Channeled Attention and Stable Errors

Tristan Gagnon-Bartsch
University of Iowa

Matthew Rabin
Harvard University

Joshua Schwartzstein*
HBS

August 3, 2026

Abstract

People persist in mistaken beliefs in settings where standard explanations—high stakes, abundant data, opportunities to learn—predict that errors should be corrected. We develop a framework centered around channeled attention that explains why. When a person channels his attention through the lens of his potentially misspecified model, exposure to data is not the same as tracking the statistics necessary to tell him his model is wrong. A model is attentionally stable when the data the person deems relevant fail, in the long run, to make the model look implausible. Models that erroneously rule things out—e.g., beliefs that a variable doesn't matter or that a distinction doesn't exist—tend to be attentionally stable because they divert attention away from data that would refute the model. Conversely, models that erroneously rule things *in* are less likely to be stable. Whether an error is discovered depends on the structure of the decision problem but is not tightly linked to its stakes; arbitrarily costly errors can persist while trivially inexpensive ones may be detected. Nevertheless, no error is robust both to changes in the decision environment and to shocks that expand attention: every channeled-attention error is vulnerable to some perturbation that would surface it. We use the framework to analyze the stability of psychological biases and empirical misconceptions, as well as to explain why fresh eyes catch mistakes, why people use overly coarse rather than overly fine categorizations, and why people recognize errors in others that they miss in themselves.

*E-mails: tgagnonbartsch@uiowa.edu, matthewrabin@fas.harvard.edu, and jschwartzstein@hbs.edu. We thank Chang Liu for excellent research assistance. For helpful comments, we are grateful to Nava Ashraf, Francesca Bastianello, Max Bazerman, Katherine Coffman, Christine Exley, Erik Eyster, Drew Fudenberg, Nicola Gennaioli, Brian Hall, David Hirshleifer, Botond Kőszegi, Spencer Kwon, Giacomo Lanzani, George Loewenstein, Michael Luca, Kathleen McGinn, Sendhil Mullainathan, Andrei Shleifer, Adi Sunderam, Dmitry Taubinsky and seminar participants at BEAM, Berkeley, Boston College, Brandeis, CEU, Cornell, Dartmouth, ESSET, Harvard, LSE, McGill, Melbourne, Princeton, Stanford, University of New South Wales, University of Technology Sydney, and Yale. Gagnon-Bartsch and Rabin thank the Centre for Experimental Research on Fairness, Inequality and Rationality (FAIR) at the Norwegian School of Economics (NHH), for financial support and hospitality while conducting research on this paper.

"My experience is what I agree to attend to. Only those items which I notice shape my mind."
— William James

1 Introduction

When do errors persist? Economists usually offer one of two answers: the mistake is too small to be worth fixing, or the data needed to fix it are missing. But people make persistent mistakes even when neither holds—in settings that are high-stakes, data-rich, and revisited often. Stakes are high in financial markets, yet professional investors sell in ways that underperform selling at random (Akepanidtaworn et al., 2023). Data are abundant for firms setting prices, yet pricing strategies persistently ignore predictable patterns of demand (DellaVigna and Gentzkow, 2019; Strulov-Shlain, 2022; List et al., 2023). And the same mistakes recur year after year in agriculture and in household budgeting: farmers keep mis-sizing an input the data in front of them would correct (Hanna et al., 2014), and households in developing countries face predictable "hungry seasons" they have the opportunity to budget around (Augenblick et al., 2026). People in these settings are not merely exposed to corrective data—they act on it daily, live with its consequences, and have every reason to learn from it. Yet the errors persist. Why don't people wake up to their errors? What could jolt them awake?

This paper formalizes a simple idea. When people have a way of thinking about a problem—a model—they attend to the data the model treats as relevant and largely ignore the rest; they are not on the lookout for evidence that would contradict it. This means that exposure is not enough. A person can observe and act on the very data that would reveal his error, yet never track the statistics that would tell him his model is wrong. As William James puts it in the opening quote, what shapes the mind is not the information one is exposed to but the information one has *attended to* through the lens of his model.

We develop a framework for assessing when a person fails to notice that his model of the world is misspecified—for when his model is *attentionally stable*, in the sense that the data he deems worth consulting never, in the long run, make the model look implausible relative to an alternative. Errors that rule things out—a variable that supposedly doesn't matter, a distinction that supposedly doesn't exist, an effect that supposedly can't be large—tend to be attentionally stable, because they lead the person to neglect data that would expose them. Errors that rule things *in* are less likely to be attentionally stable. A person who actively entertains a richer alternative, even a wrong one, gathers data to evaluate it that incidentally undercuts the model he started with. Moreover, whether an error is caught turns on the structure of the decision problem, but not tightly on its stakes.

Our framework builds on the rational inattention literature pioneered by Sims (2003), which analyzes agents who optimally weigh the expected costs and benefits of attention. But building on

Schwartzstein (2014), our model of *subjectively* rational inattention departs from this literature in two interrelated ways. First, we assume that an agent's perceived benefits from attention—and how he interprets what he notices—are based on his potentially misspecified model. Second, we study a different question: when will repeated experience lead a person to reject a false model?

Section 2 introduces our formal framework. We study agents who are Bayesian, but with misspecified priors π over the structure of the world. Our approach covers a broad array of context-specific empirical misconceptions (e.g., investor misperceptions as in Barberis et al., 1998) and psychological biases (e.g., naivete about self-control problems as in O'Donoghue and Rabin, 1999). We assume that an agent will employ a "sufficient attentional strategy" (or "SAS"), where he coarsens available information in a way that he *perceives* as sufficient for maximizing long-run payoffs. Our main emphasis is on "minimal" SASs, in which the agent attends to nothing beyond what he finds useful. Theories both help select which variables to attend to and determine which distinctions count as meaningful variables in the first place.[1]

Section 3 then introduces a criterion for evaluating when someone will discover his mistaken model. A model π is attentionally unstable relative to an alternative λ if, in the long run, the data the person notices become overwhelmingly more likely under that alternative than under his original model. Notably, it is not the absolute unlikeliness of outcomes within the person's model that causes him to abandon it. A person doesn't abandon the view that a coin is fair after 1,000 flips just from noticing that any one of the $2^{1,000}$ sequences is highly unlikely; if he notices 1,000 heads in a row, however, he will be drawn to the obvious alternative. In most of our analysis we take λ to be the correct model, but we allow for a more general λ to highlight the role of alternative explanations in our framework. We describe an erroneous model π as attentionally stable in a context if there exists a SAS where the noticed data do not eventually reveal the model to be wrong relative to the true model. To isolate the role of attention, we assume the person has access to rich enough data to discover his mistake under full attention. Channeled attention then determines whether he ever asks the questions of the data that would reveal the mistake.

To illustrate, consider a simple example of an investor who overreacts to advice from an unreliable source (e.g., from people on web forums and social media). In each period t, the investor encounters a new investment opportunity, where he decides to invest at cost $c > 0$ or pass at cost 0. He learns the opportunity's outcome whether or not he invests (he can watch an asset he passed on): it is either successful ($\omega_t = 1$), yielding gross return $r > c$, or not ($\omega_t = 0$), yielding gross return 0. Given this structure, the investor decides to invest in period t if and only if he assigns a probability greater than c/r to the success of the investment.

[1] Neither believers nor disbelievers of astrology based on the positions of the moon, the planets, and Pluto have investigated the validity of the infinity of potential alternatives to astrology. If following the investment advice of tall brokers would bring you riches because your vaporological sign (determined by the shape of clouds on your 13th birthday) is "kinda puffy", then you wouldn't know it.

To potentially aid in this assessment, the investor freely receives advice from a source he thinks is reliable and integrates this signal with his correct belief that half of investment opportunities are successful, $\Pr(\omega_t = 1) = 1/2$. Absent advice from this source, investing is unattractive, $r/2 < c$, but may become attractive depending on the advice. This source provides a binary investment recommendation, $s_t \in \{0,1\}$, with precision $\theta = \Pr(s_t = \omega_t | \omega_t)$. The investor wrongly believes that his investment decisions should follow the recommendations: his prior π over θ places probability one on $\theta > c/r$, when in reality $\theta = \theta^* < c/r$. Will the investor wake up to his misspecification? He certainly possesses rich enough data to discover that the advice performs far worse than he thought possible. Indeed, he acts on the advice every period and observes investment success. However, under a minimal sufficient attentional strategy, he only sees a reason to extract the advice's *current* decision value, not *statistics* on long-run accuracy. The investor need not wake up.

This example illustrates five main points. First, and most basically, *exposure to rich data is not enough*: this failure to wake up holds despite having and acting on data rich enough to do so. It instead arises because the investor does not see a need to ask questions of these data that would reveal his model's misspecification.[2] Under his model, the long-run success rate of recommended investments must exceed c/r; in reality it converges to $\theta^* < c/r$. Thus, a simple running tally of "advice said to invest $\rightarrow$ did it pay off?" would refute his model.

Second, *what the investor misses is not a specific outcome but rather the statistical relationship* between the recommendation and investment success. We call such a relationship a *statistical gorilla*, echoing the famous study by Simons and Chabris (1999) in which subjects are tasked with counting the number of passes in a film clip of basketball players and often fail to notice a faux gorilla walking across the court. People in this classic study don't see the "gorilla" in front of them because they're channeling their attention toward counting passes. Likewise, the investor fails to wake up in this example because he misses a striking mismatch over time between the actual distribution of the outcome and its expected distribution. In every period the advice says to invest, he loses money in expectation. But no individual failure surprises him, since his model expects failures at rate $1-\theta$. Only *the frequency* of failures is impossible under his model, and the frequency is exactly what a minimal SAS gives him no reason to track.

Third, the logic of the investor's failure to notice the statistical gorilla is *independent of the stakes* of resulting mistakes (e.g., by scaling r and c by a common factor $\psi > 0$). If anything, the investor expends *more* effort than needed by attending to and following advice that in reality shouldn't influence his investment decision.

Fourth, *incidental learning is context dependent.* Modifications to the environment that provide the investor with an incentive to track statistics regarding advice quality (asking not only whether

[2]More broadly, our results are driven by the agent's perceived (lack of) benefits of attention rather than the costs. In the language of Handel and Schwartzstein (2018), foregoing useful information because it is assessed as too costly can be seen as "frictions", whereas not imagining the value of the information can be seen as "mental gaps".

the advice is accurate enough to follow but also how accurate it is) make him more likely to notice the statistical gorilla and incidentally learn his model is wrong. To illustrate, modify the example slightly so the investor has to pay an effort cost to access the advice ahead of his investment decision (e.g., reflecting how often he checks web forums or social media), but can access it after at zero cost. To fix ideas, suppose he can access the advice ahead of an investment decision with probability α by paying an effort cost $\chi(\alpha) = k\alpha^2$ for $k \geq 0$. Now the investor believes he should track the fraction of times the advice is accurate when he is uncertain how much effort to exert, which is sufficient to alert him to advice being less accurate than he thought possible.[3]

Fifth, *easier processing can hurt incidental learning.* Making it easier to process information (e.g., by reducing k in the above example) improves within-model performance but may make the person *less* likely to discover his model is wrong because it reduces the perceived need to ask further questions of the data. In the example, if k is low enough ($k \leq (\underline{\theta} r - c)/4$, where $\underline{\theta} \equiv \inf[\text{supp}(\pi)]$) then the investor thinks the advice is certainly worth collecting and no longer thinks he needs to know precisely how accurate it is.

Section 3 establishes the groundwork for generalizing the lessons of this example. When a person follows a minimal SAS, he never notices events his model thought were impossible, so the question of stability reduces to whether the actions that seem optimal under his model eventually become implausible. Section 4 studies pathways to incidental learning. Results in this section show that incidental learning typically requires tracking long-run statistics. They also characterize the within-model uncertainty that induces such tracking: entertaining a broader array of *clearly-wrong* parameters—ones the person would inevitably reject on his own—promotes discovery, matching experimental findings in Esponda et al. (2024). One implication is that the person is more likely to reject his faulty model when he refines it through personal feedback instead of having a more narrow model to begin with. Along these lines, delegation and easier processing can reduce incidental learning by narrowing the set of questions the person asks of the data. This section additionally formalizes the decoupling of detection from stakes: arbitrarily costly errors can persist while trivial ones are caught.

Section 5 moves from context-specific results to robustness. Some errors are more resistant than others to incidental learning across environments. The section studies robustness across different objectives and across shocks that expand attention, both temporary and permanent. While we show no error is robust in both senses, providing a limit to the persistence of errors, we characterize classes of error that are robust in one sense or the other. For example, errors that invite neglect (e.g., dogmatically believing a variable does not matter or that certain patterns are impossible) tend to be more robust to different objectives than errors that draw attention (e.g., believing a variable matters when it doesn't or that impossible patterns are possible).

[3]Letting $\bar{\theta}_t \equiv \mathbb{E}_t[\theta]$, the investor chooses to have access to the advice with probability $\alpha_t = (1/4k)(\bar{\theta}_t r - c)$ in an interior solution (true when $\bar{\theta}_t \in (c/r, (c+4k)/r)$). Assuming the support of π is non-degenerate and concentrated in the interior-solution interval, the investor's minimal SAS tracks $\bar{\theta}_t$.

Section 6 turns to further applications of our results. First, we show how our results imply a reason why fresh eyes catch mistakes, shedding light on why people benefit from outside opinions and why organizations value outside consultants. Second, following the fresh eyes logic, we offer an explanation for how people may recognize errors in others even when they don't recognize those errors in themselves. Thinking (correctly) that we understand ourselves better than we understand others leads us to neglect aspects of our own behavior that we pay attention to in others'. Third, we consider the role of theory in scientific progress. Our framework highlights the crucial role of theory in reducing the enormous complexity of the world into variables worth focusing on, but also how this comes at a cost: science can become stuck with overly coarse theories. Section 7 concludes by discussing limitations and extensions of our analysis.

Our framework sits at the intersection of work on attention and the persistence of mistaken beliefs. A prominent line of psychological research highlights the surprising degree to which seemingly conspicuous stimuli may go unnoticed (see Dehaene, 2014 and Chater, 2018 for reviews). The "inattentional blindness" to gorillas in Simons and Chabris (1999) is not limited to novice observation: Drew et al. (2013) illustrated that many experienced radiologists failed to notice images of gorillas superimposed on lung x-rays they were screening for cancerous nodules even when the gorillas were 48 times larger than the average nodule that they were looking for, and even when eye-tracking demonstrated that they looked directly at the gorilla.

Beyond simply showing that conspicuous stimuli may go unnoticed, the psychological research emphasizes that a person's allocation of attention often reflects their goals (see, e.g., Chun et al., 2011 and Gazzaley and Nobre, 2012 for reviews). This theme has been embraced, clarified, and formalized by recent models of rational inattention by Sims (2003), Woodford (2012), Caplin and Dean (2015), Matějka and McKay (2015), Payzan-LeNestour and Woodford (2022) and others. The notion of sparsity following Gabaix (2014) has likewise elaborated on the calculus of optimal attention.[4] Although we (crucially) differ in allowing attention to be guided by misspecified models, our framework is in this tradition of goal-directed attention.

At the same time, we abstract from other mechanisms that shape attention. Some research emphasizes attentional capture, salience, and other stimulus-driven forces that can draw attention independently of instrumental value (see, e.g., Bordalo et al., 2022 for a review targeted to economists and Li and Camerer, 2022 for recent evidence). Those mechanisms are clearly important, and economic models have used them to explain over-weighting of salient or surprising features (e.g., Bordalo et al., 2012, 2013, 2020). Our aim is different: we isolate how top-down theories channel attention, complementing more recent work by Bordalo et al. (2026). This helps explain why some unexpected events feel surprising while many others pass unnoticed: only a small subset are framed by

[4]Unlike these other papers, Gabaix (2014) allows for an agent to have priors in new situations that are incorrect, but assumes the agent's attentional strategy is guided by the true model.

the person's theory as relevant objects of notice.[5]

Our contribution is distinct from several other explanations for persistent errors, which we return to in Section 3. People may fail to learn because the right data are unavailable (e.g., Ba, 2026), because they do not experiment (e.g., Gittins, 1979; Fudenberg and Levine, 1993), because they misprocess the evidence they do see (e.g., Rabin and Schrag, 1999; Bénabou and Tirole, 2016), or because they have little incentive to adopt a better model even after recognizing one (e.g., Fudenberg and Lanzani, 2023). We isolate a different mechanism: people can have the data, be capable of processing it, and be willing to act on it if they noticed it, yet still fail to learn because their mistaken model directs attention away from the very evidence that would overturn it.

2 Setup and Channeled Attention

2.1 Environment

Consider a person updating his beliefs over a parameter $\theta \in \Theta$ that influences the distribution of payoff-relevant outcomes. Parameter θ might be a feature of the person's surroundings (e.g., the distribution of an asset's returns) or measure the extent of his biases (e.g., naivete over present bias as in O'Donoghue and Rabin, 1999).

As depicted in Figure 1, each period $t = 1, 2, \ldots$ is structured as follows: the person (i) receives a signal $s_t \in S_t$ correlated with θ, (ii) takes an action $x_t \in X_t$, and (iii) realizes an outcome, or "resolution", $r_t \in R_t$ correlated with θ. We call the data generated in round t, $y_t := (r_t, s_t)$, an "observable"; note that y_t provides information about θ. At the end of each period t, the person earns a payoff $u_t(x_t, y_t | h^t)$, which may depend on the current action, observable, and history, h^t. History h^t comprises all data that could be observed at the time of period t's action:

$$h^t := (s_t, y_{t-1}, x_{t-1}, y_{t-2}, x_{t-2}, \ldots, y_1, x_1).$$

Let H^t be the set of all histories up to period t, and let $H := \cup_{t=1}^{\infty} H^t$. Since h^t includes the period-t signal (for reasons that will be useful below), we additionally let $h^t(\neg s_t)$ denote all data prior to period t (i.e., h^t excluding s_t).

Unless otherwise noted, we assume that Θ and each $Y_t := R_t \times S_t$ are finite-valued, each X_t is

[5] Attempting to drive north from Paris, you'll be jolted awake if you see a sign that you're entering Marseille, yet you could only get there by also mindlessly driving by Provençal villages you've never heard of (and thus couldn't expect to see). Or, in studies like the invisible gorilla, any video will contain an infinite array of unlikely events in addition to the faux gorilla. Nobody watching the video would be "surprised" by the exact directions, distances, and numbers of passes, much less the exact body and facial configurations of the players depicted. Research in perception indeed indicates that unexpected outcomes are more likely to be incidentally noticed if they share features with what people are looking out for. Most et al. (2005), for example, find that a person is much more likely to notice an unexpected black circle if instructed to attend to circles or black objects than if instructed to attend to squares or white objects.

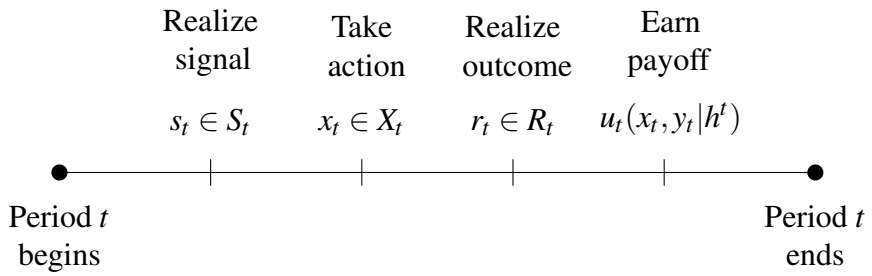

Figure 1: *Timeline of events within period t.*

compact, and $u_t(x_t, y_t|h^t)$ is continuous and bounded in x_t for all y_t and h^t. Although some of our examples relax these restrictions, they simplify our formal results.

The distribution of observables depends on θ. Denote the distribution of y_t conditional on the history and θ by $P(y_t|x_t, h^t(\neg s_t), \theta)$. Let $\pi^* \in \Delta(\Theta)$ denote the true probability distribution from which nature draws θ, and let θ^* denote the realized value. In sum, the decision environments studied in this paper are described by the tuple $(\Theta, \times_{t=1}^{\infty} X_t, \times_{t=1}^{\infty} Y_t, \times_{t=1}^{\infty} u_t, P, \pi^*)$.

A misspecified model (or "theory") is a prior belief over parameters $\pi \in \Delta(\Theta)$ such that $\pi \neq \pi^*$. Given our focus on long-run learning and our simplifying assumption that Θ is finite, the misspecified models we will consider are such that $\text{supp}(\pi) \neq \text{supp}(\pi^*)$ and, in particular, $\text{supp}(\pi^*) \not\subseteq \text{supp}(\pi)$. That is, we focus on cases where the person places no weight on some parameters that might actually occur. While this assumption may raise concerns that an agent trivially fails to learn because his model excludes the true parameter, this is not warranted; it will be clear below that, with full attention, an agent in our framework will generally discover his misspecification.[6]

[6]Many recent models, spanning a wide range of errors in both single-person and social judgments, take this "quasi-Bayesian" approach. Examples include Barberis et al. (1998) on stock-market misperceptions; Rabin (2002), Rabin and Vayanos (2010), and He (2022) on the gambler's and hot-hand fallacies; Benjamin et al. (2016) on the non-belief in the law of large numbers; Spiegler (2016) on biases in causal reasoning; and Heidhues et al. (2018) on overconfidence. Examples of biases about misreading information include Rabin and Schrag (1999) on confirmation bias; Mullainathan (2002a) on naivete about limited memory; and Gagnon-Bartsch and Bushong (2022) on attribution bias. Models of coarse or categorical thinking include Mullainathan (2002b), Jehiel (2005), Jehiel and Koessler (2008), Fryer and Jackson (2008), Mullainathan et al. (2008), and Eyster and Piccione (2013). Models that incorporate errors in reasoning about the informational content of others' behavior include Eyster and Rabin (2005), Esponda (2008), and Eyster et al. (2019); models exploring failures to understand the redundancy in such content in social-learning settings include DeMarzo et al. (2003), Eyster and Rabin (2010, 2014), Bohren (2016), Gagnon-Bartsch and Rabin (2021), Frick et al. (2020), and Bohren and Hauser (2021). Camerer et al. (2004) and Crawford and Iriberri (2007) provide models incorporating false beliefs about others' strategic reasoning. Misspecified models have also been considered in specific applications, such as firms learning about demand (Kirman, 1975; Nyarko, 1991) and macroeconomic forecasting (Sargent, 1993; Evans and Honkapohja, 2001). Further from the quasi-Bayesian approach, other models posit inconsistencies in a person's beliefs across periods. Although below we translate it to something that fits in our framework, naivete about one's future self-control falls within this broader category. The model of projection bias in Loewenstein et al. (2003) similarly posits that somebody may have systematically different beliefs about future tastes as a function of fluctuating contemporaneous tastes. Similar models of interpersonal projection bias include Madarász (2012, 2021) on information projection and Gagnon-Bartsch and Rosato (2024) and Gagnon-Bartsch et al. (2021) on taste projection.

We assume that starting with a misspecified model is the person's only mistake. He updates (when possible) according to Bayes' Rule given his prior π and chooses actions that maximize his expected lifetime utility with respect to those updated beliefs.

Unless otherwise noted, to isolate the effect of channeled attention, we maintain that the person's actions do not affect what he can learn.

Assumption 1 (Maintained assumption, unless otherwise noted)**.** The data-generating process for observables is independent of actions. Formally, for all t, $h^t \in H^t$, $x_t \in X_t$, $y_t \in Y_t$, and $\theta \in \Theta$:
(i) $P(s_t|h^t(\neg s_t),\theta) = P(s_t|y^t,\theta)$ and (ii) $P(r_t|x_t,h^t,\theta) = P(r_t|s_t,y^t,\theta)$, where $y^t := (y_{t-1},y_{t-2},\ldots,y_1)$.

Assumption 1 implies that any learning failure arises from insufficient attention, not insufficient experimentation. We also often assume that payoffs in period t are independent of the history.

Assumption 2 (Frequent but not maintained assumption)**.** For all t and $\theta \in \Theta$, the action set X_t is independent of h^t, and for all $x_t \in X_t$ and $y_t \in Y_t$, the payoff $u_t(x_t,y_t|h^t)$ is independent of $h^t \in H^t$.

Without an incentive for experimentation (Assumption 1), Assumption 2 ensures that myopically optimal actions are in fact long-run optimal. When this assumption holds, we write $u_t(x_t,y_t|h^t)$ simply as $u_t(x_t,y_t)$. In examples that relax one or both of these assumptions we directly assume that people follow myopically optimal strategies and make other assumptions that guarantee that, in the long run, they have the data to reject π in favor of π^* under full attention.

2.2 Channeled Attention

People tend to direct their attention to subjectively task-relevant information and away from a mass of other information.[7] Horowitz (2013) elegantly conveys this core truth:

> You are missing most of what is happening around you right now. ... In reading these words, you are ignoring an unthinkably large amount of information that continues to bombard all of your senses. The hum of the fluorescent lights; the ambient noise in the room; ... the constant hum of traffic or a distant lawnmower; a chirp of a bug or whine of a kitchen appliance.

To model channeled attention, we assume the person notices and remembers a coarsened form of all the available information. For each $t = 1,2,\ldots$, let N^t partition the set of histories, H^t. Following $h^t \in H^t$, the person recalls only the *noticed history*, $n^t(h^t)$: the element of N^t containing h^t. In

[7] Attention and memory do not act like cameras that faithfully record all we see, and we attend to and remember a small subset of available information. Dehaene (2014), Chater (2018), and others review the literature and highlight how goal-oriented attention reinforces the inherent constraints on attention, arguing that our "narrow channel of consciousness" is surprisingly effective at blocking out information we're not looking out for. See, e.g., Chun et al., 2011 and Gazzaley and Nobre, 2012 for reviews. However, this narrow channel is of course imperfect—a fact we return to in Sections 5 and 7.

the basic investment-advice example from the introduction, the investor sees no purpose in tracking statistics on advice accuracy; hence, $n^t(h^t)$ need not track these statistics.

Figure 2 amends the timeline of each period depicted in Figure 1 by including a stage of information coarsening. Prior to taking action x_t, the person summarizes all prior data into what he believes is a "sufficient statistic", $n^t(h^t)$. He then uses $n^t(h^t)$ to guide his action.

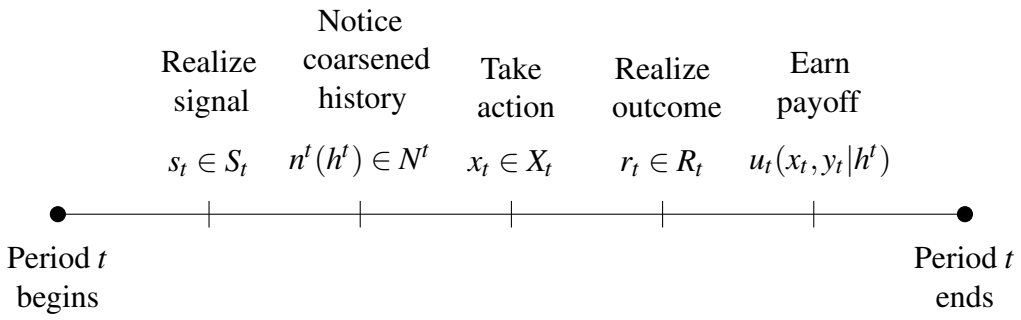

Figure 2: *Timeline of events within period t, including the coarsening of past information.*

A *noticing strategy* $\mathscr{N}$ is the full sequence of the person's noticing partitions, $\mathscr{N} = (N^1, N^2, \dots)$. This strategy specifies for each point in time what the person notices conditional on the true history. An *attentional strategy* is a pair $\phi = (\mathscr{N}, \sigma)$ combining a noticing strategy with a *behavioral strategy* $\sigma = (\sigma_1, \sigma_2, \dots)$, where $\sigma_t : N^t \to \Delta(X_t)$ maps noticed histories to actions. We assume the person automatically recalls his prior π, attentional strategy ϕ, and the current time period t. Upon observing noticed history n^t, the person's subjective likelihood of n^t given θ and ϕ is thus $P(n^t|\theta,\phi) = \sum_{h^t \in n^t} P(h^t|\theta,\phi)$, and his resulting posterior over θ is

$$\pi_t(\theta) = \frac{P(n^t|\theta,\phi)\pi(\theta)}{\sum_{\theta' \in \Theta} P(n^t|\theta',\phi)\pi(\theta')}.$$ [8]

When it does not cause confusion and avoids cumbersome notation, we suppress the dependence of likelihoods $P(\cdot|\theta,\phi)$ on ϕ.

