Learning with Misattribution of Reference Dependence

Tristan Gagnon-Bartsch
Harvard University

Benjamin Bushong[*]
Michigan State University

October 30, 2019

Abstract

This paper examines errors in learning that arise when an agent's perception of outcomes depends on how they contrast with expectations. We consider an agent who neglects how the sensation of elation or disappointment relative to expectations contributes to her overall utility, and who wrongly attributes this component of her utility to the intrinsic value of an outcome. Our model helps explain seemingly disparate evidence on belief updating in dynamic environments. We show that misattribution of reference dependence generates a contrast effect in sequential evaluations. This leads to a recency bias: the misattributor's beliefs over-weight recent experiences and under-weight earlier ones. Accordingly, a misattributor forms inflated expectations after experiencing a series of outcomes arranged in an increasing order. Turning to long-run beliefs, a loss averse misattributor will grow unduly pessimistic and undervalue prospects in proportion to their variability, leading the decision maker to reject some risky-but-optimal options. Finally, we highlight how misattribution introduces incentives for expectations management, and we show that a misattributing principal will overestimate the ability of a sophisticated party who initially suppresses expectations so as to exceed them thereafter.

JEL Classification: D83, D84, D91.
Keywords: Learning from experience, attribution bias, reference dependence, misspecified models.

[*] E-mails: gagnonbartsch@fas.harvard.edu and bbushong@msu.edu. We thank Ned Augenblick, Stefano DellaVigna, Jon Eguia, Ben Enke, Erik Eyster, Xavier Gabaix, Ben Golub, David Laibson, Devin Pope, Matthew Rabin, Gautam Rao, Josh Schwartzstein, Andrei Shleifer, Kelly Shue, Dmitry Taubinsky, Hanzhe Zhang, and seminar audiences at Berkeley, Caltech, Cornell, Harvard, Harvard Business School, LSE, the Norwegian School of Economics (NHH), Princeton, SITE, Stanford, and UCSD for comments.

1 Introduction

Learning from personal experience guides a wide range of economic decisions—it shapes, for instance, our preferences over consumer products, adoption of new technologies, and evaluations of others. But our ability to correctly learn from experience is challenged by the fact that our expectations often color the way we perceive events.[1] In this paper, we study how a person's impressions and memories are distorted when she inadvertently contrasts experiences against her expectations. Our model builds from the well-known idea that people evaluate experiences in both absolute and relative terms: the utility from an outcome depends on both its "intrinsic value" and how that value compares to expectations (e.g., Kahneman and Tversky 1979; Bell 1985; Kőszegi and Rabin 2006). When trying to learn the intrinsic value of an outcome, a person must distinguish this intrinsic ("reference-free") value from the sensation of elation or disappointment it generated. Research in psychology, however, suggests that people fail to appreciate how circumstantial factors—such as expectations—shape their experiences, and in doing so may incorrectly attribute sensations of surprise (i.e., elation or disappointment) to their intrinsic tastes.[2]

To illustrate, imagine a consumer trying a new service for the first time (e.g., a traveler flying on a new airline, or a new shopper on eBay as in the related empirical paper by Backus et al. 2018). If her experience falls short of expectations, she will feel unhappy both because of the subpar service and because this came as a 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 service, and consequently underestimate how much she would enjoy that service in the future.[3] For another example, consider a

[1] Other social sciences have long explored how prior expectations can alter a person's evaluation and memory of an experience (see, e.g., Wilson and Gilbert 2003 for a review in psychology). For instance, political scientists have argued that discrepancies between a politician's performance and citizens' expectations play a key role in how citizens perceive that politician (see, e.g., Patterson et al. 1969; Kimball and Patterson 1997; Waterman et al. 1999). Likewise, marketing has emphasized the role of expectations on perceived quality or service (see, e.g., seminal works from Oliver 1977, 1980; and Boulding et al. 1993). This research has highlighted that when outcomes deviate from expectations, a person might either assimilate that experience—interpret it in favor of their current beliefs—or contrast it—interpret the experience against their expectations. Models of confirmation bias (e.g., Rabin and Schrag 1999; Fryer, Harms, and Jackson 2018) explore the former; in this paper, we propose a mechanism that captures the latter. Finding support for both hypotheses, 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." We briefly discuss in Section 3 how both contrast and assimilation can peacefully coexist.

[2] Our model is inspired by a broader literature on other forms of misattribution demonstrating a tendency to wrongly attribute extraneous situational factors to the inherent characteristics of a good or person. For instance, Haggag and Pope (2018) demonstrate that, when assessing the value of a good, people have difficulty separating state-dependent utility caused by temporary circumstances, such as thirst, from the quality of the good. Dutton and Aron (1974) find that subjects who form opinions about people they meet for the first time exhibit judgments dependent on unrelated factors (e.g., their current state of fear or excitement). We discuss this literature at greater length in Section 2.

[3] Backus et al. (2018) find that new eBay users with higher expectations of winning an auction (measured by time

researcher collaborating with a new colleague. If the colleague contributes more than expected to the project, the researcher will feel happy both because of the project's progress and because this came as a pleasant surprise. If she misattributes this latter feeling to her partner's performance, she may recall an exaggerated perception of his contribution. As these examples suggest, surprises may distort perceived outcomes: exceeding expectations inflates perceptions, and falling short deflates them.

In this paper, we study an agent who attempts to learn the average payoff of a prospect—e.g., the average quality of a consumer service or a worker's ability—over time. After each outcome, the agent misremembers how the sensation of positive or negative surprise influenced her experienced utility and wrongly attributes this sensation to the underlying quality of that outcome. This mechanism, a form of attribution bias, provides an intuitive channel through which reference points influence memory. Furthermore, as the agent experiences outcomes over time, her perceptions become interdependent: a misinterpretation of today's outcome causes the agent to form biased expectations about tomorrow, and these biased expectations further shape the interpretation of the following outcome. We show how this process explains well-documented errors in beliefs like contrast effects and a recency bias. Moreover, our mechanism sheds light on lesser-discussed biases such as the fact that people form the most optimistic impressions after experiencing outcomes ordered in an increasing sequence. We further characterize how misattribution interacts with loss aversion to distort long-run learning. Finally, we demonstrate how misattribution (i) introduces incentives for expectations management and (ii) generates a dynamic sunk-cost fallacy.

We introduce the baseline model in Section 2. We consider a dynamic setting where in each period the agent realizes an outcome of a prospect with an unknown distribution. 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 total utility. To preview our formulation of misattribution, 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 correctly recalls her total utility, but misremembers the extent to which elation or disappointment contributed to this total. She thus infers a distorted value of each prior outcome. More specifically, 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 about the distribution of outcomes according to Bayes' Rule as if $\hat{x}$

spent in the lead) are more likely to quit using the platform if they unexpectedly lose their first auction. Specifically, bidders who surprisingly lose near the end of the auction are six percentage points more likely to exit the platform for every additional day they previously held the lead. This suggests that misattribution of disappointment can have important consequences for customer perceptions and retention, independent of material outcomes.

truly occurred.

This simple model captures several well-known ideas. First, it captures the basic concept of disconfirmation: an outcome that deviates from expectations is remembered as deviating by more than it really did. Second, our model provides an explanation that connects the "positive-negative asymmetry effect"—the notion that people's beliefs respond differentially to good and bad news—to reference-dependent preferences. If the misattributor is loss averse, then disappointments distort beliefs by more than commensurate elations, leading to an asymmetry in updating from good versus bad news.[4] Third, misattribution generates sequential contrast effects: the current outcome appears better the worse was the previous one.[5]

We begin our primary analyses in Section 3 by examining how the order in which a misattributor experiences outcomes influences her perceived value of the prospect. Consider, for example, a manager learning about the ability of a newly-hired employee whose performance determines the manager's payoffs. Even when the employee's performance is i.i.d., misattribution can generate a recency bias—recent outcomes influence beliefs more than older ones.[6] This stems from the contrast effect noted above: high initial outcomes raise expectations and cause later outcomes to be judged more harshly, while low initial outcomes lower expectations and cause later outcomes to be judged more favorably. We provide comparative statics on the strength of signals and priors that predict when a recency bias emerges. We also show that a misattributor is most optimistic about a prospect when, ceteris paribus, its outcomes are arranged in an improving order. For instance, fixing the total amount of work the employee completes, the manager is most optimistic about his ability when each of his performances is better than the last.

We extend this analysis in Section 4, where we demonstrate how misattribution distorts beliefs in the long-run. The interplay between beliefs and perceived outcomes can prevent a misattributor from reaching correct expectations despite ample experience. However, the agent will nevertheless converge to stable long-run beliefs. We characterize these steady-state beliefs and show that a mis-

[4] Baumeister et al. (2001) provide a succint definition of the positive-negative assymetry 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) captures the notion that potential losses loom large in preferences, we provide a mechanism for why *past losses* loom large in both memory and subsequent forecasts.

[5] Sequential contrast effects have been documented in numerous settings, including sequential evaluations made by teachers (Bhargava 2007) and speed daters (Bhargava and Fisman 2014). In a financial setting, Hartzmark and Shue (2018) demonstrate that prior-day earnings announcements of other firms negatively correlate with stock-price reactions to contemporaneous announcements.

[6] Such a recency bias has been documented in a range of economic decisions, such as stock-market participation and hiring decisions (Highhouse and Gallo 1997). Malmendier, Pouzo, and Vanasco (2018) and Ehling, Graniero, and Heyerdahl-Larsen (2018) incorporate an exogenous recency effect into learning models and demonstrate how it helps explain phenomena such as excessive volatility in asset prices and trend-chasing behavior. Our model endogenously generates this recency effect.

attributor's encoded outcomes follow a distribution that appears excessively variable and, due to loss aversion, negatively skewed relative to the truth. Hence, the loss-averse misattributor underestimates a prospect's mean outcome. Furthermore, increasing the true variability of the prospect causes the agent to underestimate its mean by a greater extent, as increased variability amplifies sensations of disappointment on average. These biased beliefs imply that a misatributor will too often reject beneficial yet risky prospects.

In Section 5, we relax our baseline assumption that outcomes are i.i.d., which allows us to explore two natural applications of misattribution. In our primary application, we consider how a sophisticated party can strategically manipulate a misattributor's beliefs. We analyze a career-concern setting where a misattributing (but otherwise rational) principal sequentially updates her beliefs about a worker's ability based on his output. While classical models like Holmström (1999) predict that the worker's effort inefficiently declines over time, biased evaluations introduce new incentives that oppose the classical prediction. High effort today raises the principal's expectations and causes her to judge later output more harshly, and thus a sophisticated worker may under-perform relative to the principal's expectations early in the relationship in order to exceed them later.[7] In a second application, we clarify how misattribution can generate persistent forecasting errors even if the misattributor knows the mean of a prospect. We examine an environment with autocorrelated outcomes, and show that a misattributor will consistently form overly-extrapolative forecasts of future returns. Intuitively, when today's outcome beats expectations, a misattributor exaggerates its value and—assuming positive autocorrelation—expects unreasonably high outcomes in the future. As such, the next outcome typically disappoints and the agent reverts to overly-pessimistic expectations. This pattern will continue over time: the person forms an exaggerated forecast in the direction of the most recent outcome, which leads to a subsequent reversal.[8]

Throughout most of the paper, we highlight how a misattributor exhibits systematic patterns in her beliefs despite facing an exogenous distribution of outcomes. In Section 6, we describe how to extend this baseline model to scenarios where the misattributor's own actions influence the distribution she faces. We then illustrate this extension in a stylized repeated-search problem in which a

[7] This result resembles common strategies of "expectations management" used in a variety of fields. In marketing, Kopalle and Lehmann (2006) study how a firm should optimally restrain quality expectations when consumers have preferences that depend on expectations (known as the "gap model" in that literature; see, for example, Anderson 1973, Oliver 1977, or Ho and Zheng 2004). In finance, firms commonly use a variety of mechanisms to "walk down" investors' expectations prior to earnings announcements—strategic accounting of working capital and cash flow (Burgstahler and Dichev 1997), sales (Roychowdhury 2006), or distorting analyst forecasts (e.g., Richardson, Teoh, and Wycoki 2004). Furthermore, Bartov, Givoly, and Hayn (2002) argue that such efforts to meet or beat analyst expectations could yield significant excess stock returns. Indeed, Teoh, Yang, and Zhang (2009) find that firms are rewarded for beating expectations even when they actively suppress analyst forecasts. Our model provides a plausible mechanism that helps explain why restraining expectations can effectively manipulate beliefs.

[8] Our basic prediction of extrapolative and volatile forecasts accords with a range of evidence, including Greenwood and Shleifer (2014) and Gennaioli, Ma, and Shleifer (2015) who find that investors' and managers' predictions of their future earnings exhibit forecast errors that negatively relate with past performance.

decision maker can exert costly effort each period to increase the chance of a good outcome (e.g., a consumer who spends time comparison shopping before each purchase). Exerting effort will lead a misattributor to exaggerate the value of additional effort, since high effort raises expectations and causes bad outcomes (e.g., purchases that end up being lower quality than expected) to seem even worse when they happen. As such, the agent exhibits a form of sunk-cost fallacy—the more she has already worked, the more she feels compelled to try harder going forward—and inevitably settles on inefficiently high effort.

We conclude in Section 7 by noting ways that researchers can distinguish misattribution from other biases that share similar qualitative predictions. We also present some natural extensions of our model. For instance, our model can be reframed as a bias in social learning where an observer neglects how expectations shape the experiences of others. A student reading reviews for a class may fail to appreciate that some bad ratings reflect reviewers' high expectations rather than a low-quality professor. Failing to account for others' expectations may also have important implications for how policy makers interpret surveys measuring satisfaction. For instance, James (2009) and Van Ryzin (2004) find that reported satisfaction with public services declines with increased expectations. If policy makers neglect the role of expectations in these reports, they may wrongly attribute such a decline to poor quality or changing tastes and consequently propose ill-suited reforms. Moreover, misattribution captures a common intuition regarding why informational campaigns can backfire. If agencies tout the benefits of adopting, say, health practices or agricultural technologies, perceived outcomes may be biased downward because of high expectations, leading patients or farmers to prematurely abandon the new practices.

Our paper connects to several strands of literature. First, we relate to a literature demonstrating that behavior is particularly responsive to personal experience.[9] In particular, our model predicts that a misattributor overreacts to payoff-relevant experiences—those that incite elation and disappointment—relative to those that do not. Moreover, some research in this area suggests that such experience effects arise from endogenous preference formation in response to good or bad outcomes (e.g., Thaler and Johnson 1990; Dillenberger and Rozen 2015; Imas 2016), while others suggest these effects stem from beliefs that overreact to personal experience (e.g., Malmedier and Nagel 2011 and 2016; Vissing-Jorgensen 2003). Our framework highlights that these two channels are intertwined, and may provide a way to jointly explain these seemingly disparate results.[10]

[9] For instance, personal successes and failures play an important role in IPO subscription (Kaustia and Knüpfer 2008; Chiang et al. 2011), risk taking and stock-market participation (Malmendier and Nagel 2011), insurance take-up (Gallagher 2014), college-major choice (Xia 2016), and compliance with deadlines (Haselhuhn, Pope, and Schweitzer 2012).

[10] Malmendier and Nagel (2011) document increased apparent risk aversion as investors exit the stock market in response to adverse personal experiences, which is consistent with our model's predictions. They argue this stems from biased beliefs rather than altered preferences, and later work (Malmendier and Nagel 2016) provides more direct support for the belief channel. A related literature attempts to explain such phenomena by positing that risk preferences depend

Second, we join a growing literature that explores mistaken learning when agents hold misspecified models of the world.[11] Esponda and Pouzo (2016) provide a general framework for assessing the long-run beliefs and behavior of misspecified agents. Elements of our modeling approach—in particular, our analyses of long-run beliefs—are also similar to those of Heidhues, Kőszegi, and Strack (2018). They study an agent who overestimates her ability and consequently mislearns the value of a fundamental that determines how her effort translates to output. In their model, the agent misattributes poor performance to situational factors when in reality it was driven by lower-than-expected ability. Long-run mislearning stems from a feedback loop between the agent's erroneous beliefs and her actions, which endogenously determine the distribution of outcomes she observes. A similar feedback mechanism emerges in our model, but here it stems from the interplay between erroneous beliefs and the encoding of outcomes. Thus, anomalous dynamics emerge even when the agent does not take actions that influence the distribution of outcomes she faces.

2 A Model of Misattribution of Reference-Dependent Utility

In this section we present our baseline model in which an agent experiences a series of outcomes and attempts to learn the underlying distribution. We first introduce the learning environment and then describe the agent's reference-dependent utility function. We then formalize our notion of misattribution and describe some immediate implications for beliefs. We conclude this section with a discussion of our main assumptions and empirical motivation.

2.1 Learning Environment and Reference-Dependent Preferences

Learning Environment. We focus on an agent learning about a single prospect. 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 a parameter $\theta \in \mathbb{R}$.[12] The agent is initially uncertain about θ and attempts to

directly on a decision maker's history of elations and disappointments. However, these models (e.g., Dillenberger and Rozen 2015) predict a primacy rather than recency effect: early experiences have a greater impact on behavior than later experiences. This implication stands at odds with the recency effect documented by Malmendier and Nagel (2011).

[11] In addition to related models discussed throughout the paper, examples include Eyster and Rabin (2010), Bohren (2016), Bohren and Hauser (2018), and Frick, Iijima, and Ishii (2018) on misinferring from others' behavior in social-learning contexts; Barberis, Shleifer, and Vishny (1998) and Rabin (2002) on extrapolating from small samples; Madarász (2012) on information projection; Schwartzstein (2014) on selective attention; Spiegler (2016) on biases in causal reasoning; and Nyarko (1991) and Fudenberg, Romanyuk, and Strack (2017) on experimentation with misspecified priors. Similar to our model, both 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 much different than ours. Additionally, Epstein, Noor, and Sandroni (2010) analyze the limit beliefs of an agent who under- or overreacts to information. While they demonstrate that overreaction in general can cause beliefs to converge to a false distributional parameter, we can precisely pin down these limit beliefs given our focus on a specific misspecified model.

[12] For reasons that become obvious below, we work directly with the distribution of consumption utility rather than the distribution of material outcomes. We interpret x as if it derives from a classical Bernoulli utility function, $m : \mathbb{R} \to \mathbb{R}$,

learn its value based on past realizations of x_t. To focus the analysis, our baseline model assumes $x_t = \theta + \epsilon_t$ where $\epsilon_t \sim N(0, \sigma^2)$. Unless explicitly stated, we assume the agent knows the variance $\sigma^2 > 0$, and begins with a prior belief $\theta \sim N(\theta_0, \rho^2)$. Let π_0 denote this prior distribution. Following each realization x_t, the agent updates her beliefs over θ from π_{t-1} to π_t.

Reference-Dependent Preferences. Following Kőszegi and Rabin (2006; henceforth KR), we assume the agent's overall utility has two additively separable components. The first component, "consumption utility"—introduced above as x—corresponds to the payoff traditionally studied in economics. The second component, "gain-loss utility", derives from comparing x to a reference level of utility. Following Bell (1985), we take this reference point to be the agent's expectation of x, and we consider a simple piecewise-linear specification of gain-loss utility. Specifically, if the agent believes that x is distributed according to CDF $\widehat{F}$ with a mean value denoted by $\hat{\theta}$, then

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

where parameter $\lambda \geq 1$ allows for loss aversion.[13] Thus, when holding expectation $\hat{\theta}$, the agent's total utility 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.

In the learning environment introduced above, the agent's expectations over x_t change as she updates her beliefs about the underlying distribution. Thus, the agent's total utility today depends on past outcomes through their influence on her current expectation. The agent's expectation of x_t is pinned down by her expected value of parameter θ entering round t: given belief π_{t-1}, her expectation of x_t is $\hat{\theta}_{t-1} \equiv \int_{-\infty}^{\infty} \tilde{\theta} \, \mathrm{d}\pi_{t-1}(\tilde{\theta})$, and her total utility in period t is $u\big(x_t \big| \hat{\theta}_{t-1}\big)$ as specified in Equation 2.

2.2 Misattribution of Reference-Dependent Utility

We now turn to the central assumption of our model: the agent neglects how her past experiences were influenced by reference dependence and misattributes her gain-loss utility to the prospect's underlying consumption utility.

The agent seeks to learn θ—the average consumption utility from the prospect—and her experienced utility provides a signal about θ. A rational updater faced with signal $u_t \equiv x_t + \eta n\big(x_t \big| \hat{\theta}_{t-1}\big)$

over consumption realizations $c \in \mathbb{R}$ such that $x = m(c)$. Additionally, we extend the model to multiple consumption dimensions in Appendix B.

[13] While $n(\cdot|\cdot)$ shares many similarities with Kahneman and Tversky's (1979) value function, we abstract from the other elements of prospect theory—diminishing sensitivity and probability weighting—to focus on the role of reference points and loss aversion. Extending the model to incorporate these elements is a natural direction for future research.

understands this signal is "contaminated" by a transient gain-loss term and properly accounts for this when using u_t to update her beliefs about θ. We assume that a misattributor errs in this step: she infers from u_t as if her utility function weights gains and losses by a diminished factor $\hat{\eta} \in [0, \eta)$. Hence, a misattributor treats signal u_t as if $u_t = \hat{x}_t + \hat{\eta} n\big(\hat{x}_t \big| \hat{\theta}_{t-1}\big) \equiv \hat{u}\big(\hat{x}_t \big| \hat{\theta}_{t-1}\big)$, where $\hat{\eta} < \eta$. That is, after each period the agent uses her memory of u_t along with her misspecified model of utility $\hat{u}$ to infer the consumption value she must have received. We denote this *encoded outcome* by $\hat{x}_t$ and the misinference described above implies that $\hat{x}_t$ solves

$$u\big(x_t \big| \hat{\theta}_{t-1}\big) = u_t = \hat{u}\big(\hat{x}_t \big| \hat{\theta}_{t-1}\big).^{14} \tag{3}$$

Roughly put, the agent's incorrect model of her past utility understates the degree to which u_t derives from gain-loss utility. Any gain-loss utility the decision maker fails to account for is wrongly attributed to the prospect's intrinsic consumption value. Finally, we assume the agent is unaware of her misencoding but is otherwise rational: she updates her beliefs π_{t-1} over θ according to Bayes' rule as if the encoded outcome $\hat{x}_t$ actually happened.

Our analysis of misattribution is aided by the fact that encoded outcomes take a simple form. Equation 3 implies that

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

where

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

The parameters κ^G and κ^L represent the extent that elations and disappointments, respectively, distort encoded outcomes. Intuitively, κ^G and κ^L increase in the degree of misattribution (i.e., as $\hat{\eta}$ decreases), and $\kappa^L > \kappa^G$ when the agent is loss averse (i.e., $\lambda > 1$).

The simple specification above (Equation 4) yields several immediate implications. First, it captures the basic logic of disconfirmation: an outcome that deviates from expectations is perceived as deviating by more than it really did (e.g., Anderson 1973; Oliver 1977, 1980; Geers and Lassiter 1999).

Observation 1. *Outcomes that beat expectations are distorted upward, while those that fall short are distorted downward:* If $x_t > \hat{\theta}_{t-1}$, then $\hat{x}_t > x_t$, and if $x_t < \hat{\theta}_{t-1}$, then $\hat{x}_t < x_t$.

Second, when $\lambda > 1$, disappointments and elations distort encoded outcomes—and hence beliefs—

[14] Note that $\hat{x}$ is well defined and unique: fixing any expectation $\hat{\theta}_{t-1}$, the misspecified utility function $\hat{u}$ is strictly increasing in $\hat{x}$.

asymmetrically (e.g., Skinner and Sloan 2002; Kuhnen 2015; Erkal, Gangadharan, and Koh 2019). This is consistent with the large body of evidence from psychology noting a "negativity bias" or "positive-negative asymmetry effect" in belief updating (e.g., Peeters and Czapinski 1990, Baumeister et al. 2001).

Observation 2. *Losses are misencoded by more than equivalently sized gains:* Suppose $\lambda > 1$. Consider outcomes $x^g \equiv \hat{\theta}_{t-1} + k$ and $x^l \equiv \hat{\theta}_{t-1} - k$. For any $k > 0$, $|\hat{x}^l - x^l| > |\hat{x}^g - x^g|$.

Third, misattribution generates "sequential contrast effects": fixing the value of today's outcome, its perceived value seems higher the lower was yesterday's (e.g., Bhargava 2007; Bhargava and Fisman 2014). When the previous experience lowers expectations, the current outcome is assessed against a lower benchmark and thus generates a larger elation (or a smaller disappointment).

Observation 3. *Sequential contrast effects:* $\hat{\theta}_{t-1}$ is strictly increasing in x_{t-1} and thus $\frac{\partial \hat{x}_t}{\partial x_{t-1}} < 0$.

To illustrate these observations, consider a person experimenting with a new medical treatment to reduce pain. Let x measure the effectiveness of the treatment (in utils), and suppose the patient expects $x = 50$. Imagine $x = 60$—the treatment works better than expected. To decide whether to use this treatment again, the patient infers its efficacy x from her experienced utility $u = 60 + \eta n(60|50) = 60 + \eta 10$. While she correctly recalls a pleasant experience, she fails to properly disentangle the consumption value of the treatment from the elation due to surprise. From Equation 4, the patient recalls a value $\hat{x}$ such that

$$\hat{x} = 60 + \left(\frac{\eta - \hat{\eta}}{1 + \hat{\eta}}\right) 10 > x.$$

If, for instance, $\eta = 1$ and $\hat{\eta} = 1/3$, then $\hat{x} = 65$. Contrastingly, imagine $x = 40$—the treatment works worse than expected. She then encodes a value

$$\hat{x} = 40 - \lambda \left(\frac{\eta - \hat{\eta}}{1 + \lambda\hat{\eta}}\right) 10 < x.$$

Again, if $\eta = 1$, $\hat{\eta} = 1/3$ and $\lambda = 3$, then $\hat{x} = 30$. To demonstrate a contrast effect, suppose the patient tries the treatment a second time and it yields $x_2 = 50$. Since her expectation of x_2 is increasing in the previous outcome, her perceived value on the second trial, $\hat{x}_2$, will be higher if she initially experiences $x_1 = 40$ rather than $x_1 = 60$.

Finally, our model implies a difference in learning from outcomes with direct utility consequences versus signals that do not influence payoffs. In particular, given that misattribution stems from a misunderstanding of the source of utility, those outcomes that incite sensations of elation or disappointment are prone to misencoding in our framework.[15]

[15] Relatedly, Charness and Levin (2005) find significantly greater errors in updating about the distribution of balls in

2.3 Discussion

In this section we discuss our specific modeling approaches and provide evidence motivating the underlying concept of misattribution.

Interpretation of Misencoded Outcomes. There are at least two natural interpretations for the misencoding in our model. One is that the agent observes her experienced utility u_t each period, but cannot directly observe the underlying outcome x_t. Hence, x_t must be inferred from u_t, and the misattributor errs in doing so. This is a natural interpretation in settings where a person is learning about her tastes—e.g., how much she enjoys an unfamiliar product. Another interpretation is that misattribution occurs even when the agent can observe x_t. Under this interpretation, the person forms mistaken memories based on how an experience feels relative to her expectations—e.g., a purchase that was surprisingly expensive is remembered as more costly than it really was.

Expectations as the Reference Point. 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 a natural reference point.[16] That said, our formulation of misattribution can accommodate 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. We explore reference points that depend on the agent's perception of the full distribution of outcomes (as introduced by KR) in Appendix C, which yields similar results.

