[8] Geers and Lassiter (1999) provide the following guidance for when to expect contrasts: "[I]n situations in which individuals are very motivated to gain a great deal of information (e.g., highly unpredictable situations, or situations with a great deal of subjective importance or personal interest), they should be more likely to employ finer rates of unitization [. . .] which should increase the chances for contrast effects in affective experience."

lated to expectations—seems particularly compelling. For instance, we suspect that tempering expectations—rather than hyping them—is often an effective way to enhance ex-post perceptions. A movie or a debate performance may seem worse when expectations are high, wait times may seem longer when expecting speedy service, and the weather may seem colder when expecting a warm day.

We also join a growing literature on learning among agents with misspecified models. Esponda and Pouzo (2016) provide a general framework for assessing the long-run beliefs and behavior of such agents. Elements of our modeling approach—in particular, our analyses of long-run beliefs—are also similar to those of Heidhues et al. (2018), who study how overconfidence can lead an agent to mislearn the mapping from effort to output. They consider an "active-learning" environment where the agent's actions affect the distribution of observed outcomes, and long-run mislearning arises due to a feedback loop between his erroneous beliefs and his actions. A similar feedback mechanism emerges in our model but it stems from the interplay between erroneous beliefs and the encoding of outcomes. Thus, our model shares similarities with active learning even when the agent faces exogenous data: a misattributor's current beliefs influence her *perception* of the data in much the same way that an agent's actions influence the true data in an active-learning setting.[9]

2. A model of misattribution of reference-dependent utility

Reference-Dependent Preferences. Following Kőszegi and Rabin (2006), we consider an agent whose overall utility has two additively separable components. The first component, "consumption utility", denoted by $x \in \mathbb{R}$, corresponds to the payoff traditionally studied in economics.[10] The second component, "gain-loss utility", derives from comparing x to a reference level of utility. As in Bell (1985), we take this reference point to be the agent's expectation of x, denoted by $\hat{\theta}$. We consider a simple piecewise-linear specification of gain-loss utility given by

$$n(x|\hat{\theta}) = \begin{cases} x - \hat{\theta} & \text{if} \quad x \geq \hat{\theta} \\ \lambda(x - \hat{\theta}) & \text{if} \quad x < \hat{\theta}, \end{cases} \tag{1}$$

where parameter $\lambda \geq 1$ captures loss aversion. The agent's total utility from outcome x is

$$u(x|\hat{\theta}) = x + \eta n(x|\hat{\theta}), \tag{2}$$

where $\eta > 0$ is the weight given to sensations of gain and loss relative to absolute outcomes.

[9] Heidhues, Kőszegi, and Strack (2021) study the convergence of mispecified learning with Gaussian priors and outcomes, an approach we take below. Other work on misspecified learning includes Bohren and Hauser (2021) and Frick et al. (2020) on social learning; He (2021) on the gambler's fallacy; Schwartzstein (2014) on selective attention; and Nyarko (1991) and Fudenberg et al. (2017) in experimentation settings. Other models (e.g., Eyster and Rabin, 2010 and Bohren, 2016) predict overreaction to new observations, but their underlying mechanism—a failure to account for informational redundancies in social behavior—is very different from ours. Additionally, Epstein et al. (2010) analyze the limit beliefs of an agent who under- or over-reacts to information. While they demonstrate that overreaction in general can cause beliefs to converge to a false distributional parameter, we can precisely describe these limit beliefs given our focus on a specific misspecified model.

[10] For tractability, we will work directly with the distribution of consumption utility rather than the distribution of material outcomes. We interpret x as if it derives from a standard Bernoulli utility function, $m : \mathbb{R} \to \mathbb{R}$, over consumption realizations $c \in \mathbb{R}$ such that $x = m(c)$. Appendix C.2 considers an extension to multiple consumption dimensions.

Learning Environment. The agent is initially uncertain about the distribution of consumption utility and learns about it through experience. In each period $t = 1, 2, \dots$ the agent receives consumption utility $x_t \in \mathbb{R}$ drawn independently from a distribution $F(\cdot|\theta)$ that depends on an unknown parameter $\theta \in \mathbb{R}$. To focus the analysis, we assume $x_t = \theta + \epsilon_t$ where $\epsilon_t \sim N(0, \sigma^2)$. We assume the agent knows the variance $\sigma^2 > 0$ and begins with a prior belief $\theta \sim N(\theta_0, \rho^2)$. Letting π_t denote her subjective distribution over θ following t outcomes, $\hat{\theta}_{t-1} \equiv \int_{-\infty}^{\infty} \tilde{\theta}\, d\pi_{t-1}(\tilde{\theta})$ then denotes the agent's expectation of θ entering period t.

The agent's total utility (Equation (2)) in period t is thus $u(x_t|\hat{\theta}_{t-1}) = x_t + \eta n(x_t|\hat{\theta}_{t-1})$. Importantly, the agent's expectation, $\hat{\theta}_{t-1}$, represents a fluctuating state variable that introduces variation in her utility conditional on x_t.

Misattribution of Reference-Dependent Utility. We now turn to the central feature of our model. Motivated by the literature on attribution bias, we assume that the agent neglects how her past experience was influenced by reference dependence and misattributes this state-dependent component of her utility to the underlying consumption value, x_t. In doing so, the agent misinfers x_t based on her total experienced utility. To make this logic of misinference transparent, we focus on settings where the agent observes her total utility each period, but the underlying outcome x_t is either unobserved or not readily quantifiable—e.g., a person learning about their taste for an experience good.[11]

We formalize misattribution as follows. After experiencing $u_t \equiv x_t + \eta n\big(x_t \big| \hat{\theta}_{t-1}\big)$, an unbiased agent who is updating her beliefs about θ appropriately controls for how much the state-dependent gain-loss term influenced her total utility. In contrast, a misattributor underappreciates the influence of this state-dependent component and infers from u_t as if gains and losses were weighted by a diminished factor $\hat{\eta} \in [0, \eta)$. After each period, the agent uses her memory of u_t, along with her misspecified model, to infer the consumption value she must have received. Letting $\hat{u}\big(x \big| \hat{\theta}_{t-1}\big) = x + \hat{\eta} n\big(x \big| \hat{\theta}_{t-1}\big)$, misattributor therefore "encodes" the consumption value $\hat{x}_t$ that solves

$$u\big(x_t \big| \hat{\theta}_{t-1}\big) = u_t = \hat{u}\big(\hat{x}_t \big| \hat{\theta}_{t-1}\big). \tag{3}$$

Under our model, encoded outcomes take a simple form: Equations (2) and (3) yield

$$\hat{x}_t = \begin{cases} x_t + \kappa^G \left(x_t - \hat{\theta}_{t-1}\right) & \text{if} \quad x_t \geq \hat{\theta}_{t-1} \\ x_t + \kappa^L \left(x_t - \hat{\theta}_{t-1}\right) & \text{if} \quad x_t < \hat{\theta}_{t-1}, \end{cases} \tag{4}$$

where

$$\kappa^G \equiv \left(\frac{\eta - \hat{\eta}}{1 + \hat{\eta}}\right) \quad \text{and} \quad \kappa^L \equiv \lambda \left(\frac{\eta - \hat{\eta}}{1 + \hat{\eta}\lambda}\right). \tag{5}$$

The parameters κ^G and κ^L represent the extent that elations and disappointments, respectively, distort encoded outcomes. They will therefore be used extensively in our analysis of the misattributor's updating process. Intuitively, κ^G and κ^L increase in the degree of misattribution (i.e., as $\hat{\eta}$ decreases), and $\kappa^L > \kappa^G$ under loss aversion (i.e., $\lambda > 1$).

We assume the agent is unaware of her misencoding but is otherwise Bayesian: she updates her beliefs over θ following Bayes' Rule given her encoded outcomes. Absent misattribution

[11] We suspect that misattribution can occur even when outcomes are precisely quantifiable in the moment (e.g., prices) but imperfectly remembered. We discuss this further, along with our other modeling assumptions, in Section 6.

(i.e., $\hat{\eta} = \eta$), we have $\kappa^L = \kappa^G = 0$ and thus the model embeds standard Bayesian updating—albeit with reference-dependent preferences—as a special case. Finally, whenever we consider the agent's actions (e.g., Propositions 3 and 5), we will assume she behaves according to Equation (2) given her biased beliefs (see Section 6 for a discussion of the limited impact of this assumption and plausible alternatives).

The remainder of the paper examines the process of beliefs, $\langle\hat{\theta}_t\rangle$, that results from the encoding error in Equation (4). Here, we first note some intuitive features of our model. We also note some connections to existing empirical findings (largely outside of the field of economics). Although these connections have not been made previously, we speculate that our model may provide a plausible psychological basis for these effects.

First, our model captures the basic logic of disconfirmation: an outcome that deviates from expectations is perceived as deviating by more than it really did (as in, e.g., Anderson, 1973; Oliver, 1977, 1980; Geers and Lassiter, 1999).

Observation 1. *Disconfirmation effect:* If $x_t > \hat{\theta}_{t-1}$, then $\hat{x}_t > x_t$, and if $x_t < \hat{\theta}_{t-1}$, then $\hat{x}_t < x_t$.

This also implies that the same outcome can be perceived differently as expectations change. Second, loss aversion causes disappointment to distort encoded outcomes—and hence beliefs—more than elation, consistent with evidence on "positive-negative asymmetry effects" in belief updating (as in, e.g., Peeters and Czapinski, 1990; Baumeister et al., 2001; Kuhnen, 2015).[12]

Observation 2. *Asymmetric misencoding:* Suppose $\lambda > 1$. Consider outcomes $x^g = \hat{\theta}_{t-1} + \delta$ and $x^l = \hat{\theta}_{t-1} - \delta$. For any $\delta > 0$, $|\hat{x}^l - x^l| > |\hat{x}^g - x^g|$.

Third, misattribution generates "sequential contrast effects": fixing today's outcome, its perceived value is higher when yesterday's value was lower (as in, e.g., Bhargava, 2007; Bhargava and Fisman, 2014).

Observation 3. *Sequential contrast effect:* Fixing x_t, $\hat{x}_t$ is strictly decreasing in x_{t-1}.

Since current expectations are strictly increasing in the previous outcome, an increase in that outcome implies that today's outcome is assessed against a higher benchmark and thus generates a greater disappointment (or a smaller elation).

3. Short-run dynamics of beliefs

The observations above describe how misattribution distorts encoded outcomes; we now analyze how these distortions influence belief updating in the short-run. Although the order of outcomes is irrelevant for rational inference, the order in which a misattributor experiences outcomes alters her perceived value of the prospect. In early rounds, the logic of disconfirmation

[12] Baumeister et al. (2001) succinctly describe the positive-negative asymmetry effect: "[E]vents that are negatively valenced (e.g., losing money, being abandoned by friends, and receiving criticism) will have a greater impact on the individual than positively valenced events of the same type (e.g., winning money, gaining friends, and receiving praise)... This is probably most true in the field of impression formation, in which the positive-negative asymmetry effect has been repeatedly confirmed." While loss aversion (as in Kahneman and Tversky, 1979) implies that prospective losses loom large in preferences, we provide a mechanism for why *past* losses loom large in both memory and forecasts.

(as implied by Observations 1 and 3) can lead to oscillating swings in beliefs. Over time, these swings dissipate but a "recency bias" persists: beliefs overweight recent outcomes and underweight older ones. In natural choice contexts, this bias may lead a misattributor to terminate learning too quickly.

Updating Process under Misattribution. Rational updating follows a simple rule: given prior $\hat{\theta}_{t-1}$, the estimate of θ following x_t is $(1-\alpha_t)\hat{\theta}_{t-1}+\alpha_t x_t$, where $\alpha_t \equiv 1/(t+\sigma^2/\rho^2)$ is the weight given to the most recent outcome. A misattributor's updating process follows the same rule but with the encoded outcome $\hat{x}_t$ in place of x_t. Since $\hat{x}_t = x_t + \kappa_t\big(x_t - \hat{\theta}_{t-1}\big)$ where $\kappa_t \equiv \kappa^G \mathbb{1}\{x_t > \hat{\theta}_{t-1}\} + \kappa^L \mathbb{1}\{x_t < \hat{\theta}_{t-1}\}$ (see Equation (4)), it follows that

$$\hat{\theta}_t = \underbrace{(1-\alpha_t)\hat{\theta}_{t-1}}_{\text{Direct effect}} + \alpha_t x_t + \underbrace{\alpha_t\kappa_t\big(x_t - \hat{\theta}_{t-1}\big)}_{\text{Disconfirmation effect}}. \tag{6}$$

This highlights that a misattributor's prior expectation, $\hat{\theta}_{t-1}$, has two opposing effects on her posterior, $\hat{\theta}_t$. On the one hand, the prior has a direct positive effect stemming from the usual logic of Bayesian updating. On the other hand, the prior also has a *negative* "disconfirmation" effect stemming from the contrast between the outcome and expectations.

Equation (6) reveals additional features of the updating process once rearranged as

$$\hat{\theta}_t = [1-\alpha_t(1+\kappa_t)]\hat{\theta}_{t-1} + \alpha_t(1+\kappa_t)x_t. \tag{7}$$

First, a misattributor "overreacts" to the latest outcome: she weights x_t by $\alpha_t(1+\kappa_t)$ instead of α_t. Second, we can more clearly see when the negative effect discussed above is dominant. From the definition of α_t, the posterior expectation, $\hat{\theta}_t$, is a decreasing function of the prior expectation, $\hat{\theta}_{t-1}$, when $\kappa_t > t-1+\sigma^2/\rho^2$; that is, when the encoding bias (i.e., κ_t) is sufficiently strong compared to the relative precision of beliefs entering period t (i.e., $t-1+\sigma^2/\rho^2$).[13] Intuitively, when these beliefs are imprecise, the posterior heavily weights the encoded outcome, and thus the primary effect of the prior is through the disconfirmation channel. This negative relationship between prior and posterior beliefs is a stark and novel prediction of misattribution. However, any such negative relationship is necessarily short-lived. Since the precision of beliefs increases with time, the standard positive effect of the prior eventually dominates. Specifically, once $t > t^* \equiv 1+\kappa^L-\sigma^2/\rho^2$, the posterior $\hat{\theta}_t$ is necessarily increasing in the prior, $\hat{\theta}_{t-1}$. Moreover, $t > t^*$ also guarantees that $\hat{\theta}_t$ lies between x_t and $\hat{\theta}_{t-1}$; in contrast, when $t < t^*$, $\hat{\theta}_t$ can move so strongly in the direction of x_t that the posterior "overshoots" the outcome. The value t^* therefore delineates two phases of the belief dynamics and will play an important role in our results below.[14]

Iterating Equation (7) reveals that a misattributor's expectations inconsistently weight past outcomes in two distinct ways that differ from rational updating: (i) expectations at time t generically place different weights on each of the previous outcomes; (ii) as more outcomes accrue, the weight on outcome x_τ relative to other outcomes changes. While we describe these weights below, we can readily see the preceding points by advancing Equation (7) by an additional period:

[13] The agent's belief entering period t is normally distributed with mean $\hat{\theta}_{t-1}$ and precision $[t-1+\sigma^2/\rho^2]/\sigma^2$. Thus, $t-1+\sigma^2/\rho^2$ is the precision of beliefs relative to the precision of outcomes.

[14] It is important to note that $t^* < 1$ means that there is only one phase of the dynamics. This is guaranteed when $\kappa^L < \sigma^2/\rho^2$; that is, the extent of misattribution is small relative to the precision of the prior entering the first period.

$$\hat{\theta}_{t+1} = [1-\alpha_{t+1}(1+\kappa_{t+1})][1-\alpha_t(1+\kappa_t)]\hat{\theta}_{t-1} \\ + \alpha_{t+1}\underbrace{(1+\kappa_t)[1-\alpha_t\kappa_{t+1}]}_{\text{Distortion factor on } x_t}x_t + \alpha_{t+1}\underbrace{(1+\kappa_{t+1})}_{\text{Distortion factor on } x_{t+1}}x_{t+1}. \quad (8)$$

Advancing this further reveals the distorted weights that a misattributor attaches to all past outcomes. By doing so, we can express her expectation after t rounds as a simple (mis)weighted sum of the true outcomes (see Appendix A for all proofs).

Lemma 1. *A misattributor's expectation of θ after t rounds can be expressed as*

$$\hat{\theta}_t = \beta_0^t\theta_0 + \alpha_t\sum_{\tau=1}^{t}\beta_\tau^t x_\tau, \quad (9)$$

where $\beta_0^t = \prod_{j=1}^{t}[1-\alpha_j(1+\kappa_j)]$ *and*

$$\beta_\tau^t = \begin{cases} (1+\kappa_\tau)\prod_{j=\tau}^{t-1}[1-\alpha_j\kappa_{j+1}] & \text{for} \quad \tau\in\{1,\ldots,t-1\}, \\ (1+\kappa_\tau) & \text{for} \quad \tau = t. \end{cases}$$

Since rational expectations in round t place weight α_t on each outcome, β_τ^t measures how misattribution distorts the weight on outcome x_τ relative to the rational benchmark. Hence, β_τ^t correspond to the "distortion factors" underscored in Equation (8). Appendix B.1 provides a detailed description of how these weights (and thus beliefs) evolve; our results below will emphasize the key insights.

Implications of Biased Updating. We now demonstrate three implications of this biased belief process. First, we show that for outcomes occurring after t^*, a misattributor exhibits a *recency bias*: her beliefs weight a recent gain more than a preceding gain and weight a recent loss more than a preceding loss, and she underweights an outcome by more the farther it fades into the past.

Proposition 1. *Consider expectations after t rounds, $\hat{\theta}_t$.*

1. *The weight on the most recent outcome is distorted by a factor $\beta_t^t > 1$; the weight on any earlier outcome x_τ, $\tau < t$, is distorted by a factor β_τ^t such that $\beta_\tau^t \to 0$ as $t \to \infty$.*
2. *Consider any two outcomes x_{τ_1} and x_{τ_2} that are both gains or both losses (i.e., $\kappa_{\tau_1} = \kappa_{\tau_2}$). If $\tau_2 > \tau_1 \geq t^*$, then the more recent outcome receives greater weight: $\beta_{\tau_2}^t > \beta_{\tau_1}^t > 0$.*

The intuition underlying Proposition 1 stems from the way that sequential contrast effects (Observation 3) play out over time. Namely, early outcomes have a "self limiting" influence on later beliefs. Higher expectations both increase the likelihood that subsequent outcomes are encoded as losses rather than gains, and cause subsequent disappointing (elating) outcomes to be underestimated (overestimated) by more (less). These effects dampen the positive influence of a high initial outcome on later beliefs. A similar countervailing force emerges when an initial outcome lowers early expectations. Furthermore, as t advances, an early outcome x_τ exerts this countervailing force on a larger number of subsequent outcomes, which drives the distortion factor on x_τ to zero. Both parts of the proposition together imply that the most recent outcomes eventually have the greatest influence on beliefs, regardless of the underlying parameters.[15]

[15] Part 2 of Proposition 1 focuses on outcomes in the same domain because loss aversion implies that losses influence beliefs more than gains (Observation 2). Hence, a loss in period $t-1$ may have a larger influence on beliefs than a gain

The recency bias highlighted in Proposition 1 implies that a misattributor's beliefs are perpetually too variable. Indeed, the variance of $\hat{\theta}_t$ conditional on $\hat{\theta}_{t-1}$ is always larger under misattribution than rational learning (see Appendix B.2). Next, we highlight two implications of these volatile beliefs; one for the early phase of the dynamics prior to t^*—where the current expectation may be decreasing in the previous one—and one for the more stable phase beyond t^*—where this anomalous negative relationship vanishes.

In the early phase where beliefs are relatively imprecise (i.e., $t < t^*$), $\hat{\theta}_t$ is necessarily decreasing in $\hat{\theta}_{t-1}$ if $t < \kappa^G - \sigma^2/\rho^2$ (see Equation (7)).[16] Consequently, the apparent weight that expectations place on a past outcome alternates between positive and negative as time advances according to the product in Lemma 1. This implies that beliefs exhibit predictable oscillations: when beliefs initially move in one direction, they will likely move in the opposite direction in the following round.

Proposition 2. *Consider $t < \kappa^G - \sigma^2/\rho^2$. Conditional on $\hat{\theta}_{t-1}$ and θ,* $\text{Cov}(\hat{\theta}_t, \hat{\theta}_{t+1}) < 0$.

These short-run effects evoke, for instance, the logic of the so-called *sophomore slump*: after a surprisingly good first experience, a second experience often falls short of the new, lofty expectations. This pattern arises here even when both outcomes are in fact identical. Since beliefs in these early periods "overshoot" the underlying outcome, a positive first experience can cause a misattributor to subsequently find that same outcome disappointing and thus revise her beliefs downward. This prediction—and the recency bias in Proposition 1—run contrary to alternative models of misencoded outcomes in which signals are simply exaggerated or interpreted with a pessimistic bias.[17]

The dynamics in Proposition 2 require $t^* > 1$ (since $\kappa^G - \sigma^2/\rho^2 < t^*$). It is worth noting that the constellation of parameters that give rise to $t^* > 1$ is relatively narrow. The definition of κ^G and κ^L reveal that even if $\hat{\eta} = 0$—that is, the agent is fully biased—then $t^* < 1 + \eta\lambda$, implying that t^* is small under reasonable degrees of loss aversion. Although this limits the scope of periods where Proposition 2 applies, it means that the recency bias in Proposition 1 will be observed frequently.

Once $t > t^*$, excessive variability in beliefs persists due to the recency bias in Proposition 1. This can have important choice implications; namely, it may induce the agent to terminate learning too quickly. To illustrate, we examine when a misattributor's beliefs first fall below some floor $\bar{\theta} < \theta_0$. This reflects, for example, a setting where the agent experiments with a risky prospect and follows a fixed (myopic) stopping rule where she abandons the prospect once it seems sufficiently bad.

Even if $t^* < 1$, meaning that the misattributor's posteriors are always increasing in the prior and hence never overshoot the underlying outcome, then her beliefs will cross below any thresh-

in period t. However, for any two outcomes in the same domain, the more recent one receives more weight. Moreover, when loss aversion is negligible, more recent outcomes will receive more weight regardless of their domain.

[16] This decreasing relationship also holds for $t \in (\kappa^G - \sigma^2/\rho^2, t^*)$ when the outcome in t is a loss and thus has a relatively strong distortionary effect on $\hat{x}_t$. When $t < \kappa^G - \sigma^2/\rho^2$, this negative relationship holds regardless of whether x_t comes as a gain or a loss. See Appendix B.1 for details.

[17] By considering alternative specifications for encoded outcomes, one can model such notions. For instance, one could capture "exaggerated signals" with $\hat{x}_t = \varphi x_t$ for $\varphi > 1$ and two forms of "generalized pessimism" with either (i) $\hat{x}_t = x_t - \varphi$ for $\varphi > 0$ or (ii) $\hat{x}_t = x_t$ if $x_t > 0$ and $\hat{x}_t = \varphi x_t$ for $\varphi > 1$ if $x_t < 0$. Importantly, unlike our model, each of these variants generates encoded outcomes that are truly i.i.d. Thus, they fail to generate negative covariance in beliefs or a recency bias, and they fail to predict the tactics for expectations management that we highlight in Section 5.

old $\bar{\theta}$ faster than her unbiased counterpart. To formalize, let $\langle\theta_t\rangle$ denote the process of rational Bayesian mean beliefs, and let $S^B_{\bar{\theta}} = \min\{t \geq 1 | \theta_t < \bar{\theta}\}$ and $S^M_{\bar{\theta}} = \min\{t \geq 1 | \hat{\theta}_t < \bar{\theta}\}$ be stopping times indicating when the Bayesian and misattributive belief processes first fall below $\bar{\theta}$, respectively.