We assume attentional costs are negligible, and thus the decision-maker ignores some data only when he perceives it as useless for guiding decisions. We therefore call his attentional strategy "sufficient" with respect to his theory π if he filters out only data that π deems irrelevant for decisions.

Definition 1. An attentional strategy $\phi = (\mathscr{N}, \sigma)$ is a *sufficient attentional strategy (SAS) given* π if, under π, the person expects to do no worse by following ϕ than he would by following any other

[8] This formula applies when n^t has positive probability under π—the relevant case for much of our analysis (see Proposition 2 below)—and the posterior is arbitrary otherwise.

attentional strategy $\tilde{\phi}$. Under Assumption 2, sufficiency of $\mathscr{N}$ amounts to

$$\max_{x \in X_t} \mathbb{E}_\pi[u_t(x,y)|n^t(h^t)] = \max_{x \in X_t} \mathbb{E}_\pi[u_t(x,y)|h^t]$$

for all t and $h^t \in H^t$ that occur with positive probability under (π, σ).

When following a sufficient attentional strategy (SAS), the person believes that his expected payoffs are identical to those he would earn if he noticed the complete history. Fixing π, there are typically many SASs in a given environment. Although our definition of a SAS does not require a person to ignore data he deems useless, we focus primarily on the "minimal" case where all seemingly extraneous information is ignored. Say that $\tilde{\mathscr{N}}$ *coarsens* $\mathscr{N}$ if for all t, the partition $\tilde{N}^t$ coarsens N^t and at least one of these coarsenings is strict.

Definition 2. Given π, a SAS $(\mathscr{N}, \sigma)$ is *minimal* if there does not exist another SAS that can be obtained by coarsening $\mathscr{N}$.

Minimal attentional strategies are consistent with our interpretation that the person faces small costs of attention and allocates his attention efficiently.[9]

To compare beliefs that arise under a minimal SAS to those that arise when paying full attention, let $\mathscr{N}_F$ be the *full-attention noticing strategy*, where $n^t_F(h^t) := \{h^t\}$ for all $h^t \in H$. Under $\mathscr{N}_F$, the person perfectly distinguishes the history in every period. A *full-attention SAS* is one that follows $\mathscr{N}_F$.

We conclude this section by highlighting a few different ways that a person's noticing strategy $\mathscr{N}$ may filter the observable data. First, $\mathscr{N}$ may neglect signals, s_t, as in Schwartzstein (2014). In this case, $n^t(h^t)$ merges all histories that are identical *except* for their signal realizations. For example, a household investor may neglect analyst affiliation in interpreting investment recommendations for what they predict about future stock performance (Malmendier and Shanthikumar, 2007), or a doctor may neglect predictors of heart attacks in evaluating patients (Mullainathan and Obermeyer, 2022).

Second, $\mathscr{N}$ may neglect outcomes, r_t. In this case, $n^t(h^t)$ merges all histories that are identical *except* for their outcomes. For example, a household investor who is confident that an actively managed mutual fund will outperform a passively managed fund over 5 years may neglect their realized relative performance, or a consumer who autopays their utility bills and thinks they know the rough amount they pay each month may fail to notice they're paying more than they thought possible.

[9]There may be multiple minimal SASs for a given π and environment; Proposition 1, below, highlights a prominent one. Different minimal SASs can potentially lead to different long-run beliefs. To illustrate, consider a person who thinks that each of M doctors makes the same recommendation given a fixed set of symptoms. There are thus M minimal SASs, each taking the following form: the person follows the advice of Doctor $m \in \{1, \ldots, M\}$ and ignores the others. If advice actually varies across doctors, the patient's (in his mind, inconsequential) choice of who to follow will determine his long-run beliefs.

The previous two categories don't require that all signals or outcomes are ignored. They also include cases where signals or outcomes consist of several distinct variables and some of those variables are ignored while others are noticed. For instance, a farmer might pay attention to some types of inputs and ignore others.

Third, $\mathscr{N}$ may neglect statistics. A venture fund manager may not systematically track the performance of companies she passed on, even if she recognizes individual successes and failures. An analyst may notice given instances of who a judge chooses to jail without tracking how those choices relate to details of the defendants' faces (Ludwig and Mullainathan, 2024). In these cases, a person may notice the observables that occur in the current period, but ignore certain statistics by failing to aggregate those observables across periods. This category also includes scenarios where a person tracks some statistics of the data but not all. For instance, a person may track the average of two variables while ignoring their correlation.

Fourth, and related to the third possibility, $\mathscr{N}$ may coarsely represent outcomes. While neglecting statistics entails coarsely recalling past outcomes, a person may coarsely notice outcomes *as they see them*. For example, a loan officer may only notice an applicant's credit range (e.g., Excellent, Very Good, etc.) without internalizing the precise score. A shopper may ignore the final digits of a price while remaining sensitive to the leading digits. A person checking sports news might notice whether a team won or lost without carefully noticing the exact score.[10]

The four ways of filtering data above differ in how psychologically plausible they are, and our examples differ in which kind of filtering they invoke. Neglecting statistics—failing to aggregate data the person sees individually—is widely documented. The "statistical gorilla" interpretation we emphasize throughout falls into this category, and so do most of our running examples presented below. Neglecting signals is also psychologically natural, especially when the signal requires deliberate action to be observed (e.g., checking analyst affiliations, opening a brokerage statement). Neglecting outcomes is natural when outcomes are remote or experienced indirectly (e.g., the autopaid utility bill), but less plausible for outcomes that are immediate and salient. Finally, while cases of coarsely representing outcomes seem realistic (as in the examples above), other cases are psychologically implausible, for example, where a person wrongly categorizes a seemingly impossible outcome (e.g., being hit in the face with a ball when he steps outside) as something entirely different (e.g., retrieving the mail). To understand the extent to which these implausible cases of coarse representation affect our results, Section 5 analyzes refinements to noticing strategies that incorporate attentional capture, where a person cannot help but attend to certain event features as they happen. Our results imply

[10]This discussion highlights a potentially subjective distinction between coarsely representing and neglecting variables. The distinction is clear if variables and dimensions are given ex ante, independent of a person's π. But perhaps not otherwise. In the latter case, an analyst could add a new variable to the environment that is a coarse version of an existing variable (e.g., extend the environment to specify that a person observes both the precise credit score and the credit range) and say that the person ignores the fine version of the variable yet attends to the coarse version.

that such refinements only have the potential to modify conclusions about the attentional stability of wrong models that have the structure of this (getting hit) example, where they place *zero* probability on truly-possible outcomes. We further discuss refinements to noticing strategies in Section 7 and Appendix A.

2.3 Examples

This section provides examples that we will return to when illustrating results.

Example 1 (Interpreting Others)**.** We first consider a set of examples of (mis)learning from others, generalizing the investment advice illustration from the introduction. A person may learn from news sources, social media connections, neighbors, investment advisors, and so on. To fix ideas, continue to consider an investor evaluating investment opportunities as they arise. In each period t, the investor faces a new opportunity—independent of those that came before—and the optimal action is denoted by $\omega_t \in \{0,1\}$. The investor starts with a uniform prior over ω_t, but receives advice from his sources and a private signal (e.g., his own intuition). There are I sources who may offer advice. Let $m_t^i \in \{0,1\}$ be the action recommended by source i and m_t^0 be the investor's private signal. The investor's overall signal is $s_t = (m_t^0, s_t^1, \ldots, s_t^I)$ where $s_t^i \in \{m_t^i, \varnothing\}$; to capture settings where advice is sometimes unavailable (as in the modified example in the introduction), $s_t^i = \varnothing$ indicates that the investor cannot see i's recommendation before taking his action in period t. The parameter θ dictates the joint distribution over the private signal and recommendations conditional on the true optimal action, $P^M(m^0, \ldots, m^I | \omega, \theta)$. Thus, θ determines how the investor interprets the advice. The investor then chooses $x_t \in \{0,1\}$. After taking his action, the investor can observe the true optimal action and the action recommended by each advisor: $r_t = (\omega_t; m_t^1, \ldots, m_t^I)$. This ensures ample feedback to learn the correct interpretation of advice (Assumption 1 holds).

This setup accommodates various errors in learning from others (see Eyster, 2019 for a review). We'll return below in particular to people's tendency to neglect correlations in others' opinions (as in DeMarzo et al., 2003; Eyster and Rabin, 2010, 2014; Enke and Zimmermann, 2017).

Example 2 (Production Problem)**.** We next describe a production problem. To capture several relevant applications, we first introduce the problem abstractly. Consider a producer learning about the relationship between the inputs used in period t, (w_t, z_t), and the resulting net output v_t. Net output is given by $v_t = \alpha_t(w_t, z_t) \cdot \beta_t(w_t, z_t)$, where $\alpha_t(w,z) > 0$ is the known value of a successful production process and $\beta_t(w,z) \in \{0,1\}$ indicates whether it's successful. The producer forms beliefs over the parameter $\theta(w,z) := \Pr(\beta_t(w,z) = 1)$, which represents the (stationary) likelihood of success given the inputs. The sets of feasible inputs W and Z are assumed to be finite. The producer has a prior π over $\theta := (\theta(w,z)_{w \in W, z \in Z})$, which specifies the possible relationships between inputs and output.

We consider a production context where inputs cannot be perfectly set, for example because measurement is noisy. Following Matějka and McKay (2015), we model the producer selecting a distribution over inputs. Suppose there is an outside option o with known expected output normalized to zero. In each period, the producer chooses a distribution $x_t \in \Delta((W \times Z) \cup \{o\})$ at cost $\chi \cdot D(x_t \| q^d)$, where D is the Kullback-Leibler (KL) divergence and q^d is some default distribution; $\chi > 0$ captures the cost of precisely setting inputs. For simplicity, we assume q^d is uniform, which implies that the optimal choice of x_t results in a logit choice rule (e.g., Matějka and McKay, 2015). More specifically, assume the producer chooses x_t to maximize the expectation of $v(w,z) - \chi \cdot D(x_t \| q^d)$ given his current belief π_t over θ with $\alpha_t(w_t, z_t) \equiv 1$. Letting $\mathbb{E}_t$ denote expectations given π_t, the subjectively optimal x_t satisfies

$$x_t(w,z) = \frac{e^{\mathbb{E}_t[v|w,z]/\chi}}{1 + \sum_{w',z'} e^{\mathbb{E}_t[v|w',z']/\chi}}. \tag{1}$$

After taking action x_t, the producer can observe the realized inputs (w_t, z_t), whether production was successful (β_t), and the output (v_t). There are no signals in this example and it relaxes Assumption 1.

This setup captures many applications involving erroneous beliefs about production functions. To take one example, Hanna et al. (2014) study seaweed farming where the yield (v) depends on the distance between planted seaweed pods (w) and the size of the pods (z). They provide evidence that farmers believed pod size (z) to not influence net output: their model π was dogmatic that $\theta(w,z)$ is constant in z for each value of w when, in truth, $\theta^*(w,z)$ varies in z for each w. In this example, producers (farmers) neglected the importance of an input variable. In other examples, with $\alpha_t(w_t, z_t)$ perhaps differing from 1, producers wrongly think they know the right level of an input to set (π is dogmatic that a particular level of w and/or z maximizes the expectation of v_t). To illustrate, for decades players, coaches, and general managers in the National Basketball Association (NBA) promoted or followed suboptimal shooting strategies that prioritized 2-point shots over 3-point shots, despite clear evidence that the latter yielded significantly higher expected points than long ("mid-range") versions of the former (Goldsberry, 2019; Thaler, 2020; Schwartzstein and Malhotra, 2021). In other examples, producers may wrongly think they know the parametric function that links inputs to outputs. For example, List et al. (2023) show that rideshare platform Lyft long followed pricing algorithms that assumed a smooth model of demand when demand in fact responded discontinuously to changes in the left digits of prices.

Example 3 (Mispredicting Statistical Processes)**.** We can also consider predictions about an external statistical process, such as analyst predicting whether an asset's earnings will go up or down. As in Barberis et al. (1998), suppose the analyst wrongly thinks that changes in earnings are autocorrelated when in reality earnings follow a random walk. Let $\theta_{u,u}$ ($\theta_{u,d}$) be the probability that earnings will go up in the current period conditional on going up (down) in the previous period. The analyst's

model π over $\theta = (\theta_{u,u}, \theta_{u,d})$ is such that $\theta_{u,u} \neq \theta_{u,d}$: he believes the current change depends on the previous one, when in truth $\theta^*_{u,u} = \theta^*_{u,d}$. In each period, the analyst makes either a "buy" or "sell" recommendation and can then observe the current change in earnings (r_t).

Example 4 (Mispredicting Behavior)**.** Next take a canonical problem from behavioral economics: a person mispredicts his future behavior. For a concrete example, consider a person with naive present bias (O'Donoghue and Rabin, 1999, 2001) who makes decisions about visiting the gym, as in DellaVigna and Malmendier (2006). In each period t, the person chooses whether to visit the gym (or sign-up at price m if he isn't a member). Going to the gym has an immediate effort cost c_t and a delayed health benefit $b > 0$. The person is a (β, δ) discounter (Laibson, 1997) with $\delta = 1$ and discounts future costs or benefits by a factor $\beta < 1$. Similar to Eliaz and Spiegler (2006), we allow the person's degree of present bias to vary from day to day, with β_t capturing the value on day t. In each period t, $\beta_t = \beta < 1$ with probability θ and $\beta_t = 1$ with probability $1 - \theta$: the person sometimes feels tempted to skip the gym and θ captures how often he feels tempted. The person's degree of naivete about his self-control problem is captured by his model π over θ.[11] Suppose in truth $\theta^* = 1$—the person is always tempted to skip the gym. This person naively overestimates his self control when $1 \notin \text{supp}(\pi)$ and consequently overestimates his gym attendance. This can lead to suboptimal membership decisions whenever those decisions depend on predicted attendance rates. See Appendix C.2 for more details on this example. Other well-known examples in this category of mispredicting own behavior include suboptimal actions due to overconfidence about one's memory and failures to use valuable reminders (Ericson, 2011; Bronchetti et al., 2023).

Example 5 (Spending)**.** Finally, consider a canonical consumer problem. Following Kőszegi and Matějka (2020), suppose a consumer's indirect utility of spending on goods (indexed by m) in a category (e.g., video subscription services) equals:

$$\psi \sum_m x_m - \frac{1}{2} \sum_m x_m^2 - \eta \sum_{m \neq n} x_m x_n - \sum_m p_m x_m,$$

where $\psi \in \mathbb{R}$ is a possibly stochastic taste for the category, x_m equals the amount of good m purchased, η equals a substitutability parameter that is positive when goods within the category are substitutes and negative when they are complements, and p_m is the potentially stochastic price of good m. If the person is optimizing, then, in an interior solution, for every good m she chooses $x_m = \psi - 2\eta \sum_{n \neq m} x_n - p_m$.

The person may err in several ways in this setting. She might systematically underestimate the price of a good—for instance, by neglecting add-on fees (Gabaix and Laibson, 2006)—misperceiving

[11] To ensure Assumption 1 holds, we assume the person can observe his current cost and degree of present bias regardless of his decisions; i.e., $s_t = (c_t, \beta_t)$.

$\hat{p}_m = p_m - \theta_m$. She might similarly underestimate her consumption, such as by failing to account for automatic subscription renewals, misperceiving consumption as $\hat{x}_m = x_m - \theta_m$. She might overlook the existence of cheaper substitutes, effectively acting as though the price of good j were $\hat{p}_j = p_j + \theta_j \approx \infty$. She might overestimate how much she will use a product—for example by overestimating ψ—misperceiving $\hat{\psi} = \psi + \theta$. Finally, she might neglect the shadow price of money, failing to account for spending needs in other categories (Augenblick et al., 2026), and thus misperceive the price as $\hat{p}_m = p_m - \theta$.

3 Attentional Stability

In this section, we first formalize our criterion for whether an attention-channeling decision-maker will reject his erroneous model. We then provide some initial results that help assess when this criterion is met.

3.1 The Stability Criterion

When will somebody following a SAS notice that his model of the world π is false? Roughly, we say that π is *attentionally unstable* relative to "lightbulb" model $\lambda \in \Delta(\Theta)$ (which could itself be false) if the noticed data are infinitely more likely under the lightbulb than the prior. Otherwise, we say π is *attentionally stable* relative to λ.[12]

Definition 3. A model π is *attentionally unstable* with respect to λ and SAS $\phi = (\mathcal{N}, \sigma)$ if when following SAS ϕ, there is a positive probability under π^* that either (i) the limit inferior as $t \to \infty$ of the ratio $P(n^t|\pi,\phi)/P(n^t|\lambda,\phi)$ equals 0 or (ii) the ratio $P(n^t|\pi,\phi)/P(n^t|\lambda,\phi)$ equals 0 in some finite period t, where $P(n^t|\tilde{\pi},\phi) = \sum_{\theta'} P(n^t|\theta',\phi)\tilde{\pi}(\theta')$ for all $\tilde{\pi} \in \Delta(\Theta)$. Otherwise, π is *attentionally stable* with respect to λ and ϕ. If the latter case holds when $\lambda = \pi^*$, we call ϕ a *stable attentional strategy (StAS)* given π.

In the special case of a full-attention SAS (where a person continually notices everything), a model π is unstable if it seems excessively unlikely relative to the lightbulb model λ when *all* available data is actively used to assess the relative likelihood of π versus λ. Under a minimal SAS, our interpretation is as follows: we ask when only the relevant data under π is enough to alert the person that his model is misspecified. If this selected data—which was noticed with the sole purpose of making optimal decisions given π—happens to reveal in the long or short run that π is overwhelmingly implausible relative to λ, then π is attentionally unstable.

[12] If both models assign zero probability to the noticed data, we treat the likelihood ratio as positive, since that data provides no relative evidence in favor of λ. This case is only relevant when $\lambda \neq \pi^*$.

Indeed, we refer to π being attentionally unstable as a case of *incidental* learning precisely because the data that ultimately disprove the model were noticed for reasons orthogonal to testing it. Put differently, we do not interpret the person as actively noticing aspects of the data as he collects it for the purpose of distinguishing π from λ. After all, he sees no reason to question π.[13] We also do not interpret the person as actively invested in the question of whether λ provides a better explanation for noticed data n^t than π after he collects the data; he does not seek out data beyond n^t (Section 5 analyzes cases where he seeks out such data). Rather, incidental learning happens when the data the person views as relevant and notices *under* π allows him to recognize that there's an alternative λ that better explains those data.

While our framework allows for any λ, most of our analysis assumes λ is the correct model. Accordingly, when we assess the attentional stability of π without reference to a particular alternative lightbulb model, we implicitly take $\lambda = \pi^*$. Focusing on $\lambda = \pi^*$ pins down our analysis and most closely connects our criterion of attentional stability to more classical notions of self-confirming equilibrium (Fudenberg and Levine, 1993). In general self-confirming equilibrium frameworks such as Dekel et al. (2004), an incorrect conjecture is part of a self-confirming equilibrium if and only if *exogeneously* available feedback (given stipulated equilibrium strategies) fails to disconfirm the conjecture because the observed distribution over feedback matches its conjectured distribution. This test is equivalent to asking if there fails to exist a better explanation of the observed distribution: if the observed distribution matches, there cannot exist a better explanation; if it doesn't match, then there exists a better explanation. Our test is similar but *endogenizes* the feedback representation through channeled attention. The noticing strategy determines which features and cross-period statistics of the available data enter the feedback against which the model is assessed.

This relationship raises the question of whether we can dispense with the alternative model λ and instead more closely parallel self-confirming equilibrium by deeming a model unstable when the long-run empirical distribution of noticed data fails to match the predicted distribution under π. This is possible in some cases, notably in stationary environments where what's noticed tracks empirical frequencies of finite feedback variables.[14] But outside these cases, including in the investment advice example from the introduction, the concepts differ. A long-run empirical distribution of some feedback variable may exist from an analyst's perspective yet not be tracked by the person's attentional

[13] We make two assumptions that bias our analysis in favor of classifying a model as attentionally *unstable*. First, we implicitly assume the person is aware that he selectively notices and recalls information when he assesses the likelihood of π relative to λ. Second, π is attentionally unstable if there is a *positive probability* that the selectively-noticed data makes π seem implausible relative to λ. This probabilistic definition takes a stand on instability in cases where $P(n^t|\pi)/P(n^t|\lambda)$ converges to 0 under some infinite histories, but not others. To give an example, suppose a ball is drawn from an urn each period. All of the balls in the urn have the same color, but that color is unknown ex ante—it may be red, blue, or yellow. If the person's model π posits that the urn only has two possible colors—red or blue—then $P(n^t|\pi)/P(n^t|\lambda)$ may converge to 0 when facing a yellow urn but not a red or blue one. If there is any chance that the noticed data wakes the person up, then we deem π attentionally unstable.

[14] More formally, let q_θ denote the distribution of a finite feedback variable under parameter θ and q^* its true distri-

strategy. For example, a person who notices only the current advisor recommendation but not its historical accuracy is exposed to—and acts on—repeated feedback without representing the frequency statistic that would disconfirm his model. Our likelihood-ratio criterion in Definition 3 remains well defined on the sequence of noticed histories and extends the self-confirmation logic to environments in which channeled attention does not support a conventional empirical-distribution test.

Continuing the parallel to self-confirming equilibrium, our framework is likewise agnostic on what happens after a model is rejected. Taking $\lambda = \pi^*$ potentially gives a misleading impression that if a person discovers his model is wrong, then he necessarily abandons it in favor of the true model. But if π is attentionally unstable with respect to π^*, then it is also attentionally unstable with respect to the infinite array of models that explain reality better than π. Thus, the most we can say when data generated and filtered through the lens of the model disconfirms the model is that the model is unstable. The dynamics following a "lightbulb moment" where π is deemed unstable—and specifying which model a person adopts after rejecting π—is *not* part of our formal analysis.

In the interest of presentational parsimony, probabilistic statements and definitions are with respect to the true data-generating process; we therefore take π^* to be degenerate on some θ^*.[15] Continuing the purpose of Assumptions 1 and 2 to make the effects of channeled attention clear, we also focus on models that are *identifiably wrong*, meaning they would be rejected with full attention.

Definition 4. Model π is *identifiably wrong* if when following a full-attention SAS, either of the following occur with probability one under π^*: (i) $\liminf_{t\to\infty} P(n^t|\pi,\phi)/P(n^t|\pi^*,\phi) = 0$ or (ii) $P(n^t|\pi,\phi)/P(n^t|\pi^*,\phi) = 0$ in some finite period t.

We only consider models that are identifiably wrong. This implies that, with full attention, a person would eventually reject a model that results in costly errors. With channeled attention, on the other hand, costly errors can persist.[16]

---

bution. Standard arguments imply

$$t^{-1}\log\left[P(n^t|\pi)/P(n^t|\pi^*)\right] \longrightarrow -\min_{\theta\in\text{supp}(\pi)} D_{KL}(q^*||q_\theta).$$

Hence, model π is stable with respect to π^* if and only if some parameter π considers possible delivers exactly the true distribution of noticed feedback, as in the feedback-consistency condition of self-confirming equilibrium. This observation also clarifies the relationship to Berk-Nash equilibrium (Esponda and Pouzo, 2016). Berk-Nash requires beliefs to concentrate on subjectively possible parameter values that minimize the KL divergence from observed feedback, even when the minimum divergence is positive. Our stability criterion is stronger in the frequency-revealing case with $\lambda = \pi^*$, where it requires the minimum to be zero.

[15] Equivalently, we could make probabilistic statements with respect to a realized parameter value, θ^*, that was drawn by nature from π^*. Whether lightbulb model λ is degenerate or not becomes relevant for stability only when we violate our assumption that Θ is finite. Suppose, for instance, that the person's theory about the bias of a coin is uniform on $[0,1]$ in a situation where we believe there is real uncertainty over the bias. If the coin happens to be biased 0.55, we do not want the person to deem his uncertain-prior model unstable merely because a dogmatic prior of 0.55 would have designated the realized outcome as more likely. By contrast, in the more realistic scenario where we posit a true model in which the coin is certainly unbiased, we are comfortable saying that the uncertain $[0,1]$ theory is unstable.

[16] In Appendix B.1, we characterize stability under full attention in stationary environments. We note that a model π

3.2 Basic Results

Our first proposition simplifies the test for attentional stability and characterizes a minimal sufficient attentional strategy for any model π. (Proofs of all propositions are gathered in Appendix D.) Let $X_t^*(h^t, \pi) \subseteq X_t$ denote the set of subjectively optimal actions given h^t under π if h^t has positive probability under π. If h^t has zero probability under π, let $X_t^*(h^t, \pi) = X_t^*(\tilde{h}^t, \pi)$ for some $\tilde{h}^t$ that has positive probability under π. Further, to simplify the presentation of Proposition 1, assume that $X_t^*(h^t, \pi)$ is a singleton unless otherwise noted.[17]

Proposition 1. *1. There exists a minimal SAS given π with the following form: in each period t, the person notices $X_t^*(h^t, \pi)$ and nothing more.*

2. There exists a minimal SAS given π that is a stable attentional strategy given π if $\liminf_{t\to\infty} P(X_t^*(h^t, \pi)|\pi, \phi)/P(X_t^*(h^t, \pi)|\pi^*, \phi) > 0$ with probability one under π^* when the person follows the SAS ϕ from Part 1.

3. There exists a minimal SAS given π that is not *a stable attentional strategy given π if $\liminf_{t\to\infty} P(X_t^*(h^t, \pi)|\pi, \phi)/P(X_t^*(h^t, \pi)|\pi^*, \phi) = 0$ with positive probability under π^* when the person follows the SAS ϕ from Part 1.*

Part 1 of Proposition 1 shows that it is sufficient for a person to simply query the history each period, asking "what set of actions are optimal for me to take today?" Such a SAS is also minimal since the person believes that she would take a suboptimal action with positive probability if she noticed anything coarser than the set of optimal actions. Parts 2 and 3 then draw out implications for checking whether there is a stable attentional strategy given π—the answer simply reduces to checking whether the current optimal action becomes implausible over time.

This result is best thought of as a sort of revelation principle: it simplifies checking for stable attentional strategies. To illustrate, return to Example 2 on production from Section 2.3, and follow Hanna et al. (2014) by considering a seaweed farmer who believes that yield (v) depends on the distance between pods (w) but not pod size (z). Suppose the farmer is uncertain about the optimal value of w: the support of π includes different values of $\theta = (\theta(w)_{w\in W})$, which specifies the chance of a successful yield conditional on w. His SAS will notice details of empirical success rates conditional on w but will ignore z. Additionally, since the farmer believes z is irrelevant, he will not invest in precisely setting this variable; z will be distributed according to the default (i.e., uniform) in each

is attentionally unstable with respect to λ under full attention if it explains observations worse than λ (in the Kullback-Leibler sense), reflecting results known at least since Berk (1966). This observation has two immediate implications: (i) π is attentionally stable with respect to π^* and a full-attention SAS (i.e., π is *not* identifiably wrong) if and only if there exists some $\theta \in \text{supp}(\pi)$ that induces the same distribution over observables as θ^*; and (ii) any π that assigns positive probability to θ^* is attentionally stable with respect to any $\lambda \in \Delta(\Theta)$ and a full-attention SAS.

[17] Our results do not hinge on this, but assuming uniqueness avoids cumbersome notation. We distinguish between the cases where π does or does not assign positive probability to h^t for reasons that become clear around Proposition 2.

period. The intuition from the theoretical analysis in Schwartzstein (2014) and Hanna et al. (2014) is that the belief that input z does not matter may be self confirming. If a person thinks z does not matter, he does not pay attention to the relationship between z and the output and then need not learn that z in fact does matter.[18]

Proposition 1 helps refine this intuition and see that the details of the farmer's misspecified model π (beyond neglecting the importance of z) matter for assessing attentional stability. First, suppose π accommodates the true marginal relationship between v and w given a uniform distribution of z. Thus, the support of π includes $\hat{\theta}$ such that $\hat{\theta}(w) = \frac{1}{|Z|}\sum_{z'} \theta^*(w,z')$ for all w. From Equation 1, the farmer's optimal action (i.e., distribution over inputs) in period t requires him to separately distinguish the empirical success rate for each value of w to form $\mathbb{E}_t[v|w,z] = \mathbb{E}_t[v|w]$, where the equality follows from π neglecting the role of z. Within his model, this expectation converges to $\frac{1}{|Z|}\sum_{z'} \mathbb{E}_{\theta^*}[v|w,z']$, where $\mathbb{E}_{\theta^*}$ denotes expectations under the true model. His actions similarly converge to $\hat{x}$ with $\hat{x}(w,z) = e^{\mathbb{E}_{\theta^*}[v|w]/\chi} / \left[1 + |Z|\sum_{w'} e^{\mathbb{E}_{\theta^*}[v|w']/\chi}\right]$. The actions in this example are sufficiently fine-tuned to precisely reveal the farmer's long-run conditional expectations (and thus the empirical frequencies of success given w). Yet the expectations he reaches are entirely consistent with his model. The farmer correctly learns the relationship between v and w while remaining ignorant about the influence of z. By Part 2 of Proposition 1, π is attentionally stable under this SAS, replicating the intuition in Schwartzstein (2014) and Hanna et al. (2014).