Restriction to Normally-Distributed Outcomes and Priors. We assume normally-distributed outcomes primarily to streamline the exposition and analysis. Many of our qualitative results extend beyond this specific case. For example, Observations 1 and 2 clearly hold for any distributional assumptions, and Observation 3 requires only a mild assumption that the agent's current expectation is increasing in the preceding outcome. Our specific results related to order effects in Sections 3 and 5 require only that the outcome and prior distributions are symmetric and quasi-concave (i.e., unimodal). These assumptions guarantee that a rational agent's updated estimate of θ falls between her previous estimate and the most recent observation.[17] Additionally, while we leverage our normality assumptions in Section 4 to establish that beliefs converge, our comparative statics on the steady-state belief hold for any distribution of outcomes. Finally, our general formulation of misattribution

an urn when participants observe a sample of draws that have payoff consequences relative to the case in which these signals have no payoff consequences. Their experiment suggests that the affect induced by payments is a critical factor in deviations from Bayesian updating.

[16] 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, Kirchsteiger, and Peitz 2015; against are Heffetz and List 2014; Gneezy et al. 2017; Goette, Harms, and Sprenger 2019). There is additional evidence of expectations-based reference points from the field, spanning labor supply among taxi drivers (Crawford and Meng 2011; Thakral and To 2018), platform exit from an online marketplace (Backus et al. 2018), domestic violence resulting from unexpected football losses (Card and Dahl 2011), and decisions in game shows and sports (Post et al. 2008; Pope and Schweitzer 2011; Markle et al. 2015).

[17] See Chambers and Healy (2012) for a complete characterization of when expectations "update toward the signal".

is portable in that it can be readily applied to any distributional environment, and we relax some of our baseline assumptions in Section 5.[18]

Economics Literature on Misattribution. As previously noted, our model builds on concepts—namely disconfirmation and misattribution—studied across a variety of disciplines (e.g., psychology, political science, marketing). In this section, we discuss the economics literature on misattribution, highlighting evidence in support of the basic underlying concept.

Our companion paper (Bushong and Gagnon-Bartsch 2019) provides evidence supporting of the basic notion of disconfirmation noted in Observation 1. In that paper's main experiment, participants completed one of two unfamiliar tasks: a neutral task or the neutral 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 a treatment group determined their task by flipping a coin just before working, while participants in a control group faced no uncertainty over their task. Hence, the task assignment involved either a positive or negative surprise for the treatment group but no surprise for the control group. A day 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 control group, the treatment group exhibited greater WTW on the neutral task and decreased WTW on the unpleasant task. When the stakes were highest, the WTW of treatment participants was roughly 20% higher (neutral task) and 25% lower (unpleasant task) than the WTW of control participants. These results suggest that the mere fact that some outcomes were previously associated with sensations of surprise influenced participants' ex post evaluations of those outcomes. This is consistent with misattribution of elation and disappointment to the intrinsic enjoyment of the task.

Other forms of misattribution discussed in the economics literature resemble the "fundamental attribution error" or "correspondence bias" in psychology (e.g., Ross 1977; Gilbert and Malone 1995), where transient situational factors are incorrectly attributed to underlying, stable characteristics of a person or good. For example, Haggag and Pope (2018) show that experimental participants valued an unfamiliar drink more when they first experienced it while thirsty.[19] Additionally, they find that frequent patrons of an amusement park whose most recent visit was during good weather are more likely to return. In two papers, Simonsohn (2007, 2010) explores the effect of a transient shock (weather) on

[18] Although we focus on learning about $F(\cdot|\theta)$ where θ is the mean outcome of the prospect, our notion of misattribution easily extends to learning about other distributional parameters. Under such an analysis, encoded outcomes $\hat{x}_t$ would still follow from Equation 3, but the reference point would be given by $\widehat{\mathbb{E}}_{t-1}[x_t]$, where $\widehat{\mathbb{E}}_{t-1}$ is with respect to the agent's posterior distribution of outcomes following signals $(\hat{x}_1, \ldots, \hat{x}_{t-1})$. Updating about the distributional parameters would then follow the appropriate Bayesian approach as if $\hat{x}_t$ had occurred.

[19] Haggag and Pope (2018) predict that misattributors tend to underestimate the payoff difference between outcomes in two different states. This stands in contrast to our predictions (see, e.g., the discussion in Section 6.1). Furthermore, unlike mistakes driven by misattribution of reference dependence, biased forecasts in Haggag and Pope's formulation vanish with experience. These distinctions stem from the fact that Haggag and Pope rule out complementaries where past experiences influence current consumption utility. Reference dependence introduces this complementarity, as past experiences form the reference point against which current consumption is evaluated.

the subsequent behavior of prospective college students and admissions officers. Simonsohn (2007) demonstrates that applicants with particularly strong academic qualities were evaluated more positively by admissions officers when the weather on the evaluation day was poor. Simonsohn (2010) shows that incoming freshmen are more likely to matriculate at an academically rigorous school when the weather on their visit day was cloudy versus sunny. The author interprets both results as a form of attribution bias. Relatedly, a series of papers show that CEOs (Bertrand and Mullainathan 2003) and politicians (Wolfers 2007; Cole, Healy, and Werker 2012) are rewarded for luck as if it were wrongly attributed to skill—a result that has been faithfully replicated in the lab (see, e.g., Brownback et al. 2019; Erkal, Gangadharan and Koh 2019). Our model shares a common intuition with these forms of misattribution: transient sensations (elation and disappointment in our case) are misattributed.

3 Order Effects in Belief Updating

In this section, we explore how misattribution distorts learning in the short-run. In particular, we highlight how the order in which a misattributor experiences outcomes influences her perceived value of the prospect. This leads to both a "recency bias", wherein beliefs overweight recent outcomes and underweight older ones, and an "increasing-order bias" wherein fixing the outcomes she faces, the agent forms the highest estimate of the prospect's value following an increasing sequence of outcomes. Both of these biases arise even if the agent does not exhibit loss aversion.

To fix ideas, consider a manager assessing the ability, θ, of a newly hired employee over the course of T periods. Each period, the employee generates an i.i.d. benefit $x_t = \theta + \epsilon_t$ for the manager. Although the order of outcomes is irrelevant for rational updating about θ, the misattributor's beliefs crucially depend on the sequencing, as early performances set the expectations against which later performances are assessed. For example, suppose the employee is productive every day in her first week except one bad day. If that bad day comes first, it will lower the manager's expectations and the remaining days will seem surprisingly productive. Alternatively, if the bad day comes last—after the manager has developed high expectations—it will seem surprisingly unproductive. Even though the two sequences are permutations of the same outcomes, the fact that one generates subsequent gains whereas the other ends with a loss will cause the misattributor to reach different final beliefs.

To formalize these intuitions, we first describe how a misattributor's beliefs about θ depart from rational beliefs. Rational updating follows a simple recursive rule: the estimate of θ following x_t is $\alpha_t x_t + (1 - \alpha_t)\hat{\theta}^r_{t-1}$, where $\hat{\theta}^r_{t-1}$ denotes the rational estimate entering t. The weight given to the most recent outcome, $\alpha_t \equiv 1/(t + \frac{\sigma^2}{\rho^2})$, is notably increasing in the variance of the prior relative to the variance in outcomes, ρ^2/σ^2—this ratio naturally measures the informativeness of outcomes.

Since a misattributor naively treats her misencoded outcomes as the true outcomes when learning, her updating mirrors the rational rule: following *encoded* outcome $\hat{x}_t$, she updates her prior estimate

$\hat{\theta}_{t-1}$ to reach posterior $\hat{\theta}_t = \alpha_t \hat{x}_t + (1-\alpha_t)\hat{\theta}_{t-1}$, where α_t is the same weight used by a rational agent. However, because encoded outcomes take the form $\hat{x}_t = x_t + \kappa_t\big(x_t - \hat{\theta}_{t-1}\big)$ where $\kappa_t \equiv \kappa^G \mathbb{1}\{x_t \geq \hat{\theta}_{t-1}\} + \kappa^L \mathbb{1}\{x_t < \hat{\theta}_{t-1}\}$ (Equation 4), the misattributor reaches a biased estimate

$$\hat{\theta}_t = \alpha_t(1+\kappa_t)x_t + [1 - \alpha_t(1+\kappa_t)]\hat{\theta}_{t-1}. \tag{6}$$

Equation 6 immediately reveals that a misattributor "overreacts" to the latest outcome: the rational estimate weights x_t by α_t but the biased one weights it by $\alpha_t(1+\kappa_t)$.[20] Moreover, $\hat{\theta}_t$ assigns the wrong weight to each previously experienced outcome. Iterating Equation 6, we can express a misattributor's expectation after T rounds as a (mis)weighted sum of the true outcomes (see Appendix A for all proofs):

Lemma 1. *Following any sequence $(x_1, \ldots, x_T) \in \mathbb{R}^T$, a misattributor forms an estimate*

$$\hat{\theta}_T = \xi_0^T \theta_0 + \alpha_T \sum_{t=1}^{T} \xi_t^T x_t, \tag{7}$$

where

$$\xi_t^T = \begin{cases} \prod_{j=1}^{T}[1 - \alpha_j(1+\kappa_j)] & \textit{if} \quad t = 0, \\ (1+\kappa_t)\prod_{j=t}^{T-1}[1 - \alpha_j \kappa_{j+1}] & \textit{if} \quad t \in \{1, \ldots, T-1\}, \\ (1+\kappa_T) & \textit{if} \quad t = T. \end{cases}$$

The rational estimate following $(x_1, \ldots, x_T)$ is given by Equation 7 with all $\kappa_t = 0$, and hence it assigns an equal weight of α_T to each outcome. Therefore, ξ_t^T measures the weight a misattributor assigns to outcome x_t relative to the rational weight.

Lemma 1 clearly demonstrates that a misattributor differentially weights evidence depending on when it arrives. We can further elaborate some specific properties of these order effects when a misattributor's beliefs obey a basic monotonicity property of Bayesian updating: the posterior mean $\hat{\theta}_T$ is an increasing function of each outcome x_t for $t \leq T$.

Definition 1. Let x_{-t}^T denote $(x_1, \ldots, x_T)$ excluding the t^{th} element. Beliefs are *monotonic* if for all $T \geq 1$, all $t \leq T$, and all $x_{-t}^T \in \mathbb{R}^{T-1}$, $\hat{\theta}_T$ is strictly increasing in x_t conditional on x_{-t}^T.

Monotonicity is equivalent to $\kappa^L < 1 + \sigma^2/\rho^2$; that is, the degree of misencoding is not too severe (i.e., $\kappa^L \geq \kappa^G$ is sufficiently small) or outcomes are not too informative (i.e., ρ^2/σ^2 is sufficiently small). This ensures that for all $t \geq 1$, the weight ξ_t^T in Equation 7 is positive.

[20] More formally, consider both a misattributing and rational learner who share a common prior expectation θ_0. Following outcome $x \in \mathbb{R}$, $|\hat{\theta} - \theta_0| \geq |\hat{\theta}^r - \theta_0|$, where $\hat{\theta}$ and $\hat{\theta}^r$ denote the biased and rational posterior estimates of θ, respectively. Furthermore, the misattributor's reaction $|\hat{\theta} - \theta_0|$ is decreasing in $\hat{\eta}$: she overreacts more as the extent of misattribution increases.

If beliefs are monotonic, then a misattributor exhibits a *recency bias*: her beliefs weight a recent gain more than any preceding gain and weight a recent loss more than any preceding loss.

Proposition 1. *Consider any sequence $(x_1, \ldots, x_T) \in \mathbb{R}^T$ and suppose beliefs are monotonic. For any two outcomes x_t and x_τ that are both gains or both losses (i.e., $\kappa_t = \kappa_\tau$):*

1. *A misattributor's final expectation, $\hat{\theta}_T$, places greater weight on the more recent outcome: $\xi_t^T > \xi_\tau^T$ if and only if $t > \tau$.*

2. *Expectations overweight recent outcomes and underweight early outcomes: $\xi_T^T > 1$, and for all $t < T$, $\xi_t^T \to 0$ as $T \to \infty$.*

3. *The recency bias is stronger when outcomes are more informative: ξ_t^T / ξ_τ^T is increasing in ρ^2/σ^2.*

Although Proposition 1 focuses on outcomes that fall in the same domain (i.e., both gains or both losses), this restricted focus is relevant only when the agent is loss averse. In that case, losses influence beliefs more than gains (Observation 2), and it is thus possible for a loss in period $t-1$ to have a larger influence on beliefs than a gain in period t. However, for any two outcomes that fall in the same domain, the more recent outcome has a greater influence on beliefs than the earlier one. If the agent is not loss averse, this caveat is irrelevant and Proposition 1 applies to *any* two outcomes regardless of their domain.

The intuition underlying Proposition 1 stems from the way that sequential contrast effects (introduced in Observation 3) play out over several rounds of updating. Namely, early outcomes have a "self limiting" influence on final beliefs: a high initial outcome raises expectations, which causes subsequent outcomes to be underestimated by more (or overestimated by less). This dampens the positive influence of the high initial outcome on final beliefs. Similarly, a low initial outcome depresses expectations, which causes subsequent outcomes to be *over*estimated by more (or underestimated by less). This dampens the negative influence of the low initial outcome. Furthermore, as the horizon T increases, an early outcome x_t exerts this countervailing force on a larger number of subsequent outcomes, which pushes the weight on x_t (relative to rational updating) to zero.

Additionally, this recency bias is stronger when outcomes are more informative. In such cases, early outcomes have a greater impact on expectations, which accentuates the subsequent contrast effect with later outcomes: when an initial low outcome has a stronger negative impact on expectations, then later outcomes will be biased upward by a greater amount. This heightens the self-limiting effect of early outcomes described above.

Finally, Proposition 1 reveals an important distinction between our model and reference dependence *without* misattribution. With reference-dependent preferences, an agent's experienced utility fluctuates depending on her expectations. Absent misattribution, these fluctuations would not cause the agent to form biased beliefs; however, misattribution introduces an avenue through which the agent's reference points distort her forecasts.

A recency bias suggests that a misattributor will form a higher estimate of θ when her best experiences happen near the end of the horizon. To explore this idea, suppose the agent will experience an arbitrary set of outcomes $\mathcal{X} = \{x_1, \ldots, x_T\}$. Among all permutations of $\mathcal{X}$, which sequence maximizes a misattributor's perception of θ? We can easily address this question when the misattributor's beliefs obey a convexity property of Bayesian updating: the posterior mean $\hat{\theta}_t$ always falls between the true outcome x_t and her prior $\hat{\theta}_{t-1}$.

Definition 2. Beliefs are *convex in period* t if, given any prior estimate $\hat{\theta}_{t-1} \in \mathbb{R}$ and any $x_t \in \mathbb{R}$, there exists $\tilde{\alpha} \in [0, 1]$ such that $\hat{\theta}_t = \tilde{\alpha} x_t + (1 - \tilde{\alpha})\hat{\theta}_{t-1}$. Beliefs *convex* if they are convex in period t for all $t \geq 1$.

Convexity is a somewhat stronger form of monotonicity and is equivalent to the posterior $\hat{\theta}_t$ being an increasing function of $\hat{\theta}_{t-1}$. Convexity holds whenever $\kappa^L < \sigma^2/\rho^2$—accordingly, convexity, like monotonicity, holds when outcomes are not too informative about θ.[21]

If beliefs are convex, then a misattributor exhibits an *increasing-order bias*: her perception of θ is maximized by the sequence with an increasing profile of outcomes.

Proposition 2. *Consider any set of T distinct outcomes, $\mathcal{X}$. If beliefs are convex, then among all possible orderings of the outcomes in $\mathcal{X}$, the misattributor's estimate $\hat{\theta}_T$ is highest following the sequence in which the elements are ordered from least to greatest.*

Our final short-run result shows how the increasing-order bias described in Proposition 2 extends when we relax the assumption of convex beliefs. To provide the weakest sufficient conditions, we consider a context with just two rounds. In this case, so long as one outcome beats initial expectations, the agent's estimate of θ is necessarily maximized when the misattributor receives the better outcome last. Even if both outcomes fall short of initial expectations, this result remains if loss aversion is not too strong.

Corollary 1. *Consider any $a, b \in \mathbb{R}$ such that $a > b$. Let $\hat{\theta}_2^d$ denote the mean belief following the decreasing sequence (a, b), and let $\hat{\theta}_2^i$ denote that following the increasing sequence (b, a).*

1. *If $a > \theta_0$, then $\hat{\theta}_2^i > \hat{\theta}_2^d$.*

[21] Rational beliefs in this setting are always convex given that θ and ϵ_t are independently drawn from normal distributions. Equation 6 reveals that when beliefs are not convex, then $\hat{\theta}_t$ strongly overweights x_t and *negatively* weights the prior (i.e., the contrast effect more than offsets the positive effect of a higher prior expectation on the agent's posterior). Note that even if beliefs are not initially convex, they eventually become convex when the horizon is sufficiently long: as the total number of outcomes accumulates, each additional outcome has a diminished influence on beliefs and there accordingly exists a period $t^* = \lfloor 1 + \kappa^L - \sigma^2/\rho^2 \rfloor$ beyond which beliefs are convex in each period $t > t^*$. In this case, the results below that rely on convexity will hold beyond period t^*.

2. *If both $a, b < \theta_0$ and beliefs are not convex, then there exists a threshold $\bar{\lambda} > 2$ such that $\hat{\theta}_2^i < \hat{\theta}_2^d$ for some $a, b \in \mathbb{R}$ only if $\lambda > \bar{\lambda}$.*

Regardless of whether beliefs are convex, either $a > \theta_0$ or $\lambda < 2$ guarantees that $\hat{\theta}_2^i > \hat{\theta}_2^d$. Furthermore, even with relatively strong loss aversion ($\lambda > 2$) the range of values (a, b) for which the *decreasing* sequence maximizes beliefs is limited.[22] In light of this caveat, Part 1 of the corollary provides a simple sufficient condition for our increasing-sequence bias that may be useful for empirical tests.

The results above accord with evidence that, fixing the outcomes a person faces, they tend to both (i) prefer increasing sequences (e.g., Loewenstein and Prelec 1993); and (ii) retrospectively form the most optimistic evaluations of an episode thereafter. Speaking to the latter point, Ross and Simonson (1991) allow participants to sample two video games and find that willingness to pay for the pair is significantly higher for those who sampled the better game second. Similarly, Haisley and Loewenstein (2011) demonstrate that advertising promotions that utilize giveaways are most effective when sequenced in increasing order of value—that is, the high-value promotional item is given last. Our model provides a lens through which we can understand both the preference for increasing sequences and the subsequent positive evaluations demonstrated in such papers. Reference dependence absent misattribution—with the assumption that the agent believes her reference point will be her past experiences—provides an explanation for an ex-ante preference for increasing sequences. And the altered ex-post evaluations in both Ross and Simonson (1991) and Haisley and Loewenstein (2011) are consistent with misattribution.[23]

Our results in this section may also help explain the well-known "end effect": people retrospectively form inflated impressions of experiences that end on a high note, as shown by Kahneman et al. (1993). In that paper, participants experienced a novel painful episode (putting their hand in frigid water for a fixed amount of time) and were asked how willing they were to repeat the experience. Participants were in one of two conditions: (1) 60 seconds of cold water; or (2) 60 seconds of cold water followed by 30 seconds of slightly warmer water. This second condition made the total experience longer than the first condition but the episode ended on a relatively good note. Surprisingly, participants were more willing to repeat the episode in the *second* condition, despite its longer duration. This is consistent with our model, which suggests that an agent's unpleasant early experience leads

[22] This happens only if both outcomes come as a loss relative to the person's prior—$b < a < \theta_0$—and b is sufficiently close to a (see Equation A.8 for a precise condition). To provide an intuition, suppose beliefs are not convex and hence excessively react to new outcomes. If b is sufficiently close to a, then b is perceived as a gain when experienced after a: beliefs become so pessimistic after the initial loss a that the truly worse outcome b feels like a gain. Furthermore, if losses distort beliefs sufficiently more than gains (i.e., $\lambda > \bar{\lambda}$), then $\hat{\theta}_2$ is maximized (roughly) by minimizing experienced losses. Thus, because the first outcome necessarily comes as a loss while the second comes as a gain, losses are minimized when the better outcome happens first.

[23] Several authors suggest that such assessments stem from the loose idea of "adaptation and subsequent contrast", which has a similar intuition as our formal model (see, e.g., Tversky and Griffin 1990; Loewenstein and Prelec 1993; Baumgartner, Sujan, and Padgett 1997).

her to form a pessimistic view of the episode. Thus, adding a still-bad-but-better experience to the end of the sequence can—when contrasted against her initial experience—help improve her overall impression of the episode.[24]

As noted in the introduction, our order effects seemingly stand at odds with confirmation bias, wherein new evidence is wrongly interpreted as conforming to one's expectations (e.g., Rabin and Schrag 1999; Fryer, Harms, and Jackson 2018). Under confirmation bias, outcomes deviating from expectations are encoded as closer to expectations, which implies that early outcomes are overweighted relative to later outcomes. Although misattribution makes the opposite prediction, the two mechanisms are not mutually exclusive. Indeed, empirical tests of order effects in belief updating find support for both confirmatory and recency effects (see Hogarth and Einhorn 1992 for a meta analysis and Geers and Lassiter 1999 for a specific test). Which effect prevails seems to depend on the nature of the learning problem: confirmatory effects tend to dominate as evidence becomes more ambiguous and difficult to interpret.

4 Long-Run Beliefs: Pessimism over Risky Prospects

Continuing the setup above, we now describe how errors in beliefs can persist following ample experience with the prospect. Although the misattributor places excess weight on recent outcomes (as shown in Proposition 1), her beliefs about θ eventually converge. Absent loss aversion, the agent's long-run beliefs about the mean outcome reflect the underlying truth, although the outcomes she encodes are more variable than the underlying outcomes really are. When the agent is loss averse, her long-run beliefs about the average outcome are biased downward: the agent becomes pessimistic. This pessimism increases in proportion to the prospect's true underlying variance. These distortions in beliefs imply that the agent may reject risky-but-optimal prospects.

We seek to establish convergence 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 $\hat{\theta}$. To formalize this notion, consider a function $\Delta : \mathbb{R} \to \mathbb{R}$ where $\Delta(\hat{\theta})$ is the deviation between $\hat{\theta}$ and the expected value of the encoded outcome assuming the agent holds expectation $\hat{\theta}$. Given our expression for the encoded outcome in Equation 4:

$$\Delta(\hat{\theta}) \equiv \mathbb{E}\left[x_t + \kappa_t(x_t - \hat{\theta}_{t-1}) \middle| \hat{\theta}_{t-1} = \hat{\theta}\right] - \hat{\theta}, \tag{8}$$

[24] The end effect is one-half of the peak-end heuristic in remembered utility. This other half—peak—refers to the fact that the highest valence experience (either good or bad) is given outsize weight in retrospective evaluations. Our model may also speak to this: given that peak deviates from expectations, misattribution implies that it would be overweighted in final beliefs. Kahneman et al. (1993) interpret their evidence as support of duration neglect—the tendency to fail to properly integrate utility over time.

where the expectation is conditional on the true parameters governing x_t, (θ, σ). A steady-state belief is thus a zero of Δ.

We now show that the misattributor's sequence of mean beliefs, $\langle\hat{\theta}_t\rangle$, converges to a *unique* steady-state belief, which we denote by $\hat{\theta}_\infty$. Moreover, we characterize how $\hat{\theta}_\infty$ depends on the true distributional parameters (θ, σ), the misattributor's underlying preferences, and the extent of her bias.

Proposition 3. *Consider a prospect with mean θ and variance $\sigma^2 > 0$.*

1. *There is a unique steady-state mean belief, $\hat{\theta}_\infty$. The sequence of a misattributor's expectations, $\langle\hat{\theta}_t\rangle$, converges almost surely to $\hat{\theta}_\infty$.*

2. *In the steady-state, the misattributor forms pessimistic beliefs about the mean, $\hat{\theta}_\infty \leq \theta$, and this inequality is strict if and only if $\lambda > 1$. If $\lambda > 1$, then beliefs $\hat{\theta}_\infty$ are strictly decreasing in the true variance, σ^2.*

3. *Additional comparative statics: $\hat{\theta}_\infty$ is strictly decreasing in the degree of reference dependence (η) and loss aversion (λ), and $\hat{\theta}_\infty$ is strictly increasing in the degree to which the agent accounts for reference dependence ($\hat{\eta}$).*

Part 1 of Proposition 3 shows that there is a unique root of Δ that characterizes $\hat{\theta}_\infty$. Specifically, that root is the value of $\hat{\theta}$ that solves the following equation:

$$\hat{\theta} = \theta - kF(\hat{\theta}|\theta)\left\{\hat{\theta} - \frac{\int_{-\infty}^{\hat{\theta}} xf(x|\theta)dx}{F(\hat{\theta}|\theta)}\right\} \quad \text{where} \quad k \equiv \frac{(\lambda-1)(\eta-\hat{\eta})}{(1+\eta)(1+\hat{\eta}\lambda)}, \tag{9}$$

and where, recalling Section 2, $F(\cdot|\theta)$ is the CDF of outcomes conditional on the true mean and $f(\cdot|\theta)$ is the associated PDF. Thus, the downward bias in $\hat{\theta}_\infty$ is proportional to the average encoded loss in the steady-state (i.e., the expression in braces) scaled by the likelihood of such a loss.

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 a misattributor's reference point and thus influence the current encoded outcome. Hence, following Heidhues, Kőszegi, and Strack (2019) and Esponda and Pouzo (2016), we use techniques from stochastic-approximation theory to establish convergence. The details of this analysis are discussed in the proof.

Part 2 of Proposition 3 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 immediate that such pessimistic expectations will persist, since pessimistic beliefs simultaneously imply that the misattributor will experience more pleasant surprises. The steady-state belief $\hat{\theta}_\infty$ balances these two forces:

a misattributor underestimates θ to such an extent that the resulting excess of positive surprises exactly offsets the downward bias stemming from loss aversion. Furthermore, the steady-state belief $\hat{\theta}_\infty$ is stable in the sense that once beliefs are near $\hat{\theta}_\infty$, the resulting (mis)encoded outcomes push the misattributor's expectations toward $\hat{\theta}_\infty$. Specifically, if her expectations were to move below $\hat{\theta}_\infty$, then she would experience an increased rate of elations that drive her expectations back up. Conversely, if her expectations were to move above $\hat{\theta}_\infty$, then the increased rate of disappointments would push her expectations back down.

Additionally, Part 2 of Proposition 3 shows that greater variability in the underlying distribution of outcomes causes a misattributor to underestimate θ by a larger amount. That is, a misattributor develops more pessimistic beliefs about prospects that are riskier. Increased variance implies that the agent experiences 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. For example, if a misattributing consumer faces two logistics companies (e.g., UPS and FedEx) with identical mean delivery times, she will come to believe the more variable company typically takes longer. As we show below, such mistakes can lead to poor decisions when balancing risk and return, as the agent is systematically more biased against riskier prospects.

Finally, Part 2 of Proposition 3 also shows that if the misattributor is not loss averse (i.e., $\lambda = 1$), then she correctly learns the prospect's mean outcome, θ. It is straightforward to see that correct expectations (i.e., $\hat{\theta}_\infty = \theta$) correspond to the unique steady-state belief in this case. When expecting θ, outcomes exceeding expectations are, on average, overestimated by the same extent that outcomes missing expectations are underestimated. Thus, the bias in encoded outcomes is symmetric about the misattributor's expectations and does not impart systematic drift in her beliefs—when holding correct expectations, encoded outcomes will also have a mean equal to θ.

Regardless of whether the agent is loss averse or not, the steady-state distribution of encoded outcomes has greater *variance* than the true distribution.[25] Although this paper focuses exclusively on learning the mean outcome, a misattributor would also overestimate the variance in outcomes within a richer model that allows for updating over σ^2.[26] In this case, the misattributor would mislearn the distribution of x_t even without loss aversion. Furthermore, such an extension only reinforces our conclusion that learning from experience causes a misattriubtor to undervalue a risky prospect: overestimating σ^2 lowers the prospect's perceived value for any risk-averse agent.