Proposition 3. *Suppose $t^* < 1$ and consider any $\bar{\theta} < \theta_0$. With probability one, the misattributor's stopping time, $S^M_{\bar{\theta}}$, is weakly less than the rational stopping time, $S^B_{\bar{\theta}}$, and it is strictly less with positive probability.*

Good outcomes may temporarily inflate a misattributor's beliefs above the Bayesian ones, but subsequent bad outcomes causing the Bayesian beliefs to dip below $\bar{\theta}$ will cause the misattributor's beliefs to crash even harder. Thus, when the Bayesian beliefs cross below $\bar{\theta}$, the misattributor's beliefs must do so too, if they haven't already. Although the misattributor's beliefs obey some Bayesian-like properties when $t^* < 1$, misattribution can still cause an undue propensity to quit learning about prospects that initially seem worthwhile. This clarifies our model's connection to our introductory example (Backus et al., 2022) wherein users on eBay are more likely to quit the platform after a surprising loss. Those authors note that our form of attribution bias may be a plausible psychological basis for their assumptions on how users' expectations depend on past experience. More speculatively, this may speak to, for instance, Kremer, Rao, and Schilbach's (2019) observation that overreaction to discouraging personal experience underlies the familiar pattern of adoption then abandonment of beneficial technologies. Those authors convey a variety of explanations for this pattern including overinference from small samples and base-rate neglect; likewise Hanna et al. (2014) suggest selective attention may be the cause. The proposition above suggests that misattribution may offer a compelling alternative.

4. Long-run beliefs and pessimism over risky prospects

We now examine the limiting properties of the belief process described above. In particular, we highlight how errors in beliefs can persist despite ample experience with the prospect. Although the misattributor places excess weight on recent outcomes (Proposition 1), her beliefs about θ eventually converge. However, these long-run beliefs do not converge to the truth: they are biased downward by the agent's loss aversion. This pessimism increases in proportion to the prospect's underlying variability. Furthermore, this bias implies that the agent will too often reject risky-but-worthwhile prospects even after seemingly sufficient experimentation.

We seek to establish that the misattributor's sequence of mean beliefs $\langle\hat{\theta}_t\rangle$ converges to a *steady-state belief*, $\hat{\theta}$, that is consistent with the encoded data it generates. That is, when holding expectation $\hat{\theta}$, the average encoded outcome is equal to $\hat{\theta}$. Thus, a steady-state belief solves

$$\hat{\theta} = \mathbb{E}\left[x_t + \kappa_t(x_t - \hat{\theta}_{t-1}) \middle| \hat{\theta}_{t-1} = \hat{\theta}\right], \tag{10}$$

where $\mathbb{E}[\cdot]$ is with respect to the true data generating process.

In fact, $\langle\hat{\theta}_t\rangle$ converges to a unique steady-state belief, which we denote by $\hat{\theta}_\infty$. We show this below, where we also characterize how $\hat{\theta}_\infty$ depends on the true distributional parameters and the misattributor's underlying preferences.

Proposition 4. *Fixing θ and σ^2, there is a unique steady-state belief, $\hat{\theta}_\infty$, and $\langle\hat{\theta}_t\rangle$ converges almost surely to $\hat{\theta}_\infty$. Furthermore:*

1. *The steady-state belief underestimates the true mean, $\hat{\theta}_\infty \le \theta$, and this inequality is strict if and only if $\lambda > 1$.*
2. *Comparative statics: If $\lambda > 1$, then $\hat{\theta}_\infty$ is strictly decreasing in the variance (σ^2) and the degrees of reference dependence (η) and loss aversion (λ).*

Proposition 4 shows that there is a unique solution to Equation (10), and that the process indeed converges to this value. Although outcomes are truly i.i.d., convergence does not follow directly from a basic law of large numbers because *encoded* outcomes are serially correlated: prior outcomes shift the agent's reference point, thereby influencing the current encoded outcome. Accordingly, we follow Heidhues et al. (2021) and use techniques from stochastic-approximation theory (along with the fact that outcomes are normally distributed) to establish convergence. The details of this analysis are discussed in the proof.

Part 1 of Proposition 4 shows that a loss-averse misattributor forms pessimistic beliefs over time. Intuitively, loss aversion causes the agent to encode a distribution of outcomes that is negatively skewed relative to the true distribution—she underestimates bad experiences more than she overestimates good ones. While loss aversion drives down perceptions of θ, it is not ex-ante obvious that such pessimistic expectations will persist. Reference dependence generates a force that acts against any resultant pessimism since such beliefs will generate more frequent pleasant surprises. This tension can be seen more explicitly from Equation (10), which implies that $\hat{\theta}_\infty$ is characterized by the solution to the following equation:

$$\hat{\theta} = \theta - \underbrace{k \Pr\left(x_t < \hat{\theta}\right)\left(\hat{\theta} - \mathbb{E}\left[x_t \middle| x_t < \hat{\theta}\right]\right)}_{\text{Downward Bias}} \quad \text{where} \quad k \equiv \frac{(\lambda - 1)(\eta - \hat{\eta})}{(1+\eta)(1+\hat{\eta}\lambda)}. \tag{11}$$

Equation (11) highlights that the bias is proportional to the size of the average encoded loss in the steady-state scaled by the likelihood of such a loss. If the agent became too pessimistic, losses wouldn't occur; if she approached accurate expectations, her beliefs would be immediately pushed back down. Thus, it follows that $\hat{\theta}_\infty$ is biased downward and these resultant beliefs are inherently stable. Specifically, if the agent's expectations were to move below (above) $\hat{\theta}_\infty$, then she would experience an increased rate of elations (disappointments) that drive her expectations back up (down).

It is worth briefly noting that even with $\lambda = 1$, the steady-state distribution of encoded outcomes has greater variance than the true distribution. Although we focus on learning about θ alone, this force could lead a misattributor to overestimate the variance of outcomes if she were also uncertain about σ^2. In this case, she would mislearn the distribution of x_t even without loss aversion.[18]

Part 2 of Proposition 4 shows that greater variability in the distribution of outcomes causes the misattributor to underestimate θ by a larger amount. Hence, she develops more pessimistic beliefs about prospects that are riskier. Intuitively, increased variance generates greater sensations of elation and disappointment. And since loss aversion implies that such gain-loss utility is negative on average, encoded outcomes tend to decrease in σ^2. Fig. 1 uses a simulated sequence of outcomes to depict both results, showing the path of beliefs and the density of perceived outcomes for two values of σ^2. These results imply, for example, that a client assessing the typical

[18] In Appendix B.3, we sketch an example where the agent learns about both θ and σ^2. Under misattribution, the agent's long-run beliefs overestimate σ^2 and underestimate θ as in Proposition 4.

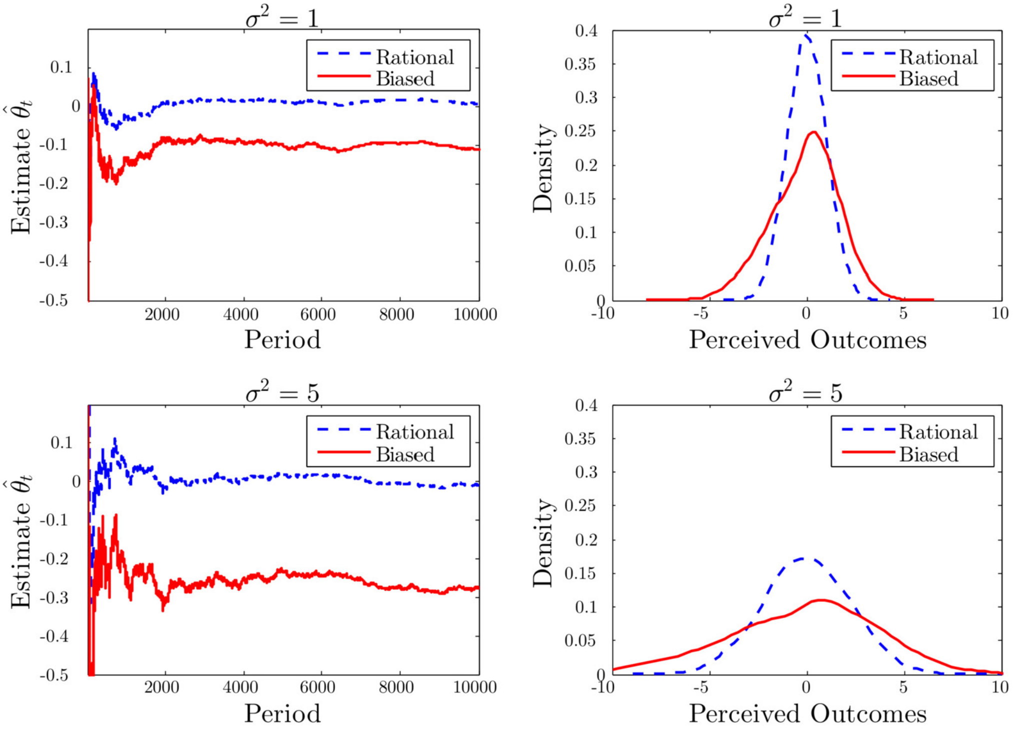

Fig. 1. The top-left panel depicts a simulated path of beliefs $\langle \hat{\theta}_t \rangle$ for both a rational and biased agent. The top-right panel shows the true and perceived density of outcomes. The simulation assumes $\theta = 0$, $\sigma^2 = 1$, $\eta = 1$, $\lambda = 3$, and $\hat{\eta} = 1/3$. The bottom panels are analogous but assume $\sigma^2 = 5$.

speed of service from two firms will conclude that the more variable firm is slower even when they are, on average, the same.

We now more formally consider how a misattributor's long-run bias can harm her decisions. Let $v(\tilde{\theta}, \sigma) \equiv \int_{-\infty}^{\infty} u(x|\tilde{\theta}) f(x|\tilde{\theta}) dx$ denote the agent's expected (per-period) utility from the prospect assuming she is confident that the mean is $\tilde{\theta}$. Accordingly, $v(\theta, \sigma)$ is the agent's valuation of the prospect if she holds correct beliefs, while $v(\hat{\theta}_\infty, \sigma)$ is the agent's valuation under long-run (mis)learning with misattribution. Proposition 4 implies that $v(\hat{\theta}_\infty, \sigma) \leq v(\theta, \sigma)$: a misattributor tends to undervalue the prospect as a result of learning from experience. Thus, if presented with an alternative known to have value $w \in \mathbb{R}$, the agent will wrongly select that alternative over the prospect whenever $w \in \big(v(\hat{\theta}_\infty, \sigma), v(\theta, \sigma)\big)$. In this sense, the distance $v(\theta, \sigma) - v(\hat{\theta}_\infty, \sigma)$ provides a measure of how much misattribution may harm decisions regarding the prospect.

Building on our result that $\hat{\theta}_\infty$ is decreasing in the variance of outcomes (Proposition 4, Part 2), our next result reveals that $v(\theta, \sigma) - v(\hat{\theta}_\infty, \sigma)$ is unboundedly increasing in σ. To highlight the choice implications of this, we consider two prospects that would be valued the same under rational learning, and show that the extent to which the misattributor undervalues the riskier of the two can be arbitrarily large. Let $\hat{\theta}_\infty(\theta, \sigma)$ denote the steady-state belief in terms of the underlying distributional parameters.

Proposition 5. *Consider $\lambda > 1$ and fix θ' and σ'. Consider any alternative parameters $\theta > \theta'$ and $\sigma > \sigma'$ such that $v(\theta, \sigma) = v(\theta', \sigma')$. Then $v\big(\hat{\theta}_\infty(\theta, \sigma), \sigma\big)$ is strictly decreasing in σ with* $\lim_{\sigma \to \infty} v\big(\hat{\theta}_\infty(\theta, \sigma), \sigma\big) = -\infty$.

As detailed in the proof, this result stems from the fact that (i) $v(\hat{\theta}, \sigma)$ is linear in both $\hat{\theta}$ and σ given the agent's reference-dependent preferences, and (ii) the magnitude of underestimation, $\theta - \hat{\theta}_\infty(\theta, \sigma)$, is increasing in σ but independent of θ.

The proposition suggests an excessive bias against risk. Consider two prospects, A and B, such that A has a higher mean and variance than B yet the two yield the same expected utility under correct beliefs. Under a misattributor's steady-state beliefs, she wrongly expects A to yield a lower utility than B, and this discrepancy can be arbitrarily large when the variance of A is sufficiently high. Such biased learning may help explain why individuals tend to excessively avoid risk based on their personal experiences, as shown in the field by Malmendier and Nagel (2011) and more directly in the lab by Shin (2021).

Both Proposition 3 and 5 highlight a central implication of misattribution: a tendency to abandon worthwhile prospects when learning from experience. In the short-run, overly volatile beliefs may cause the misattributor to abandon a prospect too quickly (Proposition 3). And even if she were to experiment indefinitely, her pessimistic long-run expectations can have a similar detrimental effect (Proposition 5).

5. Expectations management

Managing expectations is a common practice in many domains. Politicians often walk down expectations before debates, friends manage hype before introducing others to a new experience, and firms sometimes restrain the quality expectations of consumers or clients.[19] Why are these techniques effective at increasing ex-post impressions? Our model suggests that restraining expectations to subsequently beat them can inflate a misattributor's posterior perceptions. Specifically, we show how and when manipulating (i) prior expectations, (ii) the order of outcomes, or (iii) the outcomes themselves (via budget-neutral transfers) can boost a misattributor's posterior beliefs.

We consider a "designer" who seeks to maximize a misattributing agent's perception of θ (e.g., the average quality or speed of a firm's service) after a fixed number of periods, T. Ex-ante, the designer and agent share a common prior $\theta \sim N(\theta_0, \rho^2)$. For simplicity and ease of exposition, we assume the designer understands the agent's biased updating process, but the agent does not consider the designer's motives to practice expectations management. To justify this assumption, we focus on settings where, if the agent did not suffer misattribution, the designer's actions would have either no influence or a negative influence on beliefs.[20] Accordingly, the misattributor takes outcomes at face value without strategic consideration and treats them as

[19] The idea of firms limiting expectations is commonly acknowledged (e.g., Anderson, 1973). For example, Kopalle and Lehmann (2006) and Ho and Zheng (2004) discuss how firms restrain expectations about product quality and delivery times, respectively.

[20] Since a misattributor is naive about her bias, she believes that her updating process mirrors that of an unbiased agent and thus neglects the designer's incentive to manipulate her. Furthermore, this "designer" need not be a literal actor in the model and can be considered as a rhetorical device: the results below simply demonstrate which sequences of outcomes (or which priors) maximize posterior beliefs, regardless of whether they were actively manipulated or a result of chance.

i.i.d. draws from $N(\theta, \sigma^2)$. Such an agent updates her beliefs precisely as in previous sections, and our applications here directly extend our earlier results.

First, we consider a designer who can manipulate the agent's prior expectation before a single outcome is realized. Suppose that before x_1 is realized, the designer can induce the agent to adopt a belief $\theta \sim N(\theta_0 - c, \rho^2)$ for some $c \geq 0$. This is a simple way to capture the idea of "walking down" the agent's expectations.[21] The designer aims to maximize the agent's expected posterior estimate of θ following x_1. That is, he chooses c to maximize $\mathbb{E}[\hat{\theta}_1] = \alpha_1 \mathbb{E}[\hat{x}_1(c)] + (1 - \alpha_1)\big(\theta_0 - c\big)$, where $\hat{x}_1(c)$ is the agent's encoded value of x_1 when expecting $\theta_0 - c$, and $\mathbb{E}[\cdot]$ is w.r.t. the designer's prior. It is immediate that the designer would prefer $c = 0$ if the agent were unbiased. However, with misattribution, $\hat{x}_1(c)$ is increasing in c and thus walking down expectations may be beneficial. The optimal value of c is decreasing in the relative precision of the induced prior, σ^2/ρ^2. When $\sigma^2/\rho^2 < \kappa^G$, the designer has an incentive to walk down beliefs as much as possible. When $\sigma^2/\rho^2 > \kappa^G$, the optimal value of c decreases in σ^2/ρ^2 until it reaches zero at $\sigma^2/\rho^2 = (\kappa^G + \kappa^L)/2$, and it remains there for all higher values of σ^2/ρ^2. (See Appendix B.4 for a formal statement and proof.)

The intuition is straightforward. The benefit of walking down expectations comes from framing the outcome as a gain; the cost comes from lowering the prior. As discussed in Section 3, the former effect dominates when the agent's prior is sufficiently uninformative, but any potential benefit diminishes when the prior is more precise. This suggests that the designer may prefer settings with weak priors, as this provides greater scope to manipulate beliefs.

Even when the prior is relatively precise and thus simply "walking down" initial expectations is not effective (i.e. $\sigma^2/\rho^2 > \kappa^L$), misattribution still enables alternative tactics for persuasion when there are multiple periods. Consider a setting in which the designer can arrange the order of some fixed set of $T \geq 2$ outcomes. Building on the recency bias in Proposition 1, we show that a misattributor in this setting exhibits an *increasing-order bias*: her perception of θ following a fixed set of outcomes is the highest when she experiences those outcomes in an increasing order.

Proposition 6. *Consider any set of T distinct outcomes, $\mathcal{X}$. If $\sigma^2/\rho^2 > \kappa^L$, then among all possible orderings of the outcomes in $\mathcal{X}$, the misattributor's posterior expectation $\hat{\theta}_T$ is highest following the sequence in which the elements are ordered from least to greatest.*

This result follows from Proposition 1: since early outcomes influence beliefs less than later outcomes, the designer prefers to restrain early outcomes in order to beat expectations thereafter.[22] Indeed, prior research suggests that increasing sequences lead to higher ex-post impressions. For instance, Ross and Simonson (1991) had participants sample two video games and find that willingness to pay for the pair was significantly higher for those who sampled the better game last. Similarly, Haisley and Loewenstein (2011) demonstrate that advertising promotions that

[21] For example, the designer may offer guidance explicitly intended to lower expectations. Ho and Zheng (2004) discuss how firms may use "maximal delivery time" to temper expectations. Kopalle and Lehmann (2006) offer numerous examples from advertising where firms "undersell" to manage customers' expectations. And in many political settings, surrogates attempt to temper expectations before debates. We abstract from the various ways these beliefs could be induced to simply show that lowering prior expectations (however it is done) can sometimes boost the agent's posterior perceptions. Furthermore, we take the variance of this "induced" prior to be the same as the "unmanipulated" prior, but this is inconsequential for the result below.

[22] Appendix B.5 considers how this result extends for imprecise priors (i.e., $\sigma^2/\rho^2 < \kappa^L$). With $T = 2$, if one outcome beats initial expectations, then the agent's estimate of θ is maximized when she receives the better outcome last, regardless of the underlying parameters.

utilize giveaways are most effective when in increasing order of value; i.e., when they give the highest-value item last.

Of course, firms can often do more than manipulate the order of outcomes. Instead of arranging a *fixed set* of outcomes, how would a designer—subject to some budget constraint—optimally allocate resources across periods? For example, consider a firm with some fixed budget on experiential marketing or a consultant with constrained time to allocate over a month to a client's project. Among all such sequences that sum to a fixed value, B, which one maximizes the misattributor's final expectations? While the rational posterior would be identical under any such sequence, misattribution introduces an incentive to transfer early consumption to later rounds in order to generate higher posterior beliefs.

We derive the belief-optimizing sequences of length T. To add realism, we characterize the optimal sequence subject to a "participation constraint" mandating that the agent's beliefs exceed a floor $\bar{\theta} < 0$ in all periods $t = 1, \dots, T$. This captures, for instance, scenarios where the agent has an outside option and can exit the relationship. Accordingly, let $\mathcal{S}^T(B, \bar{\theta}) \subset \mathbb{R}^T$ denote the set of all outcome sequences of length T that satisfy the constraint $\sum_{t=1}^T x_t \leq B$. To simplify this analysis, we normalize $\theta_0 = 0$ (without loss of generality) and we focus on $B > 0$ (the proof also handles the case of $B < 0$, which we briefly discuss below).

Proposition 7. *Suppose $B > 0$ and consider the sequence $x^* \in \mathcal{S}^T(B, \bar{\theta})$ that maximizes the misattributor's posterior expectation, $\hat{\theta}_T$. If $\sigma^2/\rho^2 > \kappa^L$, then there exists a threshold $\bar{T}$ such that:*

1. *If $T \leq \bar{T}$, then $x_1^* = \cdots = x_{T-1}^* = 0$ and $x_T^* = B$.*
2. *If $T > \bar{T}$, then $x_1^* < \bar{\theta}$, $x_2^* = \cdots = x_{T-1}^* = \bar{\theta}$, and $x_T^* > B$.*

Proposition 7 highlights how concentrating gains at the end of an episode—rather than providing them incrementally—can boost final perceptions: the optimal sequence delivers a single gain, and does so at the end. For instance, a consulting firm may wish to restrain a client's expectations up front, and deliver results in one fell swoop at the end. Intuitively, providing gains early requires the designer to uphold this standard in subsequent rounds to avoid perception-harming losses. This is inefficient: maintaining this standard is costly, and it does nothing to inflate the agent's perceptions once she has come to expect it. Moreover, it is worth noting that the sequences in Proposition 7 lead a misattributing agent to reach a higher final belief than an unbiased one. Since misattribution tends to generate pessimistic beliefs under "unmanaged" learning (Section 4), this suggests that the optimal sequencing can overcome that negative bias.[23]

The path of the optimal sequence depends on the number of periods, T. With few ($T \leq \bar{T}$), it is optimal to maintain initial expectations; with more ($T > \bar{T}$), it is optimal to induce an initial loss—temporarily lowering beliefs to the floor $\bar{\theta} < 0$—in order to increase the final gain. Intuitively, the threshold $\bar{T}$ is increasing in loss aversion. If the horizon is short, an initial loss

[23] The optimal sequence provides a single gain at the end because the designer implicitly has linear costs in generating outcomes for the agent. This result would become less stark if the mapping from the designer's resources to the agent's consumption utility were concave, since the designer would have a strict preference to smooth outcomes when facing an unbiased agent. That is, if the designer could choose the sequence $(y_1, \dots, y_T)$ subject to $\sum_{t=1}^T y_t \leq B$ and $x_t = m(y_t)$ for some concave, increasing function m, then the designer would have a strict preference for $y_t = B/T$ for all t when facing an unbiased agent. With a misattributing agent, the designer would instead choose a sequence that generates an increasing profile of x_t, even though it yields a lower sum $\sum_{t=1}^T x_t$ than the smoothed profile.

still looms large in the agent's final beliefs, and it is thus optimal to avoid it. However, if the horizon is longer (or if λ is close to 1), then the designer has an incentive to drop beliefs early, thereby "banking" surplus in each subsequent round that can be used to surprise the agent later. This is optimal whenever the recency bias in Proposition 1 outweighs the overestimated loss in the first period. In this way, Proposition 7 highlights that actively walking down expectations is a robust implication of misattribution: it can prove effective even when posteriors are increasing in prior beliefs.[24]

The collection of results above speak to forms of expectations management used in diverse settings ranging from politics to marketing. While restraining expectations naturally lowers perceptions in many models, our model provides an intuition for why this simple tactic can effectively *boost* them. Relatedly, our results offer a warning against inflating expectations. This accords with evidence from Adhvaryu et al. (2020), who examine a field experiment in which an NGO improved workers' housing conditions in India. These improvements were modest but fell short of what was originally promised. Perhaps surprisingly, the authors find that workers who knew the original plans ahead of time perceived their conditions as worse than workers who were neither told about nor provided with any improvements at all. In this way, our results suggest the familiar adage: under promise and over deliver.