However, a modification of π suggests how a person might incidentally learn that z matters. Suppose that while the farmer believes that z does not matter, he also knows how well he should do if he set inputs optimally, for example if he lives near optimizing neighbors and can observe their yield. Let $\theta^*(w^*,z^*)$ denote the true frequency of a successful yield at the optimum. The farmer is confident he can reach this optimal frequency: for each $\theta \in \text{supp}(\pi)$, there exists $w \in W$ such that $\theta(w) = \theta^*(w^*,z^*)$. This model no longer accommodates the true marginal relationship between v and w, since in reality there is no w that delivers a marginal success rate of $\theta^*(w^*,z^*)$. As above, the farmer's optimal action in each period reveals the empirical success rates for each w. Hence, the farmer will continually see success rates (via his actions) that differ from the level of success he expects in the long run. By Part 3 of Proposition 1, π is not attentionally stable in this case. Notably, in these examples the z-neglecting farmer *is* more *likely to incidentally learn he is wrong when he is* less *wrong to begin with*: it's only the farmer whose model correctly reflects how well he should do at the optimum that incidentally learns his model is wrong.

Proposition 1 can also be used to assess the stability of commonly studied biases in environments often explored in the literature. For instance, consider Example 3 from above where the analyst

[18] In terms of the categorization of noticing strategies in Section 2.2, this SAS corresponds to neglecting statistics. The farmer notices the marginal relationship between v and w but ignores the joint relationship that further depends on z. Whether or not the farmer notices individual instances of z is irrelevant for this example; what matters is that he ignores statistics involving z.

misperceives the process of an asset's earnings by treating them as autocorrelated when they are actually i.i.d.. As we show in Appendix C.1, a setting where the analyst gives simple "buy" or "sell" recommendations promotes the stability of this bias. Intuitively, any isolated recommendation does not perfectly reveal the path of earnings and does not force the analyst to confront his misconception of the earnings process. This is despite the fact that a minimal SAS under the analyst's false model, π, notices events *more precisely* than a minimal SAS under the true model, π^*, because he wrongly thinks past outcomes help predict future returns. However, this increased attention is still insufficient to reject π. Or consider Example 4 where the gym-goer is naive about his self-control problem. Such naivete can be attentionally stable when the person's sole decision each day is to visit the gym (i.e., there is no ancillary decision that depends on future attendance rates). Reflecting Proposition 1, a minimal SAS in this case only distinguishes whether the current discounted benefit exceeds the current cost and ignores past data. That is, in each period a person only asks himself "do I want to go to the gym today?" He doesn't ask the additional question "if not, why?"[19] This limited data is insufficient to reveal the severity of his self-control problem, as we detail in Appendix C.2.[20]

Extending the theme of exploring the relationship between the degree to which a model is misspecified and its attentional stability, Proposition 1 also speaks to a basic question: Under situations where a person's model is so misspecified that he fails to conceptualize possible events, will he confront data he thought was impossible?

Definition 5. Model π *fails to conceptualize an event* (relative to π^*) if π assigns zero probability to some finite history of outcomes $y^t = (y_{t-1}, \ldots, y_1)$ that occurs with positive probability under π^*. Model π *conceptualizes all possible events* (relative to π^*) if π assigns positive probability to all finite histories of outcomes that occur with positive probability under π^*.

Intuitively, when a person fails to conceptualize an event, he is not on the lookout for that event.[21] Indeed, Part 1 of Proposition 1 demonstrates that the person need not confront his failure to conceptualize possible events: under the minimal SAS identified there, the person only notices actions

[19] In terms of the categorization of noticing strategies above, this SAS corresponds to coarsely representing outcomes.

[20] Proposition 1 also suggests that decision problems that require the person track his past behavior to inform his current action promote the discovery of errors. In the self-control example, we implicitly assumed the person faces a sequence of independent decisions: he chooses each day whether to visit the gym regardless of what happened before. Matters can be different if his decisions depend on what he's done before. Suppose he faces a once-and-for-all task that can be performed at his convenience. Whether he has done it yet influences his optimal action today. E.g., if the task is signing up for TSA PreCheck, then whether he has done so determines which queue he selects at the airport. In this case, his optimal action today (e.g., the queue he chooses) reveals whether he has completed the task yet. As time goes on, the person is increasingly surprised by the fact that he has not yet completed the task. And if he persistently procrastinates, he will eventually confront the fact that he has delayed much longer than explicable under his erroneous model. Thus, the SAS that tracks his optimal action will *not* be stable (as in Part 3 of Proposition 1).

[21] Conceptualizing an event is a more descriptive term for the statistical notion of absolute continuity. Recall that if π^* is absolutely continuous with respect to π, then π assigns positive probability to all events that are possible under the true model. Thus, if a model π fails to conceptualize an event that has positive probability under π^*, then π^* is not absolutely continuous with respect to π.

that he deemed plausible ex ante. It turns out that filtering out seemingly zero-probability events is a feature of *all* minimal SASs.

Proposition 2. *If ϕ is a minimal sufficient attentional strategy given π, then π assigns positive probability to all finite noticed histories that occur with positive probability under π^* and ϕ.*

Proposition 2 shows that for *any* misspecified model π, a minimal SAS always filters the data in a way that prevents the person from noticing outcomes he assumed impossible and thus from discovering his failure to conceptualize possible events. Any coarsened history he might notice will have positive probability under his model, π. Intuitively, the person sees no benefit in distinguishing events he assigns zero probability from those he assigns positive probability. In Example 5 on spending (Section 2.3), for instance, a person who is wrong but dogmatic about the total price of a good (e.g., neglecting add-on costs) need not notice he's wrong.[22]

This result shows the powerful role that channeled attention plays in potentially blocking people from noticing their erroneous models. Many alternative approaches to explaining the persistence of misspecified models, discussed in the next section, require that a person only sees data consistent with their misspecified model. Our framework, on the other hand, helps explain the many situations where wrong models persist even when readily-available data clearly refute them: When a person economically channels their attention, this is precisely the sort of data that goes unnoticed.

3.3 Alternative Stability Criteria

This section relates our attentional stability criterion to alternative approaches. The plainest and most pervasive reason beyond channeled attention why people may not recognize that their models are wrong is that it's hard to figure out everything in life, so people might lack the necessary data to distinguish their model from truth. This could be because they don't experiment enough to learn their model is wrong (e.g., Gittins (1979), Fudenberg and Levine (1993)). Or because the available data doesn't allow people to reliably distinguish between models (Ba, 2026). It may be that people employ misguided data-gathering strategies, engaging in confirmatory-search strategies per Wason (1968), or ones that overweight the immediate costs of exploration relative to more distant benefits. Or, finally, it could simply be because the flow of data is infrequent, as arguably true for big durable purchases. Our focus instead is on the many scenarios where people have sufficient data to discover their mistakes yet fail to do so, such as when fresh eyes catch mistakes.[23]

[22] In the literature on mistaken beliefs, a common fix to prevent people from confronting seemingly zero-probability events is to create environments where no such events happen. For early examples, see Barberis et al. (1998), where this condition holds in a natural way, and Rabin (2002), where features of the model are contrived to rule out such events. Our approach offers justification for another common approach: simply ignoring the issue.

[23] Of course, a broader model could combine the data representation margin of channeled attention with the data generation margin of, e.g., self-confirming equilibrium.

Additionally, people might fully attend to enough data to notice their errors, yet nevertheless analyze that data in an incorrect or incomplete way. This could be because of biases in statistical reasoning (see Benjamin, 2019 for a review) or motivated reasoning (see Bénabou and Tirole, 2016 for a review).[24] It could also be because people don't know the correct statistical tests to run, or because it's computationally challenging to discover regularities in a dataset (Aragones et al., 2005). Relative to these explanations based on barriers to processing information, we emphasize how complexity sometimes *helps* people recognize their errors by increasing engagement with feedback, and how people can be statistically naive in one domain yet become well-calibrated in a similarly complex domain (e.g., Green et al., 2019). Moreover, we emphasize how people can make persistent mistakes even when they don't involve ego or motivation.

Finally, there is a class of frameworks emphasizing that even when people access the relevant data and process it correctly, they may still continue to act on the wrong model. This could be because they anticipate insufficient benefit from changing their model. For example, their model could be "evolutionarily stable" in the sense that alternatives allow people to better explain their observations but don't yield higher payoffs than the original model (Fudenberg and Lanzani, 2023).[25] Alternatively, it could be that people don't know how to use a new model to guide decisions, perhaps because they haven't yet spent the time or attentional costs to do so (Reis, 2006). Or it could be that they don't want to acknowledge certain conclusions despite knowing they are true. Relative to this literature, we shed light on scenarios where people repeatedly make very *costly* mistakes.

Overall, our framework speaks to examples where the key constraint to people recognizing their error is that they don't notice they're making it. They have the data, they could in principle (cheaply) process it, and they would believe in and act on the conclusion if they saw it—but they don't see it. The next section takes up comparative statics: what are pathways to incidental learning?

4 Pathways of Incidental Learning

This section analyzes forces that influence the attentional stability of errors in particular contexts.[26] Section 5 then characterizes errors that are particularly robust to incidental learning across contexts.

[24] Some of these statistical errors, such as base-rate neglect (Benjamin et al., 2019) or the non-belief in the law of large numbers (Benjamin et al., 2016) predict cases where even full attention to infinite data does not lead to learning the right model. Others, such as confirmation bias (Rabin and Schrag, 1999), predict a tendency to reinforce original theories even when they're wrong. Due to motivated reasoning, people may also update in a self-serving manner to maintain an optimistic view of themselves (Bénabou and Tirole, 2002; Gottlieb, 2019; Möbius et al., 2022).

[25] Our question of when a person rejects her theory also connects to related models of paradigm shifts and "testability" (e.g., Hong et al., 2007; Ortoleva, 2012; Al-Najjar and Shmaya, 2015). While studying paradigm shifts has the flavor of analyzing when people wake up to errors, those papers do not study the interaction between waking up and inattention.

[26] The literature is often silent on whether a particular error is "local" or "global"; that is, whether a person must correct an error in one context if he notices it in another. Some evidence suggests that people in important instances fail to port their expertise across similar contexts (e.g., Green et al., 2019).

4.1 The Role of Tracking Statistics

Incentives to track statistics are often necessary for incidental learning. Such incentives may be external (e.g., performance plans) or inherent to a situation (e.g., a person needs to track whether he previously enrolled in a service, such as TSA pre-check, to know whether he can use it). In the spirit of Simons and Chabris (1999), where a potentially shocking gorilla often goes unnoticed, we use the term *statistical gorillas* to refer to statistics with the potential to refute one's model.

Definition 6. Let ϕ be a SAS given π. *Rejecting π when following ϕ requires noticing a long-run statistical gorilla* if and only if for every period t,

$$\frac{P(n^t|\pi,\phi)}{P(n^t|\pi^*,\phi)} > 0 \text{ with probability 1 when following } \phi. \tag{2}$$

Definition 6 says that rejecting π when following ϕ requires noticing a long-run statistical gorilla if the likelihood of n^t under π versus π^* remains strictly positive, except in the limit as $t \to \infty$. This includes the relevant case where a person only attends to data (outcomes) from the current period and thus does not combine data across periods. The collection of data across time that in the limit reveals π is false represents a "statistical gorilla". Of course, if rejecting π when following ϕ *does not* require noticing a long-run statistical gorilla, then ϕ is not a stable attentional strategy given π.

We now provide conditions for when noticing a statistical gorilla is required for incidental learning.

Proposition 3. *Consider a SAS ϕ given π.*

1. *If π conceptualizes all possible events (relative to π^*), then rejecting π when following ϕ requires noticing a long-run statistical gorilla.*

2. *If ϕ is a minimal SAS given π, then rejecting π when following ϕ requires noticing a long-run statistical gorilla.*

The first part of Proposition 3 says that rejecting π requires long-run attention whenever π conceptualizes all possible events. Intuitively, if the person anticipates the correct set of events, no finite event provides a "Eureka" moment that allows him to reject his model. Any rejection must instead come through the gradual accumulation of evidence.

The second part says that if the person follows a *minimal* SAS, then rejecting π requires long-run attention *regardless* of whether he conceptualizes all events or not. This is because any minimal SAS filters out seemingly impossible events (Proposition 2) and consequently eliminates the capacity for finite-event Eureka moments: events that could induce them will be ignored. The scope for rejecting one's model therefore becomes similar regardless of whether the person fails to conceptualize events or not. This underscores the importance of noticing long-run statistics for rejecting a misspecified

model under channeled attention: whenever the person fully economizes on attention (i.e., he follows a minimal SAS), the only path for discovering his mistake involves tracking statistics over time.

We next turn to factors that encourage such tracking.

4.2 The Role of Questioning Assumptions

Our next set of results clarify when and how greater uncertainty within one's model—reflecting uncertainty over the right assumptions to make—promotes incidental learning. These results hone a relatively simple intuition: errors that involve being more certain about the right assumptions are more stable than errors that involve being less certain. In the latter case, one has greater incentive to question and refine his model. Answering these questions from the data expands attention and offers more scope for discovering errors.

However, greater uncertainty does not *always* help with incidental learning. Our next proposition clarifies when uncertainty promotes or hinders the discovery of errors. The person begins with a model π, and we consider the stability of a refined model $\tilde{\pi}$ that results from endowing the person ex-ante with knowledge that some set of parameter values $\widetilde{\Theta}$ that were possible under π are in fact not true. Whether this reduction in uncertainty makes the model more or less stable depends on whether the person would have eventually learned that $\widetilde{\Theta}$ is false even without being endowed with this knowledge ex-ante.

Proposition 4. *Fix any environment* $\Gamma = (\Theta, \times_{t=1}^{\infty} X_t, \times_{t=1}^{\infty} Y_t, \times_{t=1}^{\infty} u_t, P, \pi^*)$. *Consider model* π *and a set of parameters* $\widetilde{\Theta} \subset \text{supp}(\pi)$. *Let* $\tilde{\pi} = \pi(\cdot|\text{supp}(\pi) \setminus \widetilde{\Theta})$ *be the resulting model when the parameters* $\widetilde{\Theta} \subset \text{supp}(\pi)$ *are known to be false. Suppose there is a StAS* ϕ *under* π *such that:*

1. $\widetilde{\Theta}$ *is rejected when following* ϕ*; that is,* $\Pr(\theta|n^t, \phi) \to 0$ *almost surely for every* $\theta \in \widetilde{\Theta}$.
2. ϕ *remains a SAS under* $\tilde{\pi}$ *and* $\tilde{\pi}$ *assigns positive probability to all finite noticed histories that occur under* π^* *and* ϕ.

Then ϕ *is also a StAS under* $\tilde{\pi}$. *The converse does not generally hold: the restricted model* $\tilde{\pi}$ *may admit a StAS even when the larger model does not.*

Proposition 4 identifies a form of uncertainty that helps incidental learning. It says that broadly if there is a stable attentional strategy for a model, then there is also a stable attentional strategy for a refined version of the model that ex-ante rules out parameters that become rejected under the original model's stable attentional strategy. But the converse is not true. This means that the process of rejecting parameters through experience (weakly) increases the scope for incidental learning.

However, greater uncertainty can also hurt. As we saw in the discussion of Example 2 (production problem) following Proposition 1, endowing the farmer with knowledge that he wouldn't figure out

on his own (the optimal success rate) encourages incidental learning.[27] More broadly, if there is a stable attentional strategy for a model, then there is not necessarily a stable attentional strategy for a refined version of the model that ex-ante rules out parameters that *won't* be rejected under the original model's stable attentional strategy. Entertaining parameters that are not rejected can reduce the scope for incidental learning because they offer alternative (but false) explanations for the noticed data. To summarize, eliminating uncertainty over parameters a person would eventually rule out on his own may hinder the discovery of an error because it reduces his perceived need for attention. On the other hand, eliminating uncertainty over parameters the person would not rule out may help with the discovery of an error because it limits the scope for adopting false beliefs.

Proposition 4 reflects the findings of Esponda et al. (2024), who experimentally examine the persistence of base-rate neglect among participants who face a repeated prediction problem and have access to rich feedback on the history of outcomes. Participants in a treatment where they are told some parameter values underlying the data-generating process are significantly less likely to use the observed data to correct their tendency for base-rate neglect relative to a treatment where they do not know the underlying parameters and must learn them from experience. That is, less-informed participants—those who do not receive details about the problem ex ante—pay more careful attention to the data and are thus more likely to incidentally notice their error.

4.3 The Role of Asking Questions of Available Data

With channeled attention, the availability of data has a limited effect on whether someone rejects their faulty model. The important factor is what questions the person asks of the data they have. Intuitively, environments that reduce the need to track and discern features of the data reduce the scope for incidental learning.

Consider the impact of being able to delegate decision-making to a third party or an algorithm. For example, imagine a ride-share firm that follows a pricing algorithm based on estimated consumer demand. As in Strulov-Shlain (2022) and List et al. (2023), suppose consumers suffer from "left-digit bias" and are insensitive to the cents component of the price—their demand responds discontinuously to an increase in price from, say, \$14.99 to \$15.00. The firm neglects this discontinuity and employs a pricing algorithm based on a smooth model of demand. If the firm merely follows the pricing recommendations from their flawed algorithm, they will bypass the data necessary to learn. In this case, one SAS for the firm involves only noticing each period that they're using the algorithm, which is clearly insufficient to reveal their mistake. By contrast, if the firm lacked access to the algorithm and had to carefully analyze consumer demand each period to set prices, they may instead discover

[27] In this example, the baseline model is π and $\widetilde{\Theta}$ includes all vectors $(\theta(w))_{w\in W}$ in the support of π such that $\max_{w\in W}\theta(w) \neq \theta^*(w^*,z^*)$. As we saw under Proposition 1, the farmer will reject his refined model $\tilde{\pi}$. In this case, then, greater uncertainty hinders incidental learning.

their error.

The following proposition captures these intuitions.

Proposition 5. *Fix any environment* $\Gamma = (\Theta, \times_{t=1}^{\infty} X_t, \times_{t=1}^{\infty} Y_t, \times_{t=1}^{\infty} u_t, P, \pi^*)$ *and a model* π*. Consider a modified environment identical to* Γ *aside from allowing the person to select an option each period that implements a subjectively optimal strategy: in the modified environment, the action space each period is* $X_t \cup \{d\}$*, where selecting* d *implements a subjectively optimal behavioral strategy. There is always a stable attentional strategy given* π *in the modified environment.*

This result says that if the environment is such that a person can always delegate her decision to another person or algorithm that implements her subjectively optimal strategy, then there is a stable attentional strategy. Outsourcing attention precludes incidental learning.

The logic is illustrated by the following whimsical twist on "a needle in a haystack". If a person can ask an algorithm whether the needle she's searching for is in a haystack, then she need not notice the gorilla lying in wait below the surface of the hay. If she instead has to search for it herself, then she will inevitably confront the gorilla. The benefits of delegation or using algorithms are obvious (e.g., reducing noise and effort in decision-making), but our analysis identifies an often-overlooked cost: relying on others to examine the data and answer the questions we *think* we should be asking prevents us from recognizing that we're asking the wrong questions.

The results above suggest that rich data (e.g., algorithms providing tailored recommendations) can hinder incidental learning by relieving the decision maker from scrutinizing the data. Richer feedback may similarly reduce the likelihood of discovering an error. Consider the tendency to neglect correlations in others' opinions (as in DeMarzo et al., 2003; Eyster and Rabin, 2010, 2014; Enke and Zimmermann, 2017). Building on the advice example (Example 1) from the introduction as well as Section 2.3, Appendix C.3 considers a person who seeks advice on the optimal action to take each period. Over time, he updates his beliefs about the quality of advice that each advisor provides. However, he wrongly treats their advice as conditionally independent, ignoring that they may share a common source of information or talk with each other. Whether the person discovers this error depends on the amount of feedback he receives. If he always observes whether advice was useful or not ex post, as in the introductory example, then he can learn about each advisor individually by comparing their recommendations with the realized outcomes. This strategy does not attend to correlations across advisors' advice, and thus promotes stability of his mistaken model. In contrast, if he does not always observe whether advice was useful or not, for example if the realized return on investment only becomes clear over a long period of time, then any SAS requires him to notice the correlation across others' advice: in the absence of direct feedback, the efficient way to update about the quality of an advisor is to "benchmark" their advice against other advice believed to be high quality. Thus, with limited feedback, the person is more likely to notice that the advisors' advice is correlated relative to the case with perfect feedback.

4.4 The Ambiguous Role of Costs

The previous results identified factors that help or hinder incidental learning. Our next result shows that, contrary to intuitions from the rational-inattention literature referenced in the introduction, the utility cost of an error is not one of these factors. There is not a tight link between whether a person discovers an error and its cost: for any misspecified model, there exist contexts where it is (arbitrarily) costly yet stable, and others where it is unstable despite being costless.

To formalize this context dependence, we separate the environment into two components given Assumption 1: the *outcome environment*, $(\times_{t=1}^{\infty} Y_t, \Theta, P, \pi^*)$, which describes possible distributions over outcomes, and the *choice environment*, $(\times_{t=1}^{\infty} X_t, \times_{t=1}^{\infty} u_t)$, which describes the action space and utility function. It is straightforward to see that for any outcome environment and model π, there exists *some* choice environment in which π is attentionally stable. For instance, if u_t is independent of observables, then the person has no incentive to notice data and thus π is stable.

More interestingly, there is always a choice environment in which π is both stable *and costly*. This means that the person's limiting behavior under a SAS given π is suboptimal relative to his limiting behavior under a SAS given the true model, π^*. The counterpoint is often true as well: for a large class of misspecified models π, there will exist a choice environment where π is necessarily attentionally unstable despite being costless. To state these results, we first define a costly error. Consider a SAS ϕ. For a history of observables y^t and current signal s_t, let h_ϕ^t denote the history consistent with (s_t, y^t) when following ϕ. The expected utility under ϕ given θ^* is then denoted by $\bar{u}_t(\phi|(s_t, y^t)) := \mathbb{E}_{(\theta^*,\phi)}[u_t(x,y|h_\phi^t)|(s_t, y^t)]$.

Definition 7. Let ϕ be a SAS given model π and ϕ^* be any SAS given π^*. The SAS ϕ is *costless* if $|\bar{u}_t(\phi|(s_t, y^t)) - \bar{u}_t(\phi^*|(s_t, y^t))| \to 0$ almost surely given ϕ and θ^*. When the SAS ϕ is not costless it is *costly*. When ϕ is also a stable attentional strategy, it is then a *costly stable attentional strategy*.

Proposition 6. *Consider an outcome environment* $(\times_{t=1}^{\infty} Y_t, \Theta, P, \pi^*)$ *and a model* π.

1. *There exists a choice environment with an arbitrarily costly stable attentional strategy given* π.
2. *If* π *conceptualizes all possible events (relative to* π^**) then there exists a choice environment in which every SAS given* π *is costless, yet none is stable.*

Part 1 of Proposition 6 says that for any outcome environment and wrong model, there are objectives that lead to a costly StAS. Part 2 additionally says that if that wrong model conceptualizes all truly possible events, then there also exist objectives where the error is costless yet there is *no* StAS. This result shows how channeled attention overturns a common intuition that a decision maker would discover his misspecification when it's costly yet remain ignorant when it's inconsequential.

The costly attentionally stable strategies identified in Proposition 6 can be arbitrarily costly. A costly mistake persists in our framework only when the person wrongly deems valuable data useless

and ignores it (e.g., the relationship between pod size and yield). The true value of this data—which determines cost of the person's mistake—has no influence on his decision to ignore it. Hence, no matter how great the true benefit of some data, the decision-maker may continually ignore it if his *perceived* benefit is sufficiently small.

Proposition 6 provides two messages. First, for many mistakes, whether a person discovers them or not depends on the situation. There are situations where a person does not wake up to a mistake even though it leads to poor decisions and others where the person *does* wake up to that same mistake. Second, this situation dependence is not closely tied to the cost of the mistake precisely because the person does not notice he is making it. This result aligns with experimental findings by Enke et al. (2023) showing that very high stakes don't significantly reduce certain widely documented cognitive biases (e.g., base-rate neglect and anchoring) in the settings they study. Higher stakes reduce within-model errors, but not necessarily the use of wrong models.

5 The Scope and Limits of Incidental Learning and Intervention

While the previous section showed that errors may be susceptible to incidental learning in certain situations, some classes of errors are more resistant to incidental learning across situations. This section explores two notions of robustness. First, we analyze which errors remain stable for any objective function. Second, we analyze when an error remains stable under perturbations that expand attention—either temporary shocks to what the person notices or a permanent shift in the theory guiding what he notices. We provide a result that synthesizes these two forms of robustness by showing that no erroneous model is robust in both senses. For any error there is some intervention that jolts incidental learning—an *incidental intervention*. Hence, in addition to shedding some light on the types of errors that are more likely to persist due to channeled attention, this section demonstrates a limit to this persistence: an error that is resistant to one class of shocks is fragile to another.

5.1 Errors that Are Robust Across Objectives

We first examine which types of erroneous models are stable for any objective function the person may have. To begin, we focus on environments that meet Assumption 2 and are "stationary" in the following sense.

Definition 8. The environment is *stationary* if X_t, Y_t, and u_t are independent of t and $P(y_t|x_t, h^t(\neg s_t), \theta) = P(y_t|x_t, \theta)$ for all $t \in \mathbb{N}$, $y_t \in Y_t$, $x_t \in X_t$, $h^t \in H^t$, and $\theta \in \Theta$.

When the environment is stationary, we denote the fixed action space, outcome space, and utility function without subscripts: X, Y, and u.

Definition 9. Restrict attention to stationary environments where Assumption 2 holds. Fixing the outcome environment, a model π is *stable across stationary objectives* if for any choice environment (action space X and objective $u : X \times Y \to \mathbb{R}$) there exists a stable attentional strategy given π.

Our next result considers when a model π is stable across stationary objectives. The result uses the concept of minimal sufficient statistics from probability theory (e.g., Lehmann and Casella, 1998), which depends on π's predicted probability distributions over outcomes. Given our stationarity assumption, y_t is i.i.d. conditional on θ with distribution $P(\cdot|\theta)$. Let $Y(\theta)$ denote the support of $P(\cdot|\theta)$ and let $Y(\pi) := \cup_{\theta \in \text{supp}(\pi)} Y(\theta)$. Consider a partition with elements

$$m_\pi(y) := \left\{ y' \in Y(\pi) \cup Y(\theta^*) \mid \exists g(y', y) > 0 \text{ s.t. } \forall \theta \in \text{supp}(\pi), P(y'|\theta) = P(y|\theta) g(y', y) \right\}. \quad (3)$$

In words, this partition is a *minimal sufficient statistic* for updating beliefs about θ: $m_\pi(y)$ lumps y' together with y if and only if, under π, the person updates the same way upon noticing y' or y. We analogously define minimal sufficient statistics over histories of outcomes $y^t = (y_{t-1}, y_{t-2}, \ldots, y_1)$, denoted by $m_\pi(y^t)$.[28] As a minor technicality, we assume that $P(y|\theta) \in (0,1)$ for all $(y, \theta) \in Y(\pi) \times \text{supp}(\pi)$, which ensures that the partition in (3) is fixed over time.[29]

Proposition 7. *Consider a stationary environment where Assumption 2 holds, and suppose* $P(y|\theta) \in (0,1)$ *for all* $(y, \theta) \in Y(\pi) \times \text{supp}(\pi)$. *A model* π *is stable across stationary objectives if there exists* $\theta \in \text{supp}(\pi)$ *such that with probability one*

$$\liminf_{t \to \infty} \frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)} > 0. \quad (4)$$

Roughly put, π is stable across stationary objectives if the person would not be alerted to his mistake when he pays attention to all information helpful for updating his beliefs about θ under π.[30] The logic is straightforward. The person never finds it necessary to retain anything beyond information that changes his beliefs about θ. Therefore, if noticing $m_\pi(y^t)$ does not force him to reject π, then there is *no* stationary choice environment that will.