Figure 1 utilizes a simulated sequence of outcomes to display the results above, depicting both the

[25] For a formal proof, see the discussion concluding the proof of Proposition 3 in Appendix A

[26] In particular, if beliefs about σ^2 do not influence how an outcome is misencoded, then the misattributor will perceive a steady-state variance $\hat{\sigma}^2 > \sigma^2$ equal to the variance of encoded outcomes given that the agent holds expectation $\hat{\theta}_\infty$ characterized by Proposition 3. For instance, in the canonical model where priors over θ and σ^2 follow independent Normal and Inverse-Gamma distributions, respectively (see, for example, DeGroot 1970), a misattributor's beliefs over (θ, σ) converge to $(\hat{\theta}_\infty, \hat{\sigma}_\infty)$ where $\hat{\theta}_\infty$ is the same value described in Proposition 3 and $\hat{\sigma}^2_\infty = \text{Var}(x_t + \kappa_t(x_t - \hat{\theta}_\infty))$.

long-run path of beliefs and the density of perceived outcomes for two different variances: $\sigma^2 = 1$ (top panels) and $\sigma^2 = 5$ (bottom panels).

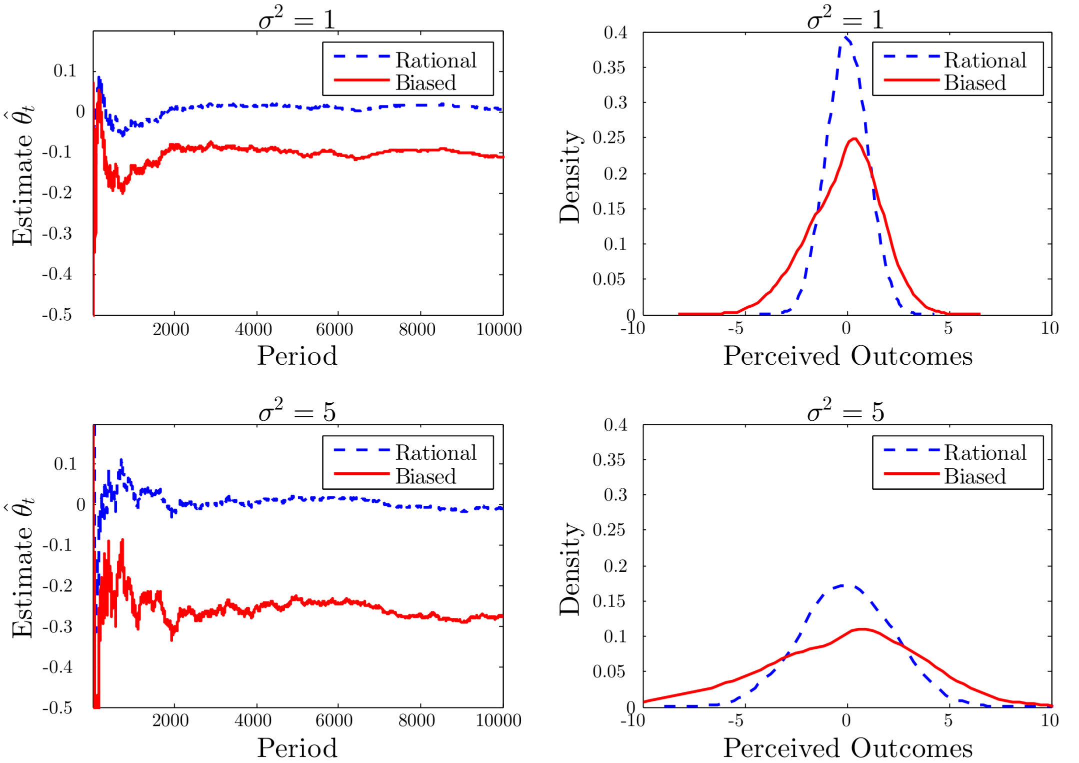

Figure 1: *The top-left panel depicts a simulated path of estimates $\langle\hat{\theta}_t\rangle$ for both a rational and biased agent. The top-right panel shows the true and perceived density of outcomes. The simulation assumes normally-distributed outcomes and priors with $\theta = 0$, $\sigma^2 = 1$, $\eta = 1$, $\lambda = 3$, and $\hat{\eta} = 1/3$. The bottom two panels are analogous except with an increased variance of $\sigma^2 = 5$.*

We now consider how the long-run beliefs described above distort a misattributor's valuation of the prospect and potentially harm her decisions. Let $v(\tilde{\theta}, \sigma)$ denote a misattributor's expected (per-period) utility from the prospect assuming she is confident that the mean is $\tilde{\theta}$:

$$v(\tilde{\theta}, \sigma) \equiv \int_{-\infty}^{\infty} \left[x + \eta n(x|\tilde{\theta})\right] f(x|\tilde{\theta}) dx. \tag{10}$$

Accordingly, $v(\theta, \sigma)$ denotes the agent's valuation when she reaches correct beliefs about the prospect.[27]

We explore how this correct-information benchmark compares to the agent's forecasted utility in the steady state, $v(\hat{\theta}_\infty, \sigma)$, given the biased beliefs characterized in Proposition 3. Given that $\hat{\theta}_\infty < \theta$, it follows that $v(\hat{\theta}_\infty, \sigma) < v(\theta, \sigma)$: a misattributor tends to undervalue risky prospects when

[27] Our formulation of v implicitly assumes the agent forecasts her utility according to her true gain-loss parameter η, which we suspect is reasonable given our interpretation that misattribution occurs in retrospect. Alternatively, one could assume that the person errs in forecasting her utility as well, which would entail substituting $\hat{\eta}$ into Equation 9 above. We do not rely on this distinction in the results below.

learning from experience. To illustrate the immediate decision consequences of this bias, consider a misattributing manger who decides to fire an employee if his valuation falls below some threshold w. There exist parameters (θ, σ) such that the manager would retain the employee under correct learning (i.e., $v(\theta, \sigma) > w$), yet fire him under misattribution (i.e., $v(\hat{\theta}_\infty, \sigma) < w$).[28]

We can further characterize which prospects foster more pessimistic beliefs and are hence undervalued by a greater extent. Building on the fact that $\hat{\theta}_\infty$ is decreasing in the true variance of outcomes (Proposition 3, Part 2), our next result shows that $v(\theta, \sigma) - v(\hat{\theta}_\infty, \sigma)$ is strictly decreasing in σ. Thus, the mistake discussed in the managerial example above is accentuated for more variable prospects. Importantly, this bias against variability is on top of the agent's intrinsic risk preferences—amplifying any existing distaste for risk—and stems from the pessimistic long-run beliefs described above. Moreover, we show that these pessimistic misperceptions of risky prospects can be arbitrarily costly. In particular, $v(\theta, \sigma) - v(\hat{\theta}_\infty, \sigma)$ is unboundedly decreasing in σ. This extreme limit suggests a misattributor may perceive arbitrarily different valuations of two prospects that in fact yield identical expected utility under correct beliefs. To help show this, let the set $\mathcal{P}(w) \equiv \{(\theta, \sigma) \mid v(\theta, \sigma) = w\}$ denote the parameter combinations (θ, σ) for which the prospect yields expected utility $w \in \mathbb{R}$ under correct beliefs.

Proposition 4. *Suppose $\lambda > 1$.*

1. *The difference between the expected utility under correct beliefs and the agent's perceived expected utility in the steady state, $v(\theta, \sigma) - v(\hat{\theta}_\infty, \sigma)$, is strictly positive and strictly increasing in σ.*

2. *Fix a constant w. For any $\hat{w} < w$, there exists a threshold $\bar{\sigma}(w, \hat{w}) > 0$ such that any prospect with $(\theta, \sigma) \in \mathcal{P}(w)$ and $\sigma > \bar{\sigma}(w, \hat{w})$ will generate a forecasted expected utility in the steady state that is strictly less than $\hat{w}$. That is, for any arbitrarily large bound $w - \hat{w}$, $\sigma > \bar{\sigma}(w, \hat{w})$ implies that the error in valuation, $v(\theta, \sigma) - v(\hat{\theta}_\infty, \sigma)$, exceeds $w - \hat{w}$.*

To illustrate, 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 made arbitrarily large by choosing a variance for A that is sufficiently high. Intuitively, the overweighting

[28] If a misattributor's biased beliefs cause her to reject a prospect that she would otherwise accept under full information—e.g., if $v(\hat{\theta}_\infty, \sigma) < w < v(\theta, \sigma)$—then the agent suffers a welfare loss. Note that $\hat{\theta}_\infty < \theta$ implies that the agent is pleasantly surprised on average, and thus her average experienced utility under incorrect beliefs (i.e., the expectation of $u(x|\hat{\theta}_\infty)$ with respect to the true distribution of x) is *higher* than the rational-learning benchmark, $v(\theta, \sigma)$. Thus, if it is optimal to accept the prospect under full information, it is also optimal to do so when holding expectation $\hat{\theta}_\infty$. That said, the misattributor does not realize she has biased beliefs. Hence, she decides whether to reject the prospect based on her forecasted valuation $v(\hat{\theta}_\infty, \sigma)$, which fails to account for the pleasant surprises she would experience as a result of her overly-pessimistic expectations.

of occasional losses swamps the benefit of a high mean when the magnitude of these losses grows large. Such biased learning may help explain why individuals tend to excessively avoid risk based on their personal experiences, as shown by Malmendier and Nagel (2011).

5 Extensions and Applications

In this section, we extend our baseline model to environments where outcomes are no longer identically distributed or independent. This allows us to examine two natural applications of misattribution where the distribution of outcomes depends on (i) the behavior of an exploitative party and (ii) past outcomes through autocorrelation.

5.1 Expectations Management and Reputation

Our first extension explores how a sophisticated party can strategically manipulate the beliefs of a misattributor. Specifically, we consider a career-concern setting where a misattributing (but otherwise rational) principal sequentially updates her beliefs about a worker's ability and offers wages based on those inferences. For instance, imagine a professor assessing her research assistant, or a homeowner assessing a contractor. Unlike the rational signal-jamming logic where high effort is interpreted as uninformative, a misattributor wrongly attributes outcomes that deviate from expectations to the worker's ability, providing a way for the worker to deceive the principal. Thus, a worker aware of the principal's misattribution faces new incentives which push against the declining effort profile predicted by both the classical model (e.g., Holmström 1999) and models of confirmation bias (e.g., Rabin and Schrag 1999). Namely, we show that the sophisticated worker follows an effort path that initially under-performs relative to the principal's expectations but consistently beats them thereafter. By initially setting the bar low, he can supply a series of elations that are interpreted as favorable signals of his ability. More generally, this exercise demonstrates how misattribution explains both the frequent use and persuasive nature of expectations management.

Following Holmström (1999), a principal (she) hires a worker (he) of unknown ability to exert effort over T periods (i.e., the principal and worker agree ex ante to a fixed-duration relationship). Each period t, the worker supplies effort $e_t \in \mathbb{R}_+$ leading to output $x_t = \theta + e_t + \epsilon_t$, where $\theta \in \mathbb{R}$ is the worker's ability. Both the principal and worker share a common prior over θ. We take x_t (net of paid wages) as the principal's consumption value in period t—she directly benefits from the worker's performance—and we maintain our baseline assumptions on θ and ϵ_t introduced in Section 2 (e.g., normally distributed and i.i.d.). Hence, the notable change to the environment is that outcomes can be strategically manipulated via the worker's effort. We assume the principal cannot directly observe effort, so she updates about the worker's ability based on her perception of x_t. These updated beliefs determine the wage the principal offers in the subsequent period. In particular, we assume that the principal pays a wage w_t at the start of each round equal to her current expectation of x_t (e.g., the market perfectly competes for the worker's labor). As such, the worker maximizes the principal's perception of his ability subject to effort costs. We assume these costs are separable across periods and given by a flow disutility function $c(\cdot)$ that is strictly increasing and convex. Thus, in each period t, the worker aims to maximize the expected value of $\sum_{k=t}^{T}[w_k - c(e_k)]$.[29]

We consider a principal who suffers misattribution while inferring the worker's ability, and a rational worker who is aware of the principal's error and attempts to exploit it. Given her mistaken perceptions of output, the principal's beliefs about θ follow the mechanics derived in Section 3. As the principal is unaware of her mistake, she neglects the worker's incentive to exploit it. Thus, she additionally develops incorrect beliefs about the worker's strategy. We close the model by assuming the principal wrongly presumes common knowledge of rationality: the principal believes the worker follows the Bayesian-Nash-Equilibrium strategy that he would play when facing a rational principal, and the principal best responds to this presumed behavior of the worker. In this case, the naive principal mispredicts the worker's effort in expectation and misattributes realized discrepancies to ability and noise. We further assume that the worker is aware of this misattribution and accordingly best responds to the principal's distorted beliefs.[30]

The sophisticated worker's optimal effort path will fall short of the principal's expectations early and then consistently beat expectations later in the relationship. Let e_t^* and e_t^r denote the worker's optimal effort in round t when facing a misattributing and rational principal, respectively.

Proposition 5. *Consider the expectations-management setting described above and assume $\lambda = 1$. If beliefs are convex, then there exists a period t^*, $1 \leq t^* < T$, such that the worker's optimal effort falls short of the rational benchmark ($e_t^* < e_t^r$) for all $t < t^*$ and exceeds this benchmark ($e_t^* > e_t^r$) for all $t \geq t^*$.*[31]

When facing a misattributing principal, early effort by the worker imposes a cost on his future selves: hard work in period one increases the principal's expectations in all future periods, which means that subsequent output will be judged more harshly. The worker therefore restrains early effort and

[29] While we abstract from discounting here to ease exposition, our proofs demonstrate that the results in this application continue to hold for any exponential discount factor $\delta \in (0, 1)$.

[30] Our qualitative predictions are robust to alternative assumptions about the principal's anticipated effort. For instance, our results extend when the principal correctly predicts the worker's effort profile despite lacking a good theory as to *why* the worker deviates from the Bayesian-Nash strategy.

[31] We restrict attention to $\lambda = 1$ and relegate the case of $\lambda > 1$ to Appendix D. Loss aversion yields qualitatively similar results but complicates the analysis.

pleasantly surprises the principal in later periods.[32,33] Of course, as in the rational case, a worker still has an incentive to provide high initial effort—first impressions remain important. However, the worker has greater incentive to maintain high effort under misattribution since he is unduly penalized for falling short of expectations. In a sense, the principal's biased evaluations impose an informal contract under which the worker is compelled to uphold any precedent he sets for high effort early in the relationship. In fact, while the misattributing worker may reduce effort in early rounds, these incentives to continually surpass expectations can lead to increased *total* effort, $\sum_{t=1}^{T} e_t^*$, relative to the rational case. Hence, misattribution can mitigate the moral-hazard problem that suppresses effort in the rational model.

Additionally, the proposition highlights that while the worker eventually beats expectations, she prefers to do so sufficiently late in the horizon to avoid setting the bar too high early on. Intuitively, the longer is the horizon of the relationship, the greater is the "externality" that early effort imposes on future selves. Indeed, the extent to which the worker under-performs at the onset is increasing in the horizon.[34]

Corollary 2. *Consider the expectations-management setting described above. If beliefs are convex, then there exists a horizon $\bar{T}$ such that $T > \bar{T}$ implies $e_1^* < e_1^r$. Furthermore, $e_1^* - e_1^r$ is strictly decreasing in T conditional on $T > \bar{T}$.*

Although we framed this application as a familiar career-concern model, the analysis directly extends to other settings where one party has an incentive to build a positive reputation. Furthermore, our results may speak to forms of expectations management used in diverse settings ranging from politics, to marketing, to finance. Politicians and firms often strategically "walk down" expectations only to later surpass them. Additionally, research from empirical finance shows that firms attempt to lower investors' expectations prior to earnings announcements. Bartov, Givoly, and Hayn (2002) demonstrate that meeting or beating analyst expectations yields significant excess stock returns. Sim-

[32] This intuition shares similarities with "ratcheting effects" studied in the literature on contracts and regulation (e.g., Freixas, Guesnerie, and Tirole 1985; Laffont and Tirole 1988). In those settings, the worker is reluctant to reveal positive private information about his efficiency early in the relationship so that he can demand higher compensation. In our setting, the worker would want to reveal positive information about θ if he could credibly do so. Misattribution, however, complicates the dynamics of revealing such information.

[33] Because Proposition 5 describes the worker's deviations from the rational path, it does not necessarily imply that e_t^* increases over time. In fact, convexity rules out an increasing effort path in this particular setting where the worker earns a reward each round. That said, misattribution can cause the worker to supply an increasing profile of effort in alternative settings even when beliefs are convex. For example, consider a setting with one payment period that follows multiple rounds of effort and evaluation. If the principal were rational, the worker would smooth his effort across rounds, since each is a perfect substitute for another in terms of the principal's posterior beliefs. In contrast, when the principal suffers misattribution, the optimal pattern of effort follows an increasing profile (see Proposition 2).

[34] As highlighted above, we assume the two parties commit ex ante to a T-period relationship and thus the worker has no additional incentive to provide a positive first impression. While allowing the principal to fire the worker would complicate the analysis, it would not change the qualitative conclusion that the worker undercuts the rational benchmark early in the relationship.

ilarly, Teoh, Yang, and Zhang (2009) show that firms are rewarded for beating expectations even when those analyst forecasts are walked down by firm guidance.[35] Although such expectations management is prevalent in a number of domains, we provide an intuitive mechanism that helps explain why this technique can effectively influence beliefs.

5.2 Over-Extrapolation and Belief Reversals in Autocorrelated Environments

This section illustrates that the type of extrapolative bias demonstrated in Section 3—where beliefs respond excessively to the most recent outcome—can in fact persist in the long run in some environments. Proposition 3 showed that when outcomes are i.i.d., the misattributor will reach a stable expectation about future outcomes. This means that the recency bias identified in Proposition 1 diminishes as the number of rounds grows large.[36] We consider a simple extension with autocorrelated outcomes to show that this vanishing recency bias is not a general phenomenon, but instead stems from the i.i.d. nature of outcomes. In particular, we show that autocorrelation induces the misattributor to continually form overly-extrapolative forecasts of future outcomes, and these forecasts exhibit a predictable error: if the agent overestimates today's outcome, she tends to *underestimate* tomorrow's outcome (and vice versa). In contrast to Proposition 3, these erroneous beliefs persist even when the misattributor is not a loss-averse.

To extend our baseline setup (Sections 2 to 4) to allow for autocorrelation, suppose $x_t = \theta + \varphi x_{t-1} + \epsilon_t$, where the parameter $\varphi \in (0, 1]$ measures the extent of autocorrelation.[37] We assume the agent knows φ and we maintain our baseline assumption that $\epsilon_t \sim N(0, \sigma^2)$. Furthermore, to simply demonstrate that autocorrelation generates persistent over-extrapolation even if updating about θ were to cease, we assume the agent knows θ and we normalize it to $\theta = 0$. Our analysis here focuses on the agent's expectation of x_t given the outcomes prior to t, which we denote by $\widehat{\mathbb{E}}_{t-1}[x_t]$. In our baseline setup from previous sections, this forecast was equal to the agent's current expectation of θ. Here, however, her forecast is simply her perception of the previous outcome scaled by the autocorrelation parameter: $\widehat{\mathbb{E}}_{t-1}[x_t] = \varphi \hat{x}_{t-1}$. As in all previous sections, the agent's reference point in round t is her expectation $\widehat{\mathbb{E}}_{t-1}[x_t]$, and her total utility from x_t is given by Equation 2 with this reference point applied.

[35] Such "expectations management" can be enacted through a number of channels, including strategic accounting of working capital and cash flow from operations (Burgstahler and Dichev 1997), real activities such as sales (Roychowdhury 2006), or through indirect channels such as managing analyst forecasts (e.g., Richardson, Teoh, and Wycoki 2004).

[36] Put differently, the misattributor's prediction of x_{t+1} after t rounds overweights the most recent outcome, x_t, by relatively less as t grows large. More precisely, the weight this expectation places on x_t relative to some preceding outcome x_{t-c} is given by Lemma 1 and is equal to $\frac{\xi_t^t}{\xi_{t-c}^t} = \frac{1+\kappa_t}{1+\kappa_{t-c}} \left(\prod_{j=t-c}^{t-1} [1 - \alpha_j \kappa_{j+1}] \right)^{-1}$. For a fixed sequence of (κ_t), this ratio is decreasing in t since $\alpha_t \to 0$ in t.

[37] We restrict attention to the case of positive autocorrelation solely for the sake of exposition. Analogous results hold for the case of $\varphi \in [-1, 0)$.

We now show that a misattributor's forecasts over-respond to the most recent outcome and are therefore excessively volatile and exhibit predictable errors. To first illustrate the intuition, consider an outcome that beats today's expectations. The misattributor will exaggerate its value and, due to the autocorrelation, expect an unreasonably high value tomorrow. These inflated expectations, however, imply that the next outcome will typically disappoint, thereby causing the misattributor to reverse course and form overly-pessimistic beliefs. This pattern will continue over time: the agent forms an exaggerated forecast in the direction of the most recent outcome, which leads to a subsequent reversal. To formalize, let $d_t \equiv \hat{x}_t - \widehat{\mathbb{E}}_{t-1}[x_t]$ denote the misattributor's forecast error realized on date t.[38]

Proposition 6. *A misattributor's forecast entering period* $t+1$ *is* $\widehat{\mathbb{E}}_t[x_{t+1}] = \varphi \hat{x}_t$, *where*

$$\hat{x}_t = (1+\kappa_t)x_t + \sum_{j=1}^{t-1}(1+\kappa_j)\left((-\varphi)^{t-j}\prod_{i=j+1}^{t}\kappa_i\right)x_j. \tag{11}$$

Hence, forecasts exhibit

1. *Excessive extrapolation and volatility:* $\widehat{\mathbb{E}}_t[x_{t+1}]$ *overweights the outcome on date* t *by a factor of* $(1+\kappa_t)$, *and conditional on* $(x_1, \ldots, x_{t-1})$, $\text{Var}\big(\widehat{\mathbb{E}}_t[x_{t+1}]\big) = (1+\kappa_t)^2\text{Var}\big(\mathbb{E}_t[x_{t+1}]\big)$.

2. *Predictable errors and reversals: Forecast errors follow a negatively-correlated process given by*
$$d_t = (1+\kappa_t)\Big\{-\varphi\left(\frac{\kappa_{t-1}}{1+\kappa_{t-1}}\right)d_{t-1} + \epsilon_t\Big\}. \tag{12}$$

The most recent outcome is overweighted by a factor of $1+\kappa_t$, implying that forecasts overreact to recent outcomes. Additionally, a misattributor's forecast wrongly depends on *all* past outcomes, while the rational forecast is independent of outcomes prior to t after conditioning on x_t. Consistent with the oscillating logic in the example above, Equation 11 reveals that the misattributor's forecast negatively weights outcomes that occurred an odd number of periods ago and positively weights those that happened an even number of periods ago.

While rational predictions generate uncorrelated forecast errors, Part 2 of Proposition 6 highlights the negative relationship between a misattributor's errors: overly optimistic forecasts are typically followed by overly pessimistic forecasts. The strength of this relationship is increasing in both the extent of misattribution and the extent of autocorrelation. Figure 2 below uses a simulated time seres to depict a misattributor's overly extrapolative forecasts (top panel) and the negative relationship in her forecast errors (bottom panel).

[38] We define the forecast error as the difference between the misattributor's encoded outcome and her expectations—this is the agent's perceived forecast error. Our prediction of a negative relationship between today's forecast error and tomorrow's holds if we alternatively define the forecast error as the difference between the true outcome and expectations.

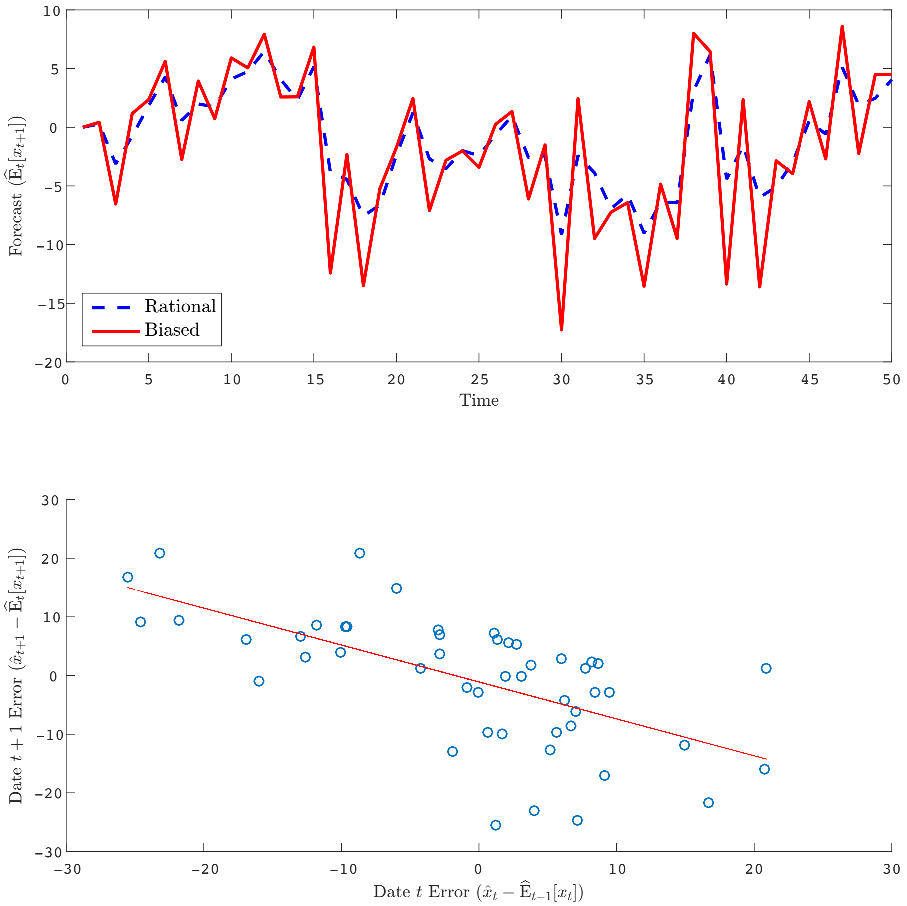

Figure 2: *The top panel displays both the rational and biased forecasts for a simulated process with* $\varphi = 0.7$*,* $\sigma = 5$*,* $\eta = 1$*,* $\lambda = 3$ *and* $\hat{\eta} = 1/3$*. Using that same data, the bottom panel depicts the negative relationship between forecast errors on date t and* $t + 1$.

Our basic prediction of overly-extrapolative and volatile forecasts accords with a range of evidence. For example, Gennaioli, Ma, and Shleifer (2015) and Greenwood and Shleifer (2014) find that managers and investors form extrapolative, volatile predictions of their future earnings and that their forecast errors are negatively correlated with past performance. While alternative models give rise to this general pattern of extrapolative beliefs and systematic reversals—e.g., Bordalo, Gennaioli, and Shleifer's (2017a) model of diagnostic expectations based on Kahneman and Tversky's (1972) representativeness heuristic—our model provides additional predictions that may help empirically disentangle these mechanisms. Namely, we predict that these patterns are more pronounced when forecasting about one's own earnings (i.e., when outcomes generate elation and disappointment) and that beliefs respond more to bad news than good.

6 Misattribution and Personal Equilibrium

In our baseline model, the misattributing agent faced an exogenous distribution of outcomes—she could not directly take actions to influence this distribution. To handle environments that endogenize the distribution—for instance, natural settings where the agent can choose between different prospects each round—we must add more structure to our framework. In particular, we must specify how the agent's strategy over actions influences her reference points. Below we sketch how to extend our baseline model in such settings, and we apply this extension to highlight an additional result: when the agent can exert effort to improve the distribution of outcomes, she will settle on inefficiently high levels of effort.