6. Discussion and conclusion

In this section, we further discuss and justify our modeling decisions. We conclude by providing some guidance for future empirical tests and extensions.

Interpretation of Misencoded Outcomes. In Section 2, we propose an intuitive interpretation for why outcomes are misencoded: the agent observes her total experienced utility following outcome x_t but does not directly observe x_t, and attribution bias then leads her to misinfer x_t from u_t. In settings where a person is learning about her tastes—e.g., how much she enjoys an unfamiliar product, a new job, etc.—it seems natural to assume that x_t is not separately observed or readily quantifiable. However, we suspect that misattribution may occur even when outcomes are salient in the moment.[25] In such settings, misattribution may happen in retrospect: the agent fails to parse the outcome from her overall memory of the experience. For instance, a purchase that was surprisingly expensive may be remembered as more costly than it really was.

Assumptions on the Agent's Forecasted Utility. Throughout the paper, we have assumed the agent forecasts her utility according to her true gain-loss parameter, η. We believe this is reasonable under the interpretation that misattribution is retrospective. Alternatively, one could assume the agent wrongly forecasts her utility as well, in which case her forecasted utility would substitute η in $u(x|\hat{\theta})$ with $\hat{\eta}$ (see Equation (2)). Our results directly related to actions do not depend on this distinction. Specifically, Proposition 5 holds for either assumption. Likewise, the interpretation of Proposition 3 remains the same so long as the biased and rational agents exhibit the same ex-ante risk attitudes.

Restriction to Normally-Distributed Outcomes and Priors. Our model assumes Gaussian outcomes and priors primarily to streamline the exposition and analysis. Many of our qualitative

[24] As shown in the proof of Proposition 7, this holds even when $B < 0$, which corresponds to the case where the sum of outcomes falls below the agent's prior. When $B < 0$, there exists $\tilde{T} > \bar{T}$ such that $T > \tilde{T}$ implies that the second sequence in Proposition 7 is optimal. When $T \leq \tilde{T}$, it is optimal to smooth losses across periods, forgoing a final gain.

[25] Indeed, in a principal-agent experiment on attribution bias, Brownback and Kuhn (2019) find that principals wrongly attribute luck to an agent's effort even when the agent's effort is perfectly observable.

results are more general. For example, Observations 1 and 2 clearly hold for any distributional assumptions, and Observation 3 only requires that the agent's expectation is increasing in the most recent outcome. Regarding the belief dynamics presented in Section 3, analogous results arise when the outcome and prior distributions are symmetric and quasi-concave, which guarantees that a rational agent's updated estimate of θ falls between her previous estimate and the most recent observation (see, e.g., Chambers and Healy, 2012 for details). Additionally, while we leverage our Gaussian assumptions in Section 4 to establish convergence, our comparative statics on the steady-state belief hold more generally.

Specification of the Gain-Loss Utility Function. We make two important assumptions about the form of reference dependence. First, for tractability, we abstract from other elements of prospect theory (i.e., diminishing sensitivity and probability weighting) and focus on a linear gain-loss function. Second, and perhaps more substantively, we assume the agent's reference point is her recent expectation about consumption utility given our focus on learning from experience. In these settings, expectations based on past outcomes seem like a natural point of comparison.[26] That said, the simple intuition of discomfirmation as highlighted in Observation 1 easily extends to alternative definitions of the reference point. Moreover, while there are several ways to model expectations-based reference points, we adopt the specification in Equation (1) primarily for tractability.[27] (Appendix C.1 provides an additional discussion on extending our model to situations where the agent's planned actions form her reference point, as in Kőszegi and Rabin, 2007.)

Concluding Thoughts. In this paper, we develop a model of learning under attribution bias and study how a novel form of this bias distorts learning from experience. This presents an opportunity for empirical work and additional extensions.

A natural avenue for empirical exploration is our prediction of belief-based contrast effects: a fixed outcome will seem better when contrasted against lower expectations. In sequential settings, we predict that contrast effects will increase in the perceived correlation between the previous outcome and the current one. Furthermore, in order to separate effects generated by our mechanism from other potential explanations—e.g. the gambler's fallacy (Chen et al., 2016)—our model suggests comparing circumstances where outcomes have utility consequences with those that don't. We predict that contrast effects will be enhanced the more that a person cares about the outcomes. This empirical strategy can also help distinguish our mechanism from other models that predict recency effects, such as base-rate neglect (e.g. Benjamin et al., 2019) or the representativeness heuristic (e.g. Bordalo et al., 2017, 2020).

As noted above, a basic feature of our model is that an agent learns differently from outcomes with utility consequences relative to other forms of information. Thus, our model may offer empirical guidance to a literature demonstrating overreaction to personal experience. Some researchers suggest that these experience effects arise from endogenous preference formation in response to good or bad outcomes (e.g., Thaler and Johnson, 1990; Dillenberger and Rozen,

[26] Several experimental studies find evidence of expectations-based reference points, though the totality of evidence is mixed (for example, favoring expectations-based reference points are Abeler et al., 2011; Ericson and Fuster, 2011; Gill and Prowse, 2012; Banerji and Gupta, 2014; Karle et al., 2015; against are Heffetz and List, 2014; Gneezy et al., 2017; Goette et al., 2019). There is additional evidence of expectations-based reference points from the field, spanning labor supply (Crawford and Meng, 2011; Thakral and To, 2021), domestic violence (Card and Dahl, 2011), and decisions in game shows and sports (Post et al., 2008; Pope and Schweitzer, 2011).

[27] Kőszegi and Rabin (2006) provide a well-known alternative specification that depends on the agent's full subjective beliefs rather than the mean belief. Our approach is far more tractable and delivers similar insights. The working version of our paper shows that our long-run results qualitatively extend under this alternative.

2015; Imas, 2016), while others suggest these effects stem from beliefs that overreact to personal experience (e.g., Malmendier and Nagel, 2011, 2016). Our model highlights that these two channels may be intertwined.

Future work could also examine the boundary between confirmation bias and misattribution. In fact, our modeling approach can accommodate a version of confirmation bias. If an agent's mistaken model posits $\hat{\eta} > \eta$ instead of $\hat{\eta} < \eta$—that is, she overcompensates for reference dependence when inferring from experienced utility—then her updating process will exhibit patterns akin to confirmation bias. Rather than exaggerating surprises, her encoded outcomes will be biased toward her current expectations. Thus, if an empirical estimate of our model suggests that $\hat{\eta} > \eta$, then confirmation bias likely dominates in that particular setting.

Finally, the form of attribution bias we consider may naturally extend to interpersonal contexts. For instance, a person reading online product reviews may fail to appreciate that a bad rating sometimes reflects the reviewer's high expectations rather than poor quality. In scenarios where consumers form their expectations based on predecessors' reviews, misattribution—that is, taking others' ratings at face value without accounting for their discrepant expectations—can hinder social learning. Additionally, these settings may provide data-rich environments to explore the empirical implications of our model. If this social misattribution occurs, we would expect ratings to demonstrate the dynamic patterns described in this paper.

Appendix A. Proofs of results in the main text

Proof of Lemma 1. The proof follows from induction on t. Note that for any $t \geq 1$ and prior estimate $\hat{\theta}_{t-1}$, a misattributor's updated belief is $\hat{\theta}_t = \alpha_t \hat{x}_t + (1-\alpha_t)\hat{\theta}_{t-1}$. From Equation (4), $\hat{x}_t = x_t + \kappa_t(x_t - \hat{\theta}_{t-1})$ where $\kappa_t = \kappa^G \mathbb{1}\{x_t > \hat{\theta}_{t-1}\} + \kappa^L \mathbb{1}\{x_t < \hat{\theta}_{t-1}\}$, and thus

$$\hat{\theta}_t = \alpha_t(1+\kappa_t)x_t + [1-\alpha_t(1+\kappa_t)]\hat{\theta}_{t-1}. \tag{A.1}$$

Turning to the induction argument, first consider the base case: since $\hat{\theta}_0 = \theta_0$, Equation (A.1) implies $\hat{\theta}_1 = \alpha_1(1+\kappa_1)x_1 + [1-\alpha_1(1+\kappa_1)]\theta_0$. Thus, letting $\beta_1^1 = (1+\kappa_1)$ and $\beta_0^1 = 1 - \alpha_1(1+\kappa_1)$ establishes the base case. Now suppose the claim holds for period $t > 1$. Substituting the expression for $\hat{\theta}_t$ implied by the claim into Equation (A.1) implies

$$\hat{\theta}_{t+1} = \alpha_{t+1}(1+\kappa_{t+1})x_{t+1} + [1-\alpha_{t+1}(1+\kappa_{t+1})]\left(\beta_0^t\theta_0 + \alpha_t \sum_{\tau=1}^{t} \beta_\tau^t x_\tau\right), \tag{A.2}$$

where $\beta_t^t = (1+\kappa_t)$, $\beta_\tau^t = (1+\kappa_\tau)\prod_{j=\tau}^{t-1}[1-\alpha_j\kappa_{j+1}]$ for $\tau = 1,\dots,t-1$, and $\beta_0^t = \prod_{j=1}^{t}[1-\alpha_j(1+\kappa_j)]$. Using the fact that $\alpha_t = \alpha_{t+1}/(1-\alpha_{t+1})$, it follows that $\alpha_t[1-\alpha_{t+1}(1+\kappa_{t+1})] = \alpha_{t+1}[1-\alpha_t\kappa_{t+1}]$. Thus, for all $\tau \geq 1$, we have $\alpha_t[1-\alpha_{t+1}(1+\kappa_{t+1})]\beta_\tau^t = \alpha_{t+1}(1+\kappa_\tau)\prod_{j=\tau}^{t}[1-\alpha_j\kappa_{j+1}]$. The expression in Equation (A.2) can therefore be written as

$$\begin{aligned}\hat{\theta}_{t+1} &= [1-\alpha_{t+1}(1+\kappa_{t+1})]\beta_0^t\theta_0 \\ &\quad + \alpha_{t+1}\left((1+\kappa_{t+1})x_{t+1} + \sum_{\tau=1}^{t}\left\{(1+\kappa_\tau)\prod_{j=\tau}^{t}[1-\alpha_j\kappa_{j+1}]\right\}x_\tau\right).\end{aligned} \tag{A.3}$$

Define $\beta_0^{t+1} = \prod_{j=1}^{t+1}[1-\alpha_j(1+\kappa_j)]$, $\beta_{t+1}^{t+1} = (1+\kappa_{t+1})$, and $\beta_\tau^{t+1} = (1+\kappa_\tau)\prod_{j=\tau}^{t}[1-\alpha_j\kappa_{j+1}]$ for $\tau = 1,\dots,t$. Equation (A.3) then reduces to $\hat{\theta}_{t+1} = \beta_0^{t+1}\theta_0 + \alpha_{t+1}\sum_{\tau=1}^{t+1}\beta_\tau^{t+1}x_\tau$, which verifies the induction step and completes the proof. □

Proof of Proposition 1. *Part 1.* From Lemma 1, we have $\beta_t^t = (1+\kappa_t) > 1$ and $\beta_\tau^t = (1+\kappa_\tau)\prod_{j=\tau}^{t-1}[1-\alpha_j\kappa_{j+1}]$. To show the latter term converges to zero in t, we must consider two cases. First, suppose that $\tau < t^*$. We can then write $\beta_\tau^t = (1+\kappa_\tau)C\prod_{j=t^*}^{t-1}[1-\alpha_j\kappa_{j+1}]$, where $C \equiv \prod_{j=\tau}^{t^*-1}[1-\alpha_j\kappa_{j+1}]$ is necessarily finite valued. Thus, $\lim_{t\to\infty}\beta_\tau^t = (1+\kappa_\tau)C\lim_{t\to\infty}\prod_{j=t^*}^{t-1}[1-\alpha_j\kappa_{j+1}]$. Note that for all $j \geq t^*$, we have $1-\alpha_j\kappa_{j+1} \in (0,1)$. Hence, $1-\alpha_j\kappa_{j+1} \leq 1-\alpha_j\kappa^G$, and thus $\lim_{t\to\infty}|\beta_\tau^t| \leq (1+\kappa_\tau)|C|\lim_{t\to\infty}\prod_{j=\tau}^{t-1}[1-\alpha_j\kappa^G]$. Since $\sum_{j=t^*}^{\infty}\alpha_j$ diverges, $\prod_{j=t^*}^{\infty}[1-\alpha_j\kappa^G]=0$. Thus, $\lim_{t\to\infty}|\beta_\tau^t|=0$. Second, suppose that $\tau \geq t^*$. Then $\lim_{t\to\infty}\beta_\tau^t = (1+\kappa_\tau)\lim_{t\to\infty}\prod_{j=\tau}^{t-1}[1-\alpha_j\kappa_{j+1}]$. Since $\tau \geq t^*$, the argument above implies $\lim_{t\to\infty}\prod_{j=\tau}^{t-1}[1-\alpha_j\kappa_{j+1}]=0$, and thus $\lim_{t\to\infty}\beta_\tau^t=0$, completing the proof of Part 1.

Part 2. Consider $\tau_2 > \tau_1 \geq t^*$. Lemma 1 then implies that $\beta_{\tau_2}^t - \beta_{\tau_1}^t = (1-\kappa_{\tau_2})\prod_{j=\tau_2}^{t-1}[1-\alpha_j\kappa_{j+1}] - (1-\kappa_{\tau_1})\prod_{j=\tau_1}^{t-1}[1-\alpha_j\kappa_{j+1}] = (1-\kappa_{\tau_2})\prod_{j=\tau_2}^{t-1}[1-\alpha_j\kappa_{j+1}]\Big(1-\prod_{j=\tau_1}^{t-1}[1-\alpha_j\kappa_{j+1}]\Big) = \beta_{\tau_2}^t\left(1-\prod_{j=\tau_1}^{t-1}[1-\alpha_j\kappa_{j+1}]\right)$, where the second equality follows from our assumption that $\kappa_{\tau_2}=\kappa_{\tau_1}$. Furthermore, since $\tau_1 \geq t^*$, we have $1-\alpha_j\kappa_{j+1}\in(0,1)$ for all $j \geq \tau_1$. Thus, $1-\prod_{j=\tau_1}^{t-1}[1-\alpha_j\kappa_{j+1}]\in(0,1)$ and hence $\beta_{\tau_2}^t - \beta_{\tau_1}^t > 0$ given that $\tau_2 > t^*$ implies $\beta_{\tau_2}^t > 0$. □

Proof of Proposition 2. Condition on $\hat{\theta}_{t-1}$, this value will not influence the covariances under consideration, and we therefore normalize $\hat{\theta}_{t-1}$ to equal zero without loss of generality. From Equation (7), we thus have

$$\begin{aligned}\text{Cov}\left(\hat{\theta}_t,\hat{\theta}_{t+1}\right) &= \text{Cov}\left(\hat{\theta}_t, \alpha_{t+1}(1+\kappa_{t+1})x_{t+1} + [1-\alpha_{t+1}(1+\kappa_{t+1})]\hat{\theta}_t\right)\\ &= \alpha_{t+1}\text{Cov}\left(\hat{\theta}_t,(1+\kappa_{t+1})x_{t+1}\right) + \text{Cov}\left(\hat{\theta}_t,[1-\alpha_{t+1}(1+\kappa_{t+1})]\hat{\theta}_t\right).\end{aligned} \tag{A.4}$$

Note that $t < \kappa^G - \sigma^2/\rho^2$ implies that $[1-\alpha_{t+1}(1+\kappa^G)] < 0$. Thus, $[1-\alpha_{t+1}(1+\kappa_{t+1})] < 0$ for any $\kappa_{t+1}\in\{\kappa^G,\kappa^L\}$ and $\text{Cov}(\hat{\theta}_t,[1-\alpha_{t+1}(1+\kappa_{t+1})]\hat{\theta}_t) < [1-\alpha_{t+1}(1+\kappa^G)]\text{Var}(\hat{\theta}_t) < 0$. Hence, to complete the proof it suffices to show that $\alpha_{t+1}\text{Cov}\big(\hat{\theta}_t,(1+\kappa_{t+1})x_{t+1}\big)$ is sufficiently small. Note that if $\kappa^L=\kappa^G$, then $(1+\kappa_{t+1})x_{t+1}$ is independent of $\hat{\theta}_t$ (since x_{t+1} is independent of $\hat{\theta}_t$ conditional on θ) and $\text{Cov}\big(\hat{\theta}_t,(1+\kappa_{t+1})x_{t+1}\big)=0$, verifying the claim of the proposition. Thus, we must further verify that $\text{Cov}\big(\hat{\theta}_t,(1+\kappa_{t+1})x_{t+1}\big)$ remains sufficiently small when $\kappa^L > \kappa^G$.

To show that the right-hand side of Equation (A.4) is negative, we expand both terms and show that any positive components from the first term are completely offset by negative components from the second term. We begin by expanding the first term. Applying Equation (7) again implies

$$\begin{aligned}\alpha_{t+1}\text{Cov}\left(\hat{\theta}_t,(1+\kappa_{t+1})x_{t+1}\right) &= \alpha_{t+1}\text{Cov}\left(\alpha_t(1+\kappa_t)x_t,(1+\kappa_{t+1})x_{t+1}\right)\\ &= \alpha_t\alpha_{t+1}\big(\text{Cov}(x_t,\kappa_{t+1}x_{t+1}) + \text{Cov}(\kappa_t x_t,\kappa_{t+1}x_{t+1})\big),\end{aligned} \tag{A.5}$$

where the second equality follows from the fact that, conditional on θ, x_t and $\kappa_t x_t$ are both independent of x_{t+1}. It will be useful to write κ_t as follows: $\kappa_t = \kappa^G + \delta_\kappa L_t$ where $\delta_\kappa \equiv \kappa^L - \kappa^G > 0$ and $L_t \equiv \mathbb{1}\{x_t < \hat{\theta}_{t-1}\}$. Using this notation, along with the conditional-independence fact used immediately above, we have $\text{Cov}(x_t,\kappa_{t+1}x_{t+1}) = \delta_\kappa\text{Cov}(x_t,L_{t+1}x_{t+1})$ and

$$\text{Cov}(\kappa_t x_t,\kappa_{t+1}x_{t+1}) = \kappa^G\delta_\kappa\text{Cov}(x_t,L_{t+1}x_{t+1}) + \delta_\kappa^2\text{Cov}(L_t x_t,L_{t+1}x_{t+1}). \tag{A.6}$$

Hence, the first term on the right-hand side of Equation (A.4) is equal to

$$\alpha_t\alpha_{t+1}\big((1+\kappa^G)\delta_\kappa \mathrm{Cov}(x_t, L_{t+1}x_{t+1}) + \delta_\kappa^2 \mathrm{Cov}(L_t x_t, L_{t+1}x_{t+1})\big). \tag{A.7}$$

We will now similarly expand the second term on the right-hand-side of Equation (A.4). Using $\hat{\theta}_t = \alpha_t(1+\kappa_t)x_t$ along with the notation introduced above, this term is equal to

$$\alpha_t^2\Bigg((1+\kappa^G)M\mathrm{Var}(x_t) + M\delta_\kappa^2\mathrm{Var}(L_t x_t) + 2(1+\kappa^G)M\delta_\kappa\mathrm{Cov}(x_t, L_t x_t)$$
$$-\alpha_{t+1}(1+\kappa^G)^2\delta_\kappa\mathrm{Cov}(x_t, L_t x_t) - \alpha_{t+1}(1+\kappa^G)\delta_\kappa^2\mathrm{Cov}(x_t, L_t L_{t+1}x_t)$$
$$-\alpha_{t+1}(1+\kappa^G)\delta_\kappa^2\mathrm{Cov}(L_t x_t, L_{t+1}x_t) - \alpha_{t+1}\delta_\kappa^3\mathrm{Cov}(L_t x_t, L_t L_{t+1}x_t)\Bigg), \tag{A.8}$$

where $M \equiv [1-\alpha_{t+1}(1+\kappa^G)] < 0$. It is immediate that each covariance in expression (A.8) is positive, and $M < 0$ therefore implies every term in (A.8) is negative. To complete the proof, we will show that the absolute value of the fourth (sixth) term of (A.8) is weakly larger than the first (second) term of expression (A.7). This will ensure that Equation (A.4) is negative, completing the proof.

Toward that end, note that conditional on θ, $\mathrm{Cov}(x_t, L_{t+1}x_t) \geq \mathrm{Cov}(x_t, L_{t+1}x_{t+1})$ since x_t and x_{t+1} are independent. Thus, to show that the absolute value of fourth term of (A.8) is weakly larger than the first term of (A.7), it suffices to show that $\alpha_t^2\alpha_{t+1}(1+\kappa^G)^2\delta_\kappa > \alpha_t\alpha_{t+1}(1+\kappa^G)\delta_\kappa \Leftrightarrow \alpha_t(1+\kappa^G) > 1$, which is true given that $[1-\alpha_{t+1}(1+\kappa^G)] < 0$ and $\alpha_t > \alpha_{t+1}$. Similarly, $\mathrm{Cov}(L_t x_t, L_{t+1}x_t) \geq \mathrm{Cov}(L_t x_t, L_{t+1}x_{t+1})$, which implies that the absolute value of sixth term of (A.8) is weakly larger than the second term of (A.7) if $\alpha_t^2\alpha_{t+1}(1+\kappa^G)\delta_\kappa^2 > \alpha_t\alpha_{t+1}\delta_\kappa^2 \Leftrightarrow \alpha_t(1+\kappa^G) > 1$, which is the same condition as above, and therefore holds. □

Proof of Proposition 3. Following any sequence of outcomes $(x_1, \ldots, x_t) \in \mathbb{R}^t$, let $\hat{\theta}_t$ and θ_t denote a misattributor's and a rational Bayesian's estimate of θ, respectively. Without loss of generality, let $\theta_0 = 0$. We first establish the following lemma.