Proposition 7 offers an intuitive way to rank the stability of a model based on its minimal sufficient statistic. If m_π is strictly coarser than the minimal sufficient statistic to distinguish π from π^*, then π is stable across stationary objectives. Models are more likely to exhibit such stability when they fail to ask questions relevant for discovering they're wrong than if they ask such questions but anticipate

[28] We define the minimal sufficient statistic in terms of the history of outcomes rather than the complete history h^t to emphasize that this object is independent of a person's objective and action space.

[29] If $P(y|\theta) \in \{0,1\}$ for some $(y, \theta) \in Y(\pi) \times \text{supp}(\pi)$, then observing y can rule out parameter values and thus shrink $\text{supp}(\pi_t)$ relative to $\text{supp}(\pi)$. This would cause $m_{\pi_t}(y)$ to vary in t.

[30] The converse is also true under an additional assumption presented in Appendix B.2 ensuring a one-to-one map between the person's beliefs over θ and his predicted distribution over outcomes.

the wrong set of answers. Dogmatic models where π is degenerate represent one extreme case of this logic: since the person thinks they have nothing to learn, the minimal sufficient statistic distinguishes no data and thus π is stable across stationary objectives.

To illustrate, consider the investment-advice example from the introduction. Let $Y = \{0,1\}$, where $y_t = 1$ denotes the outcome where the investor receives correct advice in period t (i.e., $s_t = \omega_t$) and $y_t = 0$ denotes incorrect advice. Parameter $\theta \in [0,1]$ represents the probability of correct advice in each period. Suppose the investor's theory π over values of θ assigns positive weight to θ' and $\theta'' \neq \theta'$. Then $m_\pi(0) = \{0\}$ and $m_\pi(1) = \{1\}$ since $P(1|\theta)/P(0|\theta) = \theta/(1-\theta)$ depends on θ. Furthermore, letting $k(y^t)$ denote the count of successes in y^t, $m_\pi(y^t) = \{\tilde{y}^t \mid k(\tilde{y}^t) = k(y^t)\}$ since

$$\frac{P(\tilde{y}^t|\theta)}{P(y^t|\theta)} = \frac{\theta^{k(\tilde{y}^t)}(1-\theta)^{t-1-k(\tilde{y}^t)}}{\theta^{k(y^t)}(1-\theta)^{t-1-k(y^t)}}$$

is independent of θ if and only if $k(\tilde{y}^t) = k(y^t)$. But $m_\pi(y^t)$ is a sufficient statistic for rejecting *any* π (given our assumption that its support doesn't include θ^*). This of course hinges on π placing positive probability on at least two values of θ—otherwise, $m_\pi(y^t) = Y^{t-1}$ and π is stable across stationary objectives. Models that are "merely miscalibrated" (make the right distinctions but put weight on the wrong parameter values) are less likely to be stable than models that are more fundamentally misspecified (fail to make truly important distinctions).

It is tempting to oversimplify the message of Proposition 7. One might think that simply eliminating parameters from the support of a model will make it more likely to be stable across objectives because the person has (weakly) less to learn. Surprisingly, this is not always the case, echoing the earlier message of Proposition 4. Reducing uncertainty within a model can either promote stability or hinder it. The effect on stability crucially depends on how uncertainty is reduced and its interaction with the person's attention given his objective.

These subtleties notwithstanding, we can also describe classes of errors that are and aren't stable across stationary objectives. We provide formal propositions for this categorization in Appendix B.2. For instance, we show that two classes of models that are "overly narrow" relative to the true model are stable across stationary objectives. The first includes models with an overly-narrow view of potential outcomes—they deem some truly possible outcomes as impossible. Since the person thinks he knows these values never occur, he has no incentive to follow a noticing strategy that separately distinguishes them. The second includes models that treat some truly informative signals as uninformative. Even when the person correctly perceives the support of these signals, he sees no use in noticing them.

In contrast, we show that the mirror image of the previous two classes of models—models that are "overly broad" relative to the true model—are not stable across stationary objectives. This includes models where the person (i) thinks the set of possible outcomes is larger than it really is, and (ii)

is uncertain about the likelihoods of outcomes that are, in reality, impossible. Under objectives that require the person to predict outcomes, he will eventually notice that these impossible outcomes never occur and thus reject his model. This also includes models where the person (i) thinks the set of informative signals is larger than it really is, and (ii) is uncertain about the informativeness of some signals that are truly uninformative. In trying to learn how to use these signals, the person will notice that some are in fact uninformative.

Before turning to alternative notions of robustness, we extend our definition of robustness across objectives by relaxing stationarity and allowing for history-dependent action spaces and utility functions. We will use this stronger notion below to compare various forms of robustness.

Definition 10. Fixing the outcome environment, a model π is *stable across all objectives* if for any choice environment (specifying a sequence of action spaces X_t and utility functions $u_t : X_t \times Y_t \times H^t \to \mathbb{R}$), there exists a stable attentional strategy given π.

5.2 Errors that are Robust to Expansions of Attention

We now consider when an attentionally stable model remains stable when the person is induced to pay greater attention. We consider two forms of expanded attention, which differ in where the perturbation lands.[31] The first perturbs the noticing strategy directly, holding the person's model fixed. These shocks are temporary, leading to greater attention in the moment. The second perturbs the theory that guides attention, and represents a permanent expansion of attention.

We first consider the temporary shocks, which have two interpretations. First, and our preferred interpretation in this context, they could result from momentarily entertaining an alternative theory that calls for expanded attention (e.g., a passing thought). Second, they could represent exogenous constraints on channeled attention by introducing stimuli that are impossible for the person to ignore.

Definition 11. A noticing strategy $\mathcal{N}'$ is a *temporary-noticing refinement* of $\mathcal{N}$ if (i) there is a finite set of periods T such that $N'_t \neq N_t$ if and only if $t \in T$, and (ii) N'_t is a refinement of N_t for all $t \in T$.

We say that π is fragile to temporary-noticing refinements if any existing stable attentional strategy can be eliminated by some refinement.

Definition 12. Fix the outcome and choice environment. A model π is *fragile to temporary-noticing refinements* if π admits a StAS and any such StAS can be eliminated by a temporary refinement of its noticing strategy.

[31] We focus on perturbations that involve noticing *more* than the person's original noticing strategy; the case of less attentive perturbations is trivial in terms of overturning stability.

Our next result shows that no models are *both* stable across all objectives *and* robust to temporary noticing shocks. This provides a limit on the persistence of errors due to channeled attention. Models that are quite robust in one sense are necessarily fragile in another.

Proposition 8. *Fix the outcome and choice environment. Suppose* $(\mathcal{N}, \sigma)$ *is a stable attentional strategy given* π*. Then at least one of the following must be true.*

1. *There exists a temporary refinement* $\mathcal{N}'$ *of* $\mathcal{N}$ *such that there is no stable attentional strategy given* π *under* $\mathcal{N}'$*.*

2. *There exists an alternative choice environment for which there is no stable attentional strategy given* π*.*

This result hinges on the fact that any model that is stable across all objectives must fail to conceptualize an event, making it vulnerable to shocks in attention that expose these events.

Corollary 1. *Consider an outcome environment* $(\times_{t=1}^{\infty} Y_t, \Theta, P, \pi^*)$ *and a model* π*. If* π *is stable across all objectives, then* π *fails to conceptualize an event (relative to* π^**).*

Corollary 1 builds on the results from Section 5.1 by showing that the models most robust to differing objectives are those that fail to anticipate the correct set of outcomes. Models with large omissions—failing to even realize that some event is possible—are more stable to changes in a person's objectives than models that conceptualize all possible events. In order for an error to be stable across *all* objectives, it must persist in any environment where the person has an incentive to track all patterns that he deems possible. If he's correct about which patterns are possible, he discovers any statistically identifiable mistake with such an incentive. Failing to conceptualize an event is thus necessary for π to remain immune to such incidental learning across all possible objectives. While such a failure is not strictly sufficient, there are many examples where π fails to conceptualize an event and *is* stable across all objectives.[32]

Our results on stability across objectives may shed light on how, even in data-savvy organizations made up of individuals with different objectives, errors might persist in the face of data that reveals those errors. Take the example of how Lyft for years missed signals of the benefits of more aggressive 99-cent pricing strategies:

> [O]ne might wonder why Lyft's suite of sophisticated machine-learning tools was unable to pick up such a sizeable discontinuity in demand ... [T]he institutional details of

[32]Consider a stationary outcome process with no signals and $Y = \{0, 1, 2\}$. Under the truth, $P(y_t = 1|\theta^*) = 1$. Suppose the person's model π is dogmatic about a parameter θ. If he is dogmatic that $P(y_t = 0|\theta) = 1$, then π is stable across all objectives. However, if $P(y_t = 0|\theta) = P(y_t = 2|\theta) = 1/2$ and $P(y_t = 1|\theta) = 0$, then it is easy to check that π fails to conceptualize an event but is not stable relative to an objective that rewards him for reporting the entire history correctly.

> how the machine learning system was set up at Lyft stacked the cards against finding an effect. Specifically, in order to deal with the inherently high-dimensional problem of predicting demand based on a large number of covariates, Lyft's machine learning system must impose a number of both explicit and implicit prior beliefs on the regularity of the relationship between those covariates and demand. A key restriction that is especially important for our purposes is that Lyft's machine learning systems assumed from the outset that demand curves were locally well approximated by a constant elasticity functional form conditional on covariates. This sort of regularity restriction was helpful for Lyft because the imposed smoothness of the demand curve helped stabilize the final price that the pricing algorithm produced, but it clearly biased the overall system against finding evidence of left-digit bias. Our view in light of this discussion is that these modeling choices by Lyft seemed ex ante to be a sensible approximation for helping the machine learning algorithm converge. Only ex post (i.e. given what we know from this present study) was it obvious that these modeling choices prevented Lyft from finding what turned out to be an important effect. (Page 3221 of List et al. (2023).)

In other words, Lyft's managers implicitly or explicitly thought demand would be smooth, estimated a model that assumed such smoothness, and, despite large stakes and a variety of objectives within the organization, didn't confront the data that screamed out this assumption was wrong.[33]

To shed light on pathways to learning in such cases (where part 1 of Proposition 8 applies), we next turn to the second form of expanded attention. Here, the person does not entertain an alternative model in passing, but permanently adopts it and uses it to guide his attention. We show that being vulnerable to temporary expansions of attention is equivalent to being vulnerable to perturbations of this kind toward a model π' that conceptualizes a broader set of events.

Definition 13. Fix the outcome and choice environment. A model π is *fragile to theory perturbations toward* π' if π admits a StAS but for every $\alpha \in (0,1)$ the perturbed model $\pi'' = (1-\alpha)\pi + \alpha\pi'$ admits no StAS.

Proposition 9. *Consider an outcome environment* $(\times_{t=1}^{\infty} Y_t, \Theta, P, \pi^*)$ *and a model* π*. The following are equivalent:*

1. *For any choice environment in which* π *admits a StAS,* π *is fragile to temporary-noticing refinements.*

[33] To literally map this example to a failure to conceptualize, treat a randomized pair of prices as exogeneous and a single period as a large experimental block in which the firm observes exact conversion rates from offers to rides at prices immediately below and at a dollar threshold, (q^-, q^+). Suppose every parameter in the smooth model satisfies $q^- - q^+ \leq \varepsilon$, while under the truth $q^- - q^+ = \Delta > \varepsilon$ almost surely (given π^*). Then the event $E = \{q^- - q^+ > \varepsilon\}$ has probability zero under π and probability one under π^*. Previewing how the firm eventually discovered the missed 99-cent pricing opportunity, a theory-guided analysis that separately reports whether E occurred is a temporary refinement that eliminates a stable attentional strategy.

2. *There exists a choice environment in which π is fragile to theory perturbations toward every identifiably wrong model π' that conceptualizes all possible events (relative to π^*).*

Both conditions hold if and only if π fails to conceptualize an event (relative to π^*).

This equivalency, along with Proposition 8, shows that a model that is stable across all objectives becomes susceptible to incidental learning when the model is perturbed so that it conceptualizes events. This implies that actively entertaining some forms of alternative models (i.e., those that conceptualize all events) will expand attention in a way that induces incidental learning.

Returning to the Lyft example, this is consistent with List et al. (2023)'s narrative for how they eventually discovered the large discontinuity of demand at round dollar values:

> Without a *theoretical framework* to guide this [99-cent pricing] analysis, it was difficult for Lyft to recognize that the failure to detect an effect [in earlier underpowered analyses] was due to insufficient data, not necessarily due to the absence of an effect. (Page 3221 of List et al. (2023), emphasis added.)

In line with our previous results, List et al. (2023) suggest that better theory helped Lyft discover and act on important patterns in the data that weren't conceptualized by their earlier theory. Rich data were not enough.

6 Further Applications and Principles

Fresh Eyes. Why do people benefit from outside opinions? Organizations often hire consultants in part to identify opportunities to reduce wasteful spending. Economists present papers they've worked on for years in part to receive helpful feedback. Patients and doctors seek second opinions in medical diagnoses. This application considers when and how outsiders who—relative to insiders—lack information on some dimensions of a problem will nevertheless be *more* likely to discover mistakes on other dimensions of that problem. As a recent paper on the value of diagnostic teams in medicine puts it, "fresh eyes catch mistakes" (Graber et al., 2017). This section captures and clarifies this intuition.

To study the question of when "fresh eyes" are helpful, we compare the models of an "insider" and an "outsider". The insider has a misspecified model π. The outsider is unsure if the appropriate model is π or some π'. Entertaining both possibilities, the outsider's model is $\pi'' = (1-\alpha)\pi + \alpha\pi'$ for some $\alpha \in (0,1)$. When will π be stable for the insider yet rejected by the outsider?

The interesting case here is when the insider correctly knows π' is false from the outset, while the outsider would come to learn π' is false over time. In this case, the outsider's broader model still does not encapsulate the true model—it merely includes an alternative π' that the insider already knows

is wrong.[34] Thus, we seek to answer when entertaining an additional false model, π', will help the outsider discover that π itself is misspecified.

The fact that the insider knows π' is false, while the outsider does not, captures the idea that the insider initially has superior information to the outsider. This information can induce the insider to pay less attention than the outsider. This resembles the scenario in Proposition 4. Suppose the outsider has a stable attentional strategy (given π'') that eventually rejects π'. Then whenever the insider finds this strategy sufficient, it will necessarily be stable (given π). But the converse is not true: the restricted model can be stable even when the broader model is not.

This result says that a person who correctly knows not to place weight on an erroneous model π' is less likely to reject π than a person who naively entertains both π and π'. Relative to the insider, the outsider is apt to follow a broader attentional strategy in attempt to distinguish π from π'. This endeavor can lead her to incidentally notice that both π' *and* π are false. Since the insider correctly knows that π' is false from the start, he does not engage in this additional scrutiny of the data. Interestingly, it is the outsider's task of discovering what the insider already knows that can reveal the insider's mistake.

Under channeled attention, seemingly irrelevant alternative models can have a striking effect on long-run beliefs. Under the classical Bayesian approach, incorporating alternative π' into the outsider's model would have no effect on her long-run beliefs whenever π' is unstable under full attention. Here, however, it can induce the outsider to discover her entire model is wrong, potentially leading to a large shift in beliefs. There is a value in playing devil's advocate and inspiring consideration of an alternative theory known to be false: forcing a person to disprove that alternative can teach them that their initial model is wrong as well.[35]

This result also suggests the value of turning (often implicit) assumptions into explicit hypotheses. While the insider assumes π to be true, the outsider treats π as a hypothesis relative to alternative π' and notices the necessary data to distinguish these hypotheses.[36]

Self-Unawareness. Psychologists and behavioral economists tend to focus on errors where people don't recognize mistakes in their *own* behavior or the consequences of that behavior. People are

[34] If instead π' *did* include the true model, then the outsider would trivially learn correctly.

[35] Our analysis of outsiders versus insiders may also shed light on the life-cycle of creativity. Work by psychologists and economists suggests that people tend to make radical "conceptual" innovations when they're young and more incremental "experimental" innovations when they're old (see, e.g., Galenson and Weinberg, 2000 and Weinberg and Galenson, 2005, and the cites therein). Our result here matches this pattern when we think of the young as outsiders holding π'', the old as insiders holding π, conceptual innovations as waking up to a theory being wrong, and experimental innovations as updating within a theory.

[36] This connects to a recent experimental literature that documents the value of encouraging entrepreneurs and others to think scientifically by explicitly formulating and testing hypotheses about outcomes of interest. In particular, a field experiment by Yang et al. (2025) shows that encouraging entrepreneurs to think in this way improves the accuracy of their forecasts about the revenue growth of their companies.

persistently naive about their prospective memory, their self control, and the heuristics and biases they display (see, e.g., (Bronchetti et al., 2023) for recent experimental work). How is this possible when people have such a rich set of data about themselves?

Our framework not only accommodates the possibility of persistent self-unawareness, as many examples above illustrate, but predicts a sense in which we're *less* likely to recognize mistakes in ourselves than in others, consistent with the so-called "bias-blindspot" in psychology (Pronin et al., 2002) and more recent work in experimental economics (Fedyk, 2025). The logic is similar to why fresh eyes are effective: needing to learn something (e.g., another person's preferences) that the other person already knows requires noticing more than them, which can enable incidental learning.[37]

Overall, we provide a reason why people may persistently be un-skilled and unaware of it (Kruger and Dunning, 1999): A failure to look at problems the right way compounds itself, especially when this failure concerns their own behavior that they think they understand.

The Role of Theory in Scientific Progress. Channeled attention crystallizes the role of theories and frameworks for both promoting and delaying scientific progress. On the one hand, channeled attention clarifies the necessity of theories or frameworks for organizing the enormous complexity of the world. Taking popular business-school frameworks as illustrations, there's something in common between the single-factor capital asset pricing model (CAPM) in finance, the five forces in strategy, and the 4Ps of marketing: they tell us which variables to focus on.[38]

On the other hand, our approach also clarifies how prevailing theories may channel attention away from data that are seemingly extraneous under those theories, delaying the recognition of alternative theories that better explain all the available data. Take the CAPM as an illustration. In a world where people pay attention to all information, we'd expect that professionals would, if anything, gravitate to overly complicated models with a "zoo" of factors to explain security returns, given the plethora of data available in finance (see Montiel Olea et al., 2022 for a recent formal argument along these lines). Yet the simplistic CAPM (single- and multi-factor versions) lives on in textbooks and in practice, though not in frontier finance research (e.g., Bordalo et al., 2025), perhaps in part because of channeled attention. Indeed, as discussed in Section 5.1, theories that omit factors are robustly attentionally stable, while theories that include too many factors are typically not.

We see something similar with the rise of "sufficient statistics" approaches in public finance (e.g.,

[37] Somewhat more formally, let $\Theta = \Theta^1 \times \Theta^2$, where Θ^1 corresponds to a feature of preferences and Θ^2 a feature of the environment. A person is dogmatic about his own θ^1, but an observer is uncertain about it. Our "fresh eyes" result from above then identifies cases where the observer is more likely to wake up to a misspecification than the person himself.

[38] Under the CAPM model for how markets price securities, the key variable is the sensitivity between a security's returns and the market return; in the five forces framework for assessing competitive forces in an industry, the key variables are the rivalry among existing competitors, the threat of substitute products or services, the bargaining power of suppliers, the bargaining power of buyers, and the threat of new entrants; in the 4Ps framework for analyzing marketing programs, the key variables are product, price, place, and promotion.

Chetty, 2009; Mullainathan et al., 2012) and "portable modeling" in applied (e.g., behavioral) economics (Rabin, 2013). While there are clear benefits to such approaches in terms of simplifying which statistics to estimate or how to apply models to new situations, a cost is that their application obviates the need to ask or answer questions that could incidentally alert us to misspecification. Models that assume demand curves trace out true willingness to pay (WTP) allow us to ignore data that likely correlate with WTP yet could invalidate this assumption; models of exponential discounting allow us to ignore most details in how the timing of consumption influences choices, such as the role that "now" vs. "later" plays in how people discount (e.g., Frederick et al., 2002); models of present bias focus on "now" vs. "later", but neglect how the state we're in (e.g., hungry vs. satiated) influences our decisions (e.g., Loewenstein et al., 2003); models of rational expectations allow us to sometimes neglect data on beliefs that defy rational expectations (e.g., Greenwood and Shleifer, 2014).[39]

7 Limitations, Extensions, and Further Implications

This paper develops a framework for investigating when people might notice their errors. We use this framework to partially characterize when people are more or less likely to discover their mistakes, and to investigate the stability of psychological biases and empirical misconceptions.[40]

In turning to several modifications and extensions worth considering, we note that our framework is built around a central theme that complements the emphasis of recent research on attention: a person's (potentially mistaken) prior beliefs critically influence the perceived benefits of paying attention and, consequently, the implications of limited attention. Our focus on cases where attentional costs are near zero highlights the centrality of such beliefs.[41]

This focus also stacks the deck *in favor* of a person noticing his errors, as does our study of repetitive choice contexts that provide sufficient data to identify the true model. In this sense, our at-

[39] As another example, models of errors designed for integration into economics, like all of our formal models, strip away distracting complexities so that economists can focus on the consequences of agents' errors. This makes these barriers universally non-salient to economists: without keeping in mind that the agents being studied are *not* focusing on those consequences, it can seem like such models assume agents are willfully ignoring useful information.

[40] Beyond the applications in this paper, several others have invoked our notion of channeled attention to explain the persistence of various biases and misconceptions, including overconfidence (Heidhues et al., 2018), interpersonal projection bias (Gagnon-Bartsch et al., 2021), the gambler's fallacy (He, 2022), inaccurate stereotyping in discrimination contexts (Bohren et al., 2025), and partial-equilibrium thinking (Bastianello and Fontanier, 2021).

[41] By analogy, consider a party host trying to locate an absent-minded friend who may have (yet again) run out of gas by forgetting to fill his tank. If she asks if anyone knows where her friend departed from, she might not welcome a guest interrupting to suggest that everybody *instead* try to calculate the gas mileage of the friend's car. And if another party-goer interjects "Hold on! We can't *begin* to understand where your friend is without first studying how combustion engines work!", then the host might rethink who she invites to parties as she sets off alone in search of her friend. The analogy is imperfect: one could imagine a (mean) host deeming "not here" a sufficient description of a missing guest, but it's hard to imagine an economist deeming "not the full-attention outcome" a sufficient description of the economy.

tentional stability criterion provides a "stress test": if a person does not discover his error in repetitive environments where attention is cheap, then he is unlikely to do so elsewhere.

A limitation inheres in our ambition to provide a sharp framework that researchers can broadly apply: other than our robustness results in Section 5, our analysis largely ignores how non-instrumental factors influence what draws attention. Especially when we focus on minimal attentional strategies, our framework permits agents to ignore non-instrumental data that is hard not to notice. To address this, the analyst could embed assumptions about what captures attention as a primitive, and our general framework allows such assumptions to be imposed on a case-by-case basis.[42]

Our main analysis also abstracts from interactions between attention and memory in situations where important information is not stored in databases, a focus of Schwartzstein (2014) as well as Bordalo et al. (2020). Without further restrictions on the noticing strategy, our analysis allows a person to notice today a feature of the data that emerged in the past, even if he did not previously notice it. In Appendix A we discuss refinements on SASs that incorporate assumptions on the interaction between attention and memory (e.g., requiring a person to notice statistics today in order to use them later). As we discuss there, much of our qualitative analysis remains unchanged under assumptions on memory interactions and imperfections, with the exception that some forms of imperfect memory (reflecting assumptions on imperfect encoding into memory) can actually *increase* the scope for incidental learning. In the main text, our assumption that a person is able to freely recall past information he currently deems useful makes the analysis transparent by isolating the impact of channeled attention. In our model, things go un-noticed because an agent does not see the *benefit* of noticing and remembering, which mitigates the impact that the technology of memory is likely to have on waking up.[43]

Turning to additional implications, there is a growing literature on how firms or governments could create incentives to exploit or ameliorate people's mistakes. Our framework suggests a related set of questions: how might outsiders design environments to channel people's attention in ways that prevent or encourage waking up? If the goal is to provide data to convince a person that his model π is wrong, our theory suggests that simply describing the correct model may have limited effect, even when people have an incentive to get things right and are exposed to diagnostic data. Because people don't think they need to attend to the data that would bear it out, the correct model may not be immediately compelling: merely being told the right answer does little, by itself, to redirect attention

[42] Attentional capture could be integrated into our framework by focusing solely on SASs that reflect the posited attentional capture instead of focusing on minimal SASs. Alternatively, we could preserve our focus on minimal SASs and modify the agent's payoffs so that they are rewarded for accurately reporting the object of attentional capture.

[43] In Appendix A we consider refinements on noticing strategies that increase the scope for incidental learning. But there are also realistic refinements that essentially do not change our results. For example, almost all our results for errors that conceptualize truly-important events carry over when requiring the person to notice an event when it happens in order to remember it later, yet allowing the person to freely revisit any *previously noticed* data if he currently deems it relevant.

toward that evidence.[44]

The limit just described has a revealing counterpoint. What matters for waking up is less the content of what a person is told than whether he is drawn into *actively* engaging with an alternative. Indeed, our theory predicts that such engagement helps even when the alternative is itself false. This prediction contrasts with that of standard Bayesian aggregation, where exposure to a false alternative should at best leave the agent's posterior over her own model unchanged (since the false alternative provides no positive evidence) and at worst slow his learning (by providing alternative explanations for surprising data). Channeled attention predicts instead that the act of evaluating the alternative requires the agent to attend to data he would otherwise have filtered, and that data may turn out to disprove his starting model as a by-product. An implication, developed in Section 6, is that an outsider who has strictly less information than an insider can systematically detect errors the insider has missed.

Finally, our theory highlights the importance of providing data that is relevant *within* the person's model. While many researchers find that debiasing people is particularly difficult (e.g., Soll et al., 2015), the difficulty might partially lie in the type of information debiasing campaigns choose to provide. Selecting information based solely on how much it would move people's beliefs *if it were processed* may be far less effective than targeting information that seems relevant given their biased beliefs.

References

Akepanidtaworn, Klakow, Rick Di Mascio, Alex Imas, and Lawrence Schmidt, "Selling Fast and Buying Slow: Heuristics and Trading Performance of Institutional Investors," *The Journal of Finance*, 2023, *78* (6), 3055–3098.

Al-Najjar, Nabil I. and Eran Shmaya, "Uncertainty and Disagreement in Equilibrium Models," *Journal of Political Economy*, 2015, *123*, 778 – 808.

Ali, S. Nageeb M., "Learning Self-Control," *Quarterly Journal of Economics*, 2011, *126* (2), 857–893.

Aragones, Enriqueta, Itzhak Gilboa, Andrew Postlewaite, and David Schmeidler, "Fact-Free Learning," *American Economic Review*, 2005, *95* (5), 1355–1368.

Augenblick, Ned, B. Kelsey Jack, Supreet Kaur, Felix Masiye, and Nicholas Swanson, "Retrieval Failures and Consumption Smoothing: A Field Experiment on Seasonal Hunger," Mimeo 2026.

Ba, Cuimin, "Robust Misspecified Models," *American Economic Review*, April 2026, *116* (4), 1340–79.

[44]This idea complements a different reason why people may not find the true model compelling, as highlighted by Schwartzstein and Sunderam (2021): when strategic persuaders tailor models to the data (i.e., propose models after seeing the data), they are able to propose false models that fit the data better than the true model.

Barberis, Nicholas, Andrei Shleifer, and Robert W. Vishny, "A Model of Investor Sentiment," *Journal of Financial Economics*, 1998, *49* (3), 307–343.

Bastianello, Francesca and Paul Fontanier, "Partial Equilibrium Thinking in General Equilibrium," Technical Report, Working Paper, Harvard University 2021.

Bénabou, Roland and Jean Tirole, "Self-Confidence and Personal Motivation*," *The Quarterly Journal of Economics*, 08 2002, *117* (3), 871–915.

_ and _ , "Willpower and Personal Rules," *Journal of Political Economy*, 2004, *112* (4), 848–886.

Bénabou, Roland and Jean Tirole, "Mindful economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, 2016, *30* (3), 141–164.

Benjamin, Daniel J, "Errors in Probabilistic Reasoning and Judgment Biases," *Handbook of Behavioral Economics: Applications and Foundations 1*, 2019, *2*, 69–186.

Benjamin, Daniel J., Aaron L. Bodoh-Creed, and Matthew Rabin, "Base-Rate Neglect : Foundations and Implications," Mimeo 2019.