To motivate this extension, first 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. We assume that this choice determines the agent's 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. We further assume the agent knows how her actions determine her reference point and accordingly maximizes her expected utility. This corresponds to KR's (2007) notion of "choice-acclimating personal equilibrium" aside from the fact that we do not impose rational expectations and choices are made with respect to the misattributor's biased subjective beliefs.[39]

We now sketch how to extend our baseline model more generally. 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 $F(\cdot|\theta, a_t)$. The agent begins with a prior π_0 over θ, and updates these beliefs each round conditional on her action and its resulting (mis)encoded outcome. To derive the encoded outcome, let π_{t-1} denote the agent's beliefs 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 the encoded outcome is analogous to Equation 4:

$$\hat{x}_t = \begin{cases} x_t + \kappa^G \left(x_t - \widehat{\mathbb{E}}_{t-1}[x_t|a_t] \right) & \text{if} \quad x_t \geq \widehat{\mathbb{E}}_{t-1}[x_t|a_t] \\ x_t + \kappa^L \left(x_t - \widehat{\mathbb{E}}_{t-1}[x_t|a_t] \right) & \text{if} \quad x_t < \widehat{\mathbb{E}}_{t-1}[x_t|a_t]. \end{cases} \tag{13}$$

To close the model, we assume the decision maker takes an action in each round to maximize her continuation utility (according to her true utility function) conditional on her erroneous beliefs.

[39] We share KR's intuitions that details of the environment dictate whether this is an appropriate assumption. Specifically, this solution concept is on firmest ground when there is a long period of time between decisions and outcomes. Additionally, this solution concept provides tractability.

Critically, we assume the agent chooses an action believing she encodes outcomes correctly; that is, she does not understand that she suffers misattribution. We can think of this dynamic strategy as a *biased-belief personal equilibrium*, as it extends KR's notion of personal equilibrium to the case where the agent holds erroneous expectations derived from some misspecified model.

Many of our results from previous sections still apply in this extension so long as outcomes from different actions are independent of one another. This follows from our assumption that the reference point corresponds to expectations about the chosen action. Thus, if outcomes are independent across actions, then updating about one action does not influence updating about another. Although our prior analyses focus on learning about a single prospect, our predicted patterns in updating extend in this case to learning about multiple potential prospects.[40]

6.1 Application: Escalating Effort in Repeated Tasks

As a final application and a demonstration of the extension above, we analyze a stylized repeated-search problem where, in each period, a person can exert effort to improve the likelihood of a better outcome. Imagine, for instance, a consumer researching a new purchase, or a hiring manager reviewing potential job candidates. We consider a misattributor who attempts to learn the optimal effort to exert. Importantly, she fails to appreciate how an increase in expectations resulting from her own effort will cause bad outcomes (e.g., purchases that end up being lower quality than expected) to seem even worse when they happen. As such, providing effort alters the agent's perception of bad outcomes in a way that suggests additional exertion is optimal, leading the agent to settle on an inefficiently high level of effort. Put another way, behavior exhibits a dynamic form of the sunk-cost fallacy: the more effort a person has already exerted, the more she feels compelled to try even harder going forward.

To model this scenario, we consider an agent learning about her average consumption in two distinct states, H and L, where the "high" state H provides higher consumption, on average, than the "low" one, L. The state varies each periods and is denoted by $\omega_t \in \{H, L\}$. Specifically, in each period t, $\omega_t = H$ occurs with probability $p \in (0,1)$ and $\omega_t = L$ occurs otherwise. If $\omega_t = H$, the agent earns consumption utility $x_t = \theta^H + \epsilon_t$. If $\omega_t = L$, she earns $x_t = \theta^L + \epsilon_t$ where $\theta^L < \theta^H$. Following our baseline setup, we assume that $\epsilon_t \sim N(0, \sigma^2)$ and the agent has (independent) normally distributed priors over each θ^ω; let $\hat{\theta}^\omega_{t-1}$ denote the agent's estimate of θ^ω entering round t.[41] While this setting might seem like a significant departure from previous sections, it is very simi-

[40] Independence of outcomes across actions implies that beliefs about a prospect do not update in periods where it was not chosen. Thus, this claim about how our results extend pertains only to updating in those periods in which the prospect was chosen.

[41] We assume that ω_t is observed once x_t is realized. Along with our assumption that the agent treats θ^H and θ^L as independent, this ensures that updating about each parameter θ^ω is relatively simple. Conditional on the agent's encoded outcomes, the updating rule for each θ^ω is similar to the baseline model: letting $N^\omega_t \equiv \sum_{k=1}^t \mathbb{1}\{\omega_k = \omega\}$, her estimate

lar to our baseline model (Sections 2 through 4) except the agent simultaneously learns about the means of two distinct prospects instead of one. As in all previous sections, the agent's reference point in round t is her expectation of x_t: in this setting with two possible states, this expectation is $\widehat{\mathbb{E}}_{t-1}[x_t] = p\hat{\theta}^H_{t-1} + (1-p)\hat{\theta}^L_{t-1}$. The agent's total utility from x_t is given by Equation 2 with reference point $\widehat{\mathbb{E}}_{t-1}[x_t]$ in place of $\hat{\theta}$.

Below, we assume the agent can exert effort to increase p, the chance of the good state. However, to build intuitions, we first consider learning under misattribution when p is fixed and highlight how p affects long-run beliefs. A misattributor will reach biased beliefs such that $\hat{\theta}^H_\infty > \theta^H$ and $\hat{\theta}^L_\infty < \theta^L$. To illustrate, consider the limit where consumption in each state is nearly deterministic (i.e., $\sigma \to 0$), so the agent essentially faces a binary lottery each round with a p chance of $x_t = \theta^H$ and a $1-p$ chance of $x_t = \theta^L$. If the agent were to reach correct beliefs, her reference point would be given by $p\theta^H + (1-p)\theta^L$. But this implies that a good outcome $x_t = \theta^H$ would beat expectations and consequently be encoded as better than it really was: $\hat{x}_t = \theta^H + \kappa^G\big[\theta^H - (p\theta^H + (1-p)\theta^L)\big] > \theta^H$. Similarly, a bad outcome $x_t = \theta^L$ would fall short of expectations and be encoded as worse than it really was. Rational expectations are thus unstable: when holding correct beliefs, estimates of θ^H drift upward while those of θ^L drift downward. The misattributor's biased beliefs accordingly satisfy a fixed-point condition wherein holding expectation $p\hat{\theta}^H_\infty + (1-p)\hat{\theta}^L_\infty$ causes her to encode the true outcomes θ^H and θ^L as $\hat{\theta}^H_\infty$ and $\hat{\theta}^L_\infty$, respectively.[42] Solving the system of equations implied by these conditions yields

$$\begin{aligned} \hat{\theta}^H_\infty &= \theta^H + \left(\tfrac{(1-p)\kappa^G(1+\kappa^L)}{1+p\kappa^G+(1-p)\kappa^L}\right)\big[\theta^H - \theta^L\big] \\ \hat{\theta}^L_\infty &= \theta^L - \left(\tfrac{p\kappa^L(1+\kappa^G)}{1+p\kappa^G+(1-p)\kappa^L}\right)\big[\theta^H - \theta^L\big]. \end{aligned} \tag{15}$$

The noteworthy feature of Equation 15 is that the agent becomes more biased about a state when it happens less frequently (and is hence more surprising). For example, the agent underestimates the worse outcome, $x_t = \theta^L$, by more when it is less likely (i.e., as p increases).[43]

of θ^ω after t rounds is thus

$$\hat{\theta}^\omega_t \equiv \frac{\rho^2}{N^\omega_t\rho^2 + \sigma^2}\left(\sum_{\{k \le t\,:\,\omega_k = \omega\}} \hat{x}_k\right) + \frac{\sigma^2}{N^a_t\rho^2 + \sigma^2}\theta^\omega_0.$$

[42] These conditions imply that steady-state beliefs must solve the following system of equations:

$$\begin{aligned} \hat{\theta}^H_\infty &= \theta^H + \kappa^G\big[\theta^H - \big(\hat{\theta}^H_\infty + (1-p)\hat{\theta}^L_\infty\big)\big] \\ \hat{\theta}^L_\infty &= \theta^L + \kappa^L\big[\theta^L - \big(\hat{\theta}^H_\infty + (1-p)\hat{\theta}^L_\infty\big)\big]. \end{aligned} \tag{14}$$

[43] More generally, misattribution provides an intuitive mechanism through which the probability of an outcome naturally shapes its perceived value. As such, misattribution can distort beliefs when consumers learn about goods allocated via random processes (e.g., auctions, bargaining, or scenarios where products have uncertain availability). For instance,

Having outlined how misattribution distorts beliefs in this new environment absent choice, we now allow the agent to exert effort to increase the probability of the good state, p. Formally, the person chooses this probability at the start of each period t. Her choice—denoted by $p_t \in [p_0, \bar{p}]$—comes at a cost $c(p_t - p_0)$, where $\bar{p} \in (p_0, 1)$ and $c(\cdot)$ is convex, minimized at zero, and admits a continuous marginal cost that is weakly convex with $c'(0) = 0$. We assume that $p_0 = 1/2$—that is, the default probability of the better outcome is $1/2$ when the person exerts no effort.[44]

A misattributor fails to optimize effort in this setting because she does not account for how her chosen effort—and thus her expectations—shape her perceptions of outcomes. Loosely, if the misattributor were to work to increase p, her seemingly optimal effort will eventually feel inadequate. Recall from above that increasing the chance of the good state lowers the misattributor's perception of the payoff in the bad state (see Equation 15). Hence, the agent will underestimate θ^L, and the degree to which she does so is increasing in p. Although higher expectations also cause the misattributor to overestimate θ^H by less, loss aversion ensures that perceptions of the bad state move downward by more than those of the good state. Overall, increasing p causes a misattributor to overestimate the payoff difference between the good and bad states by more. In turn, these new beliefs inspire the agent to exert greater effort and to further increase p. This pattern of escalating effort will converge to an inefficiently high steady-state level.

Proposition 7. *Consider the repeated search setting described above. If $\lambda > 1$, then a misattributor perpetually exerts excessive effort: if the full-information rational effort level, p^r, is interior ($p^r < \bar{p}$), then p_t converges almost surely to a long-run value that strictly exceeds p^r. Otherwise, if $p^r = \bar{p}$, then p_t converges almost surely to $\bar{p}$.*

Moreover, this excessive effort is costly to a misattributor: her average experienced utility at the long-run level is strictly less than what she would achieve if she correctly inferred θ^H and θ^L. The logic in this example may underlie, for instance, common intuitions that people exert excessive effort when comparison shopping. More generally, the result suggests a mechanism underlying a dynamic sunk-cost fallacy: initial effort increases the perceived value of future work, which further induces effort.

misattributors learning about a product that is at times unavailable will come to overvalue that product. Simultaneously, they will *under*value their fall-back option. Consequently, a firm may choose to limit supply when first introducing a high-quality product: those lucky enough to receive the good early may overstate its quality, thereby increasing demand later.

[44] The assumptions on $c(\cdot)$ simplify the exposition, as they guarantee a unique optimal choice each round. The assumption that $\bar{p} < 1$ reflects the idea that it is not possible to eliminate all uncertainty. Moreover, $\bar{p} < 1$ rules out paths where the person reaches $p_t = 1$ and remains there simply because she lacks feedback about θ^L. Finally, a "default" value $p_0 = 1/2$ ensures that the marginal benefit of increasing p is always positive. This rules out pathological cases where the agent actually prefers a smaller chance of the better outcome, which can happen when p_0 is sufficiently small given the risk aversion inherent in the KR solution concept.

7 Conclusion

We conclude by contrasting our approach with alternative models and by highlighting ways that future empirical work could explore the implications of our model. Additionally, we discuss two natural extensions of our framework: (1) incorporating misattribution of news utility (Kőszegi and Rabin 2009) and (2) extending misattribution to social-learning environments.

In addition to the models of mistaken learning noted in the introduction, we build on an emerging literature that examines limited or distorted memory. Wilson (2014) follows a rational approach with bounded memory and examines the optimal coarsening of information given a memory constraint. This approach yields predictions distinct from misattribution as it implies that first impressions dominate subsequent evaluations. Mullainathan (2002) provides a model of limited rationality which can generate a form of overreaction to information through memory associations. Relatedly, Bordalo, Gennaioli, and Shleifer (2017b) consider a model of selective memory in which salient events are more prone to recall. In their model, the agent correctly encodes outcomes to memory, but the context of her current choice influences which past outcomes she recalls; in contrast, the agent in our model erroneously encodes outcomes. While in some settings these two models both predict that particularly good or bad experiences “stand out”, they offer different predictions regarding an agent’s perceptions of past outcomes. Finally, our model extends a literature studying distorted beliefs in settings where an agent’s utility depends directly on those beliefs.[45]

A natural avenue for empirical exploration is our prediction of belief-based contrast effects—a fixed outcome will seem better when the previous one was worse. We predict that contrast effects increase when the perceived correlation between today’s outcome and tomorrow’s is stronger.[46] Furthermore, in order to separate effects generated by our mechanism from other potential explanations—e.g. the Gambler’s Fallacy (Chen, Moskowitz, and Shue 2016)—our model suggests comparing circumstances where outcomes have utility consequences with those that don’t. Our model predicts that contrast effects will be enhanced the more that a person cares about the outcomes she faces. This empirical strategy can also help distinguish our mechanism from base-rate neglect (e.g. Benjamin, Bodoh-Creed, and Rabin 2016) or the representativeness heuristic (e.g. Bordalo, Gennaioli, and Shleifer 2017a), which both predict recency effects and extrapolative beliefs. For instance, one could test whether investors’ forecasts are more extrapolative about companies they hold a stake in relative to those they do not. Testing whether such forecasts respond differently to losses versus gains is yet another way to distinguish our mechanism from these alternative explanations.

[45] In contrast to many models in this literature (e.g., Bénabou and Tirole 2002; Brunnermeier and Parker 2005), a misattributor does not purposefully manipulate her beliefs to maximize her belief-based utility. Rather, she forms distorted beliefs mechanically as a result of the bias in her perceptions.

[46] Indeed, Hartzmark and Shue (2018) find that contrast effects among investors stemming from prior-day earnings announcements are larger for within-sector peers than across industries.

Finally, we note two potential extensions of our model. First, throughout this paper we have omitted the notion of "news utility" (Kőszegi and Rabin 2009), in which a person experiences elations and disappointments from changes in beliefs about future consumption. News utility provides a channel for monetary outcomes to directly influence contemporaneous experienced utility, and thus incorporating misattribution of news utility would naturally extend our predictions to settings where the outcomes are monetary earnings. Moreover, this extension introduces novel comparative statics. To illustrate, consider a worker who agrees to a new position for a pre-specified amount of time, and imagine that her first encounter with the new job is worse than expected. With misattribution of news utility, her evaluation of that first experience will be worse the longer she committed to the job—that first episode creates a greater sense of disappointment about the future when the duration of her contract is longer.

Second, our model can be reframed as an interpersonal bias where an observer neglects how expectations shape the experiences of others. For instance, a person reading online reviews (e.g., Yelp) for a product 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 expectations—can hinder social learning. Additionally, these settings may provide data-rich environments to explore the empirical implications of our model. If such social misattribution occurs, we would expect ratings to demonstrate the dynamic patterns described in this paper.

References

ABELER, J., A. FALK, L. GOETTE, AND D. HUFFMAN (2011): "Reference Points and Effort Provision." *American Economic Review*, 101(2), 470–492.

ANDERSON, R. (1973): "Consumer Dissatisfaction: The Effect of Disconfirmed Expectancy on Perceived Product Performance." *Journal of Marketing Research*, 10, 38–44.

BACKUS, M., T. BLAKE, D. MASTEROV, S. TADELIS (2018): "Expectation, Disappointment, and Exit: Evidence on Reference Point Formation from an Online Marketplace." *Working Paper*.

BANERJI, A. AND N. GUPTA (2014). "Detection, Identification, and Estimation of Loss Aversion: Evidence from an Auction Experiment." *American Economic Journal: Microeconomics*, 6, 91–133.

BARBERIS, N., A. SHLEIFER, AND R. VISHNY (1998): "A Model of Investor Sentiment." *Journal of Financial Economics*, 49, 307–343.

BARTOV, E., D. GIVOLY, AND C. HAYN (2002): "The Rewards to Meeting or Beating Earnings Expectations." *Journal of Accounting and Economics,* 33(2), 173–204.

BAUMEISTER, R., C. FINKENAUER, AND K. VOHS (2001): "Bad is Stronger than Good." *Review of General Psychology*, 5(4), 323–370.

BAUMGARTNER, H., M. SUJAN, AND D. PADGETT (1997): "Patterns of Affective Reactions to Advertisements: The Integration of Moment-to-Moment Responses into Overall Judgments." *Journal of Marketing Research*, 34(2), 219–232.

BELL, D. (1985): "Disappointment in Decision Making under Uncertainty." *Operations Research*, 33(1), 1–27.

BÉNABOU, R. AND J. TIROLE (2002): "Self-Confidence and Personal Motivation." *The Quarterly Journal of Economics*, 117(3), 871–915.

BENJAMIN, D., A. BODOH-CREED, AND M. RABIN (2016): "Base-Rate Neglect: Foundations and Applications." *Working Paper*.

BERTRAND, M., AND S. MULLAINATHAN (2001): "Are CEOs Rewarded for Luck? The Ones Without Principals Are.," *Quarterly Journal of Economics*, 116(3), 901–932.

BHARGAVA, S. (2007): "Perception is Relative: Contrast Effects in the Field." *Working Paper*.

BHARGAVA, S. AND R. FISMAN (2014): "Contrast Effects in Sequential Decisions: Evidence from Speed Dating." *Review of Economics and Statistics*, 96(3), 444–457.

BOHREN, A. (2016): "Informational Herding with Model Misspecification." *Journal of Economic Theory*, 163, 222-247.

BOHREN, A. AND D. HAUSER (2018): "Social Learning with Model Misspecification: A Framework and a Robustness Result." *Working paper*.

BOULDING, W., A. KALRA, R. STAELIN, AND V. ZEITHAML (1993): "A Dynamic Process Model of Service Quality: From Expectations to Behavioral Intentions." *Journal of Marketing Research*, 30, 7—27.

BORDALO, P., N. GENNAIOLI, AND A. SHLEIFER (2017a): "Diagnostic Expectations and Credit Cycles." *Working Paper*.

BORDALO, P., N. GENNAIOLI, AND A. SHLEIFER (2017b): "Memory, Attention and Choice." *Working Paper*.

BROWNBACK, A. AND M. KUHN (2019): "Attribution Bias, Blame, and Strategic Confusion in Punishment Decisions." *Games and Economic Behavior*, Forthcoming.

BRUNNERMEIER, M.K. AND J.A. PARKER (2005): "Optimal Expectations." *American Economic Review*, 95(4), 1092–1118.

BURGSTAHLER, D. AND I. DICHEV (1997): "Earnings Management to Avoid Earnings Decreases and Losses. *Journal of Accounting and Economics*, 24, 99–126.

BUSHONG, B. AND T. GAGNON-BARTSCH (2018): "Misattribution of Reference Dependence: Evidence from Real-Effort Experiments." *Working Paper.*

CARD, D., AND G. DAHL (2011): "Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior." *Quarterly Journal of Economics*, 126(1), 103–143.

CHAMBERS, C., AND P. HEALY (2012): "Updating towards the signal." *Economic Theory*, 50, 765–786.

CHARNESS, G. AND D. LEVIN (2005): "When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect." *American Economic Review*, 95(4), 1300–1309.

CHEN, D., T. MOSKOWITZ, AND K. SHUE (2016): "Decision-Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." *Quarterly Journal of Economics*, 131(3), 1181–1241.

CHIANG, Y., D. HIRSHLEIFER, Y. QIAN, AND A. SHERMAN (2011): "Do Investors Learn From Experience? Evidence from Frequent IPO Investors." *Review of Financial Studies*, 24(5), 1560–1589.

COLE, S., A. HEALY, AND E. WERKER (2012): "Do Voters Demand Responsive Governments? Evidence from Indian Disaster Relief," *Journal of Development Economics*, 97(2), 167–181.

CRAWFORD, V., AND J. MENG (2011): "New York City Cabdrivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational- Expectations Targets for Hours and Income." *American Economic Review*, 101(5), 1912–1932.

DEGROOT, M. (1970): *Optimal Statistical Decisions*, McGraw-Hill, New York.

DILLENBERGER, D. AND K. ROZEN (2015): "History-Dependent Risk Attitude." *Journal of Economic Theory*, 157, 445–477.

DUTTON, D. AND A. AARON (1974): "Some Evidence for Heightened Sexual Attraction Under Conditions of High Anxiety." *Journal of Personality and Social Psychology*, 30, 510–517.

EHLING, P., A. GRANIERO, AND C. HEYERDAHL-LARSEN (2018): "Asset Prices and Portfolio Choice with Learning from Experience." *Review of Economic Studies*, 85(3), 1752–1780.

EPSTEIN, L., J. NOOR, AND A. SANDRONI (2010): "Non-Bayesian Learning." *The B.E. Journal of Theoretical Economics*, 10(1).

ERKAL, N., L. GANGADHARAN, AND B.H. KOH (2019): "Attribution Biases in Leadership: Is it Effort or Luck?" *Working Paper*.

ESPONDA, I. AND D. POUZO (2016): "Berk-Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models." *Econometrica*, 84(3), 1093–1130.

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

ERICSON, K. AND A. FUSTER (2011): "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments." *Quarterly Journal of Economics*, 126(4), 1879–907.

FREIXAS, X., R. GUESNERIE, AND J. TIROLE (1985): "Planning under Incomplete Information and the Ratchet Effect." *Review of Economic Studies*, 52(2), 173–191.

FRICK, M., R. IIJIMA, AND Y. ISHII (2018): "Misinterpreting Others and the Fragility of Social Learning." *Working paper*.

FRYER, R., P. HARMS, AND M. JACKSON (2018): "Updating Beliefs when Evidence is Open to Interpretation: Implications for Bias and Polarization." *Journal of the European Economics Association*, Forthcoming.

FUDENBERG, D., G. ROMANYUK, AND P. STRACK (2017): "Active Learning with a Misspecified Prior." *Theoretical Economics*, 12, 1155–1189.

GALLAGHER, J. (2014): "Learning about an Infrequent Event: Evidence from Flood Insurance Take-up in the US." *American Economic Journal: Applied Economics*, 6(3), 206–233.

GEERS, A. AND G. LASSITER (1999): "Affective Expectations and Information Gain: Evidence for Assimilation and Contrast Effects in Affective Experience." *Journal of Experimental Social Psychology*, 35(4), 394–413.

GENNAIOLI, N., Y. MA, AND A. SHLEIFER (2015): "Expectations and Investment." *NBER Macroeconomics Annual*, 30, 379–442.

GILBERT, D. AND P. MALONE (1995): "The correspondence bias." *Psychological Bulletin*, 117(1): 21–38.

GILL, D. AND V. PROWSE (2012): "A Structural Analysis of Disappointment Aversion in a Real Effort Competition." *American Economic Review*, 102(1), 469–503.

GNEEZY, U., L. GOETTE, C. SPRENGER, AND F. ZIMMERMANN (2017): "The Limits of Expectations-Based Reference Dependence." *Journal of the European Economic Association*, 15, 861–876.

GOETTE, L., A. HARMS, AND C. SPRENGER (2019): "Randomizing Endowments: An Experimental Study of Rational Expectations and Reference-Dependent Preferences." *American Economic Journal: Microeconomics*, 11(1), 185–207.

GREENWOOD, R. AND A. SHLEIFER (2014): "Expectations of Returns and Expected Returns." *Review of Financial Studies*, 27(3), 714–746.

HAGGAG, K. AND D. POPE (2018): "Attribution Bias in Economic Decision Making." *Review of Economic Studies*, Forthcoming.

HAISLEY, E. AND G. LOEWENSTEIN (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." *Journal of Marketing Research*, 48(1), 103–115.

HASELHUHN, M., D. POPE, AND M. SCHWEITZER (2012): "Size Matters (and so Does Experience): How Personal Experience with a Fine Influences Behavior." *Management Science*, 58(1), 35–51.

HARTZMARK, S. AND K. SHUE (2018): "A Tough Act to Follow: Contrast Effects in Financial Markets." *Journal of Finance*, Forthcoming.

HEFFETZ, O. AND J. LIST (2014): "Is the Endowment Effect an Expectations Effect?" *Journal of the European Economic Association*, 12, 1396–1422.

HEIDHUES, P., B. KŐSZEGI, AND P. STRACK (2018): "Unrealistic Expectations and Misguided Learning." *Econometrica*, 86(4), 1159–1214.

HEIDHUES, P., B. KŐSZEGI, AND P. STRACK (2019): "Convergence in Misspecified Learning Models with Endogenous Actions." *Working Paper*.

HIGHHOUSE, S. AND A. GALLO (1997): "Order Effects in Personnel Decision Making." *Human Performance*, 10(1), 31–46.

HO, T. AND Y. ZHENG (2004): "Setting Customer Expectations in Service Delivery: An Integrated Marketing-Operations Perspective." *Management Science*, 50(4), 479–488.

HOGARTH, R. AND H. EINHORN (1992): "Order Effects in Belief Updating: The Belief-Adjustment Model." *Cognitive Psychology*, 24, 1–55.

HOLMSTRÖM, B. (1999): "Managerial Incentive Problems: A Dynamic Perspective." *Review of Economic Studies*, 66(1), 169–182.

IMAS, A. (2016): "The Realization Effect: Risk-Taking after Realized versus Paper Losses." *American Economic Review*, 106(8), 2086–2109.

JAMES, O. (2009): "Evaluating the Expectations Disconfirmation and Expectations Anchoring Approaches to Citizen Satisfaction with Local Public Services." *Journal of Public Administration Research and Theory*, 19, 107–123.

KAHNEMAN, D., B. FREDRICKSON, C. SCHREIBER, AND D. REDELMEIER (1993): "When More Pain is Preferred to Less: Adding a Better End." *Psychological Science*, 4(6), 401–405.

KAHNEMAN, D. AND A. TVERSKY (1972): "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology*, 1972; 3(3), 430–454.

KAHNEMAN, D., AND A. TVERSKY (1979): "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 1979; 47(2), 263–291.

KAHNEMAN, D., B. FREDRICKSON, C. SCHREIBER, AND D. REDELMEIER (1993): "When More Pain Is Preferred to Less: Adding a Better End." *Psychological Science*, 4(6), 401–405.

KARLE, H., G. KIRCHSTEIGER, AND M. PEITZ (2015): "Loss Aversion and Consumption Choice: Theory and Experimental Evidence." *American Economic Journal: Microeconomics*, 7(2), 101–120.

KAUSTIA, M. AND S. KNÜPFER (2008): "Do Investors Overweight Personal Experience? Evidence from IPO Subscriptions." *Journal of Finance*, 63(6), 2679–2702.

KIMBALL, D. AND S. PATTERSON (1997): "Living Up to Expectations: Public Attitudes Toward Congress." *Journal and Politics*, 59, 701–728.

KOPALLE, P. AND D. LEHMANN (2006): "Setting Quality Expectations When Entering a Market: What Should the Promise Be?" *Marketing Science*, 25(1), 8–24.

KŐSZEGI, B., AND M. RABIN (2006): "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics*, 121(4), 1133–1165.

KŐSZEGI, B., AND M. RABIN (2007): "Reference-Dependent Risk Attitudes." *American Economic Review*, 97(4), 1047–1073.

KŐSZEGI, B., AND M. RABIN (2009): "Reference-Dependent Consumption Plans." *American Economic Review*, 99(3), 909–936.

KUSHNER, H. AND G. YIN (2003): *Stochastic Approximation and Recursive Algorithms and Applications*, Vol. 35, Springer.