Lemma A.1. *Suppose $t^* < 1$. Consider any period $t \geq 1$ and suppose that for all $i < t$, both $\hat{\theta}_i \geq \bar{\theta}$ and $\theta_i \geq \bar{\theta}$. If $\theta_t < \bar{\theta}$, then $\hat{\theta}_t < \bar{\theta}$.*

Proof of Lemma A.1: Assume $\theta_0 = 0$. Since $\theta_t = \alpha_t[x_t - \theta_{t-1}] + \theta_{t-1}$ and $\theta_{t-1} \geq \bar{\theta}$, it follows that $\theta_t < \bar{\theta}$ implies $x_t < \bar{\theta}$. Thus, $x_t < \hat{\theta}_{t-1}$ and x_t is encoded as a loss, so $\kappa_t = \kappa^L$. Hence, Equation (7) implies $\hat{\theta}_t = \alpha_t(1+\kappa^L)[x_t - \hat{\theta}_{t-1}] + \hat{\theta}_{t-1}$. We want to show $\hat{\theta}_t \leq \theta_t \Leftrightarrow (1-\alpha_t)[\hat{\theta}_{t-1} - \theta_{t-1}] \leq \alpha_t\kappa^L[\hat{\theta}_{t-1} - x_t] \Leftrightarrow$

$$[\hat{\theta}_{t-1} - \theta_{t-1}] \leq \alpha_{t-1}\kappa^L[\hat{\theta}_{t-1} - x_t], \tag{A.9}$$

where the final condition uses the fact that $\frac{\alpha_t}{1-\alpha_t} = \alpha_{t-1}$. As noted above, the right-hand side of (A.9) is positive. Our claim therefore follows immediately if $\hat{\theta}_{t-1} \leq \theta_{t-1}$; we thus assume $\hat{\theta}_{t-1} > \theta_{t-1}$ henceforth and show that the gap between these two beliefs is bounded as required by (A.9).

Let $B \equiv \sum_{i=1}^{t-1} x_i$. Note that Proposition 7 and its proof derive the upper bound on $\hat{\theta}_{t-1}$ (subject to $\sum_{i=1}^{t-1} x_i = B$ and $\hat{\theta}_i \geq \bar{\theta}$ for all $i \leq t-1$). Since $\theta_{t-1} = \alpha_{t-1}B$, we can therefore use those results to establish an upper bound on $\hat{\theta}_{t-1} - \theta_{t-1}$ as a function of θ_{t-1}, and then use that bound to verify that Condition (A.9) holds. Accordingly, the steps below will use terminology and results from the proof of Proposition 7. As such, we separately consider the case of $B > 0$ and $B < 0$ as we do there.

Case 1: $B > 0$. The proof of Proposition 7 shows $\hat{\theta}_{t-1}$ achieves its highest value (subject to the relevant constraints) following either (a) the "final gain" sequence, or (b) the "loss-gain" sequence (see that proof for details). We consider each potential maximal sequence in turn, showing that Condition (A.9) necessarily holds for either.

Case 1.a. Suppose the "final gain" sequence maximizes $\hat{\theta}_{t-1}$ (subject to the relevant constraints), which leads to a maximal value of $\hat{\theta}_{t-1}$ equal to $\alpha_{t-1}(1+\kappa^G)B = (1+\kappa^G)\theta_{t-1}$ (as argued above Equation (A.52)). Thus, $\hat{\theta}_{t-1} - \theta_t \le \kappa^G\theta_{t-1}$. Substituting this upper bound in the left-hand side of Condition (A.9), it suffices to show:

$$\kappa^G\theta_{t-1} \le \alpha_{t-1}\kappa^L[\hat{\theta}_{t-1} - x_t] \;\Leftrightarrow\; \frac{\kappa^G}{\kappa^L}\theta_{t-1} \le \alpha_{t-1}[\hat{\theta}_{t-1} - x_t]. \tag{A.10}$$

Since $B > 0$ implies that $\theta_{t-1} > 0$ and since $\kappa^G/\kappa^L \in (0, 1]$, Condition (A.10) holds if $\theta_{t-1} \le \alpha_{t-1}[\hat{\theta}_{t-1} - x_t]$. Since $\hat{\theta}_{t-1} > \theta_{t-1}$, the previous condition holds if $\theta_{t-1} \le \alpha_{t-1}[\theta_{t-1} - x_t] \Leftrightarrow \theta_{t-1} + \alpha_{t-1}(x_t - \theta_{t-1}) \le 0$. This inequality strictly holds because $(x_t - \theta_{t-1}) < 0$ together with $\alpha_{t-1} > \alpha_t$ implies that $\theta_{t-1} + \alpha_{t-1}(x_t - \theta_{t-1}) < \theta_t$, and, by assumption, $\theta_t < \bar{\theta} < 0$.

Case 1.b. Suppose the "loss-gain" sequence maximizes $\hat{\theta}_{t-1}$ (subject to the appropriate constraints), which (as shown in Equation (A.51)) implies that the maximal value of $\hat{\theta}_{t-1}$ is equal to $\alpha_{t-1}(1+\kappa^G)\left[B - \bar{\theta}\left(t - 2 + 1/\alpha_1(1+\kappa^L)\right)\right] + \bar{\theta}$, and thus

$$\hat{\theta}_{t-1} \le (1+\kappa^G)\theta_{t-1} + Q(t), \tag{A.11}$$

where $Q(t) \equiv \alpha_{t-1}(1+\kappa^G)|\bar{\theta}|\left(t - 2 + \frac{1}{\alpha_1(1+\kappa^L)}\right) + \bar{\theta}$. To show that Condition (A.9) holds, first note that we can rewrite it as

$$[1 - \alpha_{t-1}\kappa^L]\hat{\theta}_{t-1} \le \theta_{t-1} - \alpha_{t-1}\kappa^L x_t. \tag{A.12}$$

Since x_t induces $\theta_t < \bar{\theta}$, we have $x_t < \bar{x}$ where $\bar{x}$ is defined implicitly by $\alpha_t\bar{x} + (1-\alpha_t)\theta_{t-1} = \bar{\theta} \Leftrightarrow \bar{x} = \bar{\theta}/\alpha_t - \theta_{t-1}/\alpha_{t-1}$. Note that Condition (A.12) necessarily holds if it holds at $x_t = \bar{x}$ and at the maximal value of $\hat{\theta}_{t-1}$; substituting these values into (A.12) yields the following sufficient condition:

$$[1 - \alpha_{t-1}\kappa^L]Q(t) \le \Big((1+\kappa^L) - (1+\kappa^G)[1 - \alpha_{t-1}\kappa^L]\Big)\theta_{t-1} - \frac{\alpha_{t-1}}{\alpha_t}\kappa^L\bar{\theta}. \tag{A.13}$$

Note that our assumption of $t^* < 1$ ensures that $1 - \alpha_{t-1}\kappa^L \in (0, 1)$. This implies (along with $B > 0$) that the first term on right-hand side of Condition (A.13) is positive; thus, that condition holds if

$$Q(t) \le \frac{1}{1 - \alpha_{t-1}\kappa^L}\frac{\alpha_{t-1}}{\alpha_t}\kappa^L|\bar{\theta}|. \tag{A.14}$$

The right-hand side of Condition (A.14) necessarily exceeds $\kappa^L|\bar{\theta}|$, and hence it suffices to show $Q(t) \le \kappa^L|\bar{\theta}|$. This is indeed the case, as we can show that $Q(t) < \kappa^G|\bar{\theta}|$ and $\kappa^G \le \kappa^L$. From the definition of $Q(t)$, we have $Q(t) < \kappa^G|\bar{\theta}|$ iff

$$\alpha_{t-1}\left(t - 2 + \frac{1}{\alpha_1(1+\kappa^L)}\right) < 1 \;\Leftrightarrow\; t - 2 + \frac{1 + \frac{\sigma^2}{\rho^2}}{(1+\kappa^L)} < t - 1 + \frac{\sigma^2}{\rho^2}, \tag{A.15}$$

which must hold given that $\kappa^L > 0$. This completes Case 1.b and hence Case 1.

Case 2: $B < 0$. The proof of Proposition 7 shows that $\hat{\theta}_{t-1}$ achieves its highest value (subject to the appropriate constraints) following either (a) the "loss-gain" sequence, or (b) the "initial loss" sequence. We consider each potential maximal sequence in turn, showing that Condition (A.9) necessarily holds for either.

Case 2.a. Suppose the "loss-gain" sequence maximizes $\hat{\theta}_{t-1}$ (subject to the appropriate constraints). The arguments of Case 1.b apply here up to Condition (A.13), and thus we aim to show that this condition holds when $B < 0$. Since $\theta_{t-1} < \hat{\theta}_{t-1}$ by assumption, we must have $\theta_{t-1} < (1+\kappa^G)\theta_{t-1} + Q(t)$ since the right-hand side of the previous inequality is the maximal value of $\hat{\theta}_{t-1}$. Since $B < 0$ implies that $\theta_{t-1} < 0$, we thus have $|\theta_{t-1}| < Q(t)/\kappa^G$. This upper bound implies that the following condition is sufficient for Condition (A.13):

$$[1-\alpha_{t-1}\kappa^L]Q(t) \le -\left((1+\kappa^L)-(1+\kappa^G)[1-\alpha_{t-1}\kappa^L]\right)\frac{Q(t)}{\kappa^G} - \frac{\alpha_{t-1}}{\alpha_t}\kappa^L\bar{\theta}, \qquad (A.16)$$

which reduces to

$$\frac{(1+\alpha_{t-1})\kappa^L}{\kappa^G}Q(t) \le \frac{\alpha_{t-1}}{\alpha_t}\kappa^L|\bar{\theta}| \;\Leftrightarrow\; Q(t) \le \kappa^G|\bar{\theta}|, \qquad (A.17)$$

where the second inequality follows from the identity $\alpha_{t-1}/(1+\alpha_{t-1}) = \alpha_t$. Finally, the discussion around Condition (A.15) shows that $Q(t) \le \kappa^G|\bar{\theta}|$, and hence Condition (A.17) must hold.

Case 2.b. Suppose the "initial loss" sequence maximizes $\hat{\theta}_{t-1}$ (subject to the appropriate constraints), which (as shown in Equation (A.54)) implies the following bound on $\hat{\theta}_{t-1}$:

$$\hat{\theta}_{t-1} \le \frac{\alpha_1(1+\kappa^L)}{1+(t-2)\alpha_1(1+\kappa^L)}B = \frac{\alpha_1(1+\kappa^L)}{1+(t-2)\alpha_1(1+\kappa^L)}\left(\frac{\theta_t}{\alpha_{t-1}}\right). \qquad (A.18)$$

Since $\theta_{t-1} < 0$, Condition (A.18) implies that $\hat{\theta}_{t-1} < \theta_{t-1} \Leftrightarrow \frac{\alpha_1(1+\kappa^L)}{1+(t-2)\alpha_1(1+\kappa^L)} > \alpha_{t-1}$, and the previous inequality always holds given that $\kappa^L > 0$. Thus, we must have $\hat{\theta}_{t-1} \le \theta_{t-1}$ in this case. However, as argued above, our desired result (i.e., $\hat{\theta}_t < \bar{\theta}$) follows immediately when $\hat{\theta}_{t-1} \le \theta_{t-1}$. This completes Case 2 and the proof of Lemma A.1.

Let $S^B_{\bar{\theta}} = \min\{t \ge 1 | \theta_t < \bar{\theta}\}$ and $S^M_{\bar{\theta}} = \min\{t \ge 1 | \hat{\theta}_t < \bar{\theta}\}$ be stopping times for when the Bayesian and misattributor's belief processes fall below $\bar{\theta}$, respectively. Lemma A.1 shows that if $S^B_{\bar{\theta}} = t$ and $\hat{\theta}_i \ge \bar{\theta}$ for all $i = 1, \ldots, t-1$, then we necessarily have $S^M_{\bar{\theta}} = t$. This implies that $S^M_{\bar{\theta}} \le S^B_{\bar{\theta}}$ along any sample path, and hence this weak inequality holds with probability one. To complete the proof, we must show that $S^M_{\bar{\theta}} < S^B_{\bar{\theta}}$ occurs with positive probability. This can be done by considering beliefs following the first outcome: if $x_1 \in \left(\bar{\theta}/\alpha_1, \bar{\theta}/\alpha_1(1+\kappa^L)\right)$, then $\hat{\theta}_1 < \bar{\theta}$ while $\theta_1 > \bar{\theta}$. The fact that this interval has positive measure completes the proof. □

Proof of Proposition 4. We first prove that beliefs converge to a unique steady-state value.

Step One: unique steady-state belief. Let $\Delta(\hat{\theta})$ denote the deviation between $\hat{\theta}$ and the expected value of the encoded outcome assuming the agent holds expectation $\hat{\theta}$:

$$\Delta(\hat{\theta}) \equiv \mathbb{E}\left[x_t + \kappa_t(x_t - \hat{\theta}_{t-1}) \middle| \hat{\theta}_{t-1} = \hat{\theta}\right] - \hat{\theta}, \qquad (A.19)$$

where $\mathbb{E}[\cdot]$ is with respect to the true data generating process. Note that $\Delta(\hat{\theta})$ does not depend on the value of t since x_t is i.i.d. and Δ takes $\hat{\theta}_{t-1}$ as fixed. Thus, a steady-state belief is defined by $\Delta(\hat{\theta}) = 0$. We now show that there is a unique value of $\hat{\theta}$ solving $\Delta(\hat{\theta}) = 0$, and we denote this value by $\hat{\theta}_\infty$. Notice that $\mathbb{E}\big[x_t + \kappa_t(x_t - \hat{\theta}_{t-1})|\hat{\theta}_{t-1} = \hat{\theta}\big] = \theta + \kappa^G \Pr(x_t \geq \hat{\theta})(\mathbb{E}[x_t|x_t \geq \hat{\theta}] - \hat{\theta}) + \kappa^L \Pr(x_t < \hat{\theta})(\mathbb{E}[x_t|x_t < \hat{\theta}] - \hat{\theta}) = \theta - k \Pr\big(x_t < \hat{\theta}\big)\big(\hat{\theta} - \mathbb{E}[x_t|x_t < \hat{\theta}]\big)$, where

$$k \equiv \frac{\kappa^L - \kappa^G}{1 + \kappa^G} = \frac{(\lambda - 1)(\eta - \hat{\eta})}{(1 + \eta)(1 + \hat{\eta}\lambda)}. \tag{A.20}$$

Thus, a steady-state belief $\hat{\theta}$ solves

$$\Delta(\hat{\theta}) = \theta - kH(\hat{\theta}; \theta, \sigma) - \hat{\theta} = 0, \tag{A.21}$$

where $H(\hat{\theta}; \theta, \sigma) \equiv \Pr(x_t < \hat{\theta})\big(\hat{\theta} - \mathbb{E}[x_t|x_t < \hat{\theta}]\big)$ is a function of both the agent's belief, $\hat{\theta}$, and the true distributional parameters, θ and σ:

$$H(\hat{\theta}; \theta, \sigma) = \hat{\theta}\Phi\left(\frac{\hat{\theta} - \theta}{\sigma}\right) - \int_{-\infty}^{\hat{\theta}} x\frac{1}{\sigma}\phi\left(\frac{x - \theta}{\sigma}\right)dx, \tag{A.22}$$

which follows from the fact that $x_t \sim N(\theta, \sigma^2)$, and thus we can write the CDF and PDF of x_t as $\Phi((x - \theta)/\sigma)$ and $\phi((x - \theta)/\sigma)$, respectively, where Φ and ϕ are the standard-normal CDF and PDF. Note that $H(\hat{\theta}; \theta, \sigma) > 0$ for all finite $\hat{\theta}$, and $H(\hat{\theta}; \theta, \sigma)$ is a strictly increasing and continuous function of $\hat{\theta}$ with $\frac{\partial}{\partial \hat{\theta}} H(\hat{\theta}; \theta, \sigma) = \Phi\big((\hat{\theta} - \theta)/\sigma\big) > 0$. It thus follows that $\Delta(\hat{\theta})$ is a strictly decreasing and continuous function of $\hat{\theta}$ with range $\mathbb{R}$. Thus, there exists a unique, finite value $\hat{\theta}_\infty$ such that $\Delta(\hat{\theta}_\infty) = 0$.

Step Two: almost-sure convergence to the steady-state belief. Let $\hat{\theta}_\infty$ denote the unique steady-state belief that solves Equation (A.21). We now show that the sequence of beliefs $\langle\hat{\theta}_t\rangle$ converges to $\hat{\theta}_\infty$. Our convergence arguments, which rely on stochastic approximation theory, are similar to those in Esponda and Pouzo (2016) and in particular Heidhues, Kőszegi, and Strack (2021). The basic logic is as follows: while encoded outcomes are not independent ($\hat{x}_t$ is a function of $\hat{\theta}_{t-1}$, which depends on $x_1, \ldots, x_{t-1}$), they become approximately independent as t grows large and hence $\hat{\theta}_t$ changes a small amount (on average) in response to any new outcome. As such, a result from stochastic approximation theory implies that the limiting value of $\hat{\theta}_t$ is determined by the zero of Δ (Equation (A.19)), where $\Delta(\hat{\theta})$ is the average deviation of encoded outcomes from the agent's expectation, $\hat{\theta}$.

We now formalize this argument. From Equation (7), the misattributor's beliefs update according to

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \hat{\alpha}_t[x_t - \hat{\theta}_{t-1}], \tag{A.23}$$

where $\hat{\alpha}_t \equiv (1 + \kappa_t)\alpha_t$ and $\alpha_t = 1/(t + \sigma^2/\rho^2)$. This dynamic system is a special case of the one considered in Theorem 5.2.1 of Kushner and Yin (2003), who provide sufficient conditions for the convergence of such a system. Their theorem implies that $\langle\hat{\theta}_t\rangle$ converges almost surely to the unique value $\hat{\theta}_\infty$ characterized by the solution to $\Delta(\hat{\theta}) = 0$ if the following five conditions hold[28]:

A1. $\sum_{t=1}^{\infty} \hat{\alpha}_t = \infty$ and $\lim_{t\to\infty} \hat{\alpha}_t = 0$.

[28] Note that Kushner and Yin's theorem applies to cases where there may be multiple stationary points, corresponding to cases where the solution to $\Delta(\hat{\theta}) = 0$ is not unique. In our simple case where this solution is unique (as shown above), the conditions for their theorem reduce to the ones listed here (see pages 126-127 of Kushner and Yin; see also the application of this theorem in Heidhues, Kőszegi, and Strack 2021).

A2. $\sum_{t=1}^{\infty}(\hat{\alpha}_t)^2 < \infty$.

A3. $\sup_t \mathbb{E}[|\hat{x}_t - \hat{\theta}_{t-1}|^2|\theta] < \infty$, where the expectation is taken at time $t = 0$.

A4. There exists a continuous function $\bar{g} : \mathbb{R} \to \mathbb{R}$ and a sequence of random variables $\langle \xi_t \rangle$ such that $\mathbb{E}[\hat{x}_t - \hat{\theta}_{t-1}|\hat{\theta}_{t-1}] = \bar{g}(\hat{\theta}_t) + \xi_t$ and $\sum_{t=1}^{\infty} \hat{\alpha}_t |\xi_t| < \infty$ w.p. 1.

A5. There exists a continuously differentiable real-valued function h such that $\bar{g}(\hat{\theta}) = -h'(\hat{\theta})$ and $h(\hat{\theta})$ is constant on each connected subset of $\widehat{\Theta} = \{\hat{\theta} \mid \bar{g}(\hat{\theta}) = 0\}$.

We now show that Conditions A1 – A5 hold:

Condition A1. Note that

$$\sum_{t=1}^{\infty} \hat{\alpha}_t = \sum_{t=1}^{\infty} (1+\kappa_t)\alpha_t \geq (1+\kappa^G)\sum_{t=1}^{\infty}\alpha_t = (1+\kappa^G)\sum_{t=1}^{\infty}\frac{1}{t+\sigma^2/\rho^2}. \tag{A.24}$$

Since the final sum diverges to ∞, $\sum_{t=1}^{\infty}\hat{\alpha}_t$ must as well. Furthermore, it is clear that $\lim_{t\to\infty}\hat{\alpha}_t = 0$.

Condition A2. Note that

$$\sum_{t=1}^{\infty}(\hat{\alpha}_t)^2 = \sum_{t=1}^{\infty}(1+\kappa_t)^2\alpha_t^2 \leq (1+\kappa^L)^2\sum_{t=1}^{\infty}\alpha_t^2. \tag{A.25}$$

From the definition of α_t, $\sum_{t=1}^{\infty}(\alpha_t)^2 < \sum_{t=1}^{\infty} 1/t^2 < \infty$. Thus, $\sum_{t=1}^{\infty}(\hat{\alpha}_t)^2 < \infty$.

Condition A3. We must show $\sup_t \mathbb{E}[|\hat{x}_t - \hat{\theta}_{t-1}|^2|\theta] < \infty$. Note that $\hat{x}_t - \hat{\theta}_{t-1} = x_t + \kappa_t(x_t - \hat{\theta}_{t-1}) - \hat{\theta}_{t-1} = (1+\kappa_t)(x_t - \hat{\theta}_{t-1})$. Letting θ_{t-1} be the Bayesian estimate of θ following $t-1$ rounds, we have

$$\sup_t \mathbb{E}[|\hat{x}_t - \hat{\theta}_{t-1}|^2|\theta] \leq (1+\kappa^L)\sup_t \mathbb{E}[|(x_t - \theta_{t-1}) + (\theta_{t-1} - \hat{\theta}_{t-1})|^2|\theta]. \tag{A.26}$$

From Minkowski's Inequality,

$$\sqrt{\mathbb{E}[|(x_t - \theta_{t-1}) + (\theta_{t-1} - \hat{\theta}_{t-1})|^2|\theta]} \leq \sqrt{\mathbb{E}[|x_t - \theta_{t-1}|^2|\theta]} + \sqrt{\mathbb{E}[|\theta_{t-1} - \hat{\theta}_{t-1}|^2|\theta]}. \tag{A.27}$$

Since $\mathbb{E}[|x_t - \theta_{t-1}|^2|\theta]$ is finite, we need only examine the second term on the right-hand side of Equation (A.27). Using Lemma 1, we can write

$$\theta_{t-1} - \hat{\theta}_{t-1} = \alpha_{t-1}\sum_{\tau=1}^{t-1} x_\tau - \alpha_{t-1}\sum_{\tau=1}^{t-1}\beta_\tau^{t-1}x_\tau = \alpha_{t-1}\sum_{\tau=1}^{t-1}\left(1-\beta_\tau^{t-1}\right)x_\tau, \tag{A.28}$$

where β_τ^{t-1}, defined in Lemma 1, is a function of κ_j and α_j for $j \in \{\tau, \ldots, t-1\}$. Thus

$$\sqrt{\mathbb{E}[|\theta_{t-1} - \hat{\theta}_{t-1}|^2|\theta]} \leq \alpha_{t-1}\sum_{\tau=1}^{t-1}\sqrt{\mathbb{E}[|(1-\beta_\tau^{t-1})x_\tau|^2|\theta]}. \tag{A.29}$$

We now argue that for all $t \geq 2$ and all $\tau \leq t-1$, the value $|1 - \beta_\tau^{t-1}|$ is bounded from above by some finite constant L. For any t, we have $\beta_\tau^{t-1} = (1+\kappa_\tau)\prod_{j=\tau}^{t-2}[1 - \alpha_j\kappa_{j+1}]$. As described in the proof of Proposition 1 (and in Lemma B.1), for any τ, $|\beta_\tau^{t-1}|$ is decreasing in t once $t-1 > t^* = 1+\kappa^L - \sigma^2/\rho^2 < \infty$. Thus, for $\tau > t^*$, $|\beta_\tau^{t-1}| \leq 1+\kappa^L$. Furthermore, for $\tau \leq t^*$, the maximal value of β_τ^{t-1} is $\max_{i \leq t^*}(1+\kappa_\tau)\prod_{j=\tau}^{i}[1-\alpha_j\kappa_{j+1}]$, which is finite given the definition of $\alpha_j \in (0, 1)$ and $\kappa_t \in \{\kappa^G, \kappa^L\}$. Thus, it is clear that there exists a finite value L such that $|1 - \beta_\tau^{t-1}| < L$ for all $t \geq 2$ and $\tau \leq t - 1$. Thus,

$$\begin{aligned}\sqrt{\mathbb{E}[|\theta_{t-1} - \hat{\theta}_{t-1}|^2|\theta]} &\leq L\alpha_{t-1}\sum_{\tau=1}^{t-1}\sqrt{\mathbb{E}[|x_\tau|^2|\theta]} = L\alpha_{t-1}\sum_{\tau=1}^{t-1}\sqrt{\sigma^2+\theta^2} \\ &\leq L\sqrt{\sigma^2+\theta^2}, \qquad \text{(A.30)}\end{aligned}$$

where the first equality follows from the fact that $\mathbb{E}[|x_\tau|^2|\theta] = \mathrm{Var}(x_\tau) + \mathbb{E}[x_\tau|\theta]^2$. Thus, $\sqrt{\mathbb{E}[|\theta_{t-1} - \hat{\theta}_{t-1}|^2|\theta]}$ is finite as desired.