_ , Matthew Rabin, and Collin Raymond, "A Model of Nonbelief in the Law of Large Numbers," *Journal of the European Economic Association*, 2016, *14* (2), 515 – 544.

Berk, Robert H., "Limiting Behavior of Posterior Distributions when the Model is Incorrect," *The Annals of Mathematical Statistics*, 1966, *37* (1), 51 – 58.

Beshears, John, Hae Nim Lee, Katherine L. Milkman, Robert Mislavsky, and Jessica Wisdom, "Creating Exercise Habits Using Incentives: The Trade-off Between Flexibility and Routinization," *Management Science*, 2021, *67* (7), 3985–4642.

Bohren, J. Aislinn, "Informational herding with model misspecification," *Journal of Economic Theory*, 2016, *163*, 222–247.

_ and Daniel Hauser, "Learning with Heterogeneous Misspecified Models: Characterization and Robustness," *Econometrica*, 2021, *89* (6), 3025–3077.

_ , Kareem Haggag, Alex Imas, and Devin G. Pope, "Inaccurate Statistical Discrimination: An Identification Problem," *Review of Economics and Statistics*, 2025, *107* (3), 605 – 620.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience Theory of Choice Under Risk," *The Quarterly Journal of Economics*, 2012, *127* (3), 1243–1285.

_ , _ , and _ , "Salience and Consumer Choice," *Journal of Political Economy*, 2013, *121* (5), 803–843.

_ , _ , and _ , "Memory, Attention, and Choice," *Quarterly Journal of Economics*, 2020, *135* (3), 1399–1442.

_ , _ , and _ , "Salience," *Annual Review of Economics*, 2022, *14*, 521–544.

_ , _ , **Giacomo Lanzani, and Andrei Shleifer**, "A Cognitive Theory of Reasoning and Choice," *The Quarterly Journal of Economics*, 2026, *141* (3), 1921–1963.

_ , _ , **Rafael La Porta, and Andrei Shleifer**, "Finance without Exotic Risk," *Journal of Financial Economics*, 2025, *173*, 104145.

Bronchetti, Erin T, Judd B Kessler, Ellen B Magenheim, Dmitry Taubinsky, and Eric Zwick, "Is Attention Produced Optimally? Theory and Evidence from Experiments with Bandwidth Enhancements," *Econometrica*, 2023, *91* (2), 669–707.

Camerer, Colin, Teck-Hua Ho, and Juin-Kuan Chong, "A Cognitive Hierarchy Model of Games," *Quarterly Journal of Economics*, 2004, *119* (3), 861–898.

Caplin, Andrew and Mark Dean, "Revealed Preference, Rational Inattention, and Costly Information Acquisition," *American Economic Review*, 2015, *105* (7), 2183–2203.

Carrera, Mariana, Heather Royer, Mark Stehr, Justin Sydnor, and Dmitry Taubinsky, "Who Chooses Commitment? Evidence and Welfare Implications," *The Review of Economic Studies*, 2022, *89* (3), 1205–1244.

Chater, Nick, *The Mind is Flat*, Yale University Press, 2018.

Chetty, Raj, "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods," *Annual Review of Economics*, 2009, *1* (1), 451–488.

Chun, Marvin M., Julie D. Golomb, and Nicholas B. Turk-Browne, "A Taxonomy of External and Internal Attention," *Annual Review of Psychology*, 2011, *62* (1), 73–101.

Crawford, Vincent P. and Nagore Iriberri, "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?," *Econometrica*, 2007, *75* (6), 1721–1770.

Dehaene, Stanislas, *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*, Viking Press, 2014.

Dekel, Eddie, Drew Fudenberg, and David K. Levine, "Learning to Play Bayesian Games," *Games and Economic Behavior*, 2004, *46* (2), 282–303.

DellaVigna, Stefano and Matthew Gentzkow, "Uniform Pricing in US Retail Chains," *Quarterly Journal of Economics*, 2019, *134* (4), 2011–2084.

_ **and Ulrike Malmendier**, "Paying Not to Go to the Gym," *American Economic Review*, 2006, *96*, 694–719.

DeMarzo, Peter M., Dimitri Vayanos, and Jeffrey Zwiebel, "Persuasion Bias, Social Influence, and Unidimensional Opinions," *Quarterly Journal of Economics*, 2003, *118*, 909–968.

Drew, Trafton, Melissa L. H. Võ, and Jeremy M. Wolfe, "The Invisible Gorilla Strikes Again," *Psychological Science*, 2013, *24* (9), 1848 – 1853.

Eliaz, Kfir and Ran Spiegler, "Contracting with Diversely Naive Agents," *Review of Economic Studies*, 2006, *73*, 689–714.

Enke, Benjamin and Florian Zimmermann, "Correlation Neglect in Belief Formation," *Review of Economic Studies*, 12 2017, *86* (1), 313–332.

_ , Uri Gneezy, Brian Hall, David Martin, Vadim Nelidov, Theo Offerman, and Jeroen van de Ven, "Cognitive Biases: Mistakes or Missing Stakes?," *The Review of Economics and Statistics*, 07 2023, *105* (4), 818–832.

Ericson, Keith M. Marzilli, "Forgetting We Forget: Overconfidence and Memory," *Journal of the European Economic Association*, 2011, *9* (1), 43–60.

Esponda, Ignacio, "Behavioral Equilibrium in Economies with Adverse Selection," *American Economic Review*, September 2008, *98* (4), 1269–1291.

_ and Demian Pouzo, "Berk-Nash Equilibrium: A Framework for Modeling Agents With Misspecified Models," *Econometrica*, 2016, *84* (3), 1093–1130.

_ , Emanuel Vespa, and Sevgi Yuksel, "Mental Models and Learning: The Case of Base-Rate Neglect," *American Economic Review*, 2024, *114* (3), 752–782.

Evans, George W. and Seppo Honkapohja, *Learning and Expectations in Macroeconomics*, Princeton University Press, 2001.

Eyster, Erik, "Errors in Strategic Reasoning," in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, eds., *The Handbook of Behavioral Economics: Applications and Foundations*, Vol. 2, North-Holland, 2019, chapter 3, pp. 187–259.

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

_ and Matthew Rabin, "Cursed Equilibrium," *Econometrica*, 2005, pp. 1623–1672.

_ and _ , "Extensive Imitation is Irrational and Harmful," *Quarterly Journal of Economics*, 2014, *129* (4), 1861–1898.

_ and Michele Piccione, "An Approach to Asset Pricing Under Incomplete and Diverse Perceptions," *Econometrica*, 2013, *81*, 1483–1506.

_ , Matthew Rabin, and Dimitri Vayanos, "Financial Markets Where Traders Neglect the Informational Content of Prices," *The Journal of Finance*, 2019, *74* (1), 371–399.

Fedyk, Anastassia, "Asymmetric Naïveté: Beliefs About Self-Control," *Management Science*, 2025, *71* (7), 6047–6068.

Frederick, Shane, George Loewenstein, and Ted O'donoghue, "Time Discounting and Time Preference: A Critical Review," *Journal of economic literature*, 2002, *40* (2), 351–401.

Frick, Mira, Ryota Iijima, and Yuhta Ishii, "Misinterpreting Others and the Fragility of Social Learning," *Econometrica*, 2020, *88* (6), 2281–2328.

Fryer, Roland and Matthew Jackson, "A Categorical Model of Cognition and Biased Decision-Making," *BEJ Theor. Econ*, 2008, *8* (1).

Fudenberg, Drew and David K. Levine, "Self-Confirming Equilibrium," *Econometrica*, 1993, pp. 523–545.

_ **and Giacomo Lanzani**, "Which misspecifications persist?," *Theoretical Economics*, 2023, *18* (3), 1271–1315.

Gabaix, Xavier, "A Sparsity-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 2014, *129* (4), 1661–1710.

_ **and David Laibson**, "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets," *Quarterly Journal of Economics*, 2006, *121* (2), 505–540.

Gagnon-Bartsch, Tristan and Antonio Rosato, "Quality Is in the Eye of the Beholder: Taste Projection in Markets with Observational Learning," *American Economic Review*, November 2024, *114* (11), 3746–87.

_ **and Benjamin Bushong**, "Learning with misattribution of reference dependence," *Journal of Economic Theory*, 2022, *203*, 105473.

_ **and Matthew Rabin**, "Naive Social Learning, Unlearning, and Mislearning," Mimeo 2021.

_ **, Marco Pagnozzi, and Antonio Rosato**, "Projection of Private Values in Auctions," *American Economic Review*, 2021, *111* (10), 3256–3298.

_ **, Matthew Rabin, and Joshua Schwartzstein**, "Channeled Attention and Stable Errors," Mimeo 2018.

Galenson, David W and Bruce A Weinberg, "Age and the Quality of Work: The Case of Modern American Painters," *Journal of Political Economy*, 2000, *108* (4), 761–777.

Gazzaley, Adam and Anna Christina Nobre, "Top-down Modulation: Bridging Selective Attention and Working Memory," *Trends in Cognitive Sciences*, 2012, *16* (2), 129–135.

Gittins, J.C., "Bandit Processes and Dynamic Allocation Indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1979, *41* (2), 148–177.

Goldsberry, Kirk, *Sprawlball: A Visual Tour of the New Era of the NBA*, Boston, MA: Houghton Mifflin Harcourt, 2019.

Gottlieb, Daniel, "Will You Never Learn? Self Deception and Biases in Information Processing," Mimeo 2019.

Graber, Mark L, Diana Rusz, Melissa L Jones, Diana Farm-Franks, Barbara Jones, Jeannine Cyr Gluck, Dana B Thomas, Kelly T Gleason, Kathy Welte, Jennifer Abfalter et al., "The New Diagnostic Team," *Diagnosis*, 2017, *4* (4), 225–238.

Green, Etan A., Justin M. Rao, and David M. Rothschild, "A Sharp Test of the Portability of Expertise," *Management Science*, 2019, *65* (6), 2820–2831.

Greenwood, Robin and Andrei Shleifer, "Expectations of Returns and Expected Returns," *The Review of Financial Studies*, 2014, *27* (3), 714–746.

Handel, Benjamin R. and Joshua Schwartzstein, "Frictions or Mental Gaps: What's Behind the Information We (Don't) Use and When Do We Care?," *The Journal of Economic Perspectives*, 2018, *32* (1), 155–178.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein, "Learning Through Noticing: Theory and Evidence from a Field Experiment," *Quarterly Journal of Economics*, 2014, *129* (3), 1311–1353.

He, Kevin, "Mislearning from Censored Data: The Gambler's Fallacy and Other Correlational Mistakes in Optimal-Stopping Problems," *Theoretical Economics*, 2022, *17* (3), 1269–1312.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack, "Unrealistic Expectations and Misguided Learning," *Econometrica*, 2018, *86*, 1159–1214.

Hong, Harrison, Jeremy C. Stein, and Jialin Yu, "Simple Forecasts and Paradigm Shifts," *The Journal of Finance*, 2007, *62* (3), 1207–1242.

Horowitz, Alexandra, *On Looking: A Walker's Guide to the Art of Observation*, Simon and Schuster, 2013.

Jehiel, Philippe, "Analogy-Based Expectation Equilibrium," *Journal of Economic Theory*, 2005, *123* (2), 81–104.

_ **and Frédéric Koessler**, "Revisiting Games of Incomplete Information with Analogy-Based Expectations," *Games and Economic Behavior*, March 2008, *62* (2), 533–557.

Kirman, Alan, "Learning by Firms About Demand Conditions," in Richard H. Day and Theodore Groves, eds., *Adaptive Economic Models*, Academic Press, 1975, pp. 137–156.

Kőszegi, Botond and Filip Matějka, "Choice Simplification: A Theory of Mental Budgeting and Naive Diversification," *The Quarterly Journal of Economics*, 2020, *135* (2), 1153–1207.

Kruger, Justin and David Dunning, "Unskilled and Unaware of it: How Difficulties in Recognizing one's own Incompetence Lead to Inflated Self-Assessments.," *Journal of personality and social psychology*, 1999, *77* (6), 1121.

Laibson, David I., "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 1997, *112* (2), 443–478.

Lehmann, Erich L. and George Casella, *Theory of Point Estimation*, Springer New York, 1998.

Li, Xiaomin and Colin F Camerer, "Predictable Effects of Visual Salience in Experimental Decisions and Games*," *The Quarterly Journal of Economics*, 08 2022, *137* (3), 1849–1900.

List, John, Ian Muir, Devin Pope, and Gregory Sun, "Left-Digit Bias at Lyft," *The Review of Economic Studies*, 2023, *90* (6), 3186–3237.

Loewenstein, George, Ted O'Donoghue, and Matthew Rabin, "Projection Bias in Predicting Future Utility," *Quarterly Journal of Economics*, 2003, *118* (4), 1209–1248.

Ludwig, Jens and Sendhil Mullainathan, "Machine Learning as a Tool for Hypothesis Generation," *The Quarterly Journal of Economics*, 2024, *139* (2), 751–827.

Madarász, Kristóf, "Information Projection: Model and Applications," *Review of Economic Studies*, 2012, *79* (3), 961–985.

_ , "Bargaining under the Illusion of Transparency," *American Economic Review*, November 2021, *111* (11), 3500–3539.

Malmendier, Ulrike and Devin Shanthikumar, "Are Small Investors Naive About Incentives?," *Journal of Financial Economics*, 2007, *85* (2), 457–489.

Matějka, Filip and Alisdair McKay, "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model," *American Economic Review*, 2015, *105* (1), 272–298.

Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat, "Managing Self-Confidence: Theory and Experimental Evidence," *Management Science*, 2022, *68* (11), 7793–7817.

Most, Steven B., Brian J. Scholl, Erin R. Clifford, and Daniel J. Simons, "What You See Is What You Set: Sustained Inattentional Blindness and the Capture of Awareness," *Psychological Review*, 2005, *112* (1), 217–242.

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

_ , "Thinking Through Categories," 2002. Mimeo.

_ **and Ziad Obermeyer**, "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care," *The Quarterly Journal of Economics*, 2022, *137* (2), 679–727.

_ **, Joshua Schwartzstein, and Andrei Shleifer**, "Coarse Thinking and Persuasion," *Quarterly Journal of Economics*, 2008, *123* (2), 577–619.

_ **,** _ **, William J Congdon et al.**, "A Reduced-Form Approach to Behavioral Public Finance," *Annual Review of Economics*, 2012, *4* (1), 511–540.

Nyarko, Yaw, "Learning in Misspecified Models and the Possibility of Cycles," *Journal of Economic Theory*, 1991, *55* (2), 416–427.

O'Donoghue, Ted and Matthew Rabin, "Doing It Now or Later," *American Economic Review*, 1999, *89* (1), 103–124.

_ **and** _ , "Choice And Procrastination," *Quarterly Journal of Economics*, 2001, *116* (1), 121–160.

Olea, José Luis Montiel, Pietro Ortoleva, Mallesh M. Pai, and Andrea Prat, "Competing Models," *The Quarterly Journal of Economics*, 2022, *137* (4), 2419–2457.

Ortoleva, Pietro, "Modeling the Change of Paradigm: Non-Bayesian Reactions to Unexpected News," *American Economic Review*, 2012, *102* (6), 2410–2436.

Payzan-LeNestour, Elise and Michael Woodford, "Outlier Blindness: A Neurobiological Foundation for Neglect of Financial Risk," *Journal of Financial Economics*, 2022, *143* (3), 1316–1343.

Pronin, Emily, Daniel Y Lin, and Lee Ross, "The Bias Blind Spot: Perceptions of Bias in Self Versus Others," *Personality and Social Psychology Bulletin*, 2002, *28* (3), 369–381.

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

_ , "An Approach to Incorporating Psychology into Economics," *American Economic Review*, 2013, *103* (3), 617–22.

_ **and Dimitri Vayanos**, "The Gambler's and Hot-Hand Fallacies: Theory and Applications," *Review of Economic Studies*, 2010, *77*, 730–778.

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

Reis, Ricardo, "Inattentive Consumers," *Journal of monetary Economics*, 2006, *53* (8), 1761–1800.

Sargent, Thomas J., *Bounded Rationality in Macroeconomics*, Oxford University Press, 1993.

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

_ **and Adi Sunderam**, "Using Models to Persuade," *American Economic Review*, 2021, *111* (6), 276–323.

_ **and Deepak Malhotra**, "Rocket Science," *Harvard Business School Case 921-043*, 2021.

Sial, Afras, Justin Sydnor, and Dmitry Taubinsky, "Biased Memory and Perceptions of Self-Control," 2022.

Simons, Daniel J. and Christopher F. Chabris, "Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events," *Perception*, 1999, *28*, 1059 – 1074.

Sims, Christopher A., "Implications of Rational Inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690.

Soll, Jack B., Katherine L. Milkman, and John W. Payne, "A User's Guide to Debiasing," in Gideon Keren and George Wu, eds., *The Wiley Blackwell Handbook of Judgment and Decision Making*, John Wiley and Sons, 2015, chapter 33, pp. 924–951.

Spiegler, Ran, "Bayesian Networks and Boundedly Rational Expectations*," *Quarterly Journal of Economics*, 2016, *131*, 1243–1290.

Strulov-Shlain, Avner, "More Than a Penny's Worth: Left-Digit Bias and Firm Pricing," *The Review of Economic Studies*, 12 2022.

Thaler, Richard H., "The Sports Learning Curve: Why Teams Are Slow to Learn and Adapt," Panel at the MIT Sloan Sports Analytics Conference with Daryl Morey, Bill James, and Sashi Brown March 2020. Video recording available from 42 Analytics on YouTube.

Wason, Peter C, "Reasoning About a Rule," *Quarterly journal of experimental psychology*, 1968, *20* (3), 273–281.

Weinberg, Bruce A and David Galenson, "Creative Careers: The Life Cycles of Nobel Laureates in Economics," 2005.

Woodford, Michael, "Inattentive Valuation and Reference-dependent Choice," Mimeo 2012.

Yang, Mu-Jeung, Maclean Gaulin, and Nathan Seegert, "Why is Entrepreneurial Overconfidence (So) Persistent? Evidence from a Large-Scale Field Experiment," November 2025. Manuscript.

Appendix

A Attention and Memory

Throughout the text we assume that a person's memory is very flexible: any information a person doesn't currently believe she needs to notice is not top of mind; any information she currently believes she needs to notice is top of mind. In particular, she is able to access at any time anything that happened to her, whether or not she remembered or even noticed it when it first happened. There are circumstances where such assumptions are natural (e.g., if relevant information is collected and stored in an external database). There are also circumstances where these assumptions seem plainly counterfactual. Here (and in greater detail in an earlier draft of the paper, Gagnon-Bartsch et al., 2018) we explore what happens if we refine noticing strategies to account for realistic interactions between attention and memory.

Our analysis would remain unchanged if we refine noticing strategies to require that a person must notice an event as it happens in order to recall it later—e.g., a person needs to first notice a gorilla to later recall that she's seen it. While that would allow a person to turn access to that event on and off as she finds it useful, if we instead insisted that somebody can no longer access the information again if she abandons access at any point, our analysis changes in some subtle ways.

Definition A.1. A noticing strategy $\mathscr{N}$ is *memory consistent (MC)* if for all $t \in \mathbb{N}$ and $h^t \in H^t$, $\tilde{h}^t \in n^t(h^t)$ implies $(s_{t+1}, y_t, x_t; \tilde{h}^t) \in n^{t+1}((s_{t+1}, y_t, x_t; h^t))$ for all $(s_{t+1}, y_t, x_t) \in S_{t+1} \times Y_t \times X_t$.

Memory consistency ("MC") says "once unnoticed, never noticed again". That is, under memory consistency, when the agent notices a coarsening of the history today, then all further coarsenings must maintain the current one. This refinement crudely captures the idea that data is less likely to be top of mind today if it was not top of mind yesterday. Since memory consistency restricts noticing strategies, if there exists a stable attentional strategy under memory consistency, then there exists a stable attentional strategy without requiring memory consistency.[45]

In fact, memory consistency *promotes* incidental learning in some situations. To illustrate, consider a manager who in each period assigns an employee to one of two tasks, $x_t \in \{H, L\}$: "high importance" or "low importance". The manager assigns tasks based on his beliefs about the employee's ability, $\theta \in [0,1]$. The employee's output $y_t \in \{0,1\}$ is i.i.d. conditional on θ with $P(y_t = 1|\theta) = \theta$, where $y_t = 1$ denotes a "successful" job in round t and θ represents the employee's success rate. The manager has incentives to assign the "high" task on day t if and only if he currently believes the worker's success rate, θ, exceeds 50%. Suppose the support of the manager's model includes

[45] In describing the noticing strategy in Definition A.1, we implicitly assume that a person's beliefs about her noticing plans in future periods are time consistent. It is straightforward to extend our framework to handle inconsistency.

non-zero values of θ that are above and below 0.5 (i.e., his action may change as he updates his beliefs about θ), and the support has an upper bound strictly below the worker's true rate, θ^* (i.e., he is overly pessimistic about the worker). The manager need not discover this mistake without memory consistency since he can continually re-partition the full history of outcomes each round in a way that is specifically tailored to his current decision problem. In this case, the manager can follow an attentional strategy that simply notices the optimal action each round and nothing more (see Proposition 1). The optimal assignment in a single period provides but a rough sense of how often the worker has succeeded (e.g., only that his success frequency exceeds some threshold). This information alone is not enough to discover that $\theta^* \notin \text{supp}(\pi)$—he would need to further attend to and recall the worker's precise performance rate over time. Yet when h^t is always accessible, the manager has no incentive to track these seemingly superfluous statistics since his coarse attentional strategy seems sufficient.

These incentives change under memory consistency. In that case, if the manager were to notice nothing more than the optimal action today, he would no longer be able to recall the exact number of good and bad performances that happened prior. This is because the noticing partition in any later round must continue to ignore the same details that the manager ignores today. But such an attentional strategy is not sufficient, since future decisions may rely on these details despite being irrelevant today.[46] As such, a SAS under memory consistency requires the manager to notice the employee's empirical performance rate at each point in time, allowing him to both take the optimal action today and precisely update his beliefs over θ after observing more outcomes in the future. The empirical performance rate, however, will converge to $\theta^* \notin \text{supp}(\pi)$ and consequently render the manager's model attentionally unstable.

This example highlights a sense in which incidental learning under memory consistency comes from a discrepancy in the data necessary to make an optimal decision versus precisely learn parameters. Having unlimited access to historical data allows the agent to bypass details required solely for belief updating and hence limits the scope for incidental learning.

Memory consistency does not say that a person notices all information that he previously encoded. Instead, it is consistent with allowing the person to freely discard information that he previously encoded once it is no longer decision-relevant under his misspecified model. That said, there are situations where it seems likely that some data would be top of mind even when a person no longer finds it useful (e.g., immediately after an action inspired by that data). To handle such scenarios, we consider a refinement that captures the limiting situation of *automatic recall* ("AR") where a person continually notices anything that he previously noticed.

[46] For instance, assigning the high task reveals only that the employee has delivered good performance sufficiently often, but it does not reveal by *how much* the employee surpassed this benchmark. Thus, noticing only this action does not tell the manager how many subsequent bad performances he should endure before re-assigning the employee to the low task. From the manager's perspective, this attentional strategy performs worse than a strategy that pays full attention, and it is therefore not a SAS under memory consistency.

Definition A.2. A noticing strategy $\mathscr{N}$ satisfies *automatic recall (AR)* if for all $t \in \mathbb{N}$ and $h^t \in H^t$, $\tilde{h}^t \notin n^t(h^t)$ implies that $(\tilde{s}_{t+1}, \tilde{y}_t, \tilde{x}_t; \tilde{h}^t) \notin n^{t+1}((s_{t+1}, y_t, x_t; h^t))$ for all $(s_{t+1}, y_t, x_t), (\tilde{s}_{t+1}, \tilde{y}_t, \tilde{x}_t) \in S_{t+1} \times Y_t \times X_t$.

Automatic recall requires the person to distinguish the continuations of any two histories that were previously distinguished. For example, if a consumer considers a product's price when deciding whether to buy it, then automatic recall says that she always remembers this price when making future decisions. Although the combination of automatic recall and memory consistency is extreme, an earlier draft (Gagnon-Bartsch et al., 2018)—along with some of the proofs below—show that our results on when an error is attentionally stable largely continue to hold even if we impose these refinements.

B Supplemental Results

B.1 Basic Results Under Full Attention

In this section, we describe some basic results pertaining to attentional stability under a full-attention SAS. (These results were noted in Section 3; see Footnote 16). These serve as benchmarks for assessing the impact of channeled attention. For simplicity, we consider environments that are stationary (Definition 8). Given our focus on models that are identifiably wrong, π is necessarily unstable relative to π^* under full attention (by Definition 4). Nevertheless, such a π can be attentional stable relative to some alternative $\lambda \neq \pi^*$. This is possible if π explains observations as well as λ.

More precisely, Observation B.1 below shows that a model π is attentionally stable with respect to λ and a full-attention SAS if π fits observations better than λ (in the Kullback-Leibler sense) and attentionally unstable if it fits worse.[47] Given Assumption 1, let $D(\theta^* \| \lambda) := \min_{\theta \in \text{supp}(\lambda)} D(\theta^* \| \theta)$, where $D(\theta^* \| \theta)$ is the Kullback-Leibler Divergence of $P(\cdot|\theta^*)$ from $P(\cdot|\theta)$, and define $D(\theta^* \| \pi)$ analogously. Let $\Delta D(\theta^* \| \lambda, \pi) := D(\theta^* \| \lambda) - D(\theta^* \| \pi)$ be the degree to which π better explains observations than λ.

Observation B.1. *Consider a stationary environment where Assumption 2 holds, and suppose $D(\theta^* \| \lambda)$ or $D(\theta^* \| \pi)$ is finite.*

1. *Model π is attentionally stable with respect to λ and a full-attention SAS if (i) $\Delta D(\theta^* \| \lambda, \pi) > 0$ or (ii) $\Delta D(\theta^* \| \lambda, \pi) = 0$ and $\theta^* \in \text{supp}(\lambda) \cup \text{supp}(\pi)$.*

[47] The Kullback-Leibler divergence is given by

$$D(\theta^* \| \theta) = \sum_{y \in Y} P(y|\theta^*) \log \frac{P(y|\theta^*)}{P(y|\theta)}. \tag{B.1}$$

2. *Model π is attentionally unstable with respect to λ and a full-attention SAS if $\Delta D(\theta^* \| \lambda, \pi) < 0$.*

Note that if $\lambda = \pi^*$, Observation B.1 implies that π is attentionally stable under full attention precisely when $D(\theta^* \| \pi) = 0$. This is equivalent to some $\theta \in \text{supp}(\pi)$ inducing the same distribution of observables as π^*, which in turn means that π is *not* identifiably wrong. Thus, the only case in which π is stable with respect to the true model under full attention is when it is not identifiably different from π^*—the wrong model does just as well as the right one. We focus on identifiably wrong models for this reason.

B.2 Stability Across Stationary Objectives

In this section, we formalize the claims in Section 5.1 regarding classes of models that are and aren't stable across stationary objectives. Broadly speaking, these results demonstrate a sense in which models that fail to make relevant distinctions are more likely to be stable than models that make irrelevant distinctions.

Throughout this subsection, we maintain the assumptions of Proposition 7 and impose an additional assumption on the outcome environment ensuring that (1) there is a one-to-one map between the person's predicted distribution over outcomes and his beliefs over θ, and (2) updating over θ is equivalent to a count-tracking condition. Formally, let $f(\pi_t | s_t) := \sum_\theta \pi_t(\theta) P(\cdot | s_t, \theta)$ denote the predicted distribution of r_t given posterior π_t and signal s_t. Let $\{m_\pi^1, \dots, m_\pi^N\}$ enumerate the cells of the one-period partition induced by m_π. For a history of outcomes y^t, let $K_t(y^t)$ record the number of observations in each m_π^n: $K_t(y^t) := (k_t^1(y^t), \dots, k_t^N(y^t))$ where $k_t^n(y^t) := \sum_{\tau < t} \mathbf{1}\{y_\tau \in m_\pi^n\}$. We assume $f(\cdot | s_t)$ is injective and that $m_\pi(y^t)$ identifies $K_t(y^t)$.

Assumption 3. For every $t > 1$:

1. For any π_t and $\tilde{\pi}_t$ that are reachable with positive probability given π, $f(\pi_t | s_t) = f(\tilde{\pi}_t | s_t)$ implies that $\pi_t = \tilde{\pi}_t$;

2. $m_\pi(y^t) = \{\tilde{y}^t \mid K_t(\tilde{y}^t) = K_t(y^t)\}$.