KUHNEN, C. (2015): "Asymmetric Learning from Financial Information." *Journal of Finance*, 70(5), 2029–2062.

LAFFONT, J., AND J. TIROLE (1988): "The Dynamics of Incentive Contracts." *Econometrica*, 56(5), 1153–1175.

LOEWENSTEIN, G. AND D. PRELEC (1993): "Preferences for Sequences of Outcomes." *Psychological Review*, 100(1), 91–108.

LOOMES, G., AND R. SUGDEN (1986): "Disappointment and Dynamic Consistency in Choice under Uncertainty." *Review of Economic Studies*, 53(2), 271–282.

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

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

MALMENDIER, U. AND S. NAGEL (2016): "Learning from Inflation Experiences." *Quarterly Journal of Economics*, 131(1), 53-87

MALMENDIER, U., D. POUZO, AND V. VANASCO (2018): "A Theory of Experience Effects." *Working Paper*.

MARKLE, A., G. WU, R. WHITE, AND A. SACKETT (2015): "Goals as reference points in marathon running: A novel test of reference dependence," *Working Paper*.

MULLAINATHAN, S. (2002): "A Memory-Based Model of Bounded Rationality." *Quarterly Journal of Economics*, 117(3), 735–774.

NYARKO, Y. (1991): "Learning in Misspecified Models and the Possibility of Cycles." *Journal of Economic Theory*, 55, 416–427.

OLIVER, R. (1977): "Effect of Expectation and Disconfirmation of Post-Exposure Product Evaluation: An Alternative Interpretation." *Journal of Applied Psychology*, 62, 480–486.

OLIVER, R. (1980): "A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions." *Journal of Marketing Research*, 17, 460–469.

PATTERSON, S., G. BOYNTON, AND R. HEDLUND (1969): "Perceptions and Expectations of the Legislature and Support for It." *American Journal of Sociology*, 75(1), 62–76.

PEETERS, G. AND J. CZAPINSKI (1990): "Positive-Negative Asymmetry in Evaluations: The Distinction Between Affective and Informational Negativity Effects." *European Review of Social Psychology*, 1(1), 33–60.

POPE, D. AND M. SCHWEITZER (2011): "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes." *American Economic Review*, 101(1), 129–157.

POST, T., M. VAN DEN ASSEM, G. BALTUSSEN, AND R. THALER (2008): "Deal or No Deal? Decision Making under Risk in a Large-Payoff Game Show." *American Economic Review*, 98(1), 38–71.

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

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

RICHARDSON, S., S. TEOH, P. WYSOCKI (2004): "The Walk-Down to Beatable Analyst Forecasts: The Role of Equity Issuance and Insider Trading Incentives." *Contemporary Accounting Research*, 21(4), 885–924.

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, 173–220.

ROSS, W. AND I. SIMONSON (1991): "Evaluations of Pairs of Experiences: A Preference for Happy Endings." *Journal of Behavioral Decision Making*, 4, 272–282.

ROYCHOWDHURY, S. (2006) "Earnings Management Through Real Activities Manipulation." *Journal of Accounting and Economics,* 42, 335–370.

RUTLEDGE, R., N. SKANDALI, P. DAYAN, AND R. DOLAN (2014): "A Computational and Neural Model of Momentary Subjective Well-Being," *Proceedings of the National Academy of Sciences*, 111(33), 12252–12257.

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

SIMONSONHN, U. (2007): "Clouds Make Nerds Look Good: Field Evidence of the Impact of Incidental Factors on Decision Making." *Journal of Behavioral Decision Marking*, 20(2), 143–152.

SIMONSONHN, U. (2009): "Weather to Go to College." *The Economic Journal*, 120(543), 270–280.

SKINNER, D. AND R. SLOAN (2002): "Earnings Surprises, Growth Expectations, and Stock Returns or Don't Let an Earnings Torpedo Sink Your Portfolio." *Review of Accounting Studies*, 7(2), 289–312.

SPIEGLER, R. (2016): "Bayesian Networks and Boundedly Rational Expectations." *Quarterly Journal of Economics*, 131, 1243–1290.

TEOH, S., Y. YANG AND Y. ZHANG (2009): "The Earnings Numbers Game: Rewards to Walk Down and Penalties to Walk Up Of Analysts' Forecasts of Earnings." *Working Paper.*

THALER, R. AND E. JOHNSON (1990): "Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice," *Management Science*, 36, 643–660.

TVERSKY, A. AND GRIFFIN, D. (1991): "On the Dynamics of Hedonic Experience: Endowment and Contrast in Judgments of Well-Being." In Strack, F., Argyle, M. and Schwartz, N. (Eds.), *Subjective Well-Being*, Pergamon Press, 101–118.

VAN RYZIN, G. (2004): "Expectations, Performance, and Citizen Satisfaction with Urban Services." *Journal of Policy Analysis and Management*, 23, 433–448.

VISSING-JORGENSON, A. (2003): "Behavioral Finance: An Impartial Assessment or Does Irrationality Disappear with Wealth? Evidence from Expectations and Actions." In *NBER Macroeconomics Annual.*

WATERMAN, R., H. JENKINS-SMITH, AND C. SILVA (1999): "The Expectations Gap Thesis: Public Attitudes toward an Incumbent President." *Journal of Politics*, 61(4), 944–966.

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

WILSON, T. AND D. GILBERT (2003): "Affective Forecasting." *Advances in Experimental Social Psychology*, 35, pp. 345–411.

WOLFERS, J. (2007): "Are Voters Rational? Evidence from Gubernatorial Elections," *Working Paper.*

XIA, X. (2016): "Forming Wage Expectations through Learning: Evidence from College Major Choice." *Journal of Economic Behavior and Organization*, 132: 176–196.

Appendix

A Proofs of Results in the Main Text

Proof of Lemma 1.

Proof. Without loss of generality, suppose $\theta_0 = 0$. Thus, for any $t \in \{1, \ldots, T\}$, we have $\hat{\theta}_t = \alpha_t \sum_{k=1}^t \hat{x}_k = \alpha_t \sum_{k=1}^t [x_t + \kappa_t(x_t - \hat{\theta}_{t-1})]$. We use induction to prove the following: $\hat{\theta}_t = \alpha_t \sum_{k=1}^t \xi_k^t x_t$ where $\xi_k^t = (1+\kappa_k) \prod_{j=k}^{t-1} [1 - \alpha_j \kappa_{j+1}]$ for $k < t$ and $\xi_t^t = (1 + \kappa_t)$. To establish the base case, note that $\hat{\theta}_1 = \alpha_1[x_1 + \kappa_1(x_1 - \theta_0)] = \alpha_1(1+\kappa_1)x_1$. Now suppose the claim holds for period $t > 1$. Then

$$\begin{aligned}
\hat{\theta}_{t+1} &= \alpha_{t+1} \sum_{k=1}^{t+1} \hat{x}_k \\
&= \alpha_{t+1} \left\{ [(1+\kappa_{t+1})x_{t+1} - \kappa_{t+1}\hat{\theta}_t] + \frac{1}{\alpha_t}\hat{\theta}_t \right\} \\
&= \alpha_{t+1} \left\{ (1+\kappa_{t+1})x_{t+1} + [1 - \alpha_t \kappa_{t+1}] \sum_{k=1}^{t} (1+\kappa_k) \left(\prod_{j=k}^{t-1} [1 - \alpha_j \kappa_{j+1}] \right) x_k \right\} \\
&= \alpha_{t+1} \left\{ (1+\kappa_{t+1})x_{t+1} + \sum_{k=1}^{t} (1+\kappa_k) \left(\prod_{j=k}^{t} [1 - \alpha_j \kappa_{j+1}] \right) x_k \right\}.
\end{aligned}$$

Hence, the induction step holds, establishing the claim.

■

Proof of Proposition 1.

Proof. The results follow from Lemma 1. If beliefs are monotonic, then $\kappa^L < 1 + \sigma^2/\rho^2$. This implies $[1 - \alpha_t \kappa_{t+1}] \in (0, 1)$ for all $t \in \{1, \ldots, T-1\}$. Consider $t, \tau \in \{1, \ldots, T\}$ such that $\kappa_t = \kappa_\tau$.

Part 1. If $t > \tau$, then Lemma 1 implies that $\xi_t^T - \xi_\tau^T = (1-\kappa_t) \prod_{j=t}^{T-1} [1 - \alpha_j \kappa_{j+1}] - (1-\kappa_\tau) \prod_{j=\tau}^{T-1} [1 - \alpha_j \kappa_{j+1}] = (1-\kappa_t) \prod_{j=t}^{T-1} [1 - \alpha_j \kappa_{j+1}] \left(1 - \prod_{j=\tau}^{t-1} [1 - \alpha_j \kappa_{j+1}]\right) = \xi_t^T \left(1 - \prod_{j=\tau}^{t-1} [1 - \alpha_j \kappa_{j+1}]\right) > 0$. The inequality follows since $1 - \alpha_j \kappa_{j+1} \in (0, 1)$ for all relevant j. It is obvious that if instead $\tau > t$, the previous inequality would reverse.

Part 2. Also from Lemma 1, $\xi_T^T = 1 + \kappa_T > 1$ and $\lim_{T\to\infty} \xi_t^T = (1+\kappa_t) \lim_{T\to\infty} \prod_{j=t}^{T-1} [1 - \alpha_j \kappa_{j+1}] \leq (1+\kappa_t) \lim_{T\to\infty} \prod_{j=t}^{T-1} [1 - \alpha_j \kappa^G]$. Since $\sum_{j=t}^{\infty} \alpha_j$ diverges, $\prod_{j=t}^{\infty} [1 - \alpha_j \kappa^G] = 0$, completing the proof of Part 2.

Part 3. From above, if $\tau < t$, then $\xi_t^T / \xi_\tau^T = 1 / \left(\prod_{j=\tau}^{t-1} [1 - \alpha_j \kappa_{j+1}] \right)$. Note that α_j is increasing in ρ^2/σ^2, and thus $\prod_{j=\tau}^{t-1} [1 - \alpha_j \kappa_{j+1}]$ is decreasing in ρ^2/σ^2 since each term of the product is decreasing in ρ^2/σ^2. Hence ξ_t^T / ξ_τ^T is increasing in ρ^2/σ^2. ■

Proof of Proposition 2.

Proof. We begin by proving a lemma that describes the final belief after two outcomes, $\hat{\theta}_2$. The lemma shows that if beliefs are convex, then $\hat{\theta}_2$ is maximized whenever the higher outcome happens last. We then extend this preliminary result to an arbitrary number of outcomes $T > 2$.

Lemma A.1. *Consider any $a, b \in \mathbb{R}$ such that $a > b$. Let $\hat{\theta}_2^d$ denote the mean belief following the decreasing sequence (a, b), and let $\hat{\theta}_2^i$ denote that following the increasing sequence (b, a). If beliefs are convex, then $\hat{\theta}_2^i > \hat{\theta}_2^d$.*

Proof of Lemma A.1. Given our normality assumptions, the misattributor's posterior over θ after t observations is normally distributed with mean $\hat{\theta}_t$, where

$$\hat{\theta}_t = \left(\frac{\sigma^2}{t\rho^2 + \sigma^2}\right)\theta_0 + \left(\frac{\rho^2}{t\rho^2 + \sigma^2}\right)\sum_{\tau=1}^{t} \hat{x}_\tau. \tag{A.1}$$

From Equation $A.1$, we can write $\hat{\theta}_2^i = \alpha_2(\hat{b}_1^i + \hat{a}_2^i) + (1 - 2\alpha_2)\theta_0$ 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 $\kappa_1^i = \kappa^G \mathbb{1}\{b \geq \theta_0\} + \kappa^L \mathbb{1}\{b < \theta_0\}$, and $\kappa_2^i = \kappa^G \mathbb{1}\{a \geq \hat{\theta}_1^i\} + \kappa^L \mathbb{1}\{a < \hat{\theta}_1^i\}$ where $\hat{\theta}_1^i = \alpha_1(1 + \kappa_1^i)(b - \theta_0) + \theta_0$. Similarly, let $\kappa_1^d = \kappa^G \mathbb{1}\{a \geq \theta_0\} + \kappa^L \mathbb{1}\{a < \theta_0\}$, and $\kappa_2^d = \kappa^G \mathbb{1}\{b \geq \hat{\theta}_1^d\} + \kappa^L \mathbb{1}\{b < \hat{\theta}_1^d\}$ 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

$$\begin{aligned} &\kappa_1^i(b - \theta_0) + \kappa_2^i(a - \theta_0 - \alpha_1[1 + \kappa_1^i](b - \theta_0)) \\ &\qquad > \kappa_1^d(a - \theta_0) + \kappa_2^d(b - \theta_0 - \alpha_1[1 + \kappa_1^d](a - \theta_0)) \end{aligned} \tag{A.2}$$

Letting $\tilde{a} = (a - \theta_0)$ and $\tilde{b} = (b - \theta_0)$, Condition A.2 reduces to

$$\kappa_1^i \tilde{b} + \kappa_2^i(\tilde{a} - \alpha_1[1 + \kappa_1^i]\tilde{b}) > \kappa_1^d \tilde{a} + \kappa_2^d(\tilde{b} - \alpha_1[1 + \kappa_1^d]\tilde{a}). \tag{A.3}$$

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.3 reduces as follows, which is useful for checking the various cases: $\hat{\theta}_2^i > \hat{\theta}_2^d$ if and only if

$$\kappa_2^i\big(1 + \alpha_1[1 + \kappa_1^i]\big)(\tilde{a} - \tilde{b}\big) - (\kappa_2^d - \kappa_2^i)\big(\tilde{b} - \alpha_1[1 + \kappa_1^d]\tilde{a}\big) > \kappa_1^i(\tilde{a} - \tilde{b}). \tag{A.4}$$

The remainder of the proof walks through all relevant cases before imposing convexity; this analysis will be useful for Corollary 1 where we relax convexity.

<u>*Case 1:*</u> $\theta_0 < b < a$. 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 period 2. This implies $\kappa_2^i = \kappa_2^d = \kappa^G$. Hence, Condition A.4 amounts to $\kappa^G\big(1 + \alpha_1[1 + \kappa^G]\big)(\tilde{a} - \tilde{b}\big) > \kappa^G(\tilde{a} - \tilde{b})$, which is true given $\tilde{a} > \tilde{b}$.

Case 1.b. Suppose both a and b come as losses if received in period 2. This implies $\kappa_2^i = \kappa_2^d = \kappa^L$. Hence, Condition A.4 amounts to $\kappa^L\big(1+\alpha_1[1+\kappa^G]\big)(\tilde{a}-\tilde{b}\big) > \kappa^G(\tilde{a}-\tilde{b})$, which is true given $\tilde{a} > \tilde{b}$ and $\kappa^L > \kappa^G$.

Case 1.c. Suppose only a comes a gain if received in period 2. This implies $\kappa_2^i = \kappa^G$ and $\kappa_2^d = \kappa^L$. Hence, Condition A.4 amounts to $\kappa^G\big(1+\alpha_1[1+\kappa^G]\big)(\tilde{a}-\tilde{b}\big)-(\kappa^L-\kappa^G)\big(\tilde{b}-\alpha_1[1+\kappa^G]\tilde{a}\big) > \kappa^G(\tilde{a}-\tilde{b})$, which reduces to $\kappa^G\big(\alpha_1[1+\kappa^G]\big)(\tilde{a}-\tilde{b}\big)-(\kappa^L-\kappa^G)\big(\tilde{b}-\alpha_1[1+\kappa^G]\tilde{a}\big) > 0$. Since $\hat{b}_2^d$ comes as a loss, it must be that $\tilde{b}-\alpha_1[1+\kappa^G]\tilde{a} < 0$, meaning the previous condition holds.

<u>Case 2:</u> $b < a < \theta_0$. 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 period 2. This implies $\kappa_2^i = \kappa_2^d = \kappa^L$. Hence, Condition A.4 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 period 2. This implies $\kappa_2^i = \kappa_2^d = \kappa^G$. Hence, Condition A.4 amounts to $\kappa^G\big(1+\alpha_1[1+\kappa^L]\big)(\tilde{a}-\tilde{b}\big) > \kappa^L(\tilde{a}-\tilde{b})$, which holds if and only if $\kappa^G\big(1+\alpha_1[1+\kappa^L]\big) > \kappa^L$. Using the definitions of κ^G and κ^L and simplifying reveals that this condition fails only if $\lambda-1 > \alpha_1(1+\eta\lambda)$. Furthermore, for both a and b to come as gains in period 2 implies that $\alpha_1(1+\kappa^L) > 1$. Thus, there exist values of (a,b) meeting the conditions of subcase 2.b for which recency fails only if $\lambda-1 > \alpha_1(1+\eta\lambda)$ and $\alpha_1(1+\kappa^L) > 1$.

Case 2.c. Suppose only a comes as a gain if received in period 2. This implies $\kappa_2^i = \kappa^G$ and $\kappa_2^d = \kappa^L$. Hence, Condition A.4 amounts to $\kappa^G\big(1+\alpha_1[1+\kappa^L]\big)(\tilde{a}-\tilde{b}\big)-(\kappa^L-\kappa^G)\big(\tilde{b}-\alpha_1[1+\kappa^L]\tilde{a}\big) > \kappa^L(\tilde{a}-\tilde{b})$. This condition reduces to

$$\kappa^G\alpha_1(1+\kappa^L)(\tilde{a}-\tilde{b}) > (\kappa^L-\kappa^G)(1-\alpha_1(1+\kappa^L))\tilde{a}. \tag{A.5}$$

The left-hand side of Condition A.5 is always positive, while the right-hand side is positive if and only if $\alpha_1(1+\kappa^L) > 1$ (i.e., beliefs are not convex). Thus, Condition A.5 always holds if $\alpha_1(1+\kappa^L) < 0$, but may fail otherwise. To see when it fails, notice that A.5 fails when

$$\tilde{b} > \frac{1}{\kappa^G}\left(\kappa^L - \frac{\kappa^L-\kappa^G}{\alpha_1(1+\kappa^L)}\right). \tag{A.6}$$

Since this subcase assumes that $\hat{b}_2^d$ comes as a loss, it must be that $\tilde{b} < \alpha_1(1+\kappa^L)$. Hence, for there to exist a value $\tilde{b} < 0$ that falls into case 2.c and satisfies Condition A.6, we require

$$\frac{1}{\kappa^G}\left(\kappa^L - \frac{\kappa^L-\kappa^G}{\alpha_1(1+\kappa^L)}\right) > \alpha_1(1+\kappa^L). \tag{A.7}$$

Using the definitions of κ^G and κ^L, Condition A.7 is equivalent to $\lambda-1 > \alpha_1(1+\eta\lambda)$. In summary, there exist values of (a,b) meeting the conditions of subcase 2.c for which the increasing-sequence bias fails to hold (i.e., $\hat{\theta}_2^d > \hat{\theta}_2^i$) only if $\lambda-1 > \alpha_1(1+\eta\lambda)$ and $\alpha_1(1+\kappa^L) > 1$. Assuming these conditions hold, values (a, b) fail to generate an increasing-sequence bias only if

$$\frac{1}{\kappa^G}\left[\kappa^L - \frac{\kappa^L - \kappa^G}{\alpha_1(1+\kappa^L)}\right] a < b < \alpha_1(1+\kappa^L)a. \tag{A.8}$$

<u>*Case 3:*</u> $b < \theta_0 < a$. 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.3 amounts to $\kappa^L\tilde{b} + \kappa^G(\tilde{a} - \alpha_1[1+\kappa^L]\tilde{b}) > \kappa^G\tilde{a} + \kappa^L(\tilde{b} - \alpha_1[1+\kappa^G]\tilde{a}) \Leftrightarrow -\alpha_1\kappa^G[1+\kappa^L]\tilde{b} > -\alpha_1\kappa^L[1+\kappa^G]\tilde{a}$. This condition always holds given $\tilde{a} > 0 > \tilde{b}$.

In summary, the only cases where the increasing-sequence bias might not hold are 2.b and 2.c. As noted, however, $\hat{\theta}_2^i > \hat{\theta}_2^d$ in both of these ambiguous cases whenever $\alpha_1(1+\kappa^L) < 1$, i.e., beliefs are convex. ■

Proof of main proposition. 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 $\mathcal{X}$. Consider sequence $x^T \in S(\mathcal{X})$ and let $\hat{\theta}_T(x^T)$ be the misattributor's estimate following x^T. We say x^T is increasing if $x_i^T < x_{i+1}^T$ for all $i = 1, \ldots, T-1$. Toward a contradiction, suppose x^T is not increasing but $x^T = \arg\max_{\tilde{x}^T \in S(\mathcal{X})} \hat{\theta}_T(\tilde{x}^T)$. Hence, there must exist adjacent x_i^T, x_{i+1}^T such that $x_i^T > x_{i+1}^T$. Fix $\hat{\theta}_{i-1}$ entering round i. From Lemma A.1, permuting x_i^T and x_{i+1}^T would generate a higher estimate $\hat{\theta}_{i+1}$ than if the agent experiences (x_i^T, x_{i+1}^T). Hence, following this permutation, the person has a higher belief entering round $i+2$ than under the original sequence. Again from Lemma A.1, convexity implies that each $\hat{\theta}_{i+1}$ is increasing in $\hat{\theta}_{i-1}$, and hence $\hat{\theta}_T$ must increase in $\hat{\theta}_{i+1}$. Thus permuting x_i^T and x_{i+1}^T increases $\hat{\theta}_T$, implying a contradiction.

Proof of Corollary 1.

Proof. Part 1. Consider the proof of Lemma A.1. 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 $a > \theta_0$.

Part 2. Again consider the proof of Lemma A.1. Recall that $\hat{\theta}_2^d > \hat{\theta}_2^i$ is possible only in cases 2.b and 2.c, and only then when beliefs are not convex. If beliefs are not convex, then $\hat{\theta}_2^d > \hat{\theta}_2^i$ in cases 2.b and 2.c only if $\lambda > 1 + \alpha_1(1+\eta\lambda) \equiv \bar{\lambda}$. And since beliefs are not convex, $1 < \alpha_1(1+\kappa^L) < \alpha_1(1+\eta\lambda)$, implying $\bar{\lambda} > 2$. ■

Proof of Proposition 3.

Proof. Aside from establishing almost-sure convergence to the steady-state, the results of Proposition 3 do not depend on our assumption of normally-distributed outcomes. Thus, where possible, we prove the results in a somewhat more general environment. Specifically, suppose consumption utility in each period t is $x_t = \theta + \sigma z_t$, where each z_t is an i.i.d. realization of a mean-zero, unit-variance random variable Z that has support $\mathbb{R}$ and a continuously differentiable distribution F_Z (and a density denoted by f_Z). Parameters θ and σ thus denote the true mean and standard deviation of outcomes, respectively. When Z is a standard normal random variable, this more-general environment corresponds to our original baseline model.

Part 1. Step one: unique steady-state belief. Let $\Delta(\hat{\theta})$ be as defined in Equation 8, so $\Delta(\hat{\theta}) = \mathbb{E}[\hat{x}_t|\hat{\theta}_{t-1} = \hat{\theta}] - \hat{\theta}$, where $\mathbb{E}$ is with respect to the true distribution distributional parameters, (θ, σ). From Equation 4, we have $\mathbb{E}[\hat{x}_t|\hat{\theta}_{t-1} = \hat{\theta}] = \theta + \kappa^G \Pr(x_t \geq \hat{\theta})(\mathbb{E}[x_t|x_t \geq \hat{\theta}] - \hat{\theta}) + \kappa^L \Pr(x_t < \hat{\theta})(\mathbb{E}[x_t|x_t < \hat{\theta}] - \hat{\theta}) = \theta - k \Pr\left(x_t < \hat{\theta}\right)\left(\hat{\theta} - \mathbb{E}[x_t|x_t < \hat{\theta}]\right)$ where

$$k \equiv \frac{\kappa^L - \kappa^G}{1 + \kappa^G} = \frac{(\lambda - 1)(\eta - \hat{\eta})}{(1 + \eta)(1 + \hat{\eta}\lambda)}. \tag{A.9}$$

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

$$\Delta(\hat{\theta}) = \theta - kH(\hat{\theta}; \theta, \sigma) - \hat{\theta} = 0, \tag{A.10}$$

where

$$H(\hat{\theta}; \theta, \sigma) \equiv \Pr(x_t < \hat{\theta})\left(\hat{\theta} - \mathbb{E}[x_t|x_t < \hat{\theta}]\right). \tag{A.11}$$

Note that $H(\hat{\theta}; \theta, \sigma) = \hat{\theta} F_Z\left(\frac{\hat{\theta} - \theta}{\sigma}\right) - \int_{-\infty}^{\hat{\theta}} x \frac{1}{\sigma} f_Z\left(\frac{x - \theta}{\sigma}\right) dx$ and that $H(\hat{\theta}; \theta, \sigma) > 0$ for all finite $\hat{\theta}$. Furthermore, $H(\hat{\theta}; \theta, \sigma)$ is a strictly increasing function of $\hat{\theta}$:

$$\frac{\partial}{\partial \hat{\theta}} H(\hat{\theta}; \theta, \sigma) = \hat{\theta} \frac{1}{\sigma} f_Z\left(\frac{\hat{\theta} - \theta}{\sigma}\right) + F_Z\left(\frac{\hat{\theta} - \theta}{\sigma}\right) - \hat{\theta} \frac{1}{\sigma} f_Z\left(\frac{\hat{\theta} - \theta}{\sigma}\right) = F_Z\left(\frac{\hat{\theta} - \theta}{\sigma}\right) > 0. \tag{A.12}$$

Hence $\Delta(\hat{\theta})$ is a strictly decreasing function of $\hat{\theta}$ with range $\mathbb{R}$. Since $\hat{\theta}_\infty$ is defined by the solution to $\Delta(\hat{\theta}) = 0$, it therefore exists and is unique.

Step two: almost-sure convergence to the steady-state belief. Let $\hat{\theta}_\infty$ denote the unique steady-state belief that solves Equation A.10. We now show that the sequence of beliefs $\langle \hat{\theta}_t \rangle$ indeed 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 (2019). As those authors note, 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. Roughly put, the behavior of $\hat{\theta}_t$ will therefore be described by the deterministic ordinary differential equation $\hat{\theta}'(t) = \Delta(\hat{\theta}(t))$, where Δ is the average deviation of encoded outcomes assuming the agent holds expectation $\hat{\theta}(t)$ (Equation 8).

For this part—and only this part—of the proof, we return to our baseline model (with normally-distributed outcomes and priors) to apply the method noted above. Specifically, we assume $x_t = \theta + \sigma z_t$ where z_t are realizations of independent standard normal random variables, and the agent begins with a prior $\theta \sim N(\theta_0, \rho^2)$.While convergence obtains more generally whenever the conditions below are met, it is particularly straightforward to verify these conditions for the normal case given our derivation of $\hat{\theta}_t$ for normally-distributed outcomes in Section 3. From Equation 6, the misattributor's beliefs update according to

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \hat{\alpha}_t[x_t - \hat{\theta}_{t-1}], \tag{A.13}$$

where $\hat{\alpha}_t \equiv (1 + \kappa_t)\alpha_t$ and $\alpha_t = \rho^2/(t\rho^2 + \sigma^2)$. In this case, we can appeal to Theorem 5.2.1 in

Kushner and Yin (2003), who provide sufficient conditions for the convergence of dynamic systems that take this form. For our particular setting, the 4 conditions below are sufficient for $\langle\hat{\theta}_t\rangle$ to converge almost surely to the unique value $\hat{\theta}_\infty$ characterized by the solution to $\Delta(\hat{\theta}) = 0$:

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

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 $\langle\beta_t\rangle$ such that $\mathbb{E}[\hat{x}_t - \hat{\theta}_{t-1}|\hat{\theta}_{t-1}] = \bar{g}(\hat{\theta}_t) + \beta_t$ and $\sum_{t=1}^{\infty} \hat{\alpha}_t|\beta_t| < \infty$ w.p. 1.