Condition A4. We take $\bar{g} : \mathbb{R} \to \mathbb{R}$ to be the function Δ defined in Equation (A.19) and thus $\xi_t = \mathbb{E}[\hat{x}_t - \hat{\theta}_{t-1}|\hat{\theta}_{t-1}] - \Delta(\hat{\theta}_{t-1}) = 0$. As such, it is immediate that $\sum_{t=1}^{\infty}\hat{\alpha}_t|\xi_t| < \infty$ w.p. 1, as required. Furthermore, from Equation (A.19), it is straightforward that Δ (and hence $\bar{g}$) is continuous given that Φ and ϕ are continuous.

Condition A5. We define h by $h(\hat{\theta}) = -\int_{\theta_0}^{\hat{\theta}}\bar{g}(y)dy = -\int_{\theta_0}^{\hat{\theta}}\Delta(y)dy$. Given the functional form of Δ implied by Equations (A.21) and (A.22), it is clear that h is well defined, continuous, and satisfies $h'(\hat{\theta}) = -\bar{g}(\hat{\theta})$. Furthermore, given that $\bar{g}(\hat{\theta}) = \Delta(\hat{\theta})$, the fact that $\Delta(\hat{\theta}) = 0$ has a unique solution implies that the set $\widehat{\Theta} = \{\hat{\theta} \mid \bar{g}(\hat{\theta}) = 0\}$ is a singleton, and hence Condition A5 is trivially satisfied.

Next, we prove the two enumerated parts of the proposition that describe properties of $\hat{\theta}_\infty$.

Part 1. From Equation (A.21), $\hat{\theta}_\infty$ is the value of $\hat{\theta}$ that solves $\Delta(\hat{\theta}) = \theta - kH(\hat{\theta}; \theta, \sigma) - \hat{\theta} = 0$. First notice that if $\lambda = 1$, then (A.20) implies that $k = 0$, and hence $\hat{\theta}_\infty$ trivially solves $\Delta(\hat{\theta}_\infty) = \theta - \hat{\theta}_\infty = 0 \Rightarrow \hat{\theta}_\infty = \theta$. Thus, the steady-state belief matches the true parameter value when $\lambda = 1$. To show that $\hat{\theta}_\infty$ underestimates θ when $\lambda > 1$, we examine how $\hat{\theta}_\infty$ varies in λ. Since $\hat{\theta}_\infty$ solves Equation (A.21), the Implicit Function Theorem along with the fact that $\frac{\partial}{\partial\hat{\theta}}H(\hat{\theta}; \theta, \sigma) = \Phi\big((\hat{\theta} - \theta)/\sigma\big)$ implies that

$$\begin{aligned}\frac{\partial\hat{\theta}_\infty}{\partial\lambda} &= -\left(\frac{\partial\Delta(\hat{\theta})}{\partial\hat{\theta}}\right)^{-1}\frac{\partial\Delta(\hat{\theta})}{\partial\lambda}\bigg|_{\hat{\theta}=\hat{\theta}_\infty} \\ &= -\Big(1 + k\Phi\big((\hat{\theta}_\infty - \theta)/\sigma\big)\Big)^{-1}H(\hat{\theta}_\infty; \theta, \sigma)\frac{\partial k}{\partial\lambda}. \qquad \text{(A.31)}\end{aligned}$$

Recall from above that $H(\hat{\theta}_\infty; \theta, \sigma) > 0$. Furthermore, (A.20) implies that $\frac{\partial k}{\partial\lambda} > 0$ at all $\lambda \geq 1$ and $k \geq 0$ at all $\lambda \geq 1$. Hence, Equation (A.31) implies $\frac{\partial\hat{\theta}_\infty}{\partial\lambda} < 0$ at all $\lambda \geq 1$. Thus, given that $\hat{\theta}_\infty = \theta$ when $\lambda = 1$, we have $\hat{\theta}_\infty < \theta$ when $\lambda > 1$.

Part 2. The comparative static with respect to λ is established in Part 1. To show the remaining claims, we again invoke the Implicit Function Theorem as in Part 1. We begin by showing that $\lambda > 1$ implies that $\hat{\theta}_\infty$ is decreasing in σ. Analogous to Equation (A.31), we have

$$\frac{\partial\hat{\theta}_\infty}{\partial\sigma} = -\Big(1 + k\Phi\big((\hat{\theta}_\infty - \theta)/\sigma\big)\Big)^{-1}k\frac{\partial H(\hat{\theta}_\infty; \theta, \sigma)}{\partial\sigma}. \qquad \text{(A.32)}$$

From the Definition of H in Equation (A.22), we then have

$$\frac{\partial H(\hat{\theta}_\infty; \theta, \sigma)}{\partial\sigma} = \frac{\partial}{\partial\sigma}\left(\hat{\theta}_\infty\Phi\left(\frac{\hat{\theta}_\infty - \theta}{\sigma}\right) - \int_{-\infty}^{\frac{\hat{\theta}_\infty-\theta}{\sigma}}[\theta + \sigma u]\phi(u)\,du\right)$$

$$= - \int_{-\infty}^{\frac{\hat{\theta}_\infty - \theta}{\sigma}} u\phi(u)\, du, \tag{A.33}$$

which follows from differentiation under the integral sign. From Part 1, $\lambda > 1$ implies that $\hat{\theta}_\infty < \theta$, and thus the upper limit of the integral in Equation (A.33) is negative, implying that the integral itself is negative. Hence, $\lambda > 1$ implies that $\frac{\partial H(\hat{\theta}_\infty;\theta,\sigma)}{\partial\sigma} > 0$, in which case Equation (A.32) reveals that $\lambda > 1$ implies that $\frac{\partial\hat{\theta}_\infty}{\partial\sigma} < 0$.

Next, we show that $\lambda > 1$ implies that $\hat{\theta}_\infty$ is decreasing in η. As in Equation (A.31), we have

$$\frac{\partial\hat{\theta}_\infty}{\partial\eta} = -\left(\frac{\partial\Delta(\hat{\theta})}{\partial\hat{\theta}}\right)^{-1} \frac{\partial\Delta(\hat{\theta})}{\partial\eta}\bigg|_{\hat{\theta}=\hat{\theta}_\infty} = -\Big(1 + k\Phi\big((\hat{\theta}_\infty - \theta)/\sigma\big)\Big)^{-1} H(\hat{\theta}_\infty;\theta,\sigma)\frac{\partial k}{\partial\eta}. \tag{A.34}$$

As noted in Part 1, $1 + k\Phi\big((\hat{\theta}_\infty - \theta)/\sigma\big)$ and $H(\hat{\theta}_\infty;\theta,\sigma)$ are necessarily positive. Furthermore, Equation (A.20) reveals that $\lambda > 1$ implies that $\frac{\partial k}{\partial\eta} > 0$. Thus, from Equation (A.34), $\lambda > 1$ implies that $\frac{\partial\hat{\theta}_\infty}{\partial\eta} < 0$. □

Proof of Proposition 5. We first provide an expression for $v(\tilde{\theta},\sigma)$, the agent's expected (per-period) utility when she believes that outcomes are distributed $x_t \sim N(\tilde{\theta},\sigma^2)$. Let $F(\cdot|\tilde{\theta})$ denote the CDF of this distribution and let $\mathbb{E}_{\tilde{\theta}}$ denote expectations with respect to $F(\cdot|\tilde{\theta})$. Then $v(\tilde{\theta},\sigma) = \mathbb{E}_{\tilde{\theta}}[u(x|\tilde{\theta})]$. From the definition of u in Equation (2), we thus have

$$\begin{aligned} v(\tilde{\theta},\sigma) &= \mathbb{E}_{\tilde{\theta}}[x] + \eta[1 - F(\tilde{\theta}|\tilde{\theta})]\left(\mathbb{E}_{\tilde{\theta}}[x|x \geq \tilde{\theta}] - \tilde{\theta}\right) + \eta\lambda F(\tilde{\theta}|\tilde{\theta})\left(\mathbb{E}_{\tilde{\theta}}[x|x < \tilde{\theta}] - \tilde{\theta}\right) \\ &= \tilde{\theta} - \eta(\lambda - 1)F(\tilde{\theta}|\tilde{\theta})\left(\tilde{\theta} - \mathbb{E}_{\tilde{\theta}}[x|x < \tilde{\theta}]\right) = \tilde{\theta} - \eta(\lambda - 1)H(\tilde{\theta};\tilde{\theta},\sigma), \end{aligned} \tag{A.35}$$

where H is defined in Equation (A.22). From that definition, we have

$$H(\tilde{\theta};\tilde{\theta},\sigma) = \tilde{\theta}\Phi(0) - \int_{-\infty}^{0} [\tilde{\theta} + \sigma u]\phi(u)du = \sigma|\bar{z}^-|, \tag{A.36}$$

where $\bar{z}^- \equiv \int_{-\infty}^{0} u\phi(u)du < 0$. Note that $\bar{z}^-$ is a constant determined entirely by the standard-normal distribution and is hence independent of all the parameters in the model. Thus, Equation (A.35) simplifies to

$$v(\tilde{\theta},\sigma) = \tilde{\theta} - \eta(\lambda - 1)\sigma|\bar{z}^-|. \tag{A.37}$$

Having derived an expression for v, we now analyze the degree to which misattribution distorts the agent's valuation of a prospect in the steady-state. To do so, we first derive an expression for the difference between the misattributor's steady-state belief and the true mean in terms of the underlying distributional parameters, θ and σ. Let $\hat{\theta}_\infty(\theta,\sigma)$ denote the steady-state belief derived in Proposition 4 written explicitly in terms of these parameters. From Equation (A.21) in the proof of Proposition 4, note that $\hat{\theta}_\infty(\theta,\sigma)$ is the value of $\hat{\theta}$ that solves $\hat{\theta} - \theta + kH(\hat{\theta};\theta,\sigma) = 0$. Using the definition of H in Equation (A.22), we can then define the variable $\hat{z} \equiv (\hat{\theta} - \theta)/\sigma$ and rewrite $H(\hat{\theta};\theta,\sigma)$ as

$$H(\hat{\theta};\theta,\sigma)=\hat{\theta}\Phi\left(\hat{z}\right)-\int\limits_{-\infty}^{\hat{z}}[\theta+\sigma z]\phi\left(z\right)dz=\sigma\left(\hat{z}\Phi\left(\hat{z}\right)-\int\limits_{-\infty}^{\hat{z}}z\phi\left(z\right)dz\right). \tag{A.38}$$

Hence, the steady-state condition, $\hat{\theta}-\theta+kH(\hat{\theta};\theta,\sigma)=0$, is equivalent to

$$\hat{z}+k\left(\hat{z}\Phi\left(\hat{z}\right)-\int\limits_{-\infty}^{\hat{z}}z\phi\left(z\right)dz\right)=0, \tag{A.39}$$

and thus $\hat{\theta}_{\infty}(\theta,\sigma)$ is characterized by the value of $\hat{z}$ that solves Equation (A.39). Furthermore, since $\hat{\theta}_{\infty}(\theta,\sigma)$ is unique and finite, there exists a unique, finite $\hat{z}$ that solves Equation (A.39). Denote this value by z^{*}. Clearly z^{*} depends solely on Φ, ϕ, and k, and is thus independent of θ and σ. As such, $z^{*}=(\hat{\theta}_{\infty}(\theta,\sigma)-\theta)/\sigma$ implies that $\hat{\theta}_{\infty}(\theta,\sigma)=\theta+z^{*}\sigma$. Furthermore, since $\lambda>1$, Proposition 4 implies that $\hat{\theta}_{\infty}(\theta,\sigma)<\theta$ and thus $z^{*}<0$. Hence,

$$\hat{\theta}_{\infty}(\theta,\sigma)=\theta-|z^{*}|\sigma. \tag{A.40}$$

Now fix a finite parameter pair (θ',σ') and consider the set of parameters $\mathcal{P}(\theta',\sigma')\equiv\{(\theta,\sigma)\mid v(\theta,\sigma)=v(\theta',\sigma')\}$. Consider any $(\theta,\sigma)\in\mathcal{P}(\theta',\sigma')$. From Equation (A.37), we have $v(\hat{\theta}_{\infty}(\theta,\sigma),\sigma)=\hat{\theta}_{\infty}(\theta,\sigma)-\eta(\lambda-1)\sigma|\bar{z}^{-}|$. Substituting our expression for $\hat{\theta}_{\infty}(\theta,\sigma)$ from Equation (A.40) then yields

$$\begin{aligned}v(\hat{\theta}_{\infty}(\theta,\sigma),\sigma)&=\left(\theta-|z^{*}|\sigma\right)-\eta(\lambda-1)\sigma|\bar{z}^{-}|\\&=v(\theta,\sigma)-|z^{*}|\sigma=v(\theta',\sigma')-|z^{*}|\sigma,\end{aligned} \tag{A.41}$$

where the second equality follows from the definition of $v(\theta,\sigma)$ in Equation (A.37), and the third equality follows from $(\theta,\sigma)\in\mathcal{P}(\theta',\sigma')$. Since $v(\theta',\sigma')$ is fixed and finite, Equation (A.41) implies that for $(\theta,\sigma)\in\mathcal{P}(\theta',\sigma')$, we have $v(\hat{\theta}_{\infty}(\theta,\sigma),\sigma)$ strictly decreasing in σ and $\lim_{\sigma\to\infty}v(\hat{\theta}_{\infty}(\theta,\sigma),\sigma)=-\infty$. □

Proof of Proposition 6. We begin by proving a lemma that describes the posterior belief after two outcomes, which shows that if $\sigma^{2}/\rho^{2}>\kappa^{L}$, then $\hat{\theta}_{2}$ is maximized when the higher outcome happens last. We then extend this result to an arbitrary number of outcomes $T\geq 2$.

Lemma A.2. *Consider any $a,b\in\mathbb{R}$ such that $a>b$. Let $\hat{\theta}_{2}^{i}$ denote the posterior expectation following the increasing sequence (b,a), and let $\hat{\theta}_{2}^{d}$ denote it following the decreasing sequence (a,b). If $\sigma^{2}/\rho^{2}>\kappa^{L}$, then $\hat{\theta}_{2}^{i}>\hat{\theta}_{2}^{d}$.*

Proof of Lemma A.2: The misattributor's posterior expectation of θ after t observations is $\hat{\theta}_{t}=\left(\frac{\sigma^{2}}{t\rho^{2}+\sigma^{2}}\right)\theta_{0}+\left(\frac{\rho^{2}}{t\rho^{2}+\sigma^{2}}\right)\sum_{\tau=1}^{t}\hat{x}_{\tau}$, which implies $\hat{\theta}_{2}^{i}=\alpha_{2}(\hat{b}_{1}^{i}+\hat{a}_{2}^{i})+(1-2\alpha_{2})\theta_{0}$ where $\hat{b}_{1}^{i}$ and $\hat{a}_{2}^{i}$ are the encoded values of b and a respectively when facing the increasing sequence (b,a). Likewise, $\hat{\theta}_{2}^{d}=\alpha_{2}(\hat{a}_{1}^{d}+\hat{b}_{2}^{d})+(1-2\alpha_{2})\theta_{0}$, where $\hat{a}_{1}^{d}$ and $\hat{b}_{2}^{d}$ are the encoded values when facing the decreasing sequence (a,b). Let $\kappa_{1}^{i}=\kappa^{G}\mathbb{1}\{b>\theta_{0}\}+\kappa^{L}\mathbb{1}\{b<\theta_{0}\}$, and $\kappa_{2}^{i}=\kappa^{G}\mathbb{1}\{a>\hat{\theta}_{1}^{i}\}+\kappa^{L}\mathbb{1}\{a<\hat{\theta}_{1}^{i}\}$ where $\hat{\theta}_{1}^{i}=\alpha_{1}(1+\kappa_{1}^{i})(b-\theta_{0})+\theta_{0}$. Similarly, let $\kappa_{1}^{d}=\kappa^{G}\mathbb{1}\{a>\theta_{0}\}+\kappa^{L}\mathbb{1}\{a<\theta_{0}\}$, and $\kappa_{2}^{d}=\kappa^{G}\mathbb{1}\{b>\hat{\theta}_{1}^{d}\}+\kappa^{L}\mathbb{1}\{b<\hat{\theta}_{1}^{d}\}$ where $\hat{\theta}_{1}^{d}=\alpha_{1}(1+\kappa_{1}^{d})(a-\theta_{0})+\theta_{0}$.

Hence $\hat{a}_1^d = a + \kappa_1^d(a - \theta_0)$, $\hat{b}_1^i = b + \kappa_1^i(b - \theta_0)$, $\hat{a}_2^i = a + \kappa_2^i(a - \theta_0 - \alpha_1[1 + \kappa_1^i](b - \theta_0))$, and $\hat{b}_2^d = b + \kappa_2^d(b - \theta_0 - \alpha_1[1 + \kappa_1^d](a - \theta_0))$. This implies $\hat{\theta}_2^i > \hat{\theta}_2^d$ if and only if

$$\begin{aligned}\kappa_1^i(b - \theta_0) + \kappa_2^i(a - \theta_0 - \alpha_1[1 + \kappa_1^i](b - \theta_0))& \\ > \kappa_1^d(a - \theta_0) + \kappa_2^d(b - \theta_0 - \alpha_1[1 + \kappa_1^d](a - \theta_0)).& \qquad \text{(A.42)}\end{aligned}$$

Letting $\tilde{a} = (a - \theta_0)$ and $\tilde{b} = (b - \theta_0)$, Condition (A.42) reduces to

$$\kappa_1^i\tilde{b} + \kappa_2^i(\tilde{a} - \alpha_1[1 + \kappa_1^i]\tilde{b}) > \kappa_1^d\tilde{a} + \kappa_2^d(\tilde{b} - \alpha_1[1 + \kappa_1^d]\tilde{a}). \qquad \text{(A.43)}$$

There are three cases to consider depending on whether $\tilde{a}$ and $\tilde{b}$ have the same sign. When $\tilde{a}$ and $\tilde{b}$ have the same sign, then $\kappa_1^i = \kappa_1^d$ and condition (A.43) reduces as follows, which is useful for checking the various cases: $\hat{\theta}_2^i > \hat{\theta}_2^d$ if and only if

$$\kappa_2^i\big(1 + \alpha_1[1 + \kappa_1^i]\big)(\tilde{a} - \tilde{b}) - (\kappa_2^d - \kappa_2^i)\big(\tilde{b} - \alpha_1[1 + \kappa_1^d]\tilde{a}\big) > \kappa_1^i(\tilde{a} - \tilde{b}). \qquad \text{(A.44)}$$

The remainder of the proof considers all relevant cases before applying the assumption that $\sigma^2/\rho^2 > \kappa^L$); this allows us to use the analysis here to derive Corollary 1, which drops this assumption.

Case 1: $\theta_0 < b < a$. This implies $\kappa_1^i = \kappa_1^d = \kappa^G$. There are 3 sub-cases to consider:

Case 1.a. Suppose both a and b come as gains if received in $t = 2$, implying $\kappa_2^i = \kappa_2^d = \kappa^G$. Hence, Condition (A.44) amounts to $\kappa^G\big(1 + \alpha_1[1 + \kappa^G]\big)(\tilde{a} - \tilde{b}\big) > \kappa^G(\tilde{a} - \tilde{b})$, which holds given $\tilde{a} > \tilde{b}$.

Case 1.b. Suppose both a and b come as losses if received in $t = 2$, implying $\kappa_2^i = \kappa_2^d = \kappa^L$. Hence, Condition (A.44) amounts to $\kappa^L\big(1 + \alpha_1[1 + \kappa^G]\big)(\tilde{a} - \tilde{b}\big) > \kappa^G(\tilde{a} - \tilde{b})$, which holds given $\tilde{a} > \tilde{b}$ and $\kappa^L > \kappa^G$.

Case 1.c. Suppose only a comes a gain if received in $t = 2$, implying $\kappa_2^i = \kappa^G$ and $\kappa_2^d = \kappa^L$. Hence, Condition (A.44) amounts to $\kappa^G\big(1 + \alpha_1[1 + \kappa^G]\big)(\tilde{a} - \tilde{b}\big) - (\kappa^L - \kappa^G)\big(\tilde{b} - \alpha_1[1 + \kappa^G]\tilde{a}\big) > \kappa^G(\tilde{a} - \tilde{b}) \Leftrightarrow \kappa^G\big(\alpha_1[1 + \kappa^G]\big)(\tilde{a} - \tilde{b}\big) - (\kappa^L - \kappa^G)\big(\tilde{b} - \alpha_1[1 + \kappa^G]\tilde{a}\big) > 0$. Since $\hat{b}_2^d$ comes as a loss in this case, $\tilde{b} - \alpha_1[1 + \kappa^G]\tilde{a} < 0$, meaning the previous condition holds.

Case 2: $b < a < \theta_0$. This implies $\kappa_1^i = \kappa_1^d = \kappa^L$. There are 3 sub-cases to consider:

Case 2.a. Suppose both a and b come as losses if received in $t = 2$, implying $\kappa_2^i = \kappa_2^d = \kappa^L$. Hence, Condition (A.44) amounts to $\kappa^L\big(1 + \alpha_1[1 + \kappa^L]\big)(\tilde{a} - \tilde{b}\big) > \kappa^L(\tilde{a} - \tilde{b})$, which is true given $\tilde{a} > \tilde{b}$.

Case 2.b. Suppose both a and b come as gains if received in $t = 2$, implying $\kappa_2^i = \kappa_2^d = \kappa^G$. Hence, Condition (A.44) amounts to $\kappa^G\big(1 + \alpha_1[1 + \kappa^L]\big)(\tilde{a} - \tilde{b}\big) > \kappa^L(\tilde{a} - \tilde{b}) \Leftrightarrow \kappa^G\big(1 + \alpha_1[1 + \kappa^L]\big) > \kappa^L$. There are parameter values for which this condition does not hold (i.e., when $\lambda - 1 > \alpha_1(1 + \eta\lambda)$). However, for both a and b to come as gains in $t = 2$ in this case requires $\alpha_1(1 + \kappa^L) > 1$, which contradicts our assumption that $\sigma^2/\rho^2 > \kappa^L$. Thus, the condition above always holds when $\sigma^2/\rho^2 > \kappa^L$.