Part 1 says that the person can be incentivized to track the minimal sufficient statistic for updating beliefs about θ (e.g., when his payoffs are based on a strictly proper scoring rule). Part 2 says that tracking m_π is equivalent to tracking $K_t(y^t)$. Assumption 3 holds in low-dimensional canonical learning models, including the binomial setting from our investment-advice example in Section 5.1.[48] Notably, including Assumption 3 in Proposition 7 implies the converse is also true, extending it to an "if and only if" statement.

[48]This assumption rules out cross-period compression of these counts. In general, posterior beliefs identify a linear combination of these counts but not necessarily each count separately.

We first consider classes of models that are stable across stationary objectives (Definition 9). (See Appendix D.2 for proofs.)

1. Dogmatic errors. We say that π exhibits a *dogmatic error* if π places probability one on some $\theta \neq \theta^*$.

Proposition B.1. *Suppose the assumptions underlying Proposition 7 and Assumption 3 hold. If π exhibits a dogmatic error, then π is stable across stationary objectives.*

2. "Censored" models that ignore possible outcomes. We formally define "censored models" as follows.

Definition B.1. Model π is *censored* if $Y(\pi) \subset Y(\theta^*)$ and there exists $\theta \in \text{supp}(\pi)$ such that for all $y \in Y(\theta)$, $P(m_\pi(y)|\theta) = P(m_\pi(y)|y \in Y(\theta), \theta^*)$.

Proposition B.2. *Suppose the assumptions underlying Proposition 7 and Assumption 3 hold. If π is censored, then π is stable across stationary objectives.*

Note that requiring a censored model to correctly explain anticipated outcomes—i.e., for all $y \in Y(\theta)$, $P(m_\pi(y)|\theta) = P(m_\pi(y)|y \in Y(\theta), \theta^*)$—is stronger than necessary for the previous result. As we show in the proof, this extra assumption ensures that π is additionally stable across stationary objectives under automatic recall (see Definition A.2). Thus, $Y(\pi) \subset Y(\theta^*)$ alone is enough to imply that π is stable for all stationary objectives.

3. Models that neglect predictive signals. We define models with "predictor neglect" as follows.

Definition B.2. Consider an environment where $y_t = (r_t, s_t^1, \ldots, s_t^K)$ in each round, and for all $\theta \in \text{supp}(\pi) \cup \{\theta^*\}$, $P(s_t, r_t|\theta) = P(r_t|s_t, \theta)P(s_t)$ where $s_t := (s_t^1, \ldots, s_t^K)$. That is, due to uncertainty over θ, the person may be uncertain about how s predicts r, but she is certain about the frequency of s since it is independent of θ. Model π exhibits *predictor neglect* if there exists $J \in \{0, \ldots, K-1\}$ such that for all $\theta \in \text{supp}(\pi)$, $P(r_t|s_t, \theta)$ is independent of $(s_t^{J+1}, \ldots, s_t^K)$.

Proposition B.3. *Suppose the assumptions underlying Proposition 7 and Assumption 3 hold. If π exhibits predictor neglect and there exists some $\theta \in \text{supp}(\pi)$ such that $P(r|s^1, \ldots, s^J, \theta) = P(r|s^1, \ldots, s^J, \theta^*)$ for all possible $(r, s^1, \ldots, s^J)$ under θ^*, then π is stable across stationary objectives.*

Intuitively, the person feels free to ignore any information regarding the "neglected" signals $(s^{J+1}, \ldots, s^K)$. Therefore, so long as there exists some θ within the person's model that can explain the joint distribution over $(r, s^1, \ldots, s^J)$, the model is stable across stationary objectives.

Next, we consider classes of models that are not stable across stationary objectives (i.e., they are unstable for some stationary objective).

1. Uncertain models that correctly specify the set of outcomes but incorrectly specify their probabilities. Sufficient uncertainty induces incidental learning when the misspecified theory correctly predicts which outcomes are possible but incorrectly specifies the probabilities of those outcomes. The next definition describes uncertain environments where no two observations lead to the same beliefs over parameters.

Definition B.3. For any π, we say the family of distributions $\{P(\cdot|\theta)\}_{\theta\in\text{supp}(\pi)\cup\{\theta^*\}}$ satisfies the *Varying Likelihood Ratio Property (VLRP)* if for all $y,y' \in Y(\pi)$ and all $\theta,\theta' \in \text{supp}(\pi)\cup\{\theta^*\}$, $\frac{P(y|\theta)}{P(y'|\theta)} = \frac{P(y|\theta')}{P(y'|\theta')}$ if and only if $y = y'$ or $\theta = \theta'$.[49]

Whenever $\text{supp}(\pi)$ contains at least two elements, VLRP implies that the person finds it necessary to separately notice every outcome in order to learn θ; that is, $m_\pi(y)$ is a singleton for each $y \in Y(\pi)$.

Proposition B.4. *Suppose the assumptions underlying Proposition 7 and Assumption 3 hold and, in addition, VLRP holds with $Y(\pi) = Y(\theta^*)$. If $\text{supp}(\pi)$ has at least two elements and $\theta^* \notin \text{supp}(\pi)$, then π is not stable across stationary objectives.*

2. "Overly elaborate" models that anticipate too wide a range of outcomes. We define overly-elaborate models as follows.

Definition B.4. Model π is *overly elaborate* if $Y(\pi) \supset Y(\theta^*)$ and there exists $y \in Y(\pi)$ such that $m_\pi(y) \cap Y(\theta^*) = \varnothing$ with $P(m_\pi(y)|\theta) > 0$ for all $\theta \in \text{supp}(\pi)$.

Proposition B.5. *Suppose the assumptions underlying Proposition 7 and Assumption 3 hold. If π is overly elaborate, then π is not stable across stationary objectives.*

3. "Over-fit" models that assume the set of predictive signals is wider than it truly is. We provide a definition of "over-fit" models within the same class of environments in which we considered predictor neglect, above. Over-fit models can be seen as a counterpoint to those with predictor neglect.

Definition B.5. Consider the environment introduced above in Definition B.2: in each round, $y_t = (r_t, s_t^1, \ldots, s_t^K)$ and, for all $\theta \in \text{supp}(\pi)\cup\{\theta^*\}$, $P(s_t, r_t|\theta) = P(r_t|s_t,\theta)P(s_t)$ where $s_t := (s_t^1,\ldots,s_t^K)$. Model π is *over-fit* if it has the following properties:

1. There exists $J \in \{0,\ldots,K-1\}$ such that in truth $P(r_t|s_t,\theta^*)$ is independent of $(s_t^{J+1},\ldots,s_t^K)$. That is, for all $s,\tilde{s} \in S$ such that $(s^1,\ldots,s^J) = (\tilde{s}^1,\ldots,\tilde{s}^J)$ and $(s^{J+1},\ldots,s^K) \neq (\tilde{s}^{J+1},\ldots,\tilde{s}^K)$, $P(r|s,\theta^*) = P(r|\tilde{s},\theta^*)$.

[49] The VLRP condition is a generalization of the more familiar monotone likelihood ratio property (MLRP), as it does not require $\text{supp}(\pi)\cup\{\theta^*\}$ to be ordered. VLRP therefore holds for any family of distributions that satisfies (strict) MLRP.

2. For all $\theta \in \text{supp}(\pi)$, $P(r_t|s_t,\theta)$ depends on both $(s_t^1,\ldots,s_t^J)$ and $(s_t^{J+1},\ldots,s_t^K)$.

3. Signals $(s^{J+1},\ldots,s^K)$ are useful for updating under model π. That is, there exist $s,\tilde{s} \in S$ such that $(s^1,\ldots,s^J) = (\tilde{s}^1,\ldots,\tilde{s}^J)$, $(s^{J+1},\ldots,s^K) \neq (\tilde{s}^{J+1},\ldots,\tilde{s}^K)$, and $(r,\tilde{s}) \notin m_\pi((r,s))$ for some resolution r where $(r,s),(r,\tilde{s}) \in Y(\pi)$.

4. For all $\theta \in \text{supp}(\pi)$, the distribution of $m_\pi(y)$ under θ differs from its distribution under π^*.

To summarize, over-fit models are certain that some useless signals help predict outcomes, are uncertain about the extent to which they help, and contain no parameter that correctly predicts the distribution of the resulting minimal sufficient statistic (i.e., they require attention to these useless signals beyond what is necessary under the true model).

Proposition B.6. *Suppose the assumptions underlying Proposition 7 and Assumption 3 hold. If π is over-fit, then π is not stable across stationary objectives.*

Intuitively, there exist choice environments where the person seeks to learn the extent to which signals $(s^{J+1},\ldots,s^K)$ predict outcomes, and attending to these signals would eventually prove π false.

C Additional Applications

This section provides details for some of the applications discussed in main text: empirical misconceptions of asset patterns (Example 3 from Section 2.3), self control (Example 4 from 2.3), and redundancy neglect (Section 4.3).

C.1 Empirical Misconceptions of Asset Patterns

In this section, we apply our framework to assess the stability of a class of misspecified beliefs that includes the well-known model by Barberis et al. (1998) ("BSV"), which we briefly discussed in Section 3. Consider a single security that pays out 100% of its earnings as dividends. In truth, the earnings process, denoted (M_t), follows a random walk: $M_t = M_{t-1} + y_t$, where $y_t \in \{-y, y\}$ for some $y > 0$ and $M_0 = 0$. A risk-neutral representative investor holds wrong beliefs about the process, believing that it is in one of two "regimes" at any point in time: in one regime earnings are mean reverting, and in the other they trend. We show that the propensity for a market participant to eventually recognize this error will depend on the scope of her objective. In particular, an analyst who simply needs to make a buy/sell recommendation to an investor is less likely to discover the mistake than an analyst who additionally has an incentive to back up her recommendation with precise predictions about future earnings.

We consider a slightly richer misspecified model than the streamlined version presented in Example 3, and accordingly use different notation. Each agent's misspecified model posits an underlying state of the economy $\omega_t \in \{1,2\}$ that determines how y_t is drawn in round t. If $\omega_t = 1$, then y_t follows a mean-reverting Markov model; if $\omega_t = 2$, then y_t follows a trending Markov model. The presumed transition probabilities in each of the two states are given in the table below, where $\theta_L \leq 1/2 \leq \theta_H$.

Table 1: Misspecified Models of Earnings Growth (Barberis et al., 1998)

(a) The Mean-Reverting Process (Regime 1)

$\omega_t = 1$	$y_{t-1} = y$	$y_{t-1} = -y$
$y_t = y$	θ_L	$1-\theta_L$
$y_t = -y$	$1-\theta_L$	θ_L

(b) The Trending Process (Regime 2)

$\omega_t = 2$	$y_{t-1} = y$	$y_{t-1} = -y$
$y_t = y$	θ_H	$1-\theta_H$
$y_t = -y$	$1-\theta_H$	θ_H

Agents also believe the state ω_t follows a Markov process with the following transition matrix:

	$\omega_{t+1} = 1$	$\omega_{t+1} = 2$
$\omega_t = 1$	$1-\theta_1$	θ_1
$\omega_t = 2$	θ_2	$1-\theta_2$

with $\theta_1, \theta_2 \in (0,1)$. In reality, $\theta_L^* = \theta_H^* = 1/2$. An agent's misspecified model π over $\theta = (\theta_L, \theta_H, \theta_1, \theta_2)$ places no weight on parameter vectors with $\theta_L = \theta_H = 1/2$. This nests the BSV misspecification where an agent is convinced of all the transition probabilities above, and thus her misspecified model π is degenerate on some $\theta = (\theta_L, \theta_H, \theta_1, \theta_2)$ with $\theta_L < 1/2 < \theta_H$.

Earnings are observable, and agents use this information to update their beliefs about the current state and future earnings. Assuming risk-neutrality and a discount rate ρ, the equilibrium price in period t is $p_t = \mathbb{E}_t[V_t]$, where $V_t := \sum_{k=0}^{\infty} M_{t+k}/(1+\rho)^k$ is the discounted value of all future earnings and $\mathbb{E}_t$ denotes expectations conditional on $(y_{t-1}, \ldots, y_1)$.[50]

Consider an analyst who provides a buy/sell recommendation to a price-taking, risk-neutral investor. In a given period t, the analyst recommends "buy" ($x_t = 1$) if $\mathbb{E}_{\pi,t}[V_t] \geq p_t$ and "sell" ($x_t = 0$) otherwise. The analyst may additionally need to make precise predictions about future earnings to back up her buy/sell recommendation. In this case, she announces her prediction of the current period's earnings, denoted $\widehat{M}_t$, just before they are realized, and her payoff from the prediction is $-(\widehat{M}_t - M_t)^2$. We assume this is additively separable from her incentive to make an accurate recommendation. The optimal forecast is $\mathbb{E}_{\pi,t}[M_t]$.

Proposition C.1. *In the setting above, consider any identifiably wrong model π over θ. The analyst is more likely to wake up when she needs to provide predictions to back up her recommendations*

[50] The timing is such that p_t denotes the price in period t prior to the period-t earnings realization; p_t is thus based on $(y_{t-1}, \ldots, y_1)$.

than when she does not need to make such predictions: fixing π, any noticing strategy that is part of a SAS for the analyst who needs to provide predictions is also part of a SAS for the analyst who does not need to, but not vice-versa. Consequently, if there is a stable attentional strategy for the analyst who needs to provide predictions, then there is also a stable attentional strategy for the analyst who does not.

Waking up to an error depends on the scope of an individual's objective. The buy/sell recommendation requires only a coarse view of the data, thereby permitting a false belief like the BSV model to persist. On the other hand, a precise prediction about earnings requires detailed attention to patterns in the data, which is more likely to force the agent to confront inconsistencies that reveal her misspecified beliefs. Take the BSV misspecification as an example. Following the logic of Proposition 1, the analyst who only needs to make a recommendation can coarsely attend to the data so that she notices only whether she should recommend buy or sell. At any date t, her minimal noticing partition N^t contains just two elements: $n^t_B = \{h^t \in H^t \mid \mathbb{E}_{\pi,t}[V_t] \geq p_t\}$ and $n^t_S = \{h^t \in H^t \mid \mathbb{E}_{\pi,t}[V_t] < p_t\}$. In each period, she essentially notices a binary signal—"recommend buy" or "recommend sell"—and these coarse signals alone are not enough to reveal her error. By contrast, the analyst who needs to make precise predictions feels the need to more finely parse the earnings history. Her expectation of the next period's earnings is a function of the probability that the current regime is mean reverting. Letting q_t denote this probability, the analyst must partition the history of outcomes in a way that precisely distinguishes q_t. This value, however, gives a sharp sense of the sequence of earnings innovations that must have happened prior to period t. As such, the analyst must confront the history of outcomes by noticing q_t and is therefore more likely to reject the misspecified model based on her noticed data. We conjecture that there is no stable attentional strategy for an analyst who holds the misspecified BSV model and needs to offer earnings predictions.

C.2 Misunderstanding Self-Control

Why do most of us have experience using a commitment device to solve a particular self-control problem, yet display low demand for such devices overall? Why do people seem to quickly learn they have self-control problems in some situations, yet remain naive in others? Here, we provide a more detailed assessment of the stability of self-control problems, which we initially discussed in Section 3.

While some theoretical work (e.g., Ali, 2011) would seem to suggest that rational learning should correct underestimation of self-control problems, channeled attention can enable a person to persistently make this error in some situations yet wake up in others.[51] In a canonical environment, we

[51] Gottlieb (2019) integrates Bénabou and Tirole's (2004) model of willpower into Ali's (2011) model of learning about self control and shows that motivated mis-updating provides another reason why someone may fail to learn they have a self-control problem.

investigate when a person might come to realize that her self-control problem is more severe than she deemed plausible. Greater uncertainty makes her more prone to discover her error; the more dogmatic she is, the less likely she is to do so. The degree to which she thinks her uncertainty matters for behavior is also crucial: she's likely to pay attention to resolve uncertainty about her self-control problem only when she thinks this might influence her course of action (e.g., investing in a commitment device).

Consider a person who repeatedly decides whether to take an action with immediate cost and delayed benefit. To fix ideas, imagine decisions to visit the gym, as studied by DellaVigna and Malmendier (2006). At the start of each period t, the person chooses whether to buy a gym membership or—if she already has one—whether to visit the gym. A gym membership lasts for T periods (beginning the period after purchase) and costs m (paid the period after purchase). If the person goes to the gym on day t, then she also pays an immediate effort cost equal to c_t and earns benefit $b > 0$ in the future. Costs c_t are i.i.d. draws from $\mathrm{U}[0, \bar{c}]$, where $\bar{c} > b$. If the person doesn't go to the gym, she incurs no effort cost or benefit. We assume c_t is observable at the start of period t regardless of whether she goes or not (i.e., Assumption 1 holds).

Following Laibson (1997) and O'Donoghue and Rabin (1999, 2001), we consider a (β, δ) discounter with $\delta = 1$. Thus, the person discounts future costs or benefits by a factor $\beta < 1$. The Laibson and O'Donoghue-Rabin models assume the person has a constant self-control problem and is dogmatic about its degree. The model in Eliaz and Spiegler (2006), by contrast, accommodates the more realistic case where a person is uncertain about her future self-control problem. Because their model is only two periods, the formalization can accommodate both true stochastic present bias or uncertainty about a permanent degree of present bias.

Following the latter approach, we assume that the true present bias in a domain is potentially stochastic, and she is uncertain about how often her present bias is severe versus mild in the given domain. The person's discount factor fluctuates from day to day between no present bias (to simplify what we mean by "mild") and some degree of present bias. To capture naivete, we assume the person underestimates the likelihood that she will be present biased. Specifically, she believes that, in any future period t, her discount factor will be $\beta_t = \beta$ with probability q or $\beta_t = 1$ with probability $1 - q$. (Note that q was labeled as θ in the main text.) Suppose the true probability is $q^* = 1$. Thus, the person thinks she'll be tempted to avoid the gym on any given day with probability q, while in reality she is tempted with probability 1. An important special case is where the person dogmatically believes in some $\hat{q}$. The person is sophisticated when $\hat{q} = 1$; fully naive when $\hat{q} = 0$; and partially naive when $\hat{q} \in (0, 1)$. We consider a naive person with prior π over q such that $1 \notin \mathrm{supp}(\pi)$.[52]

[52] In the alternative formulation of partial naivete proposed by O'Donoghue and Rabin (2001), the person dogmatically believes that her self control is $\hat{\beta} \in [\beta, 1]$. Under that formulation, the person sees something she thought was impossible every period (i.e., β is lower than she thought possible). Since the agent in our framework has wide flexibility in how she encodes subjectively impossible events, O'Donoghue and Rabin's (2001) formulation of partial naivete is in fact more

Finally, we assume that $s_t = (c_t, \beta_t)$, so the person observes her current effort cost and degree of present bias as a signal prior to acting in period t.

Partial naivete—the belief that β_t will sometimes equal 1—is unstable under full attention since the history of costs and gym attendance will prove this theory false. Notice that a gym member visits the gym on day t if and only $\beta_t b > c_t \Leftrightarrow b > c_t/\beta_t$. Since we assume the person can observe her effort cost (c_t) and her overall desire to avoid the gym (i.e., c_t/β_t), the history of these values together will reveal that β_t never equals 1.

With channeled attention, however, π may give rise to a stable attentional strategy. To build intuition, first note that this is always true when the membership is free ($m = 0$). In this case, deciding whether to visit the gym on a given day requires the person to notice only whether her current disinclination to visit, c_t/β_t, exceeds b. She need not separately notice the precise values of c_t and β_t—she can simply ask herself if she wants to skip the gym without asking herself why (i.e., high cost vs. laziness). Furthermore, because she thinks there is nothing payoff relevant to learn, she need not remember past values of c_t/β_t or her attendance rate. Thus, when following such a SAS, the person will not notice that β_t has a distribution inconsistent with π.

When the membership is costly ($m > 0$), whether π admits a stable attentional strategy depends on how uncertain the person is about her self-control problem. In this case, the person has an incentive to collect information that helps her predict whether the membership is worthwhile. For a given point belief $\hat{q}$, the person desires the membership if $m < T\{(1-\hat{q}) \cdot \mathbb{E}[b-c|b>c]\Pr(b>c) + \hat{q} \cdot \mathbb{E}[b-c|\beta b > c]\Pr(\beta b > c)\}$, or, equivalently, if

$$m < T \cdot \left(\frac{b^2}{2\bar{c}}\right) \cdot [(1-\hat{q}) + \hat{q}\beta(2-\beta)]. \tag{C.1}$$

That is, she buys the membership if she thinks its cost is lower than the option value of being able to use the gym. This option value is high when perceived effort costs are low (i.e., low $\bar{c}$) and naivete is high (i.e., low $\hat{q}$).

It is straightforward that the person may persistently "pay not to go to the gym" under a stable attentional strategy. When Condition (C.1) holds for all $\hat{q} \in \text{supp}(\pi)$, the person thinks the membership is worthwhile no matter what. The analysis is then similar to the case where $m = 0$: if the person is certain she has sufficient self control to justify the membership, then she need not track her behavior and notice that she goes too little. Indeed, there is suggestive evidence that people not only overestimate how often they will go to the gym, but do not realize how little they went in the past (Beshears et al., 2021; Carrera et al., 2022; Sial et al., 2022).

Even when the person is initially unsure whether the membership is worthwhile (i.e., Condition C.1 holds for some $\hat{q} \in \text{supp}(\pi)$ but not all), she still need not wake up to the fact that she goes to the

likely to be attentionally stable than Eliaz and Spiegler's (2006).

gym less often than she thought possible. By Proposition 1, she need only notice whether she should buy the membership (e.g., whether Condition C.1 holds). But this does not require her to precisely notice the frequency with which she has gone to the gym.

So what would get the person to wake up? She needs incentives to *precisely* track her behavior. This could arise, for example, when facing a menu of gym memberships with different perks (e.g., access to exercise classes, different hours, different days open). We'll capture such incentives in a reduced-form way by imagining that the person needs to precisely track her willingness to pay for a gym membership—that is, she has an incentive to accurately assess her expected value of the right-hand-side of Condition (C.1). In this case, we say the person *has an incentive to track her precise willingness to pay for a gym membership*. In contrast, we say the baseline case described above (deciding to buy a fixed plan at price m) does not induce such an incentive.

Proposition C.2. *In the setting above, there is no stable attentional strategy given the person's model π if and only if both (i) π is non-degenerate with* $\text{supp}(\pi) \cap (0,1)$ *containing at least two elements (the "non-degenerate case") and (ii) the person has an incentive to track her precise willingness to pay for a gym membership.*

To summarize, waking up depends on the person having some sophistication *and* uncertainty about her self-control problems, and the incentive to precisely learn the extent of those problems to guide her actions. For intuition behind this result, first suppose that both (i) and (ii) hold. The person thinks she must notice the fraction of times that $\beta_t = \beta$ in order to distinguish between any $\hat{q}$ and $\hat{q}'$ in the support of π. In the long run, the observed fraction will be inconsistent with π, so π is attentionally unstable. If either (i) or (ii) don't hold, then the person believes there is no benefit to learning the precise extent of her self-control problem and there is a SAS under which π is attentionally stable.[53]

C.3 Redundancy Neglect

This application considers an agent who neglects the correlated nature of others' advice (as in DeMarzo et al., 2003; Eyster and Rabin, 2010, 2014; Enke and Zimmermann, 2017). We build on Example 1 from Section 2.3 where the agent (an investor) receives advice from multiple sources. In each period t, the agent encounters a new opportunity and the optimal action is $\omega_t \in \{0,1\}$. Suppose ω_t is

[53] More specifically, without an incentive to track her willingness to pay, the person's noticing partition N^t in a period t where she decides whether to buy a membership has at most two cells when following a minimal SAS: $\{h^t \in H^t \mid \mathbb{E}_t[V(\hat{q})] \geq m\}$ and $\{h^t \in H^t \mid \mathbb{E}_t[V(\hat{q})] < m\}$, where $V(\hat{q})$ denotes the right-hand side of (C.1) and $\mathbb{E}_t$ is with respect to π_t. In contrast, with an incentive to track willingness to pay, we assume the person has an incentive to truthfully state $\mathbb{E}_t[V(\hat{q})]$ for each period t in which she can buy a membership. Hence, in those periods, the person must notice how many times she has been tempted over the course of the history (including the present period since β_t is part of the signal s_t received at the start of period t), and thus N^t must have a cell for each possible realization of $\sum_{k=1}^{t} \mathbf{1}\{\beta_k < 1\}$. Since in reality $\beta_k = \beta$ in every period, the person will notice that $\sum_{k=1}^{t} \mathbf{1}\{\beta_k < 1\} = t$, which becomes increasingly improbable under π.

i.i.d. across periods with $P(\omega_t = 1) = 1/2$. The agent's objective each period is to choose $x_t \in [0,1]$ to maximize $-(x_t - \omega_t)^2$. Thus, the optimal x_t matches his subjective probability that $\omega_t = 1$. Prior to taking an action, the agent receives information about ω_t from 2 sources: the agent's signal is $s_t = (s_t^1, s_t^2) \in \{0,1\}^2$, where s_t^i is the recommended action from source i. Let $\theta_i := P(s_t^i = \omega_t | \omega_t)$ denote the precision of advice from source $i \in \{1,2\}$, and suppose $\theta_i \in \{0.5, \gamma\}$ for some $\gamma \in (0.5,1)$. That is, an information source is either uninformative (i.e., $\theta_i = 0.5$) or partially informative (i.e., $\theta_i = \gamma$).

Suppose the two information sources are in fact perfectly correlated: $s_t^1 = s_t^2$ for all t. For example, one advisor has good intuition (or access to good information), while the second one simply mimics the first. We explore the stability of a misspecified model where the agent treats these two information sources as independent, thereby thinking he receives two informative signals each period instead of one. We also describe features of the environment that can help or hinder the discovery of this error.

To specify the agent's erroneous model more precisely, let $\theta = (\theta_1, \theta_2, \theta_c)$ denote the parameter governing the agent's signals. As noted above, θ_i for $i \in \{1,2\}$ is the precision of source i. Additionally, $\theta_c \in \{0,1\}$ parameterizes the correlation in information sources: $\theta_c = 0$ denotes independence and $\theta_c = 1$ denotes perfect correlation. The agent's misspecified model π puts probability one on $\theta_c = 0$. Suppose also that under π, θ_1 and θ_2 are independent. Conditional on θ and the sequence of ω_t, suppose additionally that signals are independent across periods with $P(s_t^i = \omega | \omega_t = \omega, \theta) = \theta_i$. When $\theta_c = 0$, s_t^1 and s_t^2 are conditionally independent given ω_t. When $\omega_c = 1$, we restrict $\theta_1 = \theta_2$ and assume $s_t^1 = s_t^2$ almost surely. The true parameter is $\theta^* = (\gamma, \gamma, 1)$.

The first feature of the environment that matters for the discovery of this error is whether the agent feels compelled to learn about the precision of his information sources. For instance, an investor who seeks recommendations from different social media sources may try to learn which of those sources is more trustworthy.

The second feature concerns the feedback available to the agent. We compare two cases. We say the environment has *perfect feedback* when the agent observes the current state at the end of each period. That is, $r_t = \omega_t$ for all t. For example, the investor can readily assess whether his advisor's recommendations were good or bad. In contrast, we say the environment has *no feedback* when $r_t = \varnothing$ for all t. In such cases, we assume the agent does not observe the state nor receive utility until the end of the game. For instance, an investor makes decisions over long-term prospects that don't pay out until long after his choices and hence does not receive feedback on whether the advice was good or bad. While these two cases are extreme, they adequately reveal how less feedback can *enable* incidental learning.

Proposition C.3. *In the setting above, consider any identifiably wrong model π that puts probability one on $\theta_c = 0$ (independent signals).*

1. Perfect feedback (i.e., $r_t = \omega_t$ for all t): *There exists a stable attentional strategy given π if and only if π is dogmatic about the precision of one or both information sources.*

2. No feedback (i.e., $r_t = \varnothing$ for all t): *There exists a stable attentional strategy given π if and only if π is dogmatic about both information sources or dogmatic that one information source is uninformative.*

Proposition C.3 reveals two ways in which uncertainty enables the incidental discovery of redundancy neglect. The first part says that if the agent receives perfect feedback, then his erroneous "independent-signals" model admits a stable attentional strategy if and only if he knows the precision of one or both information sources. If he is certain about one of the sources, then he can ignore feedback about it. Under this SAS there is no way to notice the correlation between sources. However, if he is uncertain about the precision of both, then tracking how often each gives good advice will cause him to incidentally notice that their advice matches the state at an inexplicably similar rate.