We now show that Conditions A1 – A4 hold:

Condition A1. Note that

$$\sum_{t=1}^{\infty} \hat{\alpha}_t = \sum_{t=1}^{\infty} (1+\kappa_t)\alpha_t \geq (1+\kappa^G) \sum_{t=1}^{\infty} \alpha_t = (1+\kappa^G) \sum_{t=1}^{\infty} \frac{\rho^2}{t\rho^2 + \sigma^2}. \tag{A.14}$$

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

$$\sum_{t=1}^{\infty} (\hat{\alpha}_t)^2 = \sum_{t=1}^{\infty} (1+\kappa_t)^2 \alpha_t^2 \leq (1+\kappa^L)^2 \sum_{t=1}^{\infty} \alpha_t^2. \tag{A.15}$$

From the definition of α_t, $\sum_{t=1}^{\infty} (\alpha_t)^2 < \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty$. 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 rational estimate of θ following $t-1$ rounds, we have

$$\sup_t \mathbb{E}[|\hat{x}_t - \hat{\theta}_{t-1}|^2|\theta] \leq (1+\kappa^L) \sup_t \mathbb{E}[|(x_t - \theta_{t-1}) + (\theta_{t-1} - \hat{\theta}_{t-1})|^2|\theta]. \tag{A.16}$$

From Minkowski's Inequality,

$$\sqrt{\mathbb{E}[|(x_t - \theta_{t-1}) + (\theta_{t-1} - \hat{\theta}_{t-1})|^2|\theta]} \leq \sqrt{\mathbb{E}[|x_t - \theta_{t-1}|^2|\theta]} + \sqrt{\mathbb{E}[|\theta_{t-1} - \hat{\theta}_{t-1}|^2|\theta]}. \tag{A.17}$$

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.17. Using Lemma 1, we can write

$$\theta_{t-1} - \hat{\theta}_{t-1} = \alpha_{t-1} \sum_{k=1}^{t-1} x_k - \alpha_{t-1} \sum_{k=1}^{t-1} \xi_k^{t-1} x_k = \alpha_{t-1} \sum_{k=1}^{t-1} \left(1 - \xi_k^{t-1}\right) x_k, \tag{A.18}$$

where ξ_k^{t-1}, defined in Lemma 1, are functions of κ_j and α_j for $j \in \{k, \ldots, t-1\}$. Thus

$$\sqrt{\mathbb{E}[|\theta_{t-1} - \hat{\theta}_{t-1}|^2|\theta]} \leq \alpha_{t-1} \sum_{k=1}^{t-1} \sqrt{\mathbb{E}[|(1 - \xi_k^{t-1})x_k|^2|\theta]}. \tag{A.19}$$

We now argue that for all $t \geq 2$ and all $k \leq t-1$, the value $|1 - \xi_k^{t-1}|$ is bounded from above by some finite constant Q. Given that $\kappa_t \in \{\kappa^G, \kappa^L\}$ and the definition of α_j, it is clear that such a Q exists for any finite t. Thus, we need only consider the case where $t \to \infty$. In this case, we have

$$\lim_{t \to \infty} \xi_k^{t-1} = (1 + \kappa_k) \lim_{t \to \infty} \prod_{j=k}^{t-2} [1 - \alpha_j \kappa_{j+1}].$$

For sufficiently large j, $|1 - \alpha_j \kappa_{j+1}| < 1$. This means that, fixing k, there exists some $\bar{t} \geq k$ such that $|\xi_k^{t-1}|$ is decreasing in t for $t \geq \bar{t}$. Thus, given that $|1 - \xi_k^{t-1}|$ is bounded by some finite Q,

$$\begin{aligned}
\sqrt{\mathbb{E}[|\theta_{t-1} - \hat{\theta}_{t-1}|^2|\theta]} &\leq Q\alpha_{t-1} \sum_{k=1}^{t-1} \sqrt{\mathbb{E}[|x_k|^2|\theta]} \\
&= Q\alpha_{t-1} \sum_{k=1}^{t-1} \sqrt{\sigma^2 + \theta^2} \\
&= Q \frac{\rho^2}{\rho^2 + \sigma^2/(t-1)} \sqrt{\sigma^2 + \theta^2} \\
&\leq Q\sqrt{\sigma^2 + \theta^2},
\end{aligned} \tag{A.20}$$

where the first equality follows from the fact that $\mathbb{E}[|x_k|^2|\theta] = \text{Var}(x_k) + \mathbb{E}[x_k|\theta]^2$, and the second equality follows from the fact that for all $t \geq 2$, $\alpha_{t-1} = \rho^2/((t-1)\rho^2 + \sigma^2)$. 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 8 and thus $\beta_t = \mathbb{E}[\hat{x}_t - \hat{\theta}_{t-1}|\hat{\theta}_{t-1}] - \Delta(\hat{\theta}_{t-1})$. As such, it is immediate that $\sum_{t=1}^{\infty} \hat{\alpha}_t |\beta_t| < \infty$ w.p. 1, as required. Furthermore, from Equation 8, it is straightforward that $\Delta(\cdot)$ is continuous given that F_Z and f_Z are continuous.

Part 2. Let $\hat{\theta}$ denote the unique solution to Equation A.10. We first establish underestimation (i.e., $\hat{\theta} < \theta$ when $\lambda > 1$). For sake of contradiction, suppose $\lambda > 1$ and $\hat{\theta} \geq \theta$. From Equation A.10, $\hat{\theta}$ solves $\hat{\theta} + kH(\hat{\theta}; \theta, \sigma) = \theta$. Because $H(\hat{\theta}; \theta, \sigma) > 0$, $\hat{\theta} + kH(\hat{\theta}; \theta, \sigma)$ exceeds $\theta \Leftrightarrow k > 0$, which holds $\Leftrightarrow \lambda > 1$, implying a contradiction. We now turn to comparative statics. Since $\hat{\theta}$ satisfies $\Delta(\hat{\theta}) \equiv \theta - kH(\hat{\theta}; \theta, \sigma) - \hat{\theta} = 0$, the implicit function theorem implies that, for any parameter of interest $w \in \{\sigma, \eta, \hat{\eta}, \lambda\}$, we have $\frac{\partial \hat{\theta}}{\partial w} = -\left(\frac{\partial \Delta(\hat{\theta})}{\partial \hat{\theta}}\right)^{-1} \frac{\partial \Delta(\hat{\theta})}{\partial w}$. Since $\frac{\partial \Delta(\hat{\theta})}{\partial \hat{\theta}} = -k\frac{\partial}{\partial \hat{\theta}} H(\hat{\theta}; \theta, \sigma) - 1 < 0$, $\frac{\partial \hat{\theta}}{\partial w}$ has the same sign as $\frac{\partial \Delta(\hat{\theta})}{\partial w}$. To show that variance has a decreasing effect on $\hat{\theta}$, suppose $\lambda > 1$ so

$\hat{\theta} < \theta$. Then $\frac{\partial \Delta(\hat{\theta})}{\partial \sigma} = -k\frac{\partial H(\hat{\theta};\theta,\sigma)}{\partial \sigma}$, and thus

$$\begin{aligned}
\frac{\partial \Delta(\hat{\theta})}{\partial \sigma} &= -k\frac{\partial}{\partial \sigma}\left(\hat{\theta}F_Z\left(\frac{\hat{\theta}-\theta}{\sigma}\right) - \int_{-\infty}^{\frac{\hat{\theta}-\theta}{\sigma}} [\theta + \sigma u] f_Z\left(u\right) du\right) \\
&= k\left(\left(\frac{\hat{\theta}-\theta}{\sigma^2}\right)\hat{\theta} f_Z\left(\frac{\hat{\theta}-\theta}{\sigma}\right) - \left(\frac{\hat{\theta}-\theta}{\sigma^2}\right)[\theta + \sigma u] f_Z\left(u\right)\bigg|_{u=\frac{\hat{\theta}-\theta}{\sigma}} + \int_{-\infty}^{\frac{\hat{\theta}-\theta}{\sigma}} u f_Z\left(u\right) du\right) \\
&= k\int_{-\infty}^{\frac{\hat{\theta}-\theta}{\sigma}} u f_Z\left(u\right) du \\
&< 0,
\end{aligned}$$

where the second line follows from Leibniz's Rule and the final inequality holds because the value of the preceding integral is negative given $\hat{\theta} < \theta$.

Part 3. For a parameter $w \in \{\eta, \hat{\eta}, \lambda\}$, the analysis above shows that $\frac{\partial \hat{\theta}}{\partial w}$ has the same sign as $\frac{\partial \Delta(\hat{\theta})}{\partial w}$. Given that $\Delta(\hat{\theta}) = \theta - kH(\hat{\theta};\theta,\sigma) - \hat{\theta}$, it thus follows from the definition of k (Equation A.9) that $\hat{\theta}$ is decreasing in λ and η, and increasing in $\hat{\eta}$.

■

A note on the variance of encoded outcomes in the steady-state. This note shows that the variance of encoded outcomes in the steady-state exceeds the true variance in outcomes. As noted above, in the steady state encoded outcomes are given by $\hat{x}_t = x_t + \kappa_t(x_t - \hat{\theta}_\infty)$, where $\hat{\theta}_\infty$ is the solution to Equation A.10. Hence, the steady-state variance of encoded outcomes is $\text{Var}(\hat{x}_t) = \text{Var}(x_t) + \text{Var}(\kappa_t(x_t - \hat{\theta}_\infty)) + 2\text{Cov}(x_t, \kappa_t(x_t - \hat{\theta}_\infty))$. Note that $\text{Cov}(x_t, \kappa_t(x_t - \hat{\theta}_\infty)) = \mathbb{E}[x_t\kappa_t(x_t - \hat{\theta}_\infty)] - \theta\mathbb{E}[\kappa_t(x_t - \hat{\theta}_\infty)]$, where

$$\begin{aligned}
\mathbb{E}[\kappa_t(x_t - \hat{\theta}_\infty)] &= [1 - F_Z(\hat{\theta}_\infty)]\kappa^G\mathbb{E}[x_t - \hat{\theta}_\infty | x_t \geq \hat{\theta}_\infty] + F_Z(\hat{\theta}_\infty)\kappa^L\mathbb{E}[x_t - \hat{\theta}_\infty | x_t < \hat{\theta}_\infty] \\
&= \kappa^G(\theta - \hat{\theta}_\infty) + F_Z(\hat{\theta}_\infty)(\kappa^L - \kappa^G)\mathbb{E}[x_t - \hat{\theta}_\infty | x_t < \hat{\theta}_\infty], \qquad \text{(A.21)}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[x_t\kappa_t(x_t - \hat{\theta}_\infty)] &= [1 - F_Z(\hat{\theta}_\infty)]\kappa^G\mathbb{E}[x_t(x_t - \hat{\theta}_\infty) | x_t \geq \hat{\theta}_\infty] + F_Z(\hat{\theta}_\infty)\kappa^L\mathbb{E}[x_t(x_t - \hat{\theta}_\infty) | x_t < \hat{\theta}_\infty] \\
&= \kappa^G\mathbb{E}[x_t(x_t - \hat{\theta}_\infty)] + F_Z(\hat{\theta}_\infty)(\kappa^L - \kappa^G)\mathbb{E}[x_t(x_t - \hat{\theta}_\infty) | x_t < \hat{\theta}_\infty] \\
&= \kappa^G(\sigma^2 + \theta^2 - \theta\hat{\theta}_\infty) + F_Z(\hat{\theta}_\infty)(\kappa^L - \kappa^G)\mathbb{E}[x_t(x_t - \hat{\theta}_\infty) | x_t < \hat{\theta}_\infty], \qquad \text{(A.22)}
\end{aligned}$$

and the last line follows from the fact that $\sigma^2 = \text{Var}(x_t)$ and $\theta = \mathbb{E}[x_t]$. Hence,

$$\begin{aligned}
\text{Cov}(x_t, \kappa_t(x_t - \hat{\theta}_\infty)) &= \kappa^G\sigma^2 + F_Z(\hat{\theta}_\infty)(\kappa^L - \kappa^G)\mathbb{E}[x_t(x_t - \hat{\theta}_\infty) - \theta(x_t - \hat{\theta}_\infty) | x_t < \hat{\theta}_\infty] \\
&= \kappa^G\sigma^2 + F_Z(\hat{\theta}_\infty)(\kappa^L - \kappa^G)\mathbb{E}[(x_t - \theta)(x_t - \hat{\theta}_\infty) | x_t < \hat{\theta}_\infty]. \qquad \text{(A.23)}
\end{aligned}$$

Since $\hat{\theta}_\infty < \theta$ (Proposition 3, Part 2), $x_t < \hat{\theta}_\infty$ implies $x_t < \theta$, meaning the expectation in Equation A.23 is always positive. Thus, $\text{Var}(\hat{x}_t) = \text{Var}(x_t) + \text{Var}(\kappa_t(x_t - \hat{\theta}_\infty)) + 2\text{Cov}(x_t, \kappa_t(x_t - \hat{\theta}_\infty)) > \text{Var}(x_t)$.

Proof of Proposition 4.

Proof. The results of this proposition do not require our assumption of normally-distributed outcomes. Thus, we prove the results for the more general environment considered in the proof of Proposition 3. Specifically, suppose consumption utility in each period t is $x_t = \theta + \sigma z_t$, where each z_t is an i.i.d. realization of a mean-zero, unit-variance random variable Z that has support $\mathbb{R}$ and a continuously differentiable distribution F_Z (and a density denoted by f_Z).

Part 1. We first provide an expression for $v(\tilde{\theta}, \sigma)$, which assumes the agent believes outcomes are distributed according to parameters $(\tilde{\theta}, \sigma)$. Let $F(\cdot|\tilde{\theta})$ denote the CDF of x given these parameters (that is, $F(x|\tilde{\theta}) = F_Z(\frac{x-\tilde{\theta}}{\sigma})$), and let $\mathbb{E}_{\tilde{\theta}}$ denote expectations with respect to $F(\cdot|\tilde{\theta})$. Then $v(\tilde{\theta}, \sigma) = \mathbb{E}_{\tilde{\theta}}[u(x|\tilde{\theta})]$ and thus:

$$\begin{aligned} v(\tilde{\theta}, \sigma) &= \mathbb{E}_{\tilde{\theta}}[x] + \eta[1 - F(\tilde{\theta}|\tilde{\theta})] \left(\mathbb{E}_{\tilde{\theta}}[x|x \geq \tilde{\theta}] - \tilde{\theta}\right) + \eta\lambda F(\tilde{\theta}|\tilde{\theta}) \left(\mathbb{E}_{\tilde{\theta}}[x|x < \tilde{\theta}] - \tilde{\theta}\right) \\ &= \tilde{\theta} - \eta(\lambda - 1)F(\tilde{\theta}|\tilde{\theta}) \left(\tilde{\theta} - \mathbb{E}_{\tilde{\theta}}[x|x < \tilde{\theta}]\right) \\ &= \tilde{\theta} - \eta(\lambda - 1)H(\tilde{\theta}; \tilde{\theta}, \sigma), \qquad \text{(A.24)} \end{aligned}$$

where H is defined as in Equation A.11. Furthermore,

$$H(\tilde{\theta}; \tilde{\theta}, \sigma) = \tilde{\theta} F_Z(0) - \int_{-\infty}^{0} [\tilde{\theta} + \sigma u] f_Z(u) du = \sigma|\bar{z}^-|, \qquad \text{(A.25)}$$

where $\bar{z}^- \equiv \int_{-\infty}^{0} u f_Z(u) du < 0$ is a value determined entirely by the distribution of Z and hence independent of the parameters of interest. Thus

$$v(\tilde{\theta}, \sigma) = \tilde{\theta} - \eta(\lambda - 1)\sigma|\bar{z}^-|, \qquad \text{(A.26)}$$

so $v(\theta, \sigma) - v(\hat{\theta}_\infty, \sigma) = \theta - \hat{\theta}_\infty$. From Proposition 3 (Part 2), if $\lambda > 1$, then $\theta - \hat{\theta}_\infty > 0$ and $\hat{\theta}_\infty$ is strictly decreasing in σ. Thus $v(\theta, \sigma) - v(\hat{\theta}_\infty, \sigma)$ is positive and strictly increasing in σ.

Part 2. Let $\mathcal{P}(w)$ denote the set of parameter values (θ, σ) such that $v(\theta, \sigma) = w$. Fixing $(\theta, \sigma) \in \mathcal{P}(w)$, let $\hat{\theta}_\infty(\theta, \sigma)$ denote the steady-state perception of θ written explicitly as a function fo the true parameters. Thus, $v(\theta, \sigma) = w$ is the per-period expected utility under full (correct) information, and $v(\hat{\theta}_\infty(\theta, \sigma), \sigma)$ is that under the misattributor's long-run beliefs. We show that, constrained to $(\theta, \sigma) \in \mathcal{P}(w)$, $\lim_{\sigma \to \infty} v(\hat{\theta}_\infty(\theta, \sigma), \sigma) = -\infty$.

First, we show that for any prospect with $(\theta, \sigma) \in \mathcal{P}(w)$, the steady-state perceived mean $\hat{\theta}_\infty(\theta, \sigma)$ of that prospect is a linearly decreasing function of σ. Recall that $\hat{\theta}_\infty(\theta, \sigma)$ solves $\hat{\theta} - \theta + kH(\hat{\theta}; \theta, \sigma) = 0$. Since $H(\hat{\theta}; \theta, \sigma) = \hat{\theta} F_Z\left(\frac{\hat{\theta}-\theta}{\sigma}\right) - \int_{-\infty}^{\hat{\theta}} x \frac{1}{\sigma} f_Z\left(\frac{x-\theta}{\sigma}\right) dx$, we can define $\hat{z} \equiv (\hat{\theta} - \theta)/\sigma$ and rewrite $H(\hat{\theta}; \theta, \sigma)$ as

$$H(\hat{\theta}; \theta, \sigma) = \hat{\theta} F_Z(\hat{z}) - \int_{-\infty}^{\hat{z}} [\theta + \sigma z] f_Z(z)\, dz = \sigma \left(\hat{z} F_Z(\hat{z}) - \int_{-\infty}^{\hat{z}} z f_Z(z)\, dz\right). \qquad \text{(A.27)}$$

Hence, the steady-state value $\hat{\theta}_\infty(\theta, \sigma)$ is characterized by the value $\hat{z}$ that solves

$$\hat{z} + k\left(\hat{z}F_Z\left(\hat{z}\right) - \int_{-\infty}^{\hat{z}} zf_Z\left(z\right)dz\right) = 0. \tag{A.28}$$

Since $\hat{\theta}_\infty(\theta, \sigma)$ is unique and finite for all finite values (θ, σ), there exists a unique, finite $\hat{z}$ that solves Equation A.28. Denote this value by z^*. Clearly z^* depends solely on F_Z, f_Z, and k, and is thus a constant independent of θ and σ. As such, since $z^* = (\hat{\theta}_\infty(\theta, \sigma) - \theta)/\sigma$, it follows that $\hat{\theta}_\infty(\theta, \sigma) = \theta + z^*\sigma$. Furthermore, the fact that $\hat{\theta}_\infty(\theta, \sigma) < \theta$ implies that $z^* < 0$. Thus $\hat{\theta} = \theta - |z^*|\sigma$.

Now consider a prospect with $(\theta, \sigma) \in \mathcal{P}(w)$. From Equations A.24 and A.25, we have $v(\theta, \sigma) = \theta - \eta(\lambda - 1)\sigma|\bar{z}^-|$ and $v(\hat{\theta}_\infty(\theta, \sigma), \sigma) = \hat{\theta}_\infty(\theta, \sigma) - \eta(\lambda - 1)\sigma|\bar{z}^-|$, where $\bar{z}^- \equiv \int_{-\infty}^{0} uf_Z(u)du < 0$. Substituting the linear specification of $\hat{\theta}_\infty(\theta, \sigma)$ into the expression for $v(\hat{\theta}_\infty(\theta, \sigma), \sigma)$ yields

$$v(\hat{\theta}_\infty(\theta, \sigma), \sigma) = \theta - |z^*|\sigma - \eta(\lambda - 1)\sigma|\bar{z}^-| = w - |z^*|\sigma. \tag{A.29}$$

Thus $v(\hat{\theta}_\infty(\theta, \sigma), \sigma)$ diverges to $-\infty$ as $\sigma \to \infty$ along the locus of parameter values defining $\mathcal{P}(w)$. Consider any arbitrary value $\hat{w} < w$ and define $\bar{\sigma}(w, \hat{w}) \equiv (w - \hat{w})/|z^*|$. For any parameter combination $(\theta, \sigma) \in \mathcal{P}(w)$ with $\sigma > \bar{\sigma}(w, \hat{w})$, we have $v(\hat{\theta}_\infty(\theta, \sigma), \sigma) < w - |z^*|\bar{\sigma}(w, \hat{w}) = \hat{w}$.

■

Proof of Proposition 5.

Proof. We begin by deriving the wage and effort strategies of the principal and worker, respectively, for the case where the principal suffers misattribution and the worker is aware of this error. Since we assume the principal is ignorant of her bias and presumes common knowledge of rationality, she believes the worker follows the Bayesian Nash Equilibrium (BNE) strategy that he would play when facing a rational principal. Accordingly, let $\hat{e}_t$ and $\hat{\theta}_{t-1}$ denote the principal's expectation of the worker's effort and ability at the start of each round t. The principal best responds to these beliefs by offering a wage in round t equal to her expectation of output that round; that is, $w_t = \hat{\theta}_{t-1} + \hat{e}_t$. Additionally, recall that the worker knows (1) that the principal follows a misspecified updating rule due to misattribution, and (2) that the principal wrongly presumes that the worker follows the standard BNE strategy. The worker thus best responds to these erroneous beliefs held by the principal.

We now analyze Holmström's (1999) model with these particular augmentations. Unlike elsewhere in this paper, we allow for temporal discounting in order to match Holmström (1999) as closely as possible. Hence, the worker's effort in each round maximizes the sum of his discounted expected utility given discount factor $\delta \in (0, 1]$. In period t, he faces an objective function

$$U_t \equiv \sum_{\tau=t}^{T} \delta^{\tau-1}[w_\tau - c(e_\tau)] = \sum_{\tau=t}^{T} \delta^{\tau-1}[\hat{\theta}_{\tau-1} + \hat{e}_\tau - c(e_\tau)]. \tag{A.30}$$

To derive the optimal effort profile, we can isolate the part of U_t that depends on e_t. Since the principal's beliefs follow the dynamics outlined in Section 3, Lemma 1 implies $\hat{\theta}_\tau = \alpha_\tau \sum_{k=1}^{\tau} \xi_k^\tau (x_k-$

$\hat{e}_k$), where $\xi_\tau^\tau = (1+\kappa^G)$ and $\xi_k^\tau = (1+\kappa^G)\prod_{j=k}^{\tau-1}\left[1-\kappa^G\alpha_j\right]$ for all $k < t$, which yields the objective

$$-c(e_t) + \sum_{\tau=t+1}^{T} \delta^{\tau-t}\alpha_{\tau-1}\xi_t^{\tau-1}e_t = -c(e_t) + e_t(1+\kappa^G)\left\{\delta\alpha_t + \sum_{\tau=t+2}^{T}\delta^{\tau-t}\alpha_{\tau-1}\left(\prod_{j=t}^{\tau-2}[1-\kappa^G\alpha_j]\right)\right\}.$$

Letting $c_e(\cdot)$ denote the first derivative of $c(\cdot)$, the optimal effort in period t is thus $e_t^* \equiv c_e^{-1}(M_t)$, where

$$M_t \equiv (1+\kappa^G)\left\{\delta\alpha_t + \sum_{\tau=t+2}^{T}\delta^{\tau-t}\alpha_{\tau-1}\left(\prod_{j=t}^{\tau-2}[1-\kappa^G\alpha_j]\right)\right\}. \tag{A.31}$$

Given this derivation of the optimal effort path, we can now compare e_t^* to the effort provided when the principal is fully rational (and this is common knowledge). In the rational case, the worker's optimal effort in period t is $e_t^r \equiv c_e^{-1}(M_t^r)$, where $M_t^r \equiv \sum_{\tau=t+1}^{T}\delta^{\tau-t}\alpha_{\tau-1}$. Because $c(\cdot)$ is strictly convex, it follows that $e_t^* < e_t^r \Leftrightarrow M_t < M_t^r$. Notice that we can write M_t in terms of M_t^r as follows:

$$M_t = M_t^r + \kappa^G\delta\alpha_t - \sum_{\tau=t+2}^{T}\delta^{\tau-t}\alpha_{\tau-1}\left(1-(1+\kappa^G)\prod_{j=t}^{\tau-2}[1-\kappa^G\alpha_j]\right). \tag{A.32}$$

It's clear that $M_t > M_t^r$ for $t = T-1$. Furthermore, since $\prod_{j=t}^{\tau-2}[1-\kappa^G\alpha_j]$ decreases to 0 as $\tau-2-t$ grows large, a sufficiently large T implies there exists a period $t^* < T$ such that $M_t < M_t^r$ for $t < t^*$. If T is not sufficiently large, then $t^* = 1$. More formally, let $D_t \equiv M_t - M_t^r$. Equation A.32 implies

$$\begin{aligned} D_t = \kappa^G\delta\alpha_t - \delta^2\alpha_{t+1}[1-(1+\kappa^G)(1-\kappa^G\alpha_t)] \\ - \sum_{\tau=t+3}^{T}\delta^{\tau-t}\alpha_{\tau-1}\left(1-(1+\kappa^G)[1-\kappa^G\alpha_t]\prod_{j=t+1}^{\tau-2}[1-\kappa^G\alpha_j]\right) \end{aligned} \tag{A.33}$$

Hence $D_t = \delta(D_{t+1} + \kappa^G\alpha_t[1-M_{t+1}])$. Thus, $D_t > 0$ implies $D_{t+1} > 0$ so long as $D_{t+1} > -\kappa^G\alpha_t[1-M_{t+1}] = -\kappa^G\alpha_t[1-D_{t+1}-M_t^r]$, which is equivalent to $[1-\kappa^G\alpha_t]D_{t+1} > -\kappa^G\alpha_t[1-M_{t+1}^r]$. By convexity of beliefs, $[1-\kappa^G\alpha_t] > 0$, so the preceding inequality holds so long as M_t^r is sufficiently small. Since M_t^r decreases in t to a value strictly less than one, there exists a $\tilde{t}$ such that $M_t^r < 1$ for $t \geq \tilde{t}$. Thus, D_{t+1} remains positive for t sufficiently large.

■

Proof of Corollary 2.

Proof. First, the existence of $\bar{T}$ follows from the proof of Proposition 5, where we establish that there exists a value t^* such that $M_t < M_t^r$ whenever $t < t^*$ and that $t^* > 1$ when T is sufficiently large. To verify the second claim, suppose $T > \bar{T}$ and let $M_1(T)$ and $M_1^r(T)$ denote the biased and rational marginal benefit of effort in period 1 as a function of the horizon, T. From Equation A.32, $M_1^r(T+1) - M_1(T+1) > M_1^r(T) - M_1(T)$ if and only if

$$\delta^T\alpha_{T-1}\left[1-(1-\kappa^G)\prod_{j=1}^{T-1}[1-\kappa^G\alpha_j]\right] > 0, \tag{A.34}$$

which holds iff $(1-\kappa^G)\prod_{j=1}^{T-1}[1-\kappa^G\alpha_j] < 1$. If this condition fails, then $M_1(T) > M_1^r(T)$, contradicting $T > \bar{T}$. ■

Proof of Proposition 6.