Case 2.c. Suppose only a comes as a gain if received in $t = 2$, implying $\kappa_2^i = \kappa^G$ and $\kappa_2^d = \kappa^L$. Hence, Condition (A.44) amounts to $\kappa^G\big(1 + \alpha_1[1 + \kappa^L]\big)(\tilde{a} - \tilde{b}\big) - (\kappa^L - \kappa^G)\big(\tilde{b} - \alpha_1[1 + \kappa^L]\tilde{a}\big) > \kappa^L(\tilde{a} - \tilde{b}) \Leftrightarrow \kappa^G\alpha_1(1 + \kappa^L)(\tilde{a} - \tilde{b}) > (\kappa^L - \kappa^G)(1 - \alpha_1(1 + \kappa^L))\tilde{a}$. The left-hand side of the previous inequality is always positive, while the right-hand side is negative if $\alpha_1(1 + \kappa^L) < 1$. Thus, that inequality always holds under our assumption of $\sigma^2/\rho^2 > \kappa^L$.

Case 3: $b < \theta_0 < a$. This implies $\kappa_1^i = \kappa^L$, $\kappa_1^d = \kappa^G$, $\kappa_2^i = \kappa^G$, and $\kappa_2^d = \kappa^L$. Hence, Condition (A.43) amounts to $\kappa^L \tilde{b} + \kappa^G(\tilde{a} - \alpha_1[1+\kappa^L]\tilde{b}) > \kappa^G \tilde{a} + \kappa^L(\tilde{b} - \alpha_1[1+\kappa^G]\tilde{a}) \Leftrightarrow -\alpha_1\kappa^G[1+\kappa^L]\tilde{b} > -\alpha_1\kappa^L[1+\kappa^G]\tilde{a}$, which holds given $\tilde{a} > 0 > \tilde{b}$, as assumed in this case.

In summary, the only cases where $\hat{\theta}_2^i > \hat{\theta}_2^d$ might fail for some parameter values are 2.b and 2.c. However, $\hat{\theta}_2^i > \hat{\theta}_2^d$ is guaranteed in both of these cases when $\alpha_1(1+\kappa^L) < 1$, which is implied by our assumption that $\sigma^2/\rho^2 > \kappa^L$. This completes the proof of Lemma A.2.

Completing the proof of Proposition 6: Let $\mathcal{X} = \{x_1, \cdots, x_T\}$ be an arbitrary set of T distinct elements of $\mathbb{R}$. Let $S(\mathcal{X})$ be the set of all distinct sequences formed from elements of $\mathcal{X}$. For any sequence $x \in S(\mathcal{X})$, let $\hat{\theta}_T(x)$ denote the misattributor's expectation following x, and let $\hat{\theta}_t(x)$ denote the misattributor's expectation following the first t outcomes of x. We say $x = (x_1, \ldots, x_T)$ is increasing if $x_i < x_{i+1}$ for all $i = 1, \ldots, T-1$. Let $x^* = \arg\max_{x \in S(\mathcal{X})} \hat{\theta}_T(x)$. Toward a contradiction, suppose that x^* is not increasing. Hence, there must exist adjacent outcomes x_i^*, x_{i+1}^* such that $x_i^* > x_{i+1}^*$. Consider an alternative sequence x' that is identical to x^* except it permutes outcomes in i and $i+1$: $x_i' = x_{i+1}^*$ and $x_{i+1}' = x_i^*$. Note that beliefs entering round i are identical under both sequences; i.e., $\hat{\theta}_{i-1}(x^*) = \hat{\theta}_{i-1}(x')$. However, Lemma A.2 implies that the permuted sequence leads to a higher belief at the end of period $i+1$; i.e., $\hat{\theta}_{i+1}(x') > \hat{\theta}_{i+1}(x^*)$. If $i+1 = T$, then the proof is complete. Otherwise, note that the assumption $\sigma^2/\rho^2 > \kappa^L$ implies that all of the weights in Lemma 1 are always positive (see Lemma B.1 for details). This means that for any sequence $\tilde{x} \in \mathbb{R}^T$, $\hat{\theta}_T(\tilde{x})$ is strictly increasing in $\hat{\theta}_{i+1}(\tilde{x})$ when holding $(\tilde{x}_{i+2}, \ldots, \tilde{x}_T)$ fixed. Hence, $\hat{\theta}_{i+1}(x') > \hat{\theta}_{i+1}(x^*)$ implies $\hat{\theta}_T(x') > \hat{\theta}_T(x^*)$, yielding a contradiction and completing the proof. □

Proof of Proposition 7. Adopting the notation introduced in the proof of Proposition 6, let $x \in \mathbb{R}^T$ denote a generic sequence of length T, let $\hat{\theta}_T(x)$ denote the misattributor's expectation following x, and let $\hat{\theta}_t(x)$ denote this expectation following the first t outcomes of x. To streamline notation, let $\succ$ be a binary relation over sequences defined by $x \succ \tilde{x} \Leftrightarrow \hat{\theta}_T(x) > \hat{\theta}_T(\tilde{x})$.

Without loss of generality, let $\theta_0 = 0$. Although the statement of this proposition focuses on $B > 0$, our proof will consider both $B > 0$ and $B < 0$ for sake of completeness and because the case with $B < 0$ is useful for the proof of Proposition 3. Fixing x, Lemma 1 implies that $\hat{\theta}_T(x) = \alpha_T \sum_{t=1}^T \beta_t^T x_t$. We therefore aim to characterize the solution of the following problem:

$$\max_{x \in \mathbb{R}^T} \sum_{t=1}^T \beta_t^T x_t \text{ subject to } \sum_{t=1}^T x_t \leq B \text{ and } \hat{\theta}_t(x) \geq \bar{\theta} \;\; \forall t = 1, \ldots, T. \tag{A.45}$$

When analyzing this optimization problem, we must account for the fact that the weights, β_t^T, also depend on the sequence, x, since their values depend on the induced path of gains and losses (see Lemma 1). Additionally, it is clear that the budget constraint, $\sum_{t=1}^T x_t \geq B$, always binds at the optimum. We now characterize the solution to Problem (A.45), which we denote by x^*. We begin by noting three general properties of x^*, and then we describe how x^* further depends on B and T.

Property 1: x^* is weakly increasing. We first show that x^* is weakly increasing; i.e., $x_t^* \leq x_{t+1}^*$ for all $t = 1, \ldots, T-1$. This follows from Proposition 6: if x^* were not weakly increasing, then a permutation of its elements will generate a strict increase in $\hat{\theta}_T(x^*)$, leading to a contradiction. Note that this property implies that only the first outcome of x^* may be encoded as a loss; any following outcome x_t^* for $t \geq 2$ must be encoded as either a strict gain, i.e., $x_t^* > \hat{\theta}_{t-1}(x^*)$, or "neutral", which we take to mean $x_t^* = \hat{\theta}_{t-1}(x^*)$.

Property 2: x^* has at most one strict gain. Next we show that there exists at most a single $t \in \{1, \dots, T\}$ such that x_t^* represents a strict gain (i.e., $x_t^* > \hat{\theta}_{t-1}(x^*)$). We show this by contradiction: suppose there exist periods $t_2 > t_2 \geq 1$ such that $x_{t_1}^* > \hat{\theta}_{t_1-1}(x^*)$ and $x_{t_2}^* > \hat{\theta}_{t_2-1}(x^*)$. Since both of these outcomes are encoded as strict gains, Proposition 1 (along with our assumption that $\sigma^2/\rho^2 < \kappa^L \Leftrightarrow t^* < 1$) implies that the relative weights that $\hat{\theta}_T(x^*)$ places on these outcomes are such that $\beta_{t_2}^T > \beta_{t_1}^T > 0$. Thus, transferring the gain in t_1 to t_2 does not alter the budget constraint yet increases the objective in (A.45). More specifically, consider a sequence x' identical to x^* except $x'_{t_1} = x_{t_1}^* - \epsilon$ and $x'_{t_2} = x_{t_2}^* + \epsilon$. If $\epsilon \leq x_{t_1}^* - \hat{\theta}_{t_1-1}(x^*)$, then this transfer does not change how any outcomes are encoded (relative to x^*), and thus induces a sequence of weights identical to their values under x^*. However, the transfer increases Objective (A.45) by $\left(\beta_{t_2}^T - \beta_{t_1}^T\right)\epsilon > 0$, contradicting the optimality of x^*. Thus, x^* has at most one outcome encoded as a strict gain.

Property 3: any strict gain must come in the final period. Next we show that if x^* induces a strict gain, then it must come in period T. We again show this by way of contradiction. Suppose the single strict gain happens in period $t < T$. This implies that x^* induces a single upward change in beliefs, which happens in period t, requiring $x_{t+1}^* = \cdots = x_T^* = \hat{\theta}_t(x^*)$. Equation (7) then implies

$$\hat{\theta}_T(x^*) = \hat{\theta}_t(x^*) = \alpha_t(1+\kappa^G)x_t^* + [1-\alpha_T(1+\kappa^G)]\hat{\theta}_{t-1}(x^*) \tag{A.46}$$

Now consider x' such that $x'_i = x_i^*$ for all $i \leq t-1$, and $x'_i = \hat{\theta}_{t-1}(x^*)$ for each $i = t, \dots, T-1$ (note that if $t = 1$, we simply have $x'_i = \theta_0 = 0$ for all $i = 1, \dots, T-1$). The budget constraint then implies $x'_T = B - \sum_{i=1}^{T-1} x'_t$. Notice that x' is such that $\hat{\theta}_{T-1}(x') = \hat{\theta}_{t-1}(x^*)$. Equation (7) then implies that $\hat{\theta}_T(x') = \alpha_T(1+\kappa^G)x'_T + [1-\alpha_T(1+\kappa^G)]\hat{\theta}_{t-1}(x^*)$. From Equation (A.46), we then have $\hat{\theta}_T(x') > \hat{\theta}_T(x^*) \Leftrightarrow$

$$\alpha_T\left[x'_T - \hat{\theta}_{t-1}(x^*)\right] > \alpha_t\left[x_t^* - \hat{\theta}_{t-1}(x^*)\right]. \tag{A.47}$$

Notice that the budget constraint for x^* implies that $x_t^* = B - \sum_{i=1}^{t-1} x_i^* - (T-t)\hat{\theta}_t(x^*)$. Combining this with the expression for $\hat{\theta}_t(x^*)$ in Equation (A.46) yields

$$x_t^* = \frac{B - \sum_{i=1}^{t-1} x_i^* - (T-t)[1-\alpha_t(1+\kappa^G)]\hat{\theta}_{t-1}(x^*)}{1+(T-t)\alpha_t(1+\kappa^G)}. \tag{A.48}$$

Similarly, the budget constraint for x' implies

$$x'_T = B - \sum_{i=1}^{t-1} x_i^* - (T-t)\hat{\theta}_{t-1}(x^*). \tag{A.49}$$

Substituting (A.48) and (A.49) into Condition (A.47) reveals that Condition (A.47) is equivalent to

$$\alpha_T > \frac{\alpha_t}{1+(T-t)\alpha_t(1+\kappa^G)}. \tag{A.50}$$

Note that for any $t < T$, we have $\alpha_T = \alpha_t/[1+(T-t)\alpha_t]$, and thus Condition (A.50) must hold given that the right-hand side is less than $\alpha_t/[1+(T-t)\alpha_t]$ since $\kappa^G > 0$. Thus, $\hat{\theta}_T(x') > \hat{\theta}_T(x^*)$, contradicting the optimality of x^*. Hence, if x^* induces a strict gain, it must happen in period T.

Having established the three general properties of x^* above, we now turn to precisely describing x^*. In doing so, it will be helpful to define three potential types of sequences:

1. $x \in \mathbb{R}^T$ is a *final gain* sequence if $x_1 = \cdots = x_{T-1} = 0$ and $x_T = B$.
2. $x \in \mathbb{R}^T$ is an *initial loss* sequence if $x_1 < 0$ and $x_2 = \cdots = x_T = \hat{\theta}_1$.
3. $x \in \mathbb{R}^T$ is a *loss-gain* sequence if $x_1 < 0$, $x_2 = \cdots = x_{T-1} = \hat{\theta}_1$, and $x_T > \hat{\theta}_1$.

Our three general properties of x^* imply that x^* must follow one of the sequence types defined above. We now derive which of these types is optimal as a function of B and T; in doing so, we will separately handle the cases of $B > 0$ and $B < 0$.

First, note that in both cases (i.e., $B \lesseqgtr 0$), we will focus on the "loss-gain" sequence such that x_1 drops beliefs to the threshold $\bar{\theta} < 0$, meaning that the participation constraint binds. As we argue below, whenever the "loss-gain" sequence is optimal, it will be optimal to induce the biggest allowable loss (i.e., x_1 lowers beliefs to $\hat{\theta}_1 = \bar{\theta}$). Accordingly, define x^{LG} such that $x_1^{LG} = \bar{\theta}/(\alpha_1(1+\kappa^L))$, $x_2^{LG} = \cdots = x_{T-1}^{LG} = \bar{\theta}$, and $x_T^{LG} = B - \sum_{i=1}^{T-1} x_i^{LG} = B - [T - 2 - 1/\alpha_1(1+\kappa^L)]\bar{\theta}$. For any B, the final expectation following x^{LG} is $\hat{\theta}_T(x^{LG}) = \alpha_T(1+\kappa^G)x_T^{LG} + [1 - \alpha_T(1+\kappa^G)]\bar{\theta}$, and thus

$$\hat{\theta}_T(x^{LG}) = \alpha_T(1+\kappa^G)\left[B - \bar{\theta}\left(T - 1 + \frac{1}{\alpha_1(1+\kappa^L)}\right)\right] + \bar{\theta}. \tag{A.51}$$

Case 1: Suppose $B > 0$ (as in the statement of the proposition). First note that x^* cannot follow an "initial loss" sequence since this would imply a slack budget constraint given that $B > 0$. Thus, we must simply determine whether the optimal sequence is x^{LG} or the "final gain" sequence. Let x^{FG} denote the "final gain" sequence with $x_1^{FG} = \cdots = x_{T-1}^{FG} = 0$ and $x_T^{FG} = B$. Note that $\hat{\theta}_T(x^{FG}) = \alpha_T(1+\kappa^G)B$. Comparing $\hat{\theta}_T(x^{FG})$ to Equation (A.51) yields $x^{LG} \succ x^{FG} \Leftrightarrow$

$$\alpha_T(1+\kappa^G)\left(T - 1 + \frac{1}{\alpha_1(1+\kappa^L)}\right) > 1, \tag{A.52}$$

which amounts to

$$T > \bar{T}^{FG} \equiv \frac{\kappa^L(1+\kappa^G) + (\kappa^L - \kappa^G)\frac{\sigma^2}{\rho^2}}{\kappa^G(1+\kappa^L)}. \tag{A.53}$$

Taking $\bar{T} = \bar{T}^{FG}$ completes the proof of Proposition 7 as stated in the main text. Additionally, note that $\bar{T}^{FG} \geq 1$ and $\bar{T}^{FG} = 1$ only when $\kappa^L = \kappa^G$; thus, since $T \geq 2$, x^{LG} is always optimal when $B > 0$ and loss aversion is sufficiently low.

Case 2: Suppose $B < 0$. First note that x^* cannot follow a "final gain" sequence since this would imply that x^* is not weakly increasing given that $B < 0$. Thus, we must simply determine whether the optimal sequence is x^{LG} or the "initial loss" sequence. Let x^{IL} denote the "initial loss" sequence with $x_1^{IL} < 0$ and $x_2^{IL} = \cdots = x_T^{IL} = \hat{\theta}_1$. Note that $\hat{\theta}_T(x^{IL}) = \hat{\theta}_1(x^{IL}) = \alpha_1(1+\kappa^L)x_1^{IL}$. From the budget constraint, $x_1^{IL} = B - (T-1)\hat{\theta}_1(x^{IL})$, and thus

$$\hat{\theta}_T(x^{IL}) = \frac{\alpha_1(1+\kappa^L)}{1 + (T-1)\alpha_1(1+\kappa^L)}B. \tag{A.54}$$

Comparing Equations (A.54) and (A.51) yields $x^{LG} \succ x^{IL} \Leftrightarrow$

$$\alpha_T(1+\kappa^G)\left(T - 1 + \frac{1}{\alpha_1(1+\kappa^L)} - g(T)\frac{B}{\bar{\theta}}\right) > 1, \tag{A.55}$$

where $g(T) \equiv \left[1 + (T + \sigma^2/\rho^2)/(\kappa^L(T-1))\right]^{-1}$. While (A.55) tightly characterizes when $x^{LG} \succ x^{IL}$, we can also derive a useful sufficient condition for $x^{LG} \succ x^{IL}$ using the properties of $g(T)$. First note that Equations (A.55) and (A.53) imply $x^{LG} \succ x^{IL} \Leftrightarrow$

$$T > \bar{T}^{FG} + \left(\frac{1+\kappa^G}{\kappa^G}\right) g(T) B/\bar{\theta}. \tag{A.56}$$

Furthermore, note that (i) $g(T) \in (0, 1)$ for all T; (ii) $g(T)$ is strictly increasing in T; and (iii) $g(T)$ is bounded from above by $\kappa^L/(1+\kappa^L)$. Thus, $x^{LG} \succ x^{IL}$ if

$$T > \bar{T}^{IL} \equiv \bar{T}^{FG} + \frac{\kappa^L(1+\kappa^G)}{\kappa^G(1+\kappa^L)} B/\bar{\theta} = \frac{\kappa^L(1+\kappa^G)(1+B/\bar{\theta}) + (\kappa^L - \kappa^G)\frac{\sigma^2}{\rho^2}}{\kappa^G(1+\kappa^L)}. \tag{A.57}$$

Finally, we verify our claim above: if a "loss-gain" sequence is optimal, then it is optimal for x_1 to lower beliefs to an extent that the participation constraint binds; i.e., $\hat{\theta}_1 = \bar{\theta}$. To see this, notice that $\hat{\theta}_T(x^{LG})$ is strictly increasing in $\bar{\theta}$ if and only if Condition (A.52) holds. Thus, when this condition holds, we have $x^{LG} \succ \tilde{x}$ for any "loss-gain" sequence $\tilde{x}$ that induces $\hat{\theta}_1(\tilde{x}) = \cdots = \hat{\theta}_{T-1}(\tilde{x}) = \tilde{\theta} > \bar{\theta}$. Now note that if $B > 0$ and $x^{LG} \succ x^{FG}$, then Condition (A.53) holds and thus Condition (A.52) holds. Furthermore, if $B < 0$ and $x^{LG} \succ x^{IL}$, then Condition (A.56) holds, which also implies (A.52) holds. Thus, whenever the "loss-gain" sequence outperforms the other viable types of sequences, then the optimal variant of the "loss-gain" sequence involves the largest possible initial loss. □

Appendix B. Supplemental results

In this section, we provide formal details underlying some supplemental results noted in the main text.

B.1. Details on short-run belief updating

This section provides additional details on the short-run dynamics described in Section 3. In particular, we consider the relative weight that current expectations assign to a particular past outcome and describe how this weight varies over time. Standard Bayesian learning applied to our setting implies that expectations in period t weight each realized outcome identically. Misattribution, however, implies that the weight on any given outcome x_τ, $\tau < t$, will generically differ from that on any other outcome. Moreover, the way this relative weighting of x_τ evolves as t advances—both in terms of magnitude and sign—can differ across early segments the horizon. As we describe below, these different "phases" of the horizon will be defined in terms of the particular dynamic patterns they induce on the relative weights of outcomes. The boundaries of these phases are then pinned down by the bias parameters, κ^G and κ^L, and the relative precision of the prior, σ^2/ρ^2. We first derive and describe these various phases, and then conclude this section by describing some intuition on why these phases arise.

Consider some fixed outcome x_τ in period $\tau \geq 1$. We will examine the weight that expectations formed in periods $t \geq \tau$ place on x_τ. From Lemma 1, $\hat{\theta}_t = \alpha_t \sum_{\tau=1}^{t} \beta_\tau^t x_\tau + \beta_0^t \theta_0$. Hence, β_τ^t is the relative weight that $\hat{\theta}_t$ places on x_τ. Also from Lemma 1, β_τ^t evolves in t according to

$$\beta_\tau^{t+1} = \beta_\tau^t[1 - \alpha_t \kappa_{t+1}]. \tag{B.1}$$

Thus, the evolution of this weight—both in sign and magnitude—is completely determined by properties of the multiplicative term $[1 - \alpha_t \kappa_{t+1}]$. Since $\alpha_t \in (0, 1)$ is decreasing in t, note that once t is large enough that $\alpha_t \kappa_{t+1} < 1$, then β_τ^t is necessarily decreasing in t. This is the premise of the recency bias described in Proposition 1: the weight on x_τ relative to other outcomes becomes smaller as τ fades further into the past.

For small values of t, however, the evolution of β_τ^t might temporarily display other patterns. If there exist values of t such that $1 - \alpha_t \kappa_{t+1} < 0$, then β_τ^t will alternate in sign until t is large enough to ensure that $1 - \alpha_t \kappa_{t+1} > 0$. Furthermore, it is possible that β_τ^t briefly increases in t among initial rounds in which $1 - \alpha_t \kappa_{t+1} < -1$.

Thus, the complete dynamics of β_τ^t can be characterized by deriving (i) the time frame on which $1 - \alpha_t \kappa_{t+1} < 0$, which determines when the sign of the weight, $\text{sgn}(\beta_\tau^t)$, is oscillating versus constant in t; and (ii) the time frame on which $|1 - \alpha_t \kappa_{t+1}| \lesseqgtr 1$, which determines when the magnitude of the weight, $|\beta_\tau^t|$, is increasing versus decreasing in t. Before deriving these time frames, it is worth noting that these conditions depend on κ_{t+1}; that is, whether the most recent outcome is encoded as a gain or a loss. Intuitively, the extent to which an early outcome x_τ influences current expectations through the encoded value of the most recent outcome, x_{t+1}, can be smaller or larger depending on whether x_{t+1} is encoded as a gain or loss. For this reason, the relevant time frames will be defined, in part, based on the encoding of the most recent outcome.

The conditions above, along with the fact that $\alpha_t = 1/(t + \sigma^2/\rho^2)$, allow us to describe the boundaries of the various phases of the dynamics in terms of the underlying parameters. First, let $t_*^G \equiv \max\left\{\lfloor \kappa^G - \sigma^2/\rho^2 \rfloor, 0\right\}$ and $t_*^L \equiv \max\left\{\lfloor \kappa^L - \sigma^2/\rho^2 \rfloor, 0\right\}$. These values describe when the weight on x_τ will alternate signs or not as time advances. In particular, if x_{t+1} is encoded as a gain, then $\text{sgn}(\beta_\tau^{t+1}) = \text{sgn}(\beta_\tau^t)$ if and only if $t > t_*^G$; similarly, if x_{t+1} is encoded as a loss, then $\text{sgn}(\beta_\tau^{t+1}) = \text{sgn}(\beta_\tau^t)$ if and only $t > t_*^L$. Note that t_*^L is intimately tied to our definition of t^* in the main text: recall that $t > t^*$ if and only if $\hat{\theta}_t$ is necessarily increasing in $\hat{\theta}_{t-1}$; this condition is identical to $t > t_L^*$. Similarly, the case of $t^* < 1$, which we consider for some results in the main text, is equivalent to $t_L^* = 0$.