The second part says that if the agent receives no feedback, then the "independent-signals" model admits a stable attentional strategy if and only if he is dogmatic about the precision of both information sources or treats one as uninformative. If he treats both sources as potentially informative and is uncertain about the precision of at least one of them, then—unlike the case with perfect feedback—he will discover his error. This is because he will learn about the precision of the uncertain source by tracking how often the two sources give the same advice: precise advice is more likely to coincide than noisy advice. To provide some intuition, suppose an investor is dogmatic that one media outlet has high precision yet is uncertain about the precision of the other. In this case, the mere event of agreement between the two outlets would be good news about the quality of the uncertain one.

D Proofs

D.1 Proofs of Results in the Main Text

The following lemma (and remark) will be useful in establishing stability in many of the proofs to follow.

Lemma D.1. *Assume Assumptions 1 and 2 hold, and that the environment is stationary. Suppose the true parameter is θ^*, and consider $\theta, \theta' \in \Theta$ with $Y(\theta^*) \subseteq Y(\tilde{\theta})$ for $\tilde{\theta} = \theta, \theta'$. Enumerate the true support $Y(\theta^*) = \{y_1, \ldots, y_N\}$, and define*

$$\bar{Z}(\theta, \theta' | \theta^*) := \prod_{n=1}^{N} \left(\frac{P(y_n|\theta)}{P(y_n|\theta')} \right)^{P(y_n|\theta^*)}. \tag{D.1}$$

1. *If* $\bar{Z}(\theta,\theta'|\theta^*) < 1$*, then the likelihood ratio* $P(y^t|\theta)/P(y^t|\theta') \xrightarrow{\text{a.s.}} 0$*.*

2. *If* $\bar{Z}(\theta,\theta'|\theta^*) > 1$*, then* $P(y^t|\theta)/P(y^t|\theta') \xrightarrow{\text{a.s.}} \infty$*.*

3. *If* $\bar{Z}(\theta,\theta'|\theta^*) = 1$ *and* θ *or* θ' *equals* θ^**, then* $P(y^t|\theta)/P(y^t|\theta') \xrightarrow{\text{a.s.}} 1$*.*

Proof of Lemma D.1. For any $y^t = (y_{t-1},\dots,y_1)$ with $t \geq 2$, let $k_t^n(y^t)$ be the number of times outcome y_n happens prior to round t. Thus, $k_t^n(y^t) := \sum_{\tau=1}^{t-1} \mathbf{1}\{y_\tau = y_n\}$ and $k_t^n(y^t)/(t-1) \xrightarrow{\text{a.s.}} P(y_n|\theta^*)$ by the Strong Law of Large Numbers (SLLN). Then

$$\frac{P(y^t|\theta)}{P(y^t|\theta')} = \frac{\prod_{n=1}^N P(y_n|\theta)^{k_t^n(y^t)}}{\prod_{n=1}^N P(y_n|\theta')^{k_t^n(y^t)}} = \left(\frac{\prod_{n=1}^N P(y_n|\theta)^{k_t^n(y^t)/(t-1)}}{\prod_{n=1}^N P(y_n|\theta')^{k_t^n(y^t)/(t-1)}}\right)^{t-1} = (Z_t)^{t-1}, \tag{D.2}$$

where

$$Z_t := \prod_{n=1}^N \left(\frac{P(y_n|\theta)}{P(y_n|\theta')}\right)^{k_t^n(y^t)/(t-1)}. \tag{D.3}$$

If $\bar{Z}(\theta,\theta'|\theta^*) < 1$, fix any $\widetilde{Z} \in \big(\bar{Z}(\theta,\theta'|\theta^*), 1\big)$. On the event $\{\lim_{t\to\infty} Z_t = \bar{Z}(\theta,\theta'|\theta^*)\}$, which occurs with probability one by construction, there exists $T \in \mathbb{N}$ such that for all $t > T$, $Z_t < \tilde{Z} < 1$. It follows that $\limsup_{t\to\infty}(Z_t)^{t-1} \leq \lim_{t\to\infty}\big(\tilde{Z}\big)^{t-1} = 0$ on this event, and thus $(Z_t)^{t-1} \xrightarrow{\text{a.s.}} 0$. A similar argument holds for the case where $\bar{Z}(\theta,\theta'|\theta^*) > 1$. Finally, if $\bar{Z}(\theta,\theta'|\theta^*) = 1$ and θ or θ' equals θ^*, then $P(\cdot|\theta) = P(\cdot|\theta')$ by Gibb's inequality. This implies that $Z_t = 1$ for all t and thus $P(y^t|\theta)/P(y^t|\theta') = 1$ for all t. ■

Remark D.1. Note that $\ln(\bar{Z}(\theta,\theta'|\theta^*)) = D(\theta^*\|\theta') - D(\theta^*\|\theta)$, where $\bar{Z}$ is defined in Equation D.1 and D is the KL divergence defined in Equation B.1. Hence, the three conditions of Lemma D.1 are equivalent to (i) $D(\theta^*\|\theta') < D(\theta^*\|\theta)$, (ii) $D(\theta^*\|\theta') > D(\theta^*\|\theta)$, and (iii) $D(\theta^*\|\theta') = D(\theta^*\|\theta)$ and θ or θ' equals θ^*.

Proof of Proposition 1. Let $\mathscr{N}_F$ be the full-attention noticing strategy and let σ_F be a behavioral strategy such that $\phi_F = (\mathscr{N}_F, \sigma_F)$ is a SAS given π. For each h^t with positive probability under π, let $X_t^*(h^t,\pi) \subseteq X_t$ denote the set of actions to which $\sigma_F(h^t)$ assigns positive probability; by assumption, $X_t^*(h^t,\pi)$ is a singleton. For any possible h^t that has zero probability under π, take some $\hat{h}^t$ that has positive probability under π and set $X_t^*(h^t,\pi) = X_t^*(\hat{h}^t,\pi)$. Then define $\mathscr{N}$ such that for all t, $n^t(h^t) = \big\{\tilde{h}^t \in H^t \mid X_t^*(\tilde{h}^t,\pi) = X_t^*(h^t,\pi)\big\}$. Let σ be a behavioral strategy where, for all t, $\sigma_t(n^t(h^t))$ places probability one on $X_t^*(h^t,\pi)$; $\phi = (\mathscr{N},\sigma)$ is a therefore a SAS given π.

Furthermore, ϕ is minimal. To see this, consider any $\tilde{\mathscr{N}}$ that is a coarsening of $\mathscr{N}$. Then $\tilde{\mathscr{N}}$ must include some partition $\tilde{N}^t$ with an element $\tilde{n}^t$ with the following property: there exist at least two elements n^t and $\hat{n}^t$ of N^t such that $\tilde{n}^t \cap n^t \neq \varnothing$ and $\tilde{n}^t \cap \hat{n}^t \neq \varnothing$. However, by construction, all $h^t \in n^t$ prescribe the same optimal action under π, and this action is distinct from the one prescribed

by all $h^t \in \hat{n}^t$. Hence, after merging the two cells, no action is optimal at all histories in the merged cell. Any behavioral strategy measurable with respect to the coarser partition $\tilde{\mathscr{N}}$ is not part of a SAS given π since $\tilde{n}^t$ induces a suboptimal action with positive probability under π. This establishes part 1 of the result. Parts 2 and 3 then follow immediately from Part 1 together with Definition 3. ■

Proof of Proposition 2. Fix a minimal SAS $\phi = (\mathscr{N}, \sigma)$ given π. We first show the stronger claim that every cell in every finite-date noticing partition has positive probability under π and ϕ.

Toward a contradiction, suppose that for some date τ there is a cell $n \in \mathscr{N}^\tau$ with $P(n \mid \pi, \phi) = 0$. Since $P(H^\tau \mid \pi, \phi) = 1$, there exists another cell $n' \in \mathscr{N}^\tau$, $n' \neq n$. Form a new partition by replacing n and n' with $n \cup n'$ at date τ and leaving all other date partitions unchanged. Define the new behavioral strategy to agree with σ everywhere except that on $n \cup n'$ it uses the action distribution that σ assigned to n'. Because π assigns positive weight to every $\theta \in \text{supp}(\pi)$, $P(n \mid \pi, \phi) = 0$ implies $P(n \mid \theta, \phi) = 0$ for every such θ. Hence, if n' has positive subjective probability under π, the posterior and conditional decision problem on $n \cup n'$ under the coarser strategy are identical to those on n' under the original strategy. If n' also has zero probability, behavior on the union is subjectively irrelevant. The modified strategy therefore agrees with ϕ on every subjectively possible path and yields the same expected payoff under π; it is consequently a SAS given π. Moreover, the modified noticing strategy strictly coarsens $\mathscr{N}$, contradicting minimality. Thus $P(n \mid \pi, \phi) > 0$ for every cell $n \in \mathscr{N}^t$ and every finite t. In particular, every finite noticed history that occurs with positive probability under π^* and ϕ has positive probability under π and ϕ. ■

Proof of Proposition 3. Consider a SAS $\phi = (\mathscr{N}, \sigma)$ given π.

Part 1. Suppose π conceptualizes all possible events (relative to π^*). This implies that for any finite history h^t, we have $P(h^t|\pi) > 0$ and thus $P(n^t(h^t)|\pi, \phi)/P(n^t(h^t)|\pi^*, \phi) > 0$. Hence, rejecting π when following ϕ requires noticing a long-run statistical gorilla given Definition 6.

Part 2. Suppose ϕ is a minimal SAS given π. By Proposition 2, π assigns positive probability to every finite noticed history $n^t(h^t)$ that occurs with positive probability under π^* and ϕ. Thus, for any possible h^t, we have $P(n^t(h^t)|\pi, \phi)/P(n^t(h^t)|\pi^*, \phi) > 0$, which implies that rejecting π when following ϕ requires noticing a long-run statistical gorilla by Definition 6. ■

Proof of Proposition 4. Let $R = \text{supp}(\pi) \setminus \widetilde{\Theta}$. For each noticed history, Bayes' rule yields

$$\frac{P(n^t|\tilde{\pi}, \phi)}{P(n^t|\pi^*, \phi)} = \frac{\pi_t(R)}{\pi(R)} \frac{P(n^t|\pi, \phi)}{P(n^t|\pi^*, \phi)}. \tag{D.4}$$

The first factor on the right-hand side of the previous equation converges to $1/\pi(R) > 0$ and the second has a positive limit inferior. By assumption (2) of the proposition, ϕ does not distinguish any events that have positive probability under π^* but zero probability under $\tilde{\pi}$, which rules out the possibility of the likelihood ratio in D.4 reaching zero in finite time. Since ϕ is sufficient under $\tilde{\pi}$,

ϕ is a StAS given $\tilde{\pi}$. The text provides an example where the converse does not hold: the restricted model admits a StAS even when the larger model does not. ■

Proof of Proposition 5. In the original environment, let $\phi = (\mathscr{N}, \sigma)$ denote the SAS derived in the proof of Proposition 1 that, in each period t, notices only the current optimal action given π and h^t. Consider a modified environment identical to Γ aside from expanding the action space each period to $\widetilde{X}_t = X_t \cup \{d\}$, where selecting d implements the recommended action under ϕ. That is, prior to period 1, the person submits the behavioral strategy σ to a delegate or algorithm, and in any period t in which the person chooses $x_t = d$, the delegate or algorithm implements the action specified by σ_t and $n^t(h^t)$. Hence, $u_t(d, y_t|h^t) = u_t(\sigma(n^t(h^t)), y_t|h^t)$. Now consider the attentional strategy $\tilde{\phi} = (\tilde{\mathscr{N}}, \tilde{\sigma})$ where $\tilde{\mathscr{N}}$ makes no distinctions (i.e., $\tilde{N}^t = \{H^t\}$ for all t) and $\tilde{\sigma}_t(\tilde{n}^t) = d$ for all t. In the modified environment, $\tilde{\phi}$ is a SAS since it implements the same behavior as ϕ, which itself is a SAS. Furthermore, since $\tilde{\phi}$ follows the coarsest possible noticing strategy, $\tilde{\phi}$ is a stable attentional strategy given π. ■

Proof of Proposition 6. Part 1. We begin with a useful lemma.

Lemma D.2. *Let p and q be probability mass functions on a finite set C. Define*

$$G_p(c) := \frac{p(c)}{p(c)+q(c)} \quad \textit{and} \quad G_q(c) := \frac{q(c)}{p(c)+q(c)}, \tag{D.5}$$

with $G_p(c) = G_q(c) = 1/2$ when $p(c) = q(c) = 0$. Then

$$\mathbb{E}_p[G_p - G_q] = \frac{1}{2}\sum_{c\in C} \frac{(p(c)-q(c))^2}{p(c)+q(c)} \geq 0, \tag{D.6}$$

$$\mathbb{E}_q[G_q - G_p] = \frac{1}{2}\sum_{c\in C} \frac{(p(c)-q(c))^2}{p(c)+q(c)} \geq 0. \tag{D.7}$$

The inequalities are strict whenever $p \neq q$.

Proof of Lemma D.2: Using $p(c)[p(c)-q(c)] = \frac{1}{2}\big([p(c)-q(c)][p(c)+q(c)] + [p(c)-q(c)]^2\big)$ and $\sum_{c\in C}[p(c)-q(c)] = 0$, it follows that

$$\sum_{c\in C} p(c)\frac{(p(c)-q(c))}{p(c)+q(c)} = \frac{1}{2}\sum_{c\in C}\frac{(p(c)-q(c))^2}{p(c)+q(c)}. \tag{D.8}$$

The identity above is symmetric, which completes the proof of Lemma D.2.

Note that the G_x in Lemma D.2 can be interpreted as the posterior probability of model $x \in \{p, q\}$ after observing c when the two models receive equal prior weight. We will construct a choice environment where payoffs are linked to such scores.

We now turn to proving the statement of Part 1. Fix the initial signal s_1. For $t \geq 2$, define the random variable $C_t = (r_1, s_2, r_2, \ldots, r_{t-1}, s_t)$ to be the observable data following s_1 up to time of action in period t. Under Assumption 1, C_t is independent of actions. Whenever both models assign positive probability to $s_1 = s$, define $p_t(c|s) := P(C_t = c|s_1 = s, \pi)$ and $q_t(c|s) := P(C_t = c|s_1 = s, \pi^*)$. Construct the following choice environment. $X_1 = \{0, 1\}$ and $X_t = \{d\}$ for some fixed default action d. Thus, the person's only meaningful choice is in period 1. The first action is a once-and-for-all choice of which model's continuation probabilities will be scored to provide payoffs. Specifically, fix $u_1 = 0$. Let $M > 0$ be some positive scalar. For $t > 1$, if $x_1 = 0$, let

$$u_t = M \frac{p_t(C_t|s_1)}{p_t(C_t|s_1) + q_t(C_t|s_1)}, \tag{D.9}$$

and if $x_1 = 1$, let

$$u_t = M \frac{q_t(C_t|s_1)}{q_t(C_t|s_1) + p_t(C_t|s_1)}, \tag{D.10}$$

Whenever both terms in the expressions for u_t are zero, assign both actions a payoff of $M/2$. For completeness, we now specify payoffs for the cases where $s_1 = s$ has zero probability under one of the models. If $P(s_1 = s|\pi) = 0 < P(s_1 = s|\pi^*)$, then for all $t > 1$, set $u_t = 0$ if $x_1 = 0$ and $u_t = M$ if $x_1 = 1$. If $P(s_1 = s|\pi^*) = 0 < P(s_1 = s|\pi)$, then for all $t > 1$, set $u_t = M$ if $x_1 = 0$ and $u_t = 0$ if $x_1 = 1$. Note that these zero-probability conventions do not affect the optimal choice of model on its own support.

Fix $s_1 = s$ that has positive probability under π and π^*. By Lemma D.2, action $x_1 = 0$ yields a weakly greater expected payoff than $x_1 = 1$ in every future period under π, conditional on $s_1 = s$. Moreover, $x_1 = 0$ is strictly better whenever $p_t(\cdot|s) \neq q_t(\cdot|s)$ for some t. Thus action $x_1 = 0$ is optimal for every s the person deems possible. A strategy $\phi = (N, \sigma)$ that notices nothing, takes $x_1 = 0$, and $x_t = d$ in all $t > 1$ is therefore a SAS given π. Since this construction involves a noticing partition that is a singleton in each period, $N^t = \{H^t\}$ for all t, the realized noticed cell has probability one under π and π^* for all t. Thus, ϕ is a StAS.

A similar argument implies that $x_1 = 1$ yields weakly greater expected payoffs in every period under π^*, and strictly greater payoffs when $p_t(\cdot|s) \neq q_t(\cdot|s)$. Let ϕ^* be a SAS under π^* that takes $x_1 = 1$.

Now consider a realized path of outcomes under π^*. If $P(s_1|\pi) = 0$, the per-period utility gap between ϕ and ϕ^* is M for all $t > 1$. Otherwise, define

$$L_t := \frac{p_t(C_t|s_1)}{q_t(C_t|s_1)}, \tag{D.11}$$

and notice that

$$\frac{P(h^t|\pi)}{P(h^t|\pi^*)} = L_t \frac{P(s_1|\pi)}{P(s_1|\pi^*)} \tag{D.12}$$

given Assumption 1. Because π is identifiably wrong, either the ratio in D.12 becomes zero at a finite date or its limit inferior is zero almost surely. In the former case, the continuation score deficit from following ϕ rather than ϕ^* is M from that date forward. In the latter case, $\liminf_{t\to 0} L_t = 0$. Along a common observable path (s_t, y^t), the absolute per-period payoff deficit from following ϕ rather than ϕ^* is

$$|u_t(\phi^*) - u_t(\phi)| = M\frac{|q_t(C_t|s_1) - p_t(C_t|s_1)|}{q_t(C_t|s_1) + p_t(C_t|s_1)} = M\frac{|1-L_t|}{1+L_t}. \tag{D.13}$$

Along a path where $L_t \to 0$, the right-hand side of the expression above converges to $M > 0$. Thus, the stable strategy ϕ is costly. Since $M > 0$ is arbitrary, the cost can be scaled arbitrarily.

Part 2. For each period t, let $d^t = (s_t, y_{t-1}, \ldots, y_1)$ denote the pre-action history of observables up to period t. Let D^t be the set of all possible d^t under π and π^*, and take $X_t = D^t$. Define

$$u_t(x_t, y_t|h^t) = \begin{cases} 1 & \text{if} \quad x_t = d^t, \\ -1 & \text{if} \quad x_t \neq d^t. \end{cases} \tag{D.14}$$

Thus, the unique optimal action at every history is to report the complete history of observables available at the time of choice.

Fix an arbitrary SAS $\phi = (\mathscr{N}, \sigma)$ given π. If a noticed cell contained histories consistent with two distinct d^t and $\tilde{d}^t$ that have positive probability under π, then no single action could be correct following both d^t and $\tilde{d}^t$. Hence, in every period t, each noticed cell $n^t \in N^t$ contains at most one element that has positive probability under π. Now consider a history of outcomes d^t that occurs with positive probability under π^*. Because π conceptualizes all possible events (relative to π^*), $P(d^t|\pi,\phi) > 0$. Hence, the realized noticed cell conditional on d^t is consistent with no π-possible history of outcomes other than d^t. Moreover, since π conceptualizes all events, the realized notice cell is also consistent with no π^*-possible history of outcomes other than d^t. Thus, $P(n^t|\pi,\phi) = P(d^t|\pi,\phi)$ and $P(n^t|\pi^*,\phi) = P(d^t|\pi^*,\phi)$, and hence

$$\frac{P(n^t|\pi,\phi)}{P(n^t|\pi^*,\phi)} = \frac{P(d^t|\pi,\phi)}{P(d^t|\pi^*,\phi)}. \tag{D.15}$$

Under Assumption 1, d^t is independent of actions and thus the likelihood ratio in (D.15) matches the one in Definition 4 (Identifiably Wrong). Since π is identifiably wrong, its limit inferior of (D.15) is zero almost surely (it cannot be zero at a finite period t since π conceptualizes all possible events). Hence every SAS in this constructed choice environment is attentionally unstable, and no StAS given π exists.

Finally, every SAS given π reports d^t correctly under the realized path, and every SAS given π^* does the same. Both attentional strategies thus earn the same payoff at every date. Every SAS given π is therefore costless at every date, and therefore asymptotically costless as well. ■

Proof of Proposition 7. We begin by proving a similar result under memory consistency (MC) and automatic recall (AR), as defined in Definitions A.1 and A.2. The following proof of this result (Lemma D.3) makes use of stochastic noticing strategies. This extension is irrelevant for our framework in the main text without MC and AR, but it can be relevant under MC and AR. We will allow for this possibility (i.e., stochastic noticing strategies) in proving the following lemma, but we emphasize that these details are unnecessary for proving Proposition 7 as stated.

Lemma D.3. *Consider a stationary environment where Assumptions 1 and 2 hold. Suppose S is a singleton and $P(y|\theta) \in (0,1)\ \forall (y,\theta) \in Y(\pi) \times \text{supp}(\pi)$. Then the model π is stable across stationary objectives under memory consistency and automatic recall if there exists $\theta \in \text{supp}(\pi)$ such that*

$$P(m_\pi(y)|\theta) \geq P(m_\pi(y)|\theta^*)\ \forall y \in Y(\pi). \tag{D.16}$$

Proof of Lemma D.3. For any (X,u), the person believes it is sufficient to record $m_\pi(y_t)$ each period since this is sufficient for updating beliefs about θ (given S is a singleton). There are two cases to consider depending on whether the support of outcomes under the misspecified model, $Y(\pi)$, matches the true support of outcomes, $Y(\theta^*)$:

Case 1: Suppose $Y(\pi) = Y(\theta^*)$. This implies that condition (D.16) holds only if it holds with equality: $P(m_\pi(y)|\theta) = P(m_\pi(y)|\theta^*)\ \forall y \in Y(\pi)$. Under this condition, such a noticing strategy (i.e., distinguishing $m_\pi(y_t)$ each period) constitutes a stable attentional strategy given π. To see this, consider any history of outcomes $y^t \in Y(\pi)^{t-1}$, and, slightly abusing notation, denote the corresponding noticed history by $n^t = (m_\pi(y_{t-1}), \ldots, m_\pi(y_1))$. The model π is attentionally stable if for some $\theta \in \text{supp}(\pi)$, $\liminf_{t\to\infty} P(n^t|\theta)/P(n^t|\theta^*) > 0$ with probability 1 given θ^*. Since $Y(\pi)$ is finite, enumerate the elements of $m_\pi(\cdot)$ by $\{m_\pi^1, \ldots, m_\pi^N\}$. For any $y^t \in Y(\pi)^{t-1}$ with $t \geq 2$, let $k_t^n(y^t) := \sum_{\tau=1}^{t-1} \mathbf{1}\{y_\tau \in m_\pi^n\}$ denote the count of entries in y^t that are in m_π^n. Then for any $\theta \in \text{supp}(\pi)$,

$$\frac{P(n^t(y^t)|\theta)}{P(n^t(y^t)|\theta^*)} = \frac{\prod_{n=1}^N P(m_\pi^n|\theta)^{k_t^n(y^t)}}{\prod_{n=1}^N P(m_\pi^n|\theta^*)^{k_t^n(y^t)}} = \left(\frac{\prod_{n=1}^N P(m_\pi^n|\theta)^{k_t^n(y^t)/(t-1)}}{\prod_{n=1}^N P(m_\pi^n|\theta^*)^{k_t^n(y^t)/(t-1)}}\right)^{t-1}. \tag{D.17}$$

This likelihood ratio is identical to the one considered in Lemma D.1 except the "noticed" outcome space here is $\{m_\pi^1, \ldots, m_\pi^N\}$ rather than $Y(\pi)$. Since $P(m_\pi(y)|\theta) = P(m_\pi(y)|\theta^*)\ \forall y \in Y(\pi)$, the ratio in (D.17) converges to 1 in t, and hence there exists a stable attentional strategy given π irrespective of (X,u). Additionally, it is straightforward that this noticing strategy satisfies MC and AR: note that,

for all t,

$$n^t(y^t) = \{\tilde{y}^t \in Y(\pi)^t \mid m_\pi(\tilde{y}_\tau) = m_\pi(y_\tau)\ \forall \tau = 1,\ldots,t-1\}. \tag{D.18}$$

Thus, for all $y_t \in Y(\pi)$, if $\tilde{h}^t \in n^t(h^t)$, then the continuation history $(y_t, \tilde{h}^t) = (y_t, \tilde{y}_{t-1}, \ldots, \tilde{y}_1) \in n^{t+1}((y_t, h^t))$, which implies memory consistency. Furthermore, if $\tilde{h}^t \notin n^t(h^t)$, then there must exist some $\tau < t$ such that $m_\pi(\tilde{y}_\tau) \neq m_\pi(y_\tau)$. This implies that, for all $y_t, \tilde{y}_t \in Y(\pi)$, the continuation history $(y_t, \tilde{h}^t) = (y_t, \tilde{y}_{t-1}, \ldots, \tilde{y}_1) \notin n^{t+1}((y_t, h^t))$, and hence AR holds.

Case 2: Suppose $Y(\pi) \neq Y(\theta^*)$. Thus, condition (D.16) need not hold with equality. The case of equality is handled above. To handle the case without equality, suppose there exists $\theta \in \text{supp}(\pi)$ such that $P(m_\pi(y)|\theta) \geq P(m_\pi(y)|\theta^*)\ \forall y \in Y(\pi)$ with a strict inequality for some $\tilde{y} \in Y(\pi)$. As such, the support $Y(\pi)$ must exclude at least one outcome in $Y(\theta^*)$, so the set $Y^0 := Y(\theta^*) \setminus Y(\pi)$, is non-empty. Let $P^0 := \sum_{y \in Y^0} P(y|\theta^*)$. Again enumerate the elements of $m_\pi(\cdot)$ as $\{m_\pi^1, \ldots, m_\pi^N\}$. We will construct an alternative collection of sufficient statistics over $Y(\pi) \cup Y^0$ for each time period t, denoted $\left\{\tilde{m}_\pi^{(1,t)}, \ldots, \tilde{m}_\pi^{(N,t)}\right\}$ such that $P\left(\tilde{m}_\pi^{(n,t)}\middle|\theta\right) = P\left(\tilde{m}_\pi^{(n,t)}\middle|\theta^*\right)\ \forall n = 1,\ldots,N$ and $\forall t \in \mathbb{N}$. Suppose the person merges $y \in Y^0$ with elements of a partition over $Y(\pi)$ according to a randomizing device governed by discrete i.i.d. random variables z_t with support $\mathscr{Z} = \{1,\ldots,N\}$ and mass function $P(z_t = n) = [P(m_\pi^n|\theta) - P(m_\pi^n|\theta^*)]/P^0$. We augment the observation space to be $Y \times \mathscr{Z}$. Then for all t and all outcomes (y_t, z_t), define $\tilde{m}_\pi^{(n,t)}$ by

$$y_t \in \tilde{m}^{(n,t)} \Leftrightarrow \left(y_t \in m_\pi^n\right) \text{ or } \left(y_t \notin Y(\pi) \text{ and } z_t = n\right).$$

That is, y_t is lumped according to m_π if $y_t \in Y(\pi)$ and is otherwise lumped stochastically according to the randomizing device z_t. Thus, each $\tilde{m}_\pi^{(n,t)}$ is encoded with the same probability under both θ and θ^*: from the specification of $P(z_t = n)$, it follows that for all $n \in \{1,\ldots,N\}$ and all $t \in \mathbb{N}$, $P\left(\tilde{m}_\pi^{(n,t)}\middle|\theta^*\right) = P(m_\pi^n|\theta^*) + P^0 \cdot P(z_t = n) = P(m_\pi^n|\theta) = P\left(\tilde{m}_\pi^{(n,t)}\middle|\theta\right)$, where the last equality follows from the fact that the only realizations of y_t included in $\tilde{m}_\pi^{(n,t)}$ beyond those in m_π^n have probability zero under θ. Given that the distribution of noticed outcomes under $\tilde{m}_\pi(\cdot)$ is equivalent for both θ and θ^*, the proof concludes along the same lines as the case above with $Y(\pi) = Y(\theta^*)$ aside from the simple difference that each m_π^n above is replaced with the corresponding $\tilde{m}_\pi^{(n,t)}$. Furthermore, an analogous argument to that in the previous case establishes that the noticing strategy satisfies MC and AR.