Proof. Part 1. Note that $\hat{x}_t = x_t + \kappa_t\left(x_t - \widehat{\mathbb{E}}_{t-1}[x_t]\right) = x_t + \kappa_t(x_t - \varphi\hat{x}_{t-1})$. Thus, writing $\hat{x}_t$ recursively in terms of $(x_1,\ldots,x_t)$ yields

$$\begin{aligned}
\hat{x}_t &= (1+\kappa_t)x_t - \varphi\kappa_t\hat{x}_{t-1} \\
&= (1+\kappa_t)x_t - \varphi\kappa_t((1+\kappa_{t-1})x_{t-1} - \varphi\kappa_{t-1}\hat{x}_{t-2}) \\
&= (1+\kappa_t)x_t - \varphi\kappa_t(1+\kappa_{t-1})x_{t-1} + \varphi^2\kappa_t\kappa_{t-1}\hat{x}_{t-2} \\
&= (1+\kappa_t)x_t - \varphi\kappa_t(1+\kappa_{t-1})x_{t-1} + \varphi^2\kappa_t\kappa_{t-1}(1+\kappa_{t-1})x_{t-2} - \varphi^3\kappa_t\kappa_{t-1}\kappa_{t-2}\hat{x}_{t-3} \\
&\ldots \\
&= (1+\kappa_t)x_t + \sum_{j=1}^{t-1}\left((-\varphi)^{t-j}\prod_{i=j+1}^{t}\kappa_i\right)(1+\kappa_j)x_j. \qquad \text{(A.35)}
\end{aligned}$$

Hence, conditional on $(x_1,\ldots,x_{t-1})$, $\text{Var}\left(\widehat{\mathbb{E}}_t[x_{t+1}]\right) = \varphi^2\text{Var}((1+\kappa_t)x_t) > \varphi^2\text{Var}(x_t) = \text{Var}\left(\mathbb{E}_t[x_{t+1}]\right)$, where $\mathbb{E}_t[x_{t+1}]$ denotes the rational expectation.

Part 2. Let $d_t = \hat{x}_t - \widehat{\mathbb{E}}_{t-1}[x_t] = \hat{x}_t - \varphi\hat{x}_{t-1}$. Thus

$$\begin{aligned}
d_t &= (1+\kappa_t)x_t - \varphi\kappa_t\hat{x}_{t-1} - \varphi\hat{x}_{t-1} \\
&= (1+\kappa_t)(x_t - \varphi\hat{x}_{t-1}) \\
&= (1+\kappa_t)(\varphi x_{t-1} + \epsilon_t - \varphi((1+\kappa_{t-1})x_{t-1} - \varphi\kappa_{t-1}\hat{x}_{t-2})) \\
&= (1+\kappa_t)(\epsilon_t - \varphi\kappa_{t-1}(x_{t-1} - \varphi\hat{x}_{t-2})) \\
&= (1+\kappa_t)\left(\epsilon_t - \varphi\frac{\kappa_{t-1}}{1+\kappa_{t-1}}d_{t-1}\right). \qquad \text{(A.36)}
\end{aligned}$$

■

Proof of Proposition 7

Proof. As noted in Footnote 41, the updating rule for beliefs about each θ^ω is similar to the baseline model: letting $N_t^\omega \equiv \sum_{k=1}^t \mathbb{1}\{\omega_k = \omega\}$, the agent's estimate of θ^ω after t rounds is

$$\hat{\theta}_t^\omega \equiv \frac{\rho^2}{N_t^\omega\rho^2 + \sigma^2}\left(\sum_{\{k\le t\,:\,\omega_k=\omega\}}\hat{x}_k\right) + \frac{\sigma^2}{N_t^a\rho^2 + \sigma^2}\theta_0^\omega.$$

Following our assumption in the text, we focus on the case where $\sigma \to 0$. This implies that the optimal action in any given period is also myopically optimal: it is not influenced by a "forward-looking" desire to reduce uncertainty for future rounds by strategically attempting to generate more data about a particular state.[47] Denote the agent's (subjectively) optimal action in round t by p_t^*. Fixing the agent's vector of beliefs entering round t, denoted by $\hat{\theta}_{t-1} \equiv (\hat{\theta}_{t-1}^H, \hat{\theta}_{t-1}^L)$, she chooses p to maximize

$$p\hat{\theta}_{t-1}^H + (1-p)\hat{\theta}_{t-1}^L - p(1-p)\eta(\lambda-1)\left[\hat{\theta}_{t-1}^H - \hat{\theta}_{t-1}^L\right] - c(p-p_0). \tag{A.37}$$

Hence, p_t^* solves

$$\left[\hat{\theta}_{t-1}^H - \hat{\theta}_{t-1}^L\right]\Bigg(2\eta\lambda p_t^* + 1 - \eta\lambda\Bigg) = c'(p_t^* - p_0). \tag{A.38}$$

Our assumptions on the cost function are relevant here: (1) $p_0 = 1/2$ implies that the LHS of Equation A.38 is positive, and hence $c'(0) = 0$ implies that $p_t^* \in (p_0, \bar{p}]$; and (2) since the LHS of Equation A.38 is linearly increasing in p_t, $c'(\cdot)$ increasing and weakly convex generically implies that p_t^* is unique. Furthermore, from Equation A.38, it is clear that p_t^* is independent of t conditional on $\hat{\theta}_{t-1}$, so we can simply write the optimal choice in t as $p^*(\hat{\theta}_{t-1})$.

Convergence of beliefs follows from arguments similar to those used in the proof of Proposition 3 (Part 1), and we make extensive use the apparatus established there. However, there are two differences to account for. First, the agent is updating about two parameters, θ^H and θ^L. Beliefs about each θ^ω have dynamics that meet the required form, $\hat{\theta}_t^\omega = \hat{\theta}_{t-1}^\omega + \hat{\alpha}_t^\omega(\hat{x}_t - \hat{\theta}_{t-1}^\omega)$, where $\hat{\alpha}_t^\omega \equiv \mathbb{1}\{\omega_t = \omega\}\alpha_{N_t^\omega}$, N_t^ω counts the number of rounds in which ω has occurred through period t, and $\alpha_{N_t^\omega} = \rho^2/(N_t^\omega\rho^2 + \sigma^2)$. The weights $(\hat{\alpha}_t^\omega)$ correspond to the rational weights (α_τ) considered in the proof of Proposition 3, expect they only put weight on new observations in rounds in which state ω occurs. (Put differently, we are essentially considering for each $\omega \in \{H, L\}$ the dynamics of the sequence $\langle\tilde{\theta}_\tau^\omega\rangle_{\tau=1}^\infty$ defined by $\tilde{\theta}_\tau^\omega \equiv \hat{\theta}_{j^\omega(\tau)}^\omega$ where $j^\omega(\tau)$ denotes the time period in which the τ^{th} occurrence of ω happens.) Second, encoded outcomes conditional on ω, denoted by $\hat{x}^\omega$, depend on beliefs about both parameters since the agent's expected outcome in round t—and hence her reference point—is $p_t\hat{\theta}_{t-1}^H + (1-p_t)\hat{\theta}_{t-1}^L$, where $p_t = p^*(\hat{\theta}_{t-1})$. In light of these differences with respect to Proposition 3, we show that each dimension of this two-dimensional system of beliefs meets the sufficient conditions for convergence established in Proposition 3 (Part 1), and the limiting values are consequently determined by the solution of a two-dimensional system of equations analogous to the one-dimensional steady-state solution described in the proof of Proposition 3 (Part 1).

We now reverify conditions A1-A4 from Proposition 3 (Part 1) for both dimensions, $\omega \in \{H, L\}$. Conditions A1 and A2 follow immediately from the proof of Proposition 3 since for each ω, the weights $(\hat{\alpha}_t^\omega)$ are defined so that along the subsequence (τ) counting rounds in which ω occurs, $\hat{\alpha}_t^\omega$ matches the rational weight α_τ considered in Proposition 3. More precisely, since $p_t \in [1/2, \bar{p}]$ where $\bar{p} < 1$ and thus $\omega_t = \omega$ infinitely often, we have $\sum_{t=1}^\infty \hat{\alpha}_t^\omega = \sum_{t=1}^\infty \mathbb{1}\{\omega_t = \omega\}\alpha_{N_t^\omega} = \sum_{\tau=1}^\infty \alpha_\tau = \infty$, $\lim_{t\to\infty} \hat{\alpha}_t^\omega = \lim_{\tau\to\infty} \alpha_\tau = 0$, and $\sum_{t=1}^\infty (\hat{\alpha}_t^\omega)^2 = \sum_{\tau=1}^\infty (\alpha_\tau)^2 < \infty$.

Now consider condition A3. Following any action $p_t \in [0,1]$, note that $\hat{x}_t^H = x_t^H + \kappa_t(x_t^H - \bar{\theta}_{t-1}(p_t))$ where $\bar{\theta}_{t-1}(p_t) = p_t\hat{\theta}_{t-1}^H + (1-p_t)\hat{\theta}_{t-1}^L$, meaning $\hat{x}_t^H - \hat{\theta}_{t-1}^H = x_t^H + \kappa_t(x_t^H - \bar{\theta}_{t-1}(p_t)) - \hat{\theta}_{t-1}^H$, and thus

$$\begin{aligned}\hat{x}_t^H - \hat{\theta}_{t-1}^H &= (1+\kappa_t)x_t^H - (1+p_t\kappa_t)\hat{\theta}_{t-1}^H - (1-p_t)\kappa_t\hat{\theta}_{t-1}^L \\ &= (1+\kappa_t)(x_t^H - \theta_{t-1}^H) + (1-p_t)\kappa_t(\theta_{t-1}^H - \theta_{t-1}^L) \\ &\quad + (1+p_t\kappa_t)(\theta_{t-1}^H - \hat{\theta}_{t-1}^H) + (1-p_t)\kappa_t(\theta_{t-1}^L - \hat{\theta}_{t-1}^L),\end{aligned} \tag{A.39}$$

[47] We could alternatively relax the assumption that $\sigma \to 0$ and instead assume the agent is myopic.

where θ^H_{t-1} and θ^L_{t-1} are the rational beliefs entering round t. Following our verification of A3 for Proposition 3, if is clear that $\sup_t \mathbb{E}[|\hat{x}^H_t - \hat{\theta}^H_{t-1}|^2|\theta^H, \theta^L] < \infty$ after an application of Minkowski's Inequality and then noting that the expected squared absolute value of each term in the final expression of Equation A.39 is finite. In particular, $\mathbb{E}[|x^H_t - \theta^H_{t-1}|^2|\theta^H, \theta^L]$ and $\mathbb{E}[|\theta^H_{t-1} - \theta^L_{t-1}|^2|\theta^H, \theta^L]$ are both finite because they concern only the rational Bayesian estimates, and Proposition 3 (Part 1) establishes that $\mathbb{E}[|\theta^\omega_{t-1} - \hat{\theta}^\omega_{t-1}|^2|\theta^H, \theta^L]$ are finite as well. An analogous argument establishes $\sup_t \mathbb{E}[|\hat{x}^L_t - \hat{\theta}^L_t|^2|\theta^H, \theta^L] < \infty$.

Turning to condition A4, since we are considering a two-dimensional process, we must specify a function $\bar{g}^\omega : \mathbb{R}^2 \to \mathbb{R}$ for both $\omega \in \{H, L\}$. As in Proposition 3, we take $\bar{g}^\omega$ to be the expected deviation function $\Delta^\omega : \mathbb{R}^2 \to \mathbb{R}$ analogous to Equation 8. More precisely, along dimension $\omega \in \{H, L\}$, let Δ^ω be the expectation of $\hat{x}^\omega_t - \hat{\theta}^\omega_{t-1}$ conditional on $\hat{\theta}_{t-1}$ and $\omega_t = \omega$:

$$\Delta^\omega(\hat{\theta}_{t-1}) \equiv \mathbb{E}\left[\hat{x}^\omega_t(p^*(\hat{\theta}_{t-1}), \hat{\theta}_{t-1}) \middle| \theta^H, \theta^L\right] - \hat{\theta}^\omega_{t-1}, \tag{A.40}$$

where $\hat{x}^\omega_t(p^*(\hat{\theta}_{t-1}), \hat{\theta}_{t-1})$ denotes the encoded outcome conditional on $\omega_t = \omega$ given mean beliefs $\hat{\theta}_{t-1}$ entering round t and action $p^*(\hat{\theta}_{t-1})$. Thus

$$\begin{aligned}
\Delta^\omega(\hat{\theta}_{t-1}) &= \theta^\omega + \kappa^G(1 - F(\bar{\theta}_{t-1}(p^*(\hat{\theta}_{t-1}))|\theta^\omega))\Big(\mathbb{E}\big[x^\omega_t | x^\omega_t \geq \bar{\theta}_{t-1}(p^*(\hat{\theta}_{t-1}))\big] - \bar{\theta}_{t-1}(p^*(\hat{\theta}_{t-1}))\Big) \\
&+\kappa^L F(\bar{\theta}_{t-1}(p^*(\hat{\theta}_{t-1}))|\theta^\omega)\Big(\mathbb{E}\big[x^\omega_t | x^\omega_t < \bar{\theta}_{t-1}(p^*(\hat{\theta}_{t-1}))\big] - \bar{\theta}_{t-1}(p^*(\hat{\theta}_{t-1}))\Big) - \hat{\theta}^\omega_{t-1} \\
&= \theta^\omega + \kappa^G\Big(\theta^\omega - \bar{\theta}_{t-1}(p^*(\hat{\theta}_{t-1}))\Big) \\
&+ (\kappa^L - \kappa^G)F(\bar{\theta}_{t-1}(p^*(\hat{\theta}_{t-1}))|\theta^\omega)\Big(\mathbb{E}\big[x^\omega_t | x^\omega_t < \bar{\theta}_{t-1}(p^*(\hat{\theta}_{t-1}))\big] - \bar{\theta}_{t-1}(p^*(\hat{\theta}_{t-1}))\Big) - \hat{\theta}^\omega_{t-1},
\end{aligned} \tag{A.41}$$

where $F(\cdot|\theta^\omega)$ denotes the CDF of x^ω. It is clear from Equation A.38 that $p^*(\hat{\theta}_{t-1})$ is continuous in $\hat{\theta}^\omega_{t-1}$, and $\bar{\theta}_{t-1}(p^*)$ is continuous in p^*. Hence, $\Delta^\omega(\hat{\theta}_{t-1})$ is continuous in $\hat{\theta}^\omega_{t-1}$ for each $\omega \in \{H, L\}$, as required. Thus A4 holds as in Proposition 3.

The limiting beliefs $\hat{\theta}_\infty = (\hat{\theta}^H_\infty, \hat{\theta}^L_\infty)$ hence satisfy

$$\begin{bmatrix} \Delta^H(\hat{\theta}^H_\infty, \hat{\theta}^L_\infty) \\ \Delta^L(\hat{\theta}^H_\infty, \hat{\theta}^L_\infty) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Furthermore, given our assumption of $\sigma \to 0$, Δ^H and Δ^L defined by Equation A.41 reduce in this case to

$$\Delta^H(\hat{\theta}^H_\infty, \hat{\theta}^L_\infty) = \theta^H + \kappa^G(\theta^H - \bar{\theta}(p^*(\hat{\theta}_\infty))) - \hat{\theta}^H_\infty \tag{A.42}$$

$$\Delta^L(\hat{\theta}^H_\infty, \hat{\theta}^L_\infty) = \theta^L + \kappa^L(\theta^L - \bar{\theta}(p^*(\hat{\theta}_\infty))) - \hat{\theta}^L_\infty. \tag{A.43}$$

Fixing p^*, the solution to this system is given by Equation 15 in the main text. At this solution, $\hat{\theta}^H_\infty - \hat{\theta}^L_\infty > \theta^H - \theta^L$ regardless of p^*. Thus, Equation A.38 implies that the long-run action under misattribution exceeds the full-information optimal action whenever the full-information action is interior. Otherwise, the misattributor settles on the highest possible action, which coincides with the full-information action ($p^* = \bar{p}$).

■

B Misattribution with Multiple Dimensions

In this section, we extend our model of misattribution 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 KR's 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 $\hat{u}(\hat{x}|\hat{\theta}) = u(x|\hat{\theta})$. 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 (e.g., Haggag and Pope 2018), we believe this formulation provides a tractable stepping stone for empiricists.

C Misattribution with Stochastic Reference Points

In this section, we consider an alternative definition of the reference point. In particular, we consider the stochastic reference point formalized by Kőszegi and Rabin (2006). We first describe how to extend our framework to accommodate stochastic reference points, and then we present some results on an agent's long-run steady-state beliefs under misattribution for this case. These steady-state beliefs share similar features with those characterized in Proposition 3 from the main text: a loss-averse misattributor underestimates the mean of the prospect and overestimates its variance. Furthermore, she underestimates the mean in proportion to the true variance in outcomes, which again implies a greater bias against riskier prospects.

Suppose the agent believes that outcomes are distributed according to $\widehat{F}$. In KR's model, the sense of gain or loss derives from comparing outcome x with each counterfactual outcome that was possible under $\widehat{F}$. Thus, gain-loss utility is no longer given by Equation 1 and instead depends on the entire distribution of outcomes:

$$n(x|\widehat{F}) = \int_{\tilde{x}<x} (x - \tilde{x}) d\widehat{F}(\tilde{x}) + \lambda \int_{\tilde{x} \geq x} (x - \tilde{x}) d\widehat{F}(\tilde{x}). \tag{C.1}$$

Hence, outcome x is compared against each hypothetical outcome and this comparison is weighted by the probability of that hypothetical outcome.[48] Given this alternative reference point, an encoded outcome $\hat{x}$ is defined exactly as in the main text (Equation 3) except the gain-loss utility term is replaced by $n(\cdot|\widehat{F})$ defined in Equation C.1.[49]

Many of our baseline steady-state results presented in Section 4 extend when the reference point is stochastic (Equation C.1), albeit with a loss of tractability. To see why stochastic reference points complicate the analysis, note that in our baseline model in the main text, the agent's perceived distribution of outcomes in the steady state is specified entirely by her perceived mean outcome. This is because her reference point depends solely on that single moment of the outcome distribution. With stochastic reference points, however, gain-loss utility depends on the entire perceived distribution. As such, solving for the steady-state perceived distribution with stochastic reference points in general would entail finding a fixed point in the space of distributions—the steady-state distribution is such that, if believed, the person's encoded outcomes follow that distribution.

We leave a full treatment of this case to future work and focus here on the case where (i) outcomes are normally distributed such that $x_t \sim N(\theta, \sigma^2)$ and (ii) the agent fits θ and σ to the long-run distribution of encoded outcomes. That is, we analyze stead-state perceptions, denoted by $\hat{\theta}$ and $\hat{\sigma}$, such that if the agent believes the mean and variance of outcomes are $\hat{\theta}$ and $\hat{\sigma}^2$, respectively, then the distribution of encoded outcomes indeed has a mean and variance equal to $\hat{\theta}$ and $\hat{\sigma}^2$, respectively. Steady-state perceptions $\hat{\theta}$ and $\hat{\sigma}$ are thus characterized by the solution to the following system of equations:

$$\begin{aligned} \hat{\theta} &= \mathbb{E}[\hat{x}|\hat{\theta}, \hat{\sigma}] \quad &(C.2) \\ \hat{\sigma}^2 &= \text{Var}(\hat{x}|\hat{\theta}, \hat{\sigma}), \quad &(C.3) \end{aligned}$$

where $\mathbb{E}[\cdot|\hat{\theta}, \hat{\sigma}]$ and Var$(\cdot|\hat{\theta}, \hat{\sigma})$ are with respect to the true parameter values governing the distribution of x, (θ, σ), and are conditional on the person believing these parameters have values equal to $\hat{\theta}$ and $\hat{\sigma}$, respectively. Unlike our long-run analysis in the main text, the analysis here allows the agent to adjust her perception of the variance according to the encoded data. With stochastic reference points, the mean encoded outcome depends directly on the agent's perception of the variance. This interdependence does not arise with the form of reference point considered in the main text and was thus irrelevant there.

The remainder of this section analyzes the steady-state perceptions implied by Equations C.2 and C.3. In particular, we highlight that a misattributor overestimates σ (i.e., $\hat{\sigma} > \sigma$) and underestimates θ in proportion to $\hat{\sigma}$.[50]

[48] If either $\lambda = 1$ or the reference distribution $\widehat{F}$ is degenerate, then this stochastic reference-point model reduces to our baseline model (Equation 1).

[49] Since $n(x|\widehat{F})$ in Equation C.1 is strictly increasing in x conditional on $\widehat{F}$, the encoded outcome $\hat{x}$ is still well defined and unique.

[50] The latter result is similar to the comparative statics highlighted in the main text in Proposition 3. While many of our punchlines from Section 4 extend with stochastic reference points, our result that a misattributor's average experienced utility exceeds the average utility she would experience under correct beliefs (see Footnote 28) does not necessarily

As a first step, the following lemma specifies the gain-loss utility function (Equation C.1) for the case of normally-distributed outcomes. Given that the agent presumes outcomes are normally distributed, the perceived distribution $\widehat{F}$ referenced in the definition of gain-loss utility in Equation C.1 is entirely determined by the agent's perception of the mean and variance. As such, we simplify notation by writing n directly in terms of these perceptions: let $n(x|\hat{\theta},\hat{\sigma}) \equiv \int_{\tilde{x}<x}(x-\tilde{x})f(\tilde{x}|\hat{\theta},\hat{\sigma})d\tilde{x} + \lambda \int_{\tilde{x}\geq x}(x-\tilde{x})f(\tilde{x}|\hat{\theta},\hat{\sigma})d\tilde{x}$, where $f(\cdot|\hat{\theta},\hat{\sigma})$ is the PDF of a normally-distributed random variable with mean and standard deviation equal to $\hat{\theta}$ and $\hat{\sigma}$, respectively. Additionally, let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard-normal CDF and PDF, respectively.

Lemma C.1. *Suppose that the person believes $x \sim N(\hat{\theta},\hat{\sigma}^2)$. With a stochastic reference point,*

$$n(x|\hat{\theta},\hat{\sigma}) = \hat{\sigma}\lambda z - \hat{\sigma}(\lambda-1)\left[z\Phi(z)+\phi(z)\right],$$

where $z=(x-\hat{\theta})/\hat{\sigma}$.

Proof. From Equation C.1,

$$\begin{aligned} n(x|\hat{\theta},\hat{\sigma}) &= \int_{-\infty}^{x}(x-\tilde{x})f(\tilde{x}|\hat{\theta},\hat{\sigma})d\tilde{x} + \lambda\int_{x}^{\infty}(x-\tilde{x})f(\tilde{x}|\hat{\theta},\hat{\sigma})d\tilde{x} & \text{(C.4)}\\ &= \int_{-\infty}^{x}\left(\frac{x-\tilde{x}}{\hat{\sigma}}\right)\phi\left(\frac{\tilde{x}-\hat{\theta}}{\hat{\sigma}}\right)d\tilde{x} + \lambda\int_{x}^{\infty}\left(\frac{x-\tilde{x}}{\hat{\sigma}}\right)\phi\left(\frac{\tilde{x}-\hat{\theta}}{\hat{\sigma}}\right)d\tilde{x} & \text{(C.5)}\\ &= \hat{\sigma}\left(\int_{-\infty}^{z}(z-\tilde{z})\,\phi(\tilde{z})\,d\tilde{z} + \lambda\int_{z}^{\infty}(z-\tilde{z})\,\phi(\tilde{z})\,d\tilde{z}\right), & \text{(C.6)}\end{aligned}$$

where $z=\frac{x-\hat{\theta}}{\hat{\sigma}}$ and $\tilde{z}=\frac{\tilde{x}-\hat{\theta}}{\hat{\sigma}}$. Thus

$$\begin{aligned} n(x|\hat{\theta},\hat{\sigma}) &= \hat{\sigma}\left(\int_{-\infty}^{\infty}(z-\tilde{z})\,\phi(\tilde{z})\,d\tilde{z} + (\lambda-1)\int_{z}^{\infty}(z-\tilde{z})\,\phi(\tilde{z})\,d\tilde{z}\right) & \text{(C.7)}\\ &= \hat{\sigma}\left(z+(\lambda-1)z[1-\Phi(z)]-(\lambda-1)\int_{z}^{\infty}\tilde{z}\phi(\tilde{z})\,d\tilde{z}\right). & \text{(C.8)}\end{aligned}$$

Note that $\int_{z}^{\infty}\tilde{z}\phi(\tilde{z})\,d\tilde{z} = [1-\Phi(z)]\mathbb{E}[\tilde{Z}|\tilde{Z}>z]$ where $\tilde{Z}$ is a standard normal random variable. Hence $n(x|\hat{\theta},\hat{\sigma}) = \hat{\sigma}\left(z+(\lambda-1)\left(z[1-\Phi(z)]-\phi(z)\right)\right) = \hat{\sigma}\left(\lambda z-(\lambda-1)(z\Phi(z)+\phi(z))\right)$. ■

We now derive the first-moment condition, Equation C.2. To simplify matters, we focus on the case where $\hat{\eta}=0$. Hence, $\hat{x}=x+\eta n(x|\hat{\theta},\hat{\sigma})$, where $n(\cdot|\hat{\theta},\hat{\sigma})$ is given in Lemma C.1. Let $z\equiv(x-\hat{\theta})/\hat{\sigma}$ and $\bar{z}\equiv(\theta-\hat{\theta})/\hat{\sigma}$. Conditional on the agent believing in parameter values $\hat{\theta}$ and $\hat{\sigma}$, the expectation of $\hat{x}$ with respect to the true distribution, which has density $\frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)$, is

$$\mathbb{E}[x+\eta n(x|\hat{\theta},\hat{\sigma})|\hat{\theta},\hat{\sigma}] = \theta+\hat{\sigma}\eta\left(\lambda\bar{z}-(\lambda-1)\left[\int z\Phi(z)\frac{1}{\sigma}\phi((x-\theta)/\sigma)dx+\int\phi(z)\frac{1}{\sigma}\phi((x-\theta)/\sigma)dx\right]\right).$$

extend. With stochastic reference points, the agent's realized utility depends explicitly on her perceived variance. Fixing the true variance in outcomes, an agent facing normally-distributed outcomes experiences a lower utility on average when she anticipates greater variance.