Second, let $t_{**}^G \equiv \max\left\{\lfloor \kappa^G/2 - \sigma^2/\rho^2 \rfloor, 0\right\}$ and let $t_{**}^L \equiv \max\left\{\lfloor \kappa^L/2 - \sigma^2/\rho^2 \rfloor, 0\right\}$. These values describe when the weight on x_τ will increase versus decrease as t advances to $t + 1$. In particular, if x_{t+1} is encoded as a gain, then $|\beta_\tau^{t+1}| < |\beta_\tau^t|$ if and only if $t > t_{**}^G$; similarly, if x_{t+1} is encoded as a loss, then $|\beta_\tau^{t+1}| < |\beta_\tau^t|$ if and only $t > t_{**}^L$.

The following lemma summarizes the discussion above, characterizing two types of phases within the updating process: (i) those in which the signs of the weights alternate or not, and (ii) those in which the magnitudes of the weights increase versus decrease.

Lemma B.1. *Consider outcome x_τ and $t \geq \tau$. The weight β_τ^t that $\hat{\theta}_t$ assigns to x_τ has the following properties:*

1. *The direction of the weight,* $\text{sgn}(\beta_\tau^t)$, *evolves as follows:*
   (a) *Deterministic oscillation phase: if $t < t_*^G$, then* $\text{sgn}(\beta_\tau^{t+1}) \neq \text{sgn}(\beta_\tau^t)$ *and hence the directional effect of x_τ on beliefs alternates.*
   (b) *Stochastic oscillation phase: if $t \in (t_*^G, t_*^L)$, then* $\text{sgn}(\beta_\tau^{t+1}) \neq \text{sgn}(\beta_\tau^t)$ *if $\kappa_{t+1} = \kappa^L$ and* $\text{sgn}(\beta_\tau^{t+1}) = \text{sgn}(\beta_\tau^t)$ *if $\kappa_{t+1} = \kappa^G$. Hence, the directional effect of x_τ on beliefs alternates only following a loss.*
   (c) *Deterministic persistence phase: if $t > t_*^L$, then* $\text{sgn}(\beta_\tau^{t+1}) = \text{sgn}(\beta_\tau^t)$ *and hence the directional effect of x_τ on beliefs remains constant.*
2. *The magnitude of the weight, $|\beta_\tau^t|$, evolves as follows:*
   (a) *Deterministic amplification phase: if $t < t_{**}^G$, then $|\beta_\tau^{t+1}| > |\beta_\tau^t|$ and hence the weight on x_τ will grow in magnitude.*

(b) *Stochastic amplification-decay phase: if $t \in (t^G_{**}, t^L_{**})$, then $|\beta^{t+1}_\tau| > |\beta^t_\tau|$ if $\kappa_{t+1} = \kappa^L$ and $|\beta^{t+1}_\tau| < |\beta^t_\tau|$ if $\kappa_{t+1} = \kappa^G$. Hence, the weight on x_τ will grow in magnitude only following a loss.*

(c) *Deterministic decay phase: if $t > t^L_{**}$, then $|\beta^{t+1}_\tau| < |\beta^t_\tau|$ and hence the weight on x_τ will diminish in magnitude.*

It is important to note that the various phases above emerge only when the thresholds t^L_*, etc. are non-zero. If all of them are zero—which happens whenever $\kappa^L < \sigma^2/\rho^2$—then the relative weighting of every outcome is always positive and always decreasing in t. Hence, these varied phases of the dynamics exist only when κ^L and κ^G are sufficiently large relative to σ^2/ρ^2. Furthermore, it is possible that $t^L_* \geq 1$ and $t^L_{**} = 0$, but $t^L_* = 0$ necessarily implies that $t^L_{**} = 0$. Thus, a phase in which weights oscillate in sign is more likely to occur than one in which weights amplify in magnitude: the set of parameters that generate temporarily oscillating weights is a strict superset of those that generate temporarily amplifying weights.

Fig. 2 depicts the phases described in Lemma B.1 as a function of time. The unambiguous relationships between the various thresholds in Lemma B.1 give rise to a relatively clear picture of how weights evolve when there are distinct phases to the dynamics (i.e., when $\kappa^L > \sigma^2/\rho^2$). Namely, we must have $t^G_{**} \leq t^L_{**}$, $t^G_* \leq t^L_*$, $t^G_{**} \leq t^G_*$, and $t^L_{**} \leq t^L_*$. Thus, the various phases of the dynamics will unfold (roughly) as in Fig. 2. The regions of the horizon labeled above the timeline describe how the magnitudes of the weights evolve, while those labeled below describe how their signs evolve.[29] The intersections of these regions provide a complete picture of how the weight β^t_τ on a past outcome x_τ evolves as time progresses.

In the main text, we sometimes restrict focus to cases where $t > t^L_*$ (e.g., Proposition 1). The reason is twofold. First, this phase of the dynamics always exists, regardless of the parameter values. Furthermore, it is persistent in the sense that, once this phase begins, it will continue for the remainder of time. Thus, patterns in beliefs that occur beyond t^L_* (namely, a recency bias), can be viewed as a robust implication of misattribution, while patterns that may occur prior to t^L_* only emerge under particular parameter values and are fleeting when they do.[30]

We now provide intuition for why weights might oscillate in sign or increase in magnitude early in the learning process.

Intuition for why the sign of β^t_τ can oscillate in t: As x_τ increases, it will increase expectations in the subsequent round and thus decrease the encoded outcome in that round. In early rounds where posteriors are decreasing in the prior, the positive effect of x_τ on $\hat{\theta}_\tau$ will have an overall negative effect on $\hat{\theta}_{\tau+1}$. (An analogous intuition emerges as x_τ decreases, again leading to an inverse relationship between x_τ and $\hat{\theta}_{\tau+1}$.) Extending this logic forward, the effect of x_τ has an alternating effect on subsequent periods: if x_τ has a negative effect on $\hat{\theta}_{\tau+1}$, then it will have a positive effect on $\hat{\theta}_{\tau+2}$ whenever $\hat{\theta}_{\tau+2}$ is decreasing $\hat{\theta}_{\tau+1}$, and so on. This alternating logic will hold up to the highest round t such that $\hat{\theta}_t$ is decreasing in $\hat{\theta}_{t-1}$. This "highest round" is precisely the definition of t^*_G and t^*_L—these are the latest rounds in which $\hat{\theta}_t$ is decreasing in $\hat{\theta}_{t-1}$ conditional on x_t being encoded as a gain or loss, respectively. As such, the weight on x_τ can oscillate in sign up to these thresholds, but not beyond.

[29] The relationship between t^G_* and t^L_{**} is ambiguous without further specifying the parameter values. The following figure assumes $t^G_* > t^L_{**}$, but this relationship would flip with sufficiently stronger loss aversion.

[30] In fact, a recency bias starts to emerge as soon as $t > t^L_{**}$. At that point, $|\beta^t_\tau|$ is necessarily decreasing in t for the remainder of the horizon even though the sign of β^t_τ can still change up until $t^L_* > t^L_{**}$.

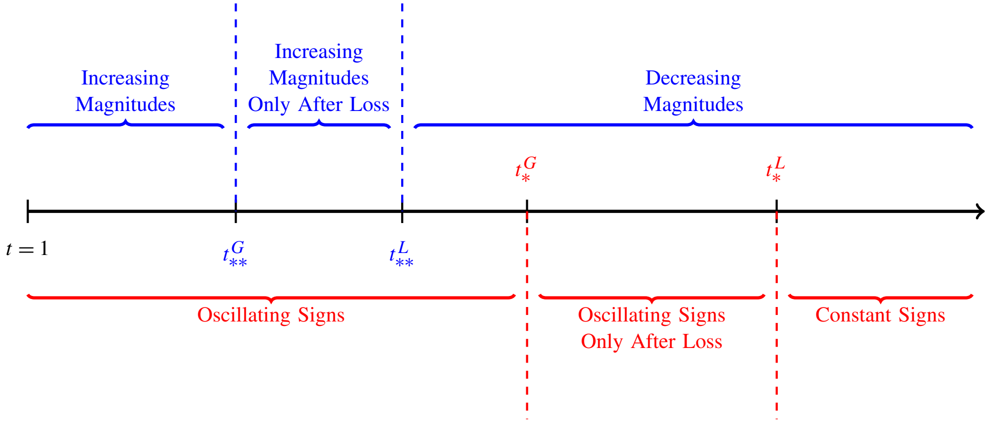

Fig. 2. The various phases of the evolving beliefs of a misattributor. The text above (below) the timeline describes how the magnitude (sign) of the weight on a past outcome evolves over time.

Intuition for why β_τ^t can increase in magnitude: Note that this pattern only arises when weights are oscillating in sign, meaning that expectations in period $t-1$ have a negative effect on expectations in period t. Given this, the bias in the encoded outcome in a previous round τ tends to have a compounding effect on the bias in later rounds. To see this, suppose x_τ comes as a gain and $\hat{x}_\tau$ is biased upward. When the next outcome comes as a loss, the upward bias in $\hat{x}_{\tau+1}$ makes that loss more severe, and thus the bias in $\hat{x}_{\tau+1}$ is larger when the bias in $\hat{x}_\tau$ is larger. In this way, the effect of x_τ on final beliefs is amplified. And this amplification effect can extend over time: amplifying the downward bias in $\hat{x}_{\tau+1}$ will cause a gain in $\tau+2$ to by exaggerated upward by a greater extent, and so on. However, this amplification effect can only emerge in the early portion of the horizon in which the current outcome has a large effect on expectations, since this large effect is necessary for $\hat{x}_\tau$ to have a large impact on $\hat{x}_{\tau+1}$, and so on. Thus, once enough outcomes have accumulated for the most recent one to have little individual impact on expectations, the amplification effect will cease. This logic underlies the threshold values t_{**}^G and t_{**}^L in Lemma B.1.

B.2. Variance of encoded outcomes and mean beliefs

In this section, we consider the variance of a misattributor's encoded outcomes and the variance of her mean beliefs.

We first show that, conditional on $\hat{\theta}_{t-1}$, the variance of the encoded outcome, $\hat{x}_t$, is greater than the variance of the true outcome, x_t. By considering $\hat{\theta}_{t-1} = \hat{\theta}_\infty$, where $\hat{\theta}_\infty$ is the steady-state value defined in Proposition 4, this result further implies that the distribution of encoded outcomes in the steady state will have greater variance than the true distribution of outcomes, as noted in the main text.

Lemma B.2. *Conditional on any* $\hat{\theta}_{t-1} \in \mathbb{R}$, $\text{Var}(\hat{x}_t) > \text{Var}(x_t)$.

Proof of Lemma B.2. Given $\hat{\theta}_{t-1}$, we have $\hat{x}_t = x_t + \kappa_t(x_t - \hat{\theta}_{t-1})$. Define the function $\tilde{n}(x|\hat{\theta}_{t-1})$ as follows: $\tilde{n}(x|\hat{\theta}_{t-1}) = \kappa^G(x - \hat{\theta}_{t-1})$ if $x \geq \hat{\theta}_{t-1}$ and $\tilde{n}(x|\hat{\theta}_{t-1}) = \kappa^L(x - \hat{\theta}_{t-1})$ if $x < \hat{\theta}_{t-1}$. Hence, $\text{Var}(\hat{x}_t) = \text{Var}(x_t) + \text{Var}(\tilde{n}(x_t|\hat{\theta}_{t-1})) + 2\text{Cov}(x_t, \tilde{n}(x_t|\hat{\theta}_{t-1}))$. It thus suffices to show that $\text{Cov}(x_t, \tilde{n}(x_t|\hat{\theta}_{t-1})) \geq 0$. Note that $\tilde{n}(x|\hat{\theta}_{t-1})$ is strictly increasing in x. Thus, $\text{Cov}(x_t, \tilde{n}(x_t|\hat{\theta}_{t-1})) \geq 0$ since the covariance of a random variable and an increasing function of that random variable is non-negative (Thorisson, 1995) provided both are square-integrable (as they are here). □

Next, we show that a misattributor's mean beliefs are excessively variable. To make this point succinctly, we show that, conditional on $(x_1, \ldots, x_{t-1})$, the variance of $\hat{\theta}_t$ exceeds the variance of the Bayesian mean belief, θ_t. Note that $\text{Var}(\hat{\theta}_t) = \text{Var}(\alpha_t \hat{x}_t + (1-\alpha_t)\hat{\theta}_{t-1}) = \alpha_t^2 \text{Var}(\hat{x}_t)$ since $\hat{\theta}_{t-1}$ is constant conditional on $(x_1, \ldots, x_{t-1})$. Similarly, $\text{Var}(\theta_t) = \alpha_t^2 \text{Var}(x_t)$. Thus, $\text{Var}(\hat{\theta}_t) > \text{Var}(\theta_t) \Leftrightarrow \text{Var}(\hat{x}_t) > \text{Var}(x_t)$, which holds due to Lemma B.2.

B.3. Extension with uncertainty over the variance of outcomes

In our baseline model, the outcome in each round is distributed $x_t \sim N(\theta, \sigma^2)$, and the agent attempts to learn θ. In this appendix, we sketch an extension where the agent is initially uncertain about both θ and σ^2. As in the main text, we analyze the agent's distorted beliefs when she uses encoded outcomes (Equation (4)) in place of true outcomes when updating her beliefs about these parameters. Misattribution in this case leads to long-run beliefs that overestimate σ^2 (regardless of the value of λ) and underestimate θ (when $\lambda > 1$, as in Proposition 4).

We focus on a familiar class of prior beliefs that admit tractable posteriors. Specifically, the agent begins with an inverse-gamma prior over the variance of outcomes: $\sigma^2 \sim IG(a_0, b_0)$.[31] As in the main text, we continue to assume that the agent has normally-distributed priors over θ: conditional on σ^2, the agent initially believes $\theta \sim N(\theta_0, \sigma^2/v_0)$, where $v_0 > 0$ is a scaling parameter known by the agent. By considering $v_0 = \sigma^2/\rho^2$ for some constant $\rho > 0$, the agent initially believes $\theta \sim N(\theta_0, \rho^2)$ exactly as in our benchmark model in the main text.

We now describe the dynamics of the agent's beliefs, starting from the first period. Following our model of misattribution in the main text, we maintain that the agent's reference point in period t is her expectation of x_t. Given the priors above, her marginal belief over θ initially follows a Student's t-distribution with mean θ_0. Thus, her reference point in $t = 1$ is θ_0, and she therefore encodes a value $\hat{x}_1$ exactly as described in Equation (4) given this expectation.

Computing the agent's updated beliefs given her encoded outcome is relatively straightforward. Since the agent's prior and posterior distributions are conjugate distributions, her posterior beliefs over σ^2 and θ are also inverse-gamma and normal, respectively, but with parameters that update as follows (see, e.g., Murphy, 2007). For any t, let $\widehat{S}_t \equiv \sum_{i=1}^t \hat{x}_i$ denote the sum of encoded outcomes through period t; we then have:

$$\hat{\theta}_t = \frac{v_0\theta_0 + \widehat{S}_t}{v_0 + t}, \tag{B.2}$$

$$v_t = v_0 + t, \tag{B.3}$$

$$a_t = a_0 + \frac{t}{2}, \tag{B.4}$$

$$b_t = b_0 + \frac{1}{2}\sum_{i=1}^{t}(\hat{x}_i - \widehat{S}_t/t)^2 + \frac{1}{2}\frac{v_0 t}{v_0 + t}(\widehat{S}_t/t - \theta_0)^2. \tag{B.5}$$

[31] The inverse-gamma PDF is $f_{IG}(y|a,b) = b^a(1/y)^{a+1}\exp(-b/y)/\Gamma(a)$, where Γ denotes the gamma function. Parameters $a_0 > 0$ and $b_0 > 0$ dictate the shape and scale of the distribution, respectively.

Furthermore, the agent's marginal belief over θ_t after t rounds is again a Student's t-distribution with mean $\hat{\theta}_t$. Hence, in any period t, $\hat{\theta}_{t-1}$ is the agent's expectation of x_t and thus $\hat{\theta}_{t-1}$ is her reference point. This means that the process of encoded outcomes in this extension follows the same process as described in Equation (4), except $\hat{\theta}_{t-1}$ is given by Equation (B.2).

Accordingly, the dynamic process governing $\hat{\theta}_t$ described above is essentially identical to the process of biased mean beliefs considered in the main text, and thus the convergence argument from Proposition 4 applies here. To see this explicitly, note that by defining $\tilde{\alpha}_t \equiv 1/(v_0 + t)$, we can write the updating rule for $\hat{\theta}_t$ in Equation (B.2) as $\hat{\theta}_t = \tilde{\alpha}_t \hat{x}_t + (1 - \tilde{\alpha}_t)\hat{\theta}_{t-1}$. Notice that $\tilde{\alpha}_t$ has the same functional form as the α_t from the main text; in fact, the two are identical when $v_0 = \sigma^2/\rho^2$. Additionally, conditional on $\hat{\theta}_{t-1}$, $\hat{x}_t$ is defined the same way here as in the main text. As such, the dynamic process for $\hat{\theta}_t$ here is equivalent to the one presented in Equation (A.23) in the proof of Proposition 4 (up to the irrelevant scaling constant v_0). The convergence argument in that proof (based on Theorem 5.2.1. of Kushner and Yin, 2003) can therefore be applied here to the process of $\hat{\theta}_t$. Thus, conditional on θ and σ^2, the limiting value of $\hat{\theta}_t$ is identical to the value $\hat{\theta}_\infty$ described in Proposition 4.

It is worth noting a key reason that beliefs about θ in this extension converge to the same value as in the main text. Given our assumed class of priors over σ^2 and θ, the agent's beliefs about σ^2 do not influence her expectation of x_t in any given round—this expectation is entirely determined by $\hat{\theta}_t$, which is a simple weighted average of the encoded outcomes. This implies that the agent's beliefs about σ^2 do not influence her encoded outcomes. Thus, the dynamic feedback process linking the agent's expectations and her encoded outcomes is independent of the agent's uncertainty over σ^2, and hence this process plays out similarly to the case in which σ^2 is known.

Finally, we can derive the agent's perceived variance by examining the agent's limiting belief over σ^2. After t rounds, the agent's expectation of σ^2 is $b_t/(a_t - 1)$, which—using Equations (B.4) and (B.5)—has the same limiting value as $\lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^t (\hat{x}_i - \widehat{S}_t/t)^2$. Since $\hat{\theta}_t$ converges a.s. to the value $\hat{\theta}_\infty$ from Proposition 4 (as argued immediately above), it follows that $\lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^t (\hat{x}_i - \widehat{S}_t/t)^2 = \lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^t (\hat{x}_i - \hat{\theta}_\infty)^2$. Note that this value is equal to the theoretical variance of encoded outcomes at the steady-state mean belief, $\hat{\theta}_\infty$. By Lemma B.2, this value strictly exceeds σ^2 under misattribution, regardless of the underlying parameters. Hence, the agent's expectation of σ^2 converges to a long-run value that exaggerates the variance of outcomes.

B.4. The optimal walk-down of expectations

Consider the first application presented in Section 5: the designer aims to maximize the agent's posterior expectation of θ following a single outcome, x_1, and the designer has the opportunity to lower the agent's expectations prior to the realization of x_1. The designer thus chooses $c \geq 0$ to maximize

$$\mathbb{E}[\hat{\theta}_1] = \alpha_1 \mathbb{E}[\hat{x}_1(c)] + (1 - \alpha_1)\big(\theta_0 - c\big), \tag{B.6}$$

where $\mathbb{E}[\cdot]$ is w.r.t. the designer's prior belief that $\theta \sim N(\theta_0, \rho^2)$, and $\hat{x}_1(c)$ is the agent's misencoded value of x_1 conditional on holding prior expectation $\theta_0 - c$. We also impose a "participation constraint" corresponding to a lower bound on the induced prior expectation; namely, we require $\theta_0 - c \geq \bar{\theta}$. Let c^* denote the value of c that maximizes Objective (B.6) subject to $c \in [0, \theta_0 - \bar{\theta}]$.

Proposition B.1. *Consider the setup above. The designer will optimally reduce the agent's expectations by c^* such that*

$$c^* = \begin{cases} 0 & \textit{if} \quad \frac{\sigma^2}{\rho^2} > \frac{\kappa^L + \kappa^G}{2}, \\ \nu \left| \Phi^{-1}\left(\frac{\frac{\sigma^2}{\rho^2} - \kappa^G}{\kappa^L - \kappa^G} \right) \right| & \textit{if} \quad \frac{\sigma^2}{\rho^2} \in \left[\kappa^G, \frac{\kappa^L + \kappa^G}{2} \right] \textit{ and } \nu \left| \Phi^{-1}\left(\frac{\frac{\sigma^2}{\rho^2} - \kappa^G}{\kappa^L - \kappa^G} \right) \right| \leq \theta_0 - \bar{\theta}, \\ \theta_0 - \bar{\theta} & \textit{if} \quad \frac{\sigma^2}{\rho^2} < \kappa^G. \end{cases}$$

Thus, lowering the agent's prior expectation is never optimal in environments where the relative precision of her prior, σ^2/ρ^2, is high. However, once this precision is sufficiently low, it is optimal to walk down her expectations, and the appropriate amount to do so increases as the precision of the prior decreases.