We now complete the proof of Proposition 7 as stated. Unlike Lemma D.3 above, Proposition 7 does not assume memory consistency and automatic recall. Thus, we now extend the proof above by dropping these two assumptions.

For any (X, u), the person believes it is sufficient to notice $m_\pi(y^t)$ and a π-sufficient coarsening of s_t each period since this is sufficient for updating beliefs about θ and for taking the optimal action. Hence, for any (X, u), noticing this data constitutes the noticing strategy for some SAS.

Additionally, by Proposition 2, we can construct the π-sufficient coarsening of s_t such that each coarsened signal has positive probability under π. By assumption, there exists $\theta \in \text{supp}(\pi)$ such that $\liminf_{t\to\infty} P(m_\pi(y^t)|\theta)/P(m_\pi(y^t)|\theta^*) > 0$ with probability 1 under θ^*. Thus, under the noticing strategy just described, $\liminf_{t\to\infty} P(n^t|\theta)/P(n^t|\theta^*) > 0$ (with probability 1 under θ^*), which implies that π admits a stable attentional strategy. Finally, since (X,u) was arbitrary, π is stable across stationary objectives. ■

Proof of Proposition 8. Step 1: If π is stable across all objectives, then π fails to conceptualize an event (relative to π^*). Toward a contradiction, suppose π does not fail to conceptualize an event (relative to π^*) and thus assigns positive probability to all finite histories h^t that occur under π^*. Consider a choice environment as in Part 2 of the proof of Proposition 6, where in each period t, the person has an incentive to accurately report the complete history of observables up to period t. Thus, as argued in that proof, any SAS $\phi = (\mathscr{N}, \sigma)$ must have the following property: for all t and all $n^t \in N^t$, π assigns positive probability to at most one element of n^t. Continuing to follow the arguments in the proof in Part 2 of Proposition 6, under such a SAS ϕ, the inferred history in each period t is precisely the realized history: since $n^t(h^t)$ includes only one history that is assigned positive probability by π, and because h^t has positive probability under π, h^t must be the sole element of n^t that the person deems possible. Consequently,

$$\frac{P(n^t(h^t)|\pi,\phi)}{P(n^t(h^t)|\pi^*,\phi)} = \frac{P(h^t|\pi)}{P(h^t|\pi^*)} \quad \text{and thus} \quad \liminf_{t\to\infty} \frac{P(n^t(h^t)|\pi,\phi)}{P(n^t(h^t)|\pi^*,\phi)} = \liminf_{t\to\infty} \frac{P(h^t|\pi)}{P(h^t|\pi^*)} = 0, \tag{D.19}$$

where the final equality follows from the assumption that π is identifiably wrong. Thus, Definition 3 deems π attentionally unstable with respect to π^* and ϕ, implying π is not stable across all objectives.

Step 2: Combining Step 1 with results from Proposition 9. To prove the statement of Proposition 8, we rely on steps from the proof of Proposition 9. Fix the outcome environment and suppose $\phi = (\mathscr{N}, \sigma)$ is a stable attentional strategy given π for some choice environment $(\times_{t=1}^\infty X_t, \times_{t=1}^\infty u_t)$. First suppose (1) does not hold. Thus, there does not exist a temporary refinement $\mathscr{N}'$ of $\mathscr{N}$ such that there is no stable attentional strategy given π under $\mathscr{N}'$. By Step 1 from the proof of Proposition 9, if π fails to conceptualize an event (relative to π^*), then π is fragile to temporary-noticing refinements. Since (1) does not hold, π is not fragile to temporary-noticing refinements and π must therefore conceptualizes all possible events (relative to π^*). Then, by Step 1 of this proof, π cannot be stable across all objectives, and hence (2) must hold. Second, suppose (2) does not hold. Thus, there does not exist an alternative choice environment $(\times_{t=1}^\infty X'_t, \times_{t=1}^\infty u'_t)$ for which there is no stable attentional strategy given π. This implies that π is stable across all objectives. By Step 1 of this proof, this implies that π fails to conceptualize an event (relative to π^*), and hence Step 1 from the proof of Proposition 9 implies that π is fragile to temporary-noticing refinements. Thus, (1) must hold. ■

Proof of Corollary 1. This follows from Step 1 of the proof of Proposition 8. ■

Proof of Proposition 9. We begin with two useful lemmas.

Lemma D.4. *Fix an outcome environment. Consider a choice environment that admits a StAS $\phi = (\mathscr{N}, \sigma)$ given π. Suppose there is a finite history of outcomes $y^t = (y_{t-1}, \ldots, y_1)$ such that $P(y^t|\pi) = 0$ and $P(y^t|\pi^*) > 0$. Then there is a temporary noticing refinement $\mathscr{N}'$ of $\mathscr{N}$ such that (i) $N'^\tau = N^\tau$ for all $\tau \neq t$ and (ii) there is no behavioral strategy σ' such that $(\mathscr{N}', \sigma')$ is a StAS given π.*

Proof of Lemma D.4: Let N'^t refine N^t so that it distinguishes the event y^t. Define the cells of N'^t as follows: for each $n^t \in N^t$, let $n'^t_O = \{h^t \mid y^t \text{ occurs under } h^t\}$ and $n'^t_C = \{h^t \mid y^t \text{ does not occur under } h^t\}$. Let N'^t_O collect all of the former class of cells under which y^t occurs under h^t. For any behavioral strategy σ' let $\phi' = (\mathscr{N}', \sigma')$ and note that Assumption 1 implies $P(y^t|\pi, \phi') = 0$ and $P(y^t|\pi^*, \phi') > 0$. Moreover, for each $n'^t_O \in N'^t_O$, $P(n'^t_O|\pi, \phi') = 0$ and $P(n'^t_O|\pi^*, \phi') > 0$. Thus, with positive probability under π^*, we have $P(n'^t_O|\pi, \phi')/P(n'^t_O|\pi^*, \phi') = 0$, which means ϕ' is attentionally unstable given π by Definition 3. This completes the proof of Lemma D.4.

Lemma D.5. *Fix an outcome environment. Consider a choice environment that admits a StAS $\phi = (\mathscr{N}, \sigma)$ given π. Suppose π conceptualizes all possible events (relative to π^*). Then every temporary noticing refinement $\mathscr{N}'$ of $\mathscr{N}$ can be paired with a behavioral strategy σ' such that $(\mathscr{N}', \sigma')$ is a StAS given π.*

Proof of Lemma D.5: Consider a refinement $\mathscr{N}'$ of $\mathscr{N}$ such that $N'^\tau = N^\tau$ for all $\tau \neq t$. For each refined cell $n'^t \in N'^t$, let $n^t \in N^t$ be its unique parent cell and define $\sigma'_t(n'^t) = \sigma_t(n^t)$. The attentional strategy $\phi' = (\mathscr{N}', \sigma')$ behaves exactly as ϕ at every history. Additionally, it induces the same probability measure over actions and observables as ϕ and thus achieves the same subjective payoffs as ϕ. Hence, ϕ' is a SAS. Let T be the finite set of dates in which N'^t refines N^t. At every $t \in T$, any $n'^t \in N'^t$ that occurs with positive probability under π^* also has positive probability under π under behavioral strategy σ'. This rules out the possibility of the likelihood ratio in Definition 3 reaching zero in finite time. For all $t > \max T$, $N'^t = N^t$ and the induced distribution over histories is the same as under ϕ. The process of likelihood ratios in Definition 3 under ϕ' therefore coincides with the one under ϕ from date $\max T$ onward and has a strictly positive limit inferior given that ϕ is a StAS given π. Hence, ϕ' is also a StAS given π, completing the proof of Lemma D.5.

We now prove the statement of Proposition 9. We prove the equivalence of (1) and (2) by showing that each of these statements is equivalent to π failing to conceptualize an event (relative to π^*).

Step 1: If π fails to conceptualize an event (relative to π^*), then in any choice environment that admits a StAS, π is fragile to temporary-noticing refinements. Consider any choice environment that admits a StAS given π. Consider any such StAS $\phi = (\mathscr{N}, \sigma)$. By Lemma D.4, there exists a temporary noticing refinement $\mathscr{N}'$ of $\mathscr{N}$ under which there exists no StAS involving $\mathscr{N}'$. Since

the choice environment and StAS were arbitrary, it follows that π is fragile to temporary-noticing refinements.

Step 2: If π is fragile to temporary-noticing refinements in any choice environment that admits a StAS, then π fails to conceptualize an event (relative to π^*). Fix a choice environment that admits a StAS $\phi = (\mathscr{N}, \sigma)$. Toward a contradiction, suppose π conceptualizes all possible events (relative to π^*). Lemma D.5 then says that for a temporary refinement $\mathscr{N}'$ of $\mathscr{N}$, one can construct an attentional strategy $\phi' = (\mathscr{N}', \sigma')$ that is a StAS given π. This contradicts π being fragile to temporary-noticing refinements in choice environments that admit a StAS.

Step 3: If π fails to conceptualize an event (relative to π^*), then there exists a choice environment where π is fragile to theory perturbations toward any identifiably wrong π' that conceptualizes all possible events (relative to π^*). We will construct a choice environment to demonstrate this fragility. Let $d^t := (s_t, y^t)$ be the observable data prior to acting in period t. Let $Z_t(\pi) = \{d^t \mid P(d^t|\pi^*) > 0 \text{ and } P(d^t|\pi) = 0\}$. The action set in each t is $X_t = \{o\} \cup Z_t(\pi)$, where o is a default action. Consider the history-dependent utility function u_t defined as follows:

$$u_t(x_t, y_t|h^t) = \begin{cases} 1 & \text{if} \quad d^t \notin Z_t(\pi) \text{and} x_t = o, \\ 1 & \text{if} \quad d^t \in Z_t(\pi) \text{and} x_t = d^t, \\ 0 & \text{otherwise.} \end{cases} \tag{D.20}$$

Thus, the person should use the default action on every history that π regards as possible and instead report the history of outcomes under histories deemed impossible under π. Under π, a singleton noticing strategy ($N^t = \{H^t\}$ for all t) paired with the constant behavioral strategy $x_t = o$ is a StAS. We now show that this StAS is fragile to the relevant theory perturbations. Let π' be identifiably wrong and conceptualize all possible events (relative to π^*). Let $\pi'' = (1-\alpha)\pi + \alpha\pi'$ for some $\alpha \in (0,1)$. Every $d^t \in Z(\pi)$ now has positive probability under π'', and the unique optimal action conditional on d^t is $x_t = d^t$. Thus, any SAS given π'' must separately distinguish each $d^t \in Z_t(\pi)$. Once d^τ occurs, then for all $t > \tau$, $d^t \in Z_t(\pi)$ and $x_t = d^t$, and hence for any SAS $\phi = (\mathscr{N}, \sigma)$ under π'',

$$\frac{P(n^t|\pi'', \phi)}{P(n^t|\pi^*, \phi)} = \frac{P(d^t|\pi'')}{P(d^t|\pi^*)}. \tag{D.21}$$

Since $P(d^t|\pi) = 0$, the previous equality implies

$$\frac{P(n^t|\pi'', \phi)}{P(n^t|\pi^*, \phi)} = \alpha \frac{P(d^t|\pi')}{P(d^t|\pi^*)}. \tag{D.22}$$

Under Assumption 1, π' being identifiably wrong implies that $\liminf_{t\to\infty} \alpha P(d^t|\pi')/P(d^t|\pi^*) = 0$; thus the arbitrary SAS ϕ is attentionally unstable given π''. Hence, in this choice environment,

π is fragile to theory perturbations toward any identifiably wrong π' that conceptualizes all events (relative to π^*).

Step 4: If there exists a choice environment where π is fragile to theory perturbations toward any identifiably wrong π' that conceptualizes all possible events (relative to π^*), then π fails to conceptualize an event (relative to π^*). Toward a contradiction, suppose π conceptualizes all possible events (relative to π^*). Then simply taking $\pi' = \pi$ implies that for any $\alpha \in (0,1)$, $\pi'' = \pi$. Hence, if ϕ is a StAS given π, it is also a StAS given π''. This contradicts the assumption that π is fragile to any theory perturbation toward a model π' that conceptualizes all possible events (relative to π^*).

Step 5: Piecing together the previous steps. We now piece together the steps above to establish the proposition as stated. Steps 1 and 2 together imply that condition (1) from Proposition 9 is equivalent to π failing to conceptualize an event (relative to π^*). Steps 3 and 4 together imply that condition (2) from Proposition 9 is equivalent to π failing to conceptualize an event (relative to π^*). Thus conditions (1) and (2) are equivalent to each other. ■

D.2 Proofs of Supplemental Results

Proof of Observation B.1. Under the full-attention noticing strategy, $P(n^t(h^t)|\pi) = P(h^t|\pi)$ and $P(n^t(h^t)|\lambda) = P(h^t|\lambda)$ for all $h^t \in H$. Suppose $D(\theta^*\|\lambda)$ and $D(\theta^*\|\pi)$ are finite (the case where one is infinite is obvious), and thus $P(h^t|\pi) > 0$ and $P(h^t|\lambda) > 0$ for all h^t in the support of $P(\cdot|\theta^*)$. Let $\Theta_\pi^{\min} = \arg\min_{\tilde{\theta}\in\text{supp}(\pi)} D(\theta^*\|\tilde{\theta})$, $\Theta_\lambda^{\min} = \arg\min_{\tilde{\theta}\in\text{supp}(\lambda)} D(\theta^*\|\tilde{\theta})$, and, for any set $\tilde{\Theta} \subseteq \Theta$, $P_\pi(h^t|\tilde{\Theta}) = \sum_{\theta'\in\tilde{\Theta}} P(h^t|\theta')\pi(\theta'|\tilde{\Theta})$, and define $P_\lambda(h^t|\tilde{\Theta})$ analogously. We can then expand $P(h^t|\pi)$ as follows:

$$\begin{aligned} P(h^t|\pi) &= P_\pi(h^t|\Theta_\pi^{\min})\cdot\pi(\Theta_\pi^{\min}) + P_\pi(h^t|\text{supp}(\pi)\setminus\Theta_\pi^{\min})\cdot(1-\pi(\Theta_\pi^{\min})) \\ &= P_\pi(h^t|\Theta_\pi^{\min})\cdot\left[\pi(\Theta_\pi^{\min}) + \frac{P_\pi(h^t|\text{supp}(\pi)\setminus\Theta_\pi^{\min})}{P_\pi(h^t|\Theta_\pi^{\min})}\cdot(1-\pi(\Theta_\pi^{\min}))\right]. \end{aligned}$$

Similarly expand $P(h^t|\lambda)$. As a result,

$$\frac{P(h^t|\pi)}{P(h^t|\lambda)} = \frac{P_\pi(h^t|\Theta_\pi^{\min})\cdot\left[\pi(\Theta_\pi^{\min}) + \frac{P_\pi(h^t|\text{supp}(\pi)\setminus\Theta_\pi^{\min})}{P_\pi(h^t|\Theta_\pi^{\min})}\cdot(1-\pi(\Theta_\pi^{\min}))\right]}{P_\lambda(h^t|\Theta_\lambda^{\min})\cdot\left[\lambda(\Theta_\lambda^{\min}) + \frac{P_\lambda(h^t|\text{supp}(\lambda)\setminus\Theta_\lambda^{\min})}{P_\lambda(h^t|\Theta_\lambda^{\min})}\cdot(1-\lambda(\Theta_\lambda^{\min}))\right]}. \tag{D.23}$$

Without loss of generality, assume $\Theta_\pi^{\min} \neq \text{supp}(\pi)$. Thus, for all $\theta \in \text{supp}(\pi)\setminus\Theta_\pi^{\min}$ and $\theta' \in \Theta_\pi^{\min}$, we have $D(\theta^*\|\theta) > D(\theta^*|\theta')$. It follows from Part 1 of Lemma D.1 that $\frac{P(h^t|\theta)}{P(h^t|\theta')} \xrightarrow{\text{a.s.}} 0$, implying

that $\frac{P_\pi(h^t|\text{supp}(\pi)\setminus\Theta_\pi^{\min})}{P_\pi(h^t|\Theta_\pi^{\min})} \xrightarrow{\text{a.s.}} 0$ (and analogously for the λ case). Thus, (D.23) implies that

$$\frac{P(h^t|\pi)}{P(h^t|\lambda)} \xrightarrow{\text{a.s.}} \frac{P_\pi(h^t|\Theta_\pi^{\min})}{P_\lambda(h^t|\Theta_\lambda^{\min})} \cdot \frac{\pi(\Theta_\pi^{\min})}{\lambda(\Theta_\lambda^{\min})}. \tag{D.24}$$

Because Θ is finite, the pairwise likelihood-ratio conclusions from Lemma D.1 extend to the finite mixtures above; the conditional prior weights do not affect the asymptotic comparison.

If $\Delta D(\theta^*\|\lambda,\pi) > 0$ (i.e., $D(\theta^*\|\pi) < D(\theta^*\|\lambda)$), then for all $\theta \in \Theta_\pi^{\min}$ and $\theta' \in \Theta_\lambda^{\min}$, we have $D(\theta^*\|\theta) = D(\theta^*\|\pi) < D(\theta^*\|\lambda) = D(\theta^*\|\theta')$. It then follows from Part 2 of Lemma D.1 that $\frac{P(h^t|\theta)}{P(h^t|\theta')} \xrightarrow{\text{a.s.}} \infty$, and thus $\frac{P_\pi(h^t|\Theta_\pi^{\min})}{P_\lambda(h^t|\Theta_\lambda^{\min})} \xrightarrow{\text{a.s.}} \infty$. Therefore, D.24 implies that $\frac{P(h^t|\pi)}{P(h^t|\lambda)} \xrightarrow{\text{a.s.}} \infty$, and hence π is attentionally stable with respect to λ and a full-attention SAS. Similarly, if $\Delta D(\theta^*\|\lambda,\pi) = 0$ with $\theta^* \in \text{supp}(\lambda) \cup \text{supp}(\pi)$, then Part 3 of Lemma D.1 implies that $\frac{P(h^t|\pi)}{P(h^t|\lambda)}$ converges a.s. to a positive constant, again implying that π is attentionally stable with respect to λ and a full-attention SAS.

Finally, if $\Delta D(\theta^*\|\lambda,\pi) < 0$, then a similar argument based on Part 1 of Lemma D.1 implies that $\frac{P_\pi(h^t|\Theta_\pi^{\min})}{P_\lambda(h^t|\Theta_\lambda^{\min})} \xrightarrow{\text{a.s.}} 0$, and thus π is attentionally unstable with respect to λ and a full-attention SAS. ■

Proof of Proposition B.1. Suppose π exhibits a dogmatic error. The definition of m_π then implies that, for any $y \in Y(\pi)$, $m_\pi(y) = Y(\pi)$. It is therefore immediate from Proposition 7 that π is stable across stationary objectives. ■

Proof of Proposition B.2. Suppose π is censored. Thus there exists $\theta \in \text{supp}(\pi)$ such that $P(m_\pi(y)|\theta) = P(m_\pi(y)|y \in Y(\theta), \theta^*)$ for all $y \in Y(\theta)$. Since $Y(\theta) \subset Y(\theta^*)$, it must be that $P(m_\pi(y)|\theta^*) \leq P(m_\pi(y)|y \in Y(\theta), \theta^*) = P(m_\pi(y)|\theta)$ for all $y \in Y(\theta)$, which implies that the sufficient condition for stability for all objectives (with memory consistency and automatic recall) from the proof of Proposition 7 holds (Condition D.16). If π is stable across stationary objectives with memory consistency and automatic recall, then it is stable across stationary objectives more generally. ■

Proof of Proposition B.3. Because π exhibits predictor neglect, the minimal sufficient statistic $m_\pi(y)$ does not distinguish observations that differ only in the neglected signals. Thus, $m_\pi(r, s^1, \ldots, s^K) = m_\pi(r, \tilde{s}^1, \ldots, \tilde{s}^K)$ whenever $(r, s^1, \ldots, s^J) = (\tilde{r}, \tilde{s}^1, \ldots, \tilde{s}^J)$. Hence, for every stationary objective, there exists a SAS with noticing strategy m_π that ignores distinctions involving only $(s^{J+1}, \ldots, s^K)$. Let m_π be the partition of $Y(\pi)$ defined in (3) and enumerate its elements as $\{m_\pi^1, \ldots, m_\pi^N\}$. Then for each $n = 1, \ldots, N$, $m_\pi(y^t)$ must record the count $k_t^n(y^t)$ of outcomes y_τ, $\tau < t$, such that $y_\tau \in m_\pi^n$. Then $\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)}$ is identical to (D.26) from the proof of Proposition B.5, below. It then follows from that proof that π is stable across stationary objectives if $P(m_\pi^n|\theta^*) = P(m_\pi^n|\theta)$ for all $n = 1, \ldots, N$. The fact that $P(r|s^1, \ldots, s^J, \theta) = P(r|s^1, \ldots, s^J, \theta^*)$ for all possible $(r, s^1, \ldots, s^J)$ under θ^* along with the

definition of m_π implies $P(m_\pi^n|\theta^*) = P(m_\pi^n|\theta)$ for all $n = 1,\ldots,N$, so π is stable across stationary objectives. ■

Proof of Proposition B.4. Suppose $\{P(\cdot|\theta)\}_{\theta\in\text{supp}(\pi)\cup\theta^*}$ satisfies VLRP. Thus, for each $y \in Y(\pi)$, there exists no $y' \in Y(\pi)$ such that $y' \neq y$ and $\frac{P(y|\theta)}{P(y'|\theta)}$ is constant in $\theta \in \text{supp}(\pi)$. This implies that for all $y \in Y(\pi)$, $m_\pi(y) = \{y\}$. Accordingly, for any $y^t \in Y(\pi)^{t-1}$ with $t \geq 2$ and all $y_n \in Y(\pi)$, $m_\pi(y^t)$ must record the count of outcomes y_τ in y^t such that $y_\tau = y_n$ (denoted by $k_t^n(y^t)$). Then

$$\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)} = \frac{\prod_{n=1}^N P(y_n|\theta)^{k_t^n(y^t)}}{\prod_{n=1}^N P(y_n|\theta^*)^{k_t^n(y^t)}} = \left(\frac{\prod_{n=1}^N P(y_n|\theta)^{k_t^n(y^t)/(t-1)}}{\prod_{n=1}^N P(y_n|\theta^*)^{k_t^n(y^t)/(t-1)}}\right)^{t-1}. \tag{D.25}$$

Hence, Lemma D.1 along with (D.25) implies that $\lim_{t\to\infty} P(m_\pi(y^t)|\theta)/P(m_\pi(y^t)|\theta^*) > 0$ with probability 1 given θ^* iff $-D(\theta^*\|\theta) \geq 0$. Since the Kullback-Leibler Divergence is non-negative, $-D(\theta^*\|\theta) \geq 0 \Leftrightarrow D(\theta^*\|\theta) = 0 \Leftrightarrow P(\cdot|\theta) = P(\cdot|\theta^*)$, which contradicts VLRP. Hence, π is not stable across stationary objectives. ■

Proof of Proposition B.5. Let m_π be the partition of $Y(\pi)$ defined in (3). Since this partition is unique and finite, enumerate its elements as $\{m_\pi^1,\ldots,m_\pi^N\}$. For any $y^t \in Y(\pi)^{t-1}$ with $t \geq 2$, $m_\pi(y^t)$ must record the count of outcomes y_τ in y^t such that $y_\tau \in m_\pi^n$ (denoted by $k_t^n(y^t)$). Then

$$\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)} = \frac{\prod_{n=1}^N P(m_\pi^n|\theta)^{k_t^n(y^t)}}{\prod_{n=1}^N P(m_\pi^n|\theta^*)^{k_t^n(y^t)}} = \left(\frac{\prod_{n=1}^N P(m_\pi^n|\theta)^{k_t^n(y^t)/(t-1)}}{\prod_{n=1}^N P(m_\pi^n|\theta^*)^{k_t^n(y^t)/(t-1)}}\right)^{t-1}. \tag{D.26}$$

The likelihood ratio (D.26) is effectively identical to the one considered in the proof of Proposition 7 (Equation D.17). Thus $\lim_{t\to\infty} P(m_\pi(y^t)|\theta)/P(m_\pi(y^t)|\theta^*) > 0$ with probability 1 given θ^* iff $D(\theta^*\|\theta) = 0$, where D in this case is the KL distance from $P_m(\cdot|\theta)$ to $P_m(\cdot|\theta^*)$ with $P_m(\cdot|\theta)$ denoting the implied probability measure over $\{m_\pi^1,\ldots,m_\pi^N\}$ given θ. Since π is overly elaborate, there exists some m_π^n such that $m_\pi^n \cap Y(\theta^*) = \varnothing$, implying $P_m(m_\pi^n|\theta) > 0$ while $P_m(m_\pi^n|\theta^*) = 0$. Finally, since $D(\theta^*\|\theta)$ is non-negative and $D(\theta^*\|\theta) = 0 \Leftrightarrow P_m(\cdot|\theta) = P_m(\cdot|\theta^*)$, and because the latter equality does not hold (as previously noted), we must have $D(\theta^*\|\theta) > 0 = D(\theta^*\|\theta^*)$. It then follows from Lemma D.1 and Remark D.1 that ratio (D.26) converges to 0 a.s. and π is therefore not stable across stationary objectives. ■

Proof of Proposition B.6. Following the setup of the proof of Proposition B.5, let m_π be the partition of $Y(\pi)$ defined in (3) and enumerate its elements as $\{m_\pi^1,\ldots,m_\pi^N\}$. Again following the proof of Proposition B.5, for any $y^t \in Y(\pi)^{t-1}$ with $t \geq 2$, $m_\pi(y^t)$ must record the count of outcomes y_τ in y^t such that $y_\tau \in m_\pi^n$ (denoted by $k^n(y^t)$), and thus $\lim_{t\to\infty} P(m_\pi(y^t)|\theta)/P(m_\pi(y^t)|\theta^*)$ (which in this case is identical to the likelihood ratio in Equation D.26) is positive with probability 1 given θ^* iff $D(\theta^*\|\theta) = 0$, where D in this case is the KL distance from $P_m(\cdot|\theta)$ to $P_m(\cdot|\theta^*)$. Note that

$D(\theta^*\|\theta) = 0$ iff $P_m(\cdot|\theta) = P_m(\cdot|\theta^*)$. By property (d), for every $\theta \in \text{supp}(\theta)$ there exists some cell m^n_π such that $P_m(m^n_\pi|\theta) \neq P_m(m^n_\pi|\theta^*)$. Hence, $P_m(\cdot|\theta) \neq P_m(\cdot|\theta^*)$ for all $\theta \in \text{supp}(\pi)$, so $D(\theta^*\|\theta) > 0$. It then follows from the likelihood-ratio argument in (D.26) that $\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)} \to 0$ almost surely for all $\theta \in \text{supp}(\pi)$. Thus, π is not stable across stationary objectives. ■

Proof of Proposition C.1. For the analyst who does not need to provide predictions, a minimal SAS contains at most two elements, $n^t_B = \{h^t \in H^t \mid \mathbb{E}_{\pi,t}[V_t] \geq p_t\}$ and $n^t_S = \{h^t \in H^t \mid \mathbb{E}_{\pi,t}[V_t] < p_t\}$. Thus, the analyst must distinguish only whether $\mathbb{E}_{\pi,t}[V_t]$ is high enough to justify buying at the current price. The analyst who needs to provide predictions, on the other hand, needs to additionally notice $\mathbb{E}_{\pi,t}[M_t]$. It is clear then that any sufficient noticing strategy for the analyst who needs to provide predictions is also a sufficient noticing strategy for the analyst who does not, but not vice-versa. This implies that if there is a stable attentional strategy for the analyst who needs to provide predictions, then there is also a stable attentional strategy for the analyst who does not need to provide predictions. ■

Proof of Proposition C.2. Consider the setup from Section C.2, and consider a model π such that $\bar{q} < 1$ where $\bar{q} = \max \text{supp}(\pi)$. There are two relevant types of periods in this application: (1) periods where the person decides whether to buy (or renew) the membership, and (2) periods where the person decides whether to visit the gym (assuming she has a membership). We assume these types of periods are mutually exclusive. (This is inconsequential but simplifies matters.)

($\Leftarrow$) Suppose (i) π is non-degenerate and (ii) the person has an incentive to track her precise willingness to pay for a gym membership. Then in each period t where the person decides whether to buy a membership, N^t has exactly $t+1$ elements—one for each possible realization of $\sum_{k=1}^t \mathbf{1}\{\beta_k = \beta\}$ under π (see the discussion in Footnote 53). In reality $\beta_k = \beta$