Let $w = \frac{x-\theta}{\sigma}$, which implies $z = a + bw$ where $a = \bar{z} = \frac{\theta-\hat{\theta}}{\hat{\sigma}}$ and $b = \frac{\sigma}{\hat{\sigma}}$. Hence,

$$\mathbb{E}[x+\eta n(x|\hat{\theta},\hat{\sigma})|\hat{\theta},\hat{\sigma}] = \theta+\hat{\sigma}\eta\Bigg(\lambda a-(\lambda-1)\Bigg[\int (a+bw)\,\Phi(a+bw)\phi(w)dw+\int \phi(a+bw)\phi(w)dw\Bigg]\Bigg).$$

Thus,

$$\mathbb{E}[x+\eta n(x|\hat{\theta},\hat{\sigma})|\hat{\theta},\hat{\sigma}] = \theta+\hat{\sigma}\eta\Bigg(\lambda a-(\lambda-1)\big[aI_1+bI_2+I_3\big]\Bigg), \tag{C.9}$$

where

$$\begin{aligned} I_1 &\equiv \int \Phi(a+bw)\phi(w)dw = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right), && \text{(C.10)}\\ I_2 &\equiv \int w\Phi(a+bw)\phi(w)dw = \frac{b}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right), && \text{(C.11)}\\ I_3 &\equiv \int \phi(a+bw)\phi(w)dw = \frac{1}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right). && \text{(C.12)}\end{aligned}$$

Hence, the first equation of the steady-state system, Equation C.2, amounts to

$$0 = a+\eta\left\{\lambda a-(\lambda-1)\left[a\Phi\left(\frac{a}{\sqrt{1+b^2}}\right)+\sqrt{1+b^2}\phi\left(\frac{a}{\sqrt{1+b^2}}\right)\right]\right\}. \tag{C.13}$$

We now turn to the second-moment equation of the steady-state system, Equation C.3. Note that $\text{Var}(\hat{x}|\hat{\theta},\hat{\sigma}) = \mathbb{E}[\hat{x}^2|\hat{\theta},\hat{\sigma}] - \mathbb{E}[\hat{x}|\hat{\theta},\hat{\sigma}]^2$, where $\mathbb{E}[\hat{x}^2|\hat{\theta},\hat{\sigma}] = \sigma^2+\theta^2+2\eta\mathbb{E}[xn(x|\hat{\theta},\hat{\sigma})|\hat{\theta},\hat{\sigma}] + \eta^2\mathbb{E}[n(x|\hat{\theta},\hat{\sigma})^2|\hat{\theta},\hat{\sigma}]$. Since $\mathbb{E}[\hat{x}|\hat{\theta},\hat{\sigma}]$ is already derived above, we next derive $\mathbb{E}[n(x|\hat{\theta},\hat{\sigma})^2|\hat{\theta},\hat{\sigma}]$ and $\mathbb{E}[xn(x|\hat{\theta},\hat{\sigma})|\hat{\theta},\hat{\sigma}]$ in turn. From Lemma C.1,

$$\begin{aligned} n(x|\hat{\theta},\hat{\sigma})^2 = \hat{\sigma}^2\big\{&\lambda^2z^2-[2\lambda(\lambda-1)]\left(z^2\Phi(z)+z\phi(z)\right)\\ &+(\lambda-1)^2(z^2\Phi(z)^2+2z\Phi(z)\phi(z)+\phi(z)^2)\big\}. \end{aligned} \tag{C.14}$$

We must take the expectation of each these terms with respect to the true distribution. To do so, we first rewrite $\mathbb{E}[n(x|\hat{\theta},\hat{\sigma})^2|\hat{\theta},\hat{\sigma}]$ in terms of several Gaussian integrals, and then evaluate those integrals.

$$\begin{aligned} \mathbb{E}[n(x|\hat{\theta},\hat{\sigma})^2|\hat{\theta},\hat{\sigma}] = \hat{\sigma}^2\big\{&\lambda^2(a^2+b^2I_4)-[2\lambda(\lambda-1)]\left(a^2I_1+2abI_2+b^2I_5+aI_3+bI_6\right)\\ &+(\lambda-1)^2(a^2I_{10}+2abI_{11}+b^2I_{12}+2(aI_8+bI_9)+I_7)\big\} \end{aligned} \tag{C.15}$$

where

$$\begin{array}{ll} I_1 = \int \Phi(a+bw)\phi(w)dw & I_2 = \int w\Phi(a+bw)\phi(w)dw\\ I_3 = \int \phi(a+bw)\phi(w)dw & I_4 = \int w^2\phi(w)dw\\ I_5 = \int w\phi(a+bw)\phi(w)dw & I_6 = \int w^2\Phi(a+bw)\phi(w)dw\\ I_7 = \int \phi(a+bw)^2\phi(w)dw & I_8 = \int \Phi(a+bw)\phi(a+bw)\phi(w)dw\\ I_9 = \int w\Phi(a+bw)\phi(a+bw)\phi(w)dw & I_{10} = \int \Phi(a+bw)^2\phi(w)dw\\ I_{11} = \int w\Phi(a+bw)^2\phi(w)dw & I_{12} = \int w^2\Phi(a+bw)^2\phi(w)dw. \end{array}$$

We now evaluate each of these integral terms. Note that I_1, I_2, and I_3 are derived above, and

$I_4 = \mathbb{E}[w^2] = \sigma^2 + \mathbb{E}[w]^2 = 1$ since w is standard normal. We now turn to the remaining terms.

I_5: Letting $f(\cdot|m,s)$ denote a generic normal PDF with mean m and standard deviation s, the following identity will be useful:

$$f(w|m_1,s_1)f(w|m_2,s_2) = Sf(w|\bar{m},\bar{s}) \tag{C.16}$$

where

$$\bar{m} = \frac{m_1 s_2 + m_2 s_1}{s_1^2 + s_2^2} \quad \text{and} \quad \bar{s} = \sqrt{\frac{s_1^2 s_2^2}{s_1^2 + s_2^2}},$$

and $S \equiv f\left(m_1 \middle| m_2, \sqrt{s_1^2 + s_2^2}\right)$ is a scaling factor. Using this identity with $\bar{m} = -a/b\left(\frac{1}{b^2}+1\right)$,

$$I_5 = \frac{1}{b}\frac{1}{\sqrt{1+\frac{1}{b^2}}}\phi\left(\frac{-a/b}{\sqrt{1+\frac{1}{b^2}}}\right)\int wf(w|\bar{m},\bar{s})dw = \frac{1}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right)\frac{-ab}{1+b^2}, \tag{C.17}$$

which follows from the fact that ϕ is symmetric: $\phi(-y) = \phi(y)$.

I_6: Using integration by parts,

$$\begin{aligned} I_6 &= \int w^2\Phi(a+bw)\phi(w)dw = \int w\Phi(a+bw)\left[w\phi(w)\right]dw \\ &= -w\Phi(a+bw)\phi(w)\Big|_{-\infty}^{\infty} + \int \phi(w)\left[bw\phi(a+bw) + \Phi(a+bw)\right]dw \\ &= b\int w\phi(a+bw)\phi(w)dw + \int \Phi(a+bw)\phi(w)dw & \text{(C.18)} \\ &= bI_5 + I_1. & \text{(C.19)} \end{aligned}$$

I_7: Using our identity for the product of two normal densities above (Equation C.16), $\phi(z)^2 = Sf(z|\bar{m},\bar{s})$, where $\bar{m} = 0$, $\bar{s} = 1/\sqrt{2}$ and $S = \phi(0)/\sqrt{2}$. Hence, $I_7 = \phi(0)\int \phi(\tilde{a} + \tilde{b}z))\phi(w)dw$, where $\tilde{a} = \sqrt{2}a$ and $\tilde{b} = \sqrt{2}b$. Thus, using the derivation of I_3:

$$I_7 = \frac{\phi(0)}{\sqrt{1+\tilde{b}^2}}\phi\left(\frac{\tilde{a}}{\sqrt{1+\tilde{b}^2}}\right) = \frac{\phi(0)}{\sqrt{1+2b^2}}\phi\left(\frac{\sqrt{2}a}{\sqrt{1+2b^2}}\right). \tag{C.20}$$

$aI_8 + bI_9$: Note that $aI_8 + bI_9 = \int (a+bw)\Phi(a+bw)\phi(a+bw)\phi(w)dw$, which can be simplified using our formula for the product of two normal density functions: $\phi(a+bw)\phi(w) = Sf(w|\bar{m},\bar{s}^2)$ where $f(\cdot|\bar{m},\bar{s}^2)$ is the PDF of a normal random variable with mean $\bar{m} = -ab/(1+b^2)$, standard deviation $\bar{s} = 1/(\sqrt{1+b^2})$, and scaling factor $S = \frac{1}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right)$. Thus

$$aI_8 + bI_9 = S\left((a+b\bar{m})\int \Phi(a+bw)f(w|\bar{m},\bar{s}^2)dw + b\bar{s}\int y\Phi(a+bw)f(w|\bar{m},\bar{s}^2)dw\right),$$

where $y = \frac{w-\bar{m}}{\bar{s}}$. Finally, using the derivation of I_1 (Equation C.10), $\int \Phi(a+bw)f(w|\bar{m},\bar{s}^2)dw = \int \Phi(a'+b'y)\phi(y)dy = \Phi\left(\frac{a'}{\sqrt{1+b'^2}}\right)$, where $a' = a + b\bar{m} = a/(1+b^2)$ and $b' = b\bar{s} = b/\sqrt{1+b^2}$.

Likewise, using the derivation of I_2 (Equation C.11), $\int y\Phi(a'+b'y)\phi(y)dy = \frac{b'}{\sqrt{1+b'^2}}\phi\left(\frac{a'}{\sqrt{1+b'^2}}\right)$. Combining all the terms above yields

$$aI_8 + bI_9 = \frac{1}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right)\left[\frac{a}{1+b^2}\Phi\left(\frac{a}{\sqrt{1+b^2}\sqrt{1+2b^2}}\right)\right.$$
$$\left.+\frac{b^2}{\sqrt{1+b^2}\sqrt{1+2b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}\sqrt{1+2b^2}}\right)\right]. \quad \text{(C.21)}$$

I_{10}: Note that

$$I_{10} = \int \Phi(a+bw)^2\phi(w)dw = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) - 2T\left(\frac{a}{\sqrt{1+b^2}}, \frac{1}{\sqrt{1+2b^2}}\right),$$

where $T(h,q) = \phi(h)\int_0^q \frac{\phi(hx)}{1+x^2}dx$ is Owen's T function.
I_{11}: Note that

$$I_{11} = \int w\Phi(a+bw)^2\phi(w)dw = \frac{2b}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right)\Phi\left(\frac{a}{\sqrt{1+b^2}\sqrt{1+2b^2}}\right).$$

I_{12}: Integration by parts yields

$$\begin{aligned} I_{12} &= \int \left[\Phi(a+bw)^2 + 2bw\Phi(a+bw)\phi(a+bw)\right]\phi(w)dw \\ &= I_{10} + 2b\int w\Phi(a+bw)\phi(a+bw)\phi(w)dw. \end{aligned} \quad \text{(C.22)}$$

The integral $\int w\Phi(a+bw)\phi(a+bw)\phi(w)dw$ was calculated in the derivation of $aI_8 + bI_9$ above. It follows that

$$I_{12} = I_{10} + \frac{2b^2}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right)\left[\frac{-a}{1+b^2}\Phi\left(\frac{a}{\sqrt{1+b^2}\sqrt{1+2b^2}}\right)\right.$$
$$\left.+\frac{1}{\sqrt{1+b^2}\sqrt{1+2b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}\sqrt{1+2b^2}}\right)\right]. \quad \text{(C.23)}$$

Finally, using I_1 through I_{12} derived above, we can write $\mathbb{E}[n(x|\hat{\theta},\hat{\sigma})^2|\hat{\theta},\hat{\sigma}]$ (Equation C.15) as

$\hat{\sigma}^2 N_2(a,b)$, where

$$\begin{aligned} N_2(a,b) &= \lambda^2(a^2+b^2) \\ &- [2\lambda(1-\lambda)]\left[(a^2+b^2)\Phi\left(\frac{a}{\sqrt{1+b^2}}\right)+a\sqrt{1+b^2}\phi\left(\frac{a}{\sqrt{1+b^2}}\right)\right] \\ &+ (\lambda-1)^2\Bigg\{(a^2+b^2)\Phi\left(\frac{a}{\sqrt{1+b^2}}\right)-2(a^2+b^2)T\left(\frac{a}{\sqrt{1+b^2}},\frac{1}{\sqrt{1+2b^2}}\right) \\ &+ 2(1+b^2)S\left[a\Phi\left(\frac{a}{\sqrt{1+b^2}\sqrt{1+2b^2}}\right)+\frac{b^2}{\sqrt{1+b^2}\sqrt{1+2b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}\sqrt{1+2b^2}}\right)\right] \\ &+ \frac{\phi(0)}{\sqrt{1+2b^2}}\phi\left(\frac{\sqrt{2}a}{\sqrt{1+2b^2}}\right)\Bigg\}, \end{aligned} \tag{C.24}$$

and $S=\frac{1}{\sqrt{1+b^2}}\phi\left(\frac{a}{\sqrt{1+b^2}}\right)$.

We now derive $\mathbb{E}[xn(x|\hat{\theta},\hat{\sigma})|\hat{\theta},\hat{\sigma}]$. Using Lemma C.1 and the change of variables introduced above,

$$\begin{aligned} xn(x|\hat{\theta},\hat{\sigma}) &= \hat{\sigma}\lambda(\theta a+[\theta b+\sigma a]w+\sigma bw^2) \\ &\quad -\hat{\sigma}(\lambda-1)(\theta a+[\theta b+\sigma a]w+\sigma bw^2)\Phi(a+bw)-\hat{\sigma}(\lambda-1)(\theta-\sigma w)\phi(a+bw). \end{aligned} \tag{C.25}$$

Taking expectations with respect to w and using the integral identities above yields

$$\mathbb{E}[xn(x|\hat{\theta},\hat{\sigma})|\hat{\theta},\hat{\sigma}]=\hat{\sigma}\left\{\lambda[\theta a+\sigma b]-(\lambda-1)\big(\theta[aI_1+bI_2+I_3]+\sigma bI_1\big)\right\}. \tag{C.26}$$

We can now use the expressions derived above to assess Var$(\hat{x}|\hat{\theta},\hat{\sigma})$. Recall from above that Var$(\hat{x}|\hat{\theta},\hat{\sigma})=\mathbb{E}[\hat{x}^2|\hat{\theta},\hat{\sigma}]-\mathbb{E}[\hat{x}|\hat{\theta},\hat{\sigma}]^2$, where $\mathbb{E}[\hat{x}^2|\hat{\theta},\hat{\sigma}]=\sigma^2+\theta^2+2\eta\mathbb{E}[xn(x|\hat{\theta},\hat{\sigma})|\hat{\theta},\hat{\sigma}]+\eta^2\mathbb{E}[n(x|\hat{\theta},\hat{\sigma})^2|\hat{\theta},\hat{\sigma}]$. From Equation C.9, we have $\mathbb{E}[\hat{x}|\hat{\theta},\hat{\sigma}]=\theta+\hat{\sigma}\eta N_1(a,b)$ where $N_1(a,b)\equiv\lambda a-(\lambda-1)[aI_1+bI_2+I_3]$. Using this expression along with Equation C.24 yields

$$\text{Var}(\hat{x}|\hat{\theta},\hat{\sigma})=\sigma^2+\theta^2+2\eta\mathbb{E}[xn(x|\hat{\theta},\hat{\sigma})|\hat{\theta},\hat{\sigma}]+\eta^2\hat{\sigma}^2N_2(a,b)-(\theta+\hat{\sigma}\eta N_1(a,b))^2, \tag{C.27}$$

Further substituting Equation C.26 into C.27 yields

$$\begin{aligned} \text{Var}(\hat{x}|\hat{\theta},\hat{\sigma}) &= \sigma^2+2\eta\hat{\sigma}(\mathbb{E}[xn(x|\hat{\theta},\hat{\sigma})|\hat{\theta},\hat{\sigma}]-N_1(a,b))+\eta^2\hat{\sigma}^2[N_2(a,b)-N_1(a,b)] \\ &= \sigma^2+2\eta\sigma^2(\lambda-(\lambda-1)I_1)+\eta^2\hat{\sigma}^2[N_2(a,b)-N_1(a,b)] \\ &= \sigma^2+2\eta\sigma^2N_0(a,b)+\eta^2\hat{\sigma}^2[N_2(a,b)-N_1(a,b)], \end{aligned} \tag{C.28}$$

where $N_0(a,b)\equiv\lambda-(\lambda-1)I_1$. Thus, the steady-state condition from Equation C.3, $\hat{\sigma}^2=$ Var$(\hat{x}|\hat{\theta},\hat{\sigma})$, is equivalent to

$$0=b^2+2\eta b^2N_0(a,b)+\eta^2[N_2(a,b)-N_1(a,b)]. \tag{C.29}$$

Finally, $\hat{\theta}$ and $\hat{\sigma}$ are implicitly defined by the values a and b that solve the system given by Equations

C.13 and C.29:

$$\begin{aligned} 0 &= a + \eta N_1(a,b) && \text{(C.30)} \\ 0 &= b^2 + 2\eta b^2 N_0(a,b) + \eta^2[N_2(a,b) - N_1(a,b)]. && \text{(C.31)} \end{aligned}$$

One can show that this system has a unique solution, which we denote by (a^*, b^*). Moreover, both of the equations in the system above depend solely on the PDF and CDF of the normal distribution and parameters η and λ; importantly, they are independent of θ and σ. The solution (a^*, b^*) therefore characterizes $\hat{\theta}$ and $\hat{\sigma}$ as follows: $\hat{\sigma} = \sigma/b^*$, which implies $\hat{\theta} = \theta - a^*\hat{\sigma} = \theta - \frac{a^*}{b^*}\sigma$. Again, since a^* and b^* are independent of σ, the perceived mean is linearly decreasing in the true variance of outcomes.

D Reputation Model with Loss Aversion

This section explores the career-concern model of Section 5.1 when the principal is loss averse ($\lambda > 1$). In this case, gains and losses affect the principal's beliefs asymmetrically. Hence, the worker's strategy must account for how his effort affects the likelihood that future outcomes will be weighted as gains versus losses. Technically speaking, the marginal effect of the worker's current effort on future wages—M_t from Equation A.31—is no longer deterministic when $\lambda > 1$. In particular, the κ_t terms are no longer all equal to κ^G as in Equation A.31, but can be either κ^G or κ^Ldepending on whether the future outcome comes as a gain or loss. Thus, the agent's optimal effort in each round must now take into account how her behavior will influence the distribution over future realizations of $\kappa_t \in \{\kappa^G, \kappa^L\}$. As argued below, this change is relatively minor and does not alter the predicted shape of the worker's effort profile in expectation.

Specifically, we derive the worker's optimal effort policy when $T = 3$, highlighting that (1) it is qualitatively similar to the case without loss aversion ($\lambda = 1$) analyzed in the main text, and (2) the intuition behind Proposition 5 continues to hold. In particular, in the three-period model, e_1 will fall below the rational benchmark, while e_2 exceeds it. Finally, we argue that these qualitative similarities between the $\lambda = 1$ and $\lambda > 1$ cases hold more generally for any $T \geq 3$.

Suppose $T = 3$, and for ease of exposition let $\delta = 1$. Note that $e_3^* = 0$. Thus, in period 1, the worker faces the following objective function: $\Pi_1 \equiv \mathbb{E}_{x_1}[\hat{\theta}_1] - c(e_1) + \mathbb{E}_{(x_1,x_2)}[\hat{\theta}_2 - c(e_2)]$. Conditional on x_2, we can write $\hat{\theta}_2$ as $\hat{\theta}_2 = \hat{\theta}_1 + \alpha_2(1+\kappa_2)[x_2 - \hat{e}_2 - \hat{\theta}_1] = \hat{\theta}_1 + \alpha_2(1+\kappa_2)d_2$, where $d_2 \equiv x_2 - \hat{e}_2 - \hat{\theta}_1$. Hence, the period-1 objective can be written as $\Pi_1 = \mathbb{E}_{(x_1,x_2)}[2\hat{\theta}_1 + \alpha_2(1+\kappa_2)d_2 - c(e_2)] - c(e_1)$. Because e_2 is a function of e_1, optimizing Π_1 with respect to e_1 requires us to first derive how e_2 depends on e_1 at the optimum.

The period-2 objective is $\Pi_2 \equiv \mathbb{E}_{x_2}[\hat{\theta}_2] - c(e_2) = \hat{\theta}_1 + \alpha_2\mathbb{E}_{x_2}[(1+\kappa_2)d_2|x_1] - c(e_2)$. Thus, e_2 must satisfy $c'(e_2) = \alpha_2\frac{\partial}{\partial e_2}\mathbb{E}_{x_2}[(1+\kappa_2)d_2|x_1]$. Note that $d_2 \sim N(\tilde{\theta}_1 + e_2 - \hat{\theta}_1 - \hat{e}_2, \sigma_1^2)$, where $\tilde{\theta}_1$ is the rational estimate of θ following x_1 and σ_1^2 is the variance of $\tilde{\theta}_1 + \epsilon_2$, which is independent of e_1. Let $p_2 \equiv \Pr(d_2 > 0|x_1) = 1 - \Phi\left(\frac{-\bar{d}_2}{\sigma_1}\right)$, where Φ is the standard normal CDF and $\bar{d}_2 = \mathbb{E}[d_2|x_1]$. Thus, $\mathbb{E}_{x_2}[(1+\kappa_2)d_2|x_1] = \bar{d}_2 + p_2\kappa^G\mathbb{E}_{x_2}[d_2|d_2 > 0, x_1] + (1-p_2)\kappa^G\mathbb{E}_{x_2}[-d_2|d_2 < 0, x_1]$. Since $\mathbb{E}_{x_2}[d_2|d_2 > 0] = \bar{d}_2 + \sigma_1\phi\left(\frac{-\bar{d}_2}{\sigma_1}\right)/p_2$ and $\mathbb{E}_{x_2}[d_2|d_2 < 0] = \bar{d}_2 - \sigma_1\phi\left(\frac{-\bar{d}_2}{\sigma_1}\right)/(1-p_2)$, it follows that $\mathbb{E}_{x_2}[(1+\kappa_2)d_2|x_1] = [1+p_2\kappa^G+(1-p_2)\kappa^L]\bar{d}_2 - (\kappa^L-\kappa^G)\sigma_1\phi\left(\frac{-\bar{d}_2}{\sigma_1}\right)$. This implies

$$\begin{aligned}\frac{\partial}{\partial e_2}\mathbb{E}_{x_2}[(1+\kappa_2)d_2|x_1] = -(\kappa^L-\kappa^G)\frac{\partial p_2}{\partial e_2}\bar{d}_2 + [1+p_2\kappa^G+(1-p_2)\kappa^L]\frac{\partial \bar{d}_2}{\partial e_2}\\ -(\kappa^L-\kappa^G)\sigma_1\phi\left(\frac{-\bar{d}_2}{\sigma_1}\right)\left(\frac{\bar{d}_2}{\sigma_1}\right)\left(\frac{-1}{\sigma_1}\right)\frac{\partial \bar{d}_2}{\partial e_2}.\end{aligned} \tag{D.1}$$

From the definition of p_2, $\frac{\partial p_2}{\partial e_2} = \phi\left(\frac{-\bar{d}_2}{\sigma_1}\right)\left(\frac{1}{\sigma_1}\right)\frac{\partial \bar{d}_2}{\partial e_2}$, and since $\frac{\partial \bar{d}_2}{\partial e_2} = 1$, Equation D.1 reduces to $\frac{\partial}{\partial e_2}\mathbb{E}_{x_2}[(1+\kappa_2)d_2|x_1] = [1+p_2\kappa^G+(1-p_2)\kappa^L]$. Hence, the first-order condition for e_2 amounts to

$$c'(e_2) = \alpha_2[1+p_2\kappa^G+(1-p_2)\kappa^L]. \tag{D.2}$$

Returning to the optimal effort choice in period 1, note that e_1 must satisfy first-order condition

$$\mathbb{E}_{x_1}\left[2\frac{\partial\hat{\theta}_1}{\partial e_1} + \frac{\partial}{\partial e_1}\mathbb{E}_{x_2}[\alpha_2(1+\kappa_2)d_2 - c(e_2)|x_1]\right] = c'(e_1). \tag{D.3}$$

Analogous to the derivation of Equation D.1, $\frac{\partial}{\partial e_1}\mathbb{E}_{x_2}[(1+\kappa_2)d_2|x_1] = [1+p_2\kappa^G+(1-p_2)\kappa^L]\frac{\partial \bar{d}_2}{\partial e_1}$, where $\frac{\partial \bar{d}_2}{\partial e_1} = \frac{\partial e_2}{\partial e_1} - \frac{\partial \hat{\theta}_1}{\partial e_1}$ given that $\bar{d}_2 = \tilde{\theta}_1 + e_2 - \hat{\theta}_1 - \hat{e}_2$. Hence, the first-order condition for e_1 amounts to

$$\mathbb{E}_{x_1}\left[2\frac{\partial\hat{\theta}_1}{\partial e_1} + \alpha_2[1+p_2\kappa^G+(1-p_2)\kappa^L]\left[\frac{\partial e_2}{\partial e_1} - \frac{\partial\hat{\theta}_1}{\partial e_1}\right] - c'(e_2)\frac{\partial e_2}{\partial e_1}\right] = c'(e_1). \tag{D.4}$$

Since the worker's choice of e_2 conditional on x_1 must satisfy Equation D.2, Equation D.4 yields

$$\mathbb{E}_{x_1}\left[2\frac{\partial\hat{\theta}_1}{\partial e_1} - \alpha_2[1+p_2\kappa^G+(1-p_2)\kappa^L]\frac{\partial\hat{\theta}_1}{\partial e_1}\right] = c'(e_1). \tag{D.5}$$

Since $\hat{\theta}_1 = \alpha_1(1+\kappa_1)d_1$ where $d_1 = x_1 - \hat{e}_1$, $\frac{\partial\hat{\theta}_1}{\partial e_1} = \alpha_1(1+\kappa_1)$. Thus, the first-oder condition for e_1 reduces further to $c'(e_1) = \alpha_1\mathbb{E}_{x_1}\left[(1+\kappa_1)(2-\alpha_2-\alpha_2[p_2\kappa^G+(1-p_2)\kappa^L])\right] = \mathbb{E}_{x_1}\left[(1+\kappa_1)(\alpha_1+\alpha_2-\alpha_1\alpha_2[p_2\kappa^G+(1-p_2)\kappa^L])\right]$, where the second equality follows from the fact that $\alpha_1(1-\alpha_2) = \alpha_2$. Note that $[p_2\kappa^G+(1-p_2)\kappa^L] = \mathbb{E}[\kappa_2|x_1]$ given the optimal policy. Hence, the first-order condition for e_1 can be written as $c'(e_1) = \mathbb{E}_{(x_1,x_2)}[(1+\kappa_1)(\alpha_1+\alpha_2[1-\kappa_2\alpha_1])]$. This first-order condition is equivalent to the one with $\lambda = 1$ (see Equation A.31) aside from the expected value over κ_t on the right-hand side: with $\lambda = 1$, each $\kappa_t = \kappa^G$ deterministically.

Given the similarity between the solutions with $\lambda = 1$ and $\lambda > 1$, the predictions of Proposition 5 continue to hold with $\lambda > 1$. We can see this directly in the analysis above. The first-order condition for e_2 requires that $c'(e_2) = \alpha_2[1+p_2\kappa^G+(1-p_2)\kappa^L]$. Effort absent misattribution, however, solves $c'(e_2^r) = \alpha_2$. Since $\alpha_2[1+p_2\kappa^G+(1-p_2)\kappa^L] > \alpha_2$, second-round effort under misattribution exceeds the rational benchmark. Contrastingly, first-round effort may fall short: effort absent misattribution solves $c'(e_1^r) = \alpha_1+\alpha_2$, while effort under misattribution solves $c'(e_1) = \mathbb{E}_{(x_1,x_2)}[(1+\kappa_1)(\alpha_1+\alpha_2[1-\kappa_2\alpha_1])]$.

One could continue this backward induction argument for arbitrary T and show that, in general, $e_t = c_e^{-1}(\mathbb{E}[M_t^l|e_t, h_{t-1}])$, where $M_t^l \equiv (1+\kappa_t)\left\{\delta\alpha_t + \sum_{\tau=t+2}^{T} \delta^{\tau-t}\alpha_{\tau-1}\left(\prod_{j=t}^{\tau-2}[1-\kappa_{j+1}\alpha_j]\right)\right\}$ and $\mathbb{E}[M_t^l|e_t, h_{t-1}]$ is the expected value of M_t^l conditional on the history, the worker's current choice, and his policy going forward. Recall that when $\lambda = 1$, $e_t = c_e^{-1}(M_t)$, where $M_t = (1+\kappa^G)\left\{\delta\alpha_t + \sum_{\tau=t+2}^{T} \delta^{\tau-t}\alpha_{\tau-1}\left(\prod_{j=t}^{\tau-2}[1-\kappa^G\alpha_j]\right)\right\}$ (see Equation A.31). Hence, the key difference between the solution with $\lambda > 1$ and the one with $\lambda = 1$ is that the expected values of κ_t in M_t^l will fall in the interval $[\kappa^G, \kappa^L]$ rather than remain constant at κ^G deterministically. This change does not alter the qualitative pattern of the effort profile in expectation.