Proof. [Proof of Proposition B.1] To derive c^*, we first derive $\mathbb{E}[\hat{x}_1(c)]$. Let $\nu = \sqrt{\rho^2 + \sigma^2}$. From the designer's perspective, $x_1 \sim N(\theta_0, \nu^2)$. Since $\hat{x}_1 = x_1 + \kappa_1(x_1 - (\theta_0 - c))$ where $\kappa_1 \equiv \kappa^G \mathbb{1}\{x_1 > \theta_0 - c\} + \kappa^L \mathbb{1}\{x_1 < \theta_0 - c\}$, we have

$$\begin{aligned} \mathbb{E}[\hat{x}_1(c)] = \theta_0 + \kappa^G \Big[1 - \Phi\Big(\big((\theta_0 - c) - \theta_0 \big)/\nu \Big) \Big] \mathbb{E}\big[x_1 - (\theta_0 - c) | x \geq \theta_0 - c \big] \\ + \kappa^L \Phi\Big(\big((\theta_0 - c) - \theta \big)/\nu \Big) \mathbb{E}\big[x_1 - (\theta_0 - c_0) | x < \theta_0 - c \big]. \end{aligned} \quad \text{(B.7)}$$

Letting $z(c) \equiv -c/\nu$ and letting Z be a standard normal random variable, the above reduces to:

$$\begin{aligned} \mathbb{E}[\hat{x}_1(c)] &= \theta_0 + \nu\kappa^G \big[1 - \Phi(z(c)) \big] \mathbb{E}\big[Z - z(c) | Z \geq z(c) \big] \\ &\quad + \nu\kappa^L \Phi(z(c)) \mathbb{E}\big[Z - z(c) | Z < z(c) \big] \\ &= \theta_0 - \nu\kappa^G \left[(1 - \Phi(z(c)))z(c) - \phi(z(c)) \right] - \nu\kappa^L \left[\Phi(z(c))z(c) + \phi(z(c)) \right] \\ &= \theta_0 + \kappa^G c - \nu(\kappa^L - \kappa^G) G(c), \end{aligned} \quad \text{(B.8)}$$

where $G(c) \equiv [\Phi(z(c))z(c) + \phi(z(c))]$. Thus, $G'(c) = -\Phi(z(c))/\nu$, and hence

$$\frac{\partial}{\partial c} \mathbb{E}[\hat{x}_1(c)] = \kappa^G + (\kappa^L - \kappa^G)\Phi(-c/\nu) > 0. \quad \text{(B.9)}$$

Given the objective in Equation (B.6), the marginal benefit of decreasing expectations is

$$\alpha_1 \frac{\partial}{\partial c} \mathbb{E}[\hat{x}_1(c)] = \frac{\rho^2}{\rho^2 + \sigma^2} \left(\kappa^G + (\kappa^L - \kappa^G)\Phi\left(\frac{-c}{\sqrt{\rho^2 + \sigma^2}} \right) \right) > 0, \quad \text{(B.10)}$$

where we've used the fact that $\nu = \sqrt{\rho^2 + \sigma^2}$; the marginal cost is $(1 - \alpha_1) = \sigma^2/(\rho^2 + \sigma^2)$. Hence, the first-order condition for c^* is

$$\frac{\rho^2}{\rho^2 + \sigma^2} \left(\kappa^G + (\kappa^L - \kappa^G)\Phi\left(\frac{-c}{\sqrt{\rho^2 + \sigma^2}} \right) \right) = \frac{\sigma^2}{\rho^2 + \sigma^2}. \quad \text{(B.11)}$$

The SOC for a maximum holds since $-(\kappa^L - \kappa^G)\phi\left(-c/\sqrt{\rho^2 + \sigma^2} \right) < 0$, and hence an interior solution for c^* is given by:

$$c^* = -\sqrt{\rho^2 + \sigma^2}\,\Phi^{-1}\left(\frac{\frac{\sigma^2}{\rho^2} - \kappa^G}{\kappa^L - \kappa^G} \right). \quad \text{(B.12)}$$

Note that the expression for c^* above can sometimes be undefined or violate the constraint that c cannot be negative. This happens whenever $(\frac{\sigma^2}{\rho^2} - \kappa^G)/(\kappa^L - \kappa^G) \notin (0, 1/2)$. In such cases, we have a corner solution. Thus, the complete solution for c^* is as follows:

$$c^* = \begin{cases} 0 & \text{if} \quad \frac{\sigma^2}{\rho^2} > \frac{\kappa^L+\kappa^G}{2}, \\ -\nu\Phi^{-1}\left(\frac{\frac{\sigma^2}{\rho^2}-\kappa^G}{\kappa^L-\kappa^G}\right) & \text{if} \quad \frac{\sigma^2}{\rho^2} \in \left[\kappa^G, \frac{\kappa^L+\kappa^G}{2}\right] \text{ and } -\nu\Phi^{-1}\left(\frac{\frac{\sigma^2}{\rho^2}-\kappa^G}{\kappa^L-\kappa^G}\right) \le \theta_0 - \bar{\theta}, \\ \theta_0 - \bar{\theta} & \text{if} \quad \frac{\sigma^2}{\rho^2} < \kappa^G. \end{cases}$$

□

B.5. *The optimal ordering of outcomes for imprecise priors*

Proposition 6 shows that if prior beliefs are relatively precise (i.e., $\sigma^2/\rho^2 > \kappa^L$), then the misattributor's posterior expectation following a fixed set of outcomes is the highest when those outcomes are experienced in an increasing order. The following corollary shows that this result directly extends for any value of σ^2/ρ^2 when $T = 2$ so long as at last one of the outcomes beats initial expectations.

Corollary 1. *Consider any $a, b \in \mathbb{R}$ such that $a > b$. Let $\hat{\theta}_2^i$ denote the misattributor's posterior expectation following the increasing sequence (b, a), and let $\hat{\theta}_2^d$ denote it following the decreasing sequence (a, b). If $a > \theta_0$, then $\hat{\theta}_2^i > \hat{\theta}_2^d$.*

Proof of Corollary 1. Consider the proof of Lemma A.2. The relevant cases given $a > \theta_0$ are Cases 1 and 3. However, in both of these cases, we have $\hat{\theta}_2^i > \hat{\theta}_2^d$ without assuming anything beyond $a > b$ and $a > \theta_0$. □

Providing a full characterization of the optimal sequencing of outcomes when $\sigma^2/\rho^2 < \kappa^L$ requires arranging outcomes according to Lemma 1 so that the largest outcomes get the largest weight. Although this is mathematically simple, it does not provide much insight; hence, we focus on $T = 2$ in Corollary 1 to deliver a crisp result.

Appendix C. Model extensions

C.1. *Misattribution and personal equilibrium*

In our baseline model, the misattributing agent faces an exogenous distribution of outcomes. To handle environments where the agent's actions influence the distribution of data, we must further specify how the agent's strategy influences her reference points, since this determines how she (mis)encodes outcomes. Extending the model is relatively straightforward when the agent's current action pins down her reference point.

To motivate this extension, consider a setting where the agent is learning about two normally-distributed prospects with unknown means, θ^A and θ^B. In each period, she must choose between a random draw from either A or B. A natural way to model this setting follows Kőszegi and Rabin's (2007) notion of "choice-acclimating personal equilibrium", but allows for the agent to hold biased expectations and make choices with respect to those biased beliefs. Following this approach, the agent's current choice determines her reference point for that period. For instance, if the agent chooses prospect A today, then her reference point is what she (currently) expects to earn from prospect A. As in Kőszegi and Rabin (2007), the agent understands that her actions will determine her reference points and accordingly takes actions to maximize her expected utility given her (potentially biased) beliefs about the two prospects.

We now describe this extension more formally. Each period t begins with a decision phase in which the agent selects an action $a_t \in \mathcal{A}$, where $\mathcal{A}$ is a compact subset of $\mathbb{R}$. As in the baseline model, the agent is initially uncertain about a distributional parameter $\theta \in \mathbb{R}^K$ for some finite $K \geq 1$. Conditional on the parameter θ and chosen action a_t, outcome $x_t \in \mathbb{R}$ is distributed according to $F(\cdot|\theta, a_t)$. For an example similar to our baseline model, imagine the agent is learning about N independent prospects; each prospect $n \in \mathcal{A} = \{1, \dots, N\}$ has normally distributed outcomes with mean θ_n and a known variance. Then $\theta = (\theta_1, \dots, \theta_N)$, and $F(\cdot|\theta, n)$ is a normal distribution with mean θ_n.

The agent begins with a prior π_0 over θ, and updates her beliefs each round conditional on her action and its resulting (mis)encoded outcome. To formalize the encoded outcome, let π_{t-1} denote the agent's belief over θ entering round t. Conditional on choice a_t, her expected outcome is $\widehat{\mathbb{E}}_{t-1}[x_t|a_t] \equiv \int_{-\infty}^{\infty} x d\widehat{F}_{t-1}(x|a_t)$ where $\widehat{F}_{t-1}(x|a_t) \equiv \int F(x|\theta, a_t) d\pi_{t-1}(\theta)$, and thus the encoded outcome is analogous to Equation (4), except $\widehat{\mathbb{E}}_{t-1}[x_t|a_t]$ takes the place of $\hat{\theta}_{t-1}$:

$$\hat{x}_t = \begin{cases} x_t + \kappa^G \left(x_t - \widehat{\mathbb{E}}_{t-1}[x_t|a_t]\right) & \text{if} \quad x_t \geq \widehat{\mathbb{E}}_{t-1}[x_t|a_t] \\ x_t + \kappa^L \left(x_t - \widehat{\mathbb{E}}_{t-1}[x_t|a_t]\right) & \text{if} \quad x_t < \widehat{\mathbb{E}}_{t-1}[x_t|a_t]. \end{cases} \tag{C.1}$$

In the example above with normal distributions, the agent's reference point in a round in which she chooses prospect n is simply her current estimate of θ_n.

Under this approach, the agent then takes an action in each round to maximize her expected utility (according to her true utility function) conditional on her erroneous beliefs. We find it natural to assume the agent chooses an action under the presumption that she encodes outcomes correctly; that is, she is naive about her attribution bias. We find it similarly natural that the agent makes decisions with respect to her true utility function given that attribution bias is a retrospective error. More generally, we can think of the concept above as a *biased-belief personal equilibrium*, as it extends Kősezgi and Rabin's notion of personal equilibrium to the case where the agent holds erroneous expectations.

It is wroth noting that many of our results from the main text apply in this extension so long as outcomes from different actions are independent of one another. This follows from the assumption that the reference point corresponds to expectations about the chosen action: if outcomes are also independent across actions, then updating about one action does not influence updating about another. For instance, in the example above with normal distributions, beliefs about the mean of each prospect would exhibit the precise properties described in Sections 3 and 4 along the sequence of periods in which that prospect is chosen.

C.2. Misattribution with multiple dimensions

Here, we discuss how to extend our model to settings where consumption utility is multidimensional. This extension requires an additional assumption on how surprises along one dimension influence encoded outcomes on other dimensions. While there are a range of plausible assumptions, we assume that the encoded outcome on one dimension depends entirely on sensations of elation or disappointment felt on that dimension. We propose this specific assumption to eliminate a potential degree of freedom and to provide a starting place for potential empirical exploration.

Following Kőszegi and Rabin's (2006) multidimensional model, suppose consumption vector $c \in \mathbb{R}^K$ generates consumption utility $x \in \mathbb{R}^K$ that is additively separable across K dimensions. Let $x = (x^1, \ldots, x^K)$ with $x^k \in \mathbb{R}$ denoting consumption utility on dimension k, and let $\widehat{F}$ denote the agent's subjective CDF over x. Define the vector $\hat{\theta} = (\hat{\theta}^1, \ldots, \hat{\theta}^K)$ such that element $\hat{\theta}^k$ denotes the expected consumption utility on dimension k according to $\widehat{F}$. The person's total utility from x given reference distribution $\widehat{F}$ is then $u(x|\hat{\theta}) = \sum_{k=1}^{K} u_k(x^k|\hat{\theta}^k)$, where $u_k(x^k|\hat{\theta}^k) \equiv x^k + \eta n(x^k|\hat{\theta}^k)$ is the total utility along dimension k and $n(x^k|\hat{\theta}^k)$ is the unidimensional gain-loss utility assumed in our baseline model (Equation (1)).

Our notion of misattribution generally extends to this setting: following outcome x and total utility level $u = u(x|\hat{\theta})$, a misattributor encodes a distorted value $\hat{x}$ that would have generated the same total utility level u if she instead had a utility function $\hat{u}(\cdot|\hat{\theta})$ that weights each gain-loss term, $n(\cdot|\hat{\theta}^k)$, by $\hat{\eta} \in [0, \eta)$. That is, the person encodes $\hat{x}$ that solves $\hat{u}(\hat{x}|\hat{\theta}) = u(x|\hat{\theta})$ as in Equation (3). To further pin down the misencoded outcome on each dimension, we assume that each $\hat{x}^k$ depends solely on gains and losses experienced on dimension k: $\hat{x}^k$ is defined by $\hat{u}_k(\hat{x}^k|\hat{\theta}^k) = \hat{x}^k + \hat{\eta} n(\hat{x}^k|\hat{\theta}^k) = x^k + \eta n(x^k|\hat{\theta}^k) = u_k(x^k|\hat{\theta}^k)$. While we suspect that the more general psychology of "attribution bias" may lead to across-dimension misencoding (see, e.g., discussions in Haggag et al., 2019), we believe this formulation provides a tractable stepping stone for future research.

References

Abeler, J., Falk, A., Goette, L., Huffman, D., 2011. Reference points and effort provision. Am. Econ. Rev. 101 (2), 470–492.

Adhvaryu, A., Nyshadham, A., Xu, H., 2020. Hostel Takeover: Living Conditions, Reference Dependence, and the Well-being of Migrant Workers. Working Paper.

Anderson, R., 1973. Consumer dissatisfaction: the effect of disconfirmed expectancy on perceived product performance. J. Mark. Res. 10, 38–44.

Backus, M., Blake, T., Masterov, D., Tadelis, S., 2022. Expectation, disappointment, and exit: evidence on reference point formation from an online marketplace. J. Eur. Econ. Assoc. 20 (1), 116–149.

Banerji, A., Gupta, N., 2014. Detection, identification, and estimation of loss aversion: evidence from an auction experiment. Am. Econ. J. Microecon. 6, 91–133.

Baumeister, R., Finkenauer, C., Vohs, K., 2001. Bad is stronger than good. Rev. Gen. Psychol. 5 (4), 323–370.

Bell, D., 1985. Disappointment in decision making under uncertainty. Oper. Res. 33 (1), 1–27.

Benjamin, D., Bodoh-Creed, A., Rabin, M., 2019. Base-Rate Neglect: Foundations and Implications. Working Paper.

Bhargava, S., 2007. Perception is Relative: Contrast Effects in the Field. Working Paper.

Bhargava, S., Fisman, R., 2014. Contrast effects in sequential decisions: evidence from speed dating. Rev. Econ. Stat. 96 (3), 444–457.

Bohren, A., 2016. Informational herding with model misspecification. J. Econ. Theory 163, 222–247.

Bohren, A., Hauser, D., 2021. Learning with model misspecification: characterization and robustness. Econometrica 89, 3025–3077.

Bordalo, P., Gennaioli, N., Shleifer, A., 2017. Diagnostic expectations and credit cycles. J. Finance 73 (1), 199–227.

Bordalo, P., Gennaioli, N., Shleifer, A., 2020. Memory, attention and choice. Q. J. Econ. 135 (3), 1399–1442.

Boulding, W., Kalra, A., Staelin, R., Zeithaml, V., 1993. A dynamic process model of service quality: from expectations to behavioral intentions. J. Mark. Res. 30, 7–27.

Brownback, A., Kuhn, M., 2019. Understanding outcome bias. Games Econ. Behav. 117, 342–360.

Bushong, B., Gagnon-Bartsch, T., 2022. Reference dependence and attribution bias: evidence from real-effort experiments. Am. Econ. J. Microecon. Forthcoming.

Card, D., Dahl, G., 2011. Family violence and football: the effect of unexpected emotional cues on violent behavior. Q. J. Econ. 126 (1), 103–143.

Chambers, C., Healy, P., 2012. Updating towards the signal. Econ. Theory 50, 765–786.

Chen, D., Moskowitz, T., Shue, K., 2016. Decision-making under the gambler's fallacy: evidence from asylum judges, loan officers, and baseball umpires. Q. J. Econ. 131 (3), 1181–1241.

Crawford, V., Meng, J., 2011. New York city cabdrivers' labor supply revisited: reference-dependent preferences with rational- expectations targets for hours and income. Am. Econ. Rev. 101 (5), 1912–1932.

Dillenberger, D., Rozen, K., 2015. History-dependent risk attitude. J. Econ. Theory 157, 445–477.

Ehling, P., Graniero, A., Heyerdahl-Larsen, C., 2018. Asset prices and portfolio choice with learning from experience. Rev. Econ. Stud. 85 (3), 1752–1780.

Epstein, L., Noor, J., Sandroni, A., 2010. Non-Bayesian learning. B. E. J. Theor. Econ. 10 (1).

Esponda, I., Pouzo, D., 2016. Berk-Nash equilibrium: a framework for modeling agents with misspecified models. Econometrica 84 (3), 1093–1130.

Eyster, E., Rabin, M., 2010. Naive herding in rich-information settings. Am. Econ. J. Microecon. 2 (4), 221–243.

Ericson, K., Fuster, A., 2011. Expectations as endowments: evidence on reference-dependent preferences from exchange and valuation experiments. Q. J. Econ. 126 (4), 1879–1907.

Frick, M., Iijima, R., Ishii, Y., 2020. Misinterpreting others and the fragility of social learning. Econometrica 88 (6), 2281–2328.

Fryer, R., Harms, P., Jackson, M., 2019. Updating beliefs when evidence is open to interpretation: implications for bias and polarization. J. Eur. Econ. Assoc. 17 (5), 1470–1501.

Fudenberg, D., Romanyuk, G., Strack, P., 2017. Active learning with a misspecified prior. Theor. Econ. 12, 1155–1189.

Geers, A., Lassiter, G., 1999. Affective expectations and information gain: evidence for assimilation and contrast effects in affective experience. J. Exp. Soc. Psychol. 35 (4), 394–413.

Gilbert, D., Malone, P., 1995. The correspondence bias. Psychol. Bull. 117 (1), 21–38.

Gill, D., Prowse, V., 2012. A structural analysis of disappointment aversion in a real effort competition. Am. Econ. Rev. 102 (1), 469–503.

Gneezy, U., Goette, L., Sprenger, C., Zimmermann, F., 2017. The limits of expectations-based reference dependence. J. Eur. Econ. Assoc. 15, 861–876.

Goette, L., Harms, A., Sprenger, C., 2019. Randomizing endowments: an experimental study of rational expectations and reference-dependent preferences. Am. Econ. J. Microecon. 11 (1), 185–207.

Haggag, K., Pope, D., Bryant-Lees, K., Bos, M., 2019. Attribution bias in consumer choice. Rev. Econ. Stud. 86 (5), 2136–2183.

Haggag, K., Patterson, R., Pope, N., Feudo, A., 2021. Attribution bias in major decisions: evidence from the United States military academy. J. Public Econ. 200, 104445.

Haisley, E., Loewenstein, G., 2011. It's not what you get but when you get it: the effect of gift sequence on deposit balances and customer sentiment in a commercial bank. J. Mark. Res. 48 (1), 103–115.

Hanna, R., Duflo, E., Greenstone, M., 2016. Up in smoke: the influence of household behavior on the long-run impact of improved cooking stoves. Am. Econ. J., Econ. Policy 8 (1), 80–114.

Hanna, R., Mullainathan, S., Schwartzstein, J., 2014. Learning through noticing: theory and evidence from a field experiment. Q. J. Econ. 129 (3), 1311–1353.

He, K., 2021. Mislearning from censored data: the gambler's fallacy in optimal-stopping problems. Theor. Econ. Forthcoming.

Heffetz, O., List, J., 2014. Is the endowment effect an expectations effect? J. Eur. Econ. Assoc. 12, 1396–1422.

Heidhues, P., Kőszegi, B., Strack, P., 2018. Unrealistic expectations and misguided learning. Econometrica 86 (4), 1159–1214.

Heidhues, P., Kőszegi, B., Strack, P., 2021. Convergence in misspecified learning models with endogenous actions. Theor. Econ. 16 (1), 73–99.

Ho, T., Zheng, Y., 2004. Setting customer expectations in service delivery: an integrated marketing-operations perspective. Manag. Sci. 50 (4), 479–488.

Hogarth, R., Einhorn, H., 1992. Order effects in belief updating: the belief-adjustment model. Cogn. Psychol. 24, 1–55.

Imas, A., 2016. The realization effect: risk-taking after realized versus paper losses. Am. Econ. Rev. 106 (8), 2086–2109.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47 (2), 263–291.

Karle, H., Kirchsteiger, G., Peitz, M., 2015. Loss aversion and consumption choice: theory and experimental evidence. Am. Econ. J. Microecon. 7 (2), 101–120.

Kimball, D., Patterson, S., 1997. Living up to expectations: public attitudes toward congress. J. Polit. 59, 701–728.

Kopalle, P., Lehmann, D., 2006. Setting quality expectations when entering a market: what should the promise be? Mark. Sci. 25 (1), 8–24.

Kőszegi, B., Rabin, M., 2006. A model of reference-dependent preferences. Q. J. Econ. 121 (4), 1133–1165.

Kőszegi, B., Rabin, M., 2007. Reference-dependent risk attitudes. Am. Econ. Rev. 97 (4), 1047–1073.

Kremer, M., Rao, G., Schilbach, F., 2019. Behavioral development economics. In: Bernheim, D., DellaVigna, S., Laibson, D. (Eds.), The Handbook of Behavioral Economics: Applications and Foundations (Volume 2). North-Holland.

Kushner, H., Yin, G., 2003. Stochastic Approximation and Recursive Algorithms and Applications, vol. 35. Springer.

Kuhnen, C., 2015. Asymmetric learning from financial information. J. Finance 70 (5), 2029–2062.

Malmendier, U., Nagel, S., 2011. Depression babies: do macroeconomic experiences affect risk-taking? Q. J. Econ. 126, 373–416.

Malmendier, U., Nagel, S., 2016. Learning from inflation experiences. Q. J. Econ. 131 (1), 53–87.

Malmendier, U., Pouzo, D., Vanasco, V., 2020. Investor experiences and financial market dynamics. J. Financ. Econ. 136, 597–622.

Murphy, K., 2007. Conjugate Bayesian Analysis of the Gaussian Distribution. Working Paper.

Nyarko, Y., 1991. Learning in misspecified models and the possibility of cycles. J. Econ. Theory 55, 416–427.

Oliver, R., 1977. Effect of expectation and disconfirmation of post-exposure product evaluation: an alternative interpretation. J. Appl. Psychol. 62, 480–486.

Oliver, R., 1980. A cognitive model of the antecedents and consequences of satisfaction decisions. J. Mark. Res. 17, 460–469.

Peeters, G., Czapinski, J., 1990. Positive-negative asymmetry in evaluations: the distinction between affective and informational negativity effects. Eur. Rev. Soc. Psychol. 1 (1), 33–60.

Pope, D., Schweitzer, M., 2011. Is tiger woods loss averse? Persistent bias in the face of experience, competition, and high stakes. Am. Econ. Rev. 101 (1), 129–157.

Post, T., van den Assem, M., Baltussen, G., Thaler, R., 2008. Deal or no deal? Decision making under risk in a large-payoff game show. Am. Econ. Rev. 98 (1), 38–71.

Rabin, M., Schrag, J., 1999. Inference by believers in the law of small numbers. Q. J. Econ. 114 (1), 37–82.

Ross, L., 1977. The intuitive psychologist and his shortcomings: distortions in the attribution process. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology. Academic Press, pp. 173–220.

Ross, W., Simonson, I., 1991. Evaluations of pairs of experiences: a preference for happy endings. J. Behav. Decis. Mak. 4, 272–282.

Schwartzstein, J., 2014. Selective attention and learning. J. Eur. Econ. Assoc. 12 (6), 1423–1452.

Shin, M., 2021. Subjective expectations, experiences, and stock market participation: evidence from the lab. J. Econ. Behav. Organ. 186, 672–689.

Thakral, N., To, T., 2021. Daily labor supply and adaptive reference points. Am. Econ. Rev. 111 (8), 2417–2443.

Thaler, R., Johnson, E., 1990. Gambling with the house money and trying to break even: the effects of prior outcomes on risky choice. Manag. Sci. 36, 643–660.

Thorisson, H., 1995. Coupling methods in probability theory. Scand. J. Stat. 22 (2), 159–182.

Waterman, R., Jenkins-Smith, H., Silva, C., 1999. The expectations gap thesis: public attitudes toward an incumbent president. J. Polit. 61 (4), 944–966.

Wilson, T., Gilbert, D., 2003. Affective forecasting. Adv. Exp. Soc. Psychol. 35, 345–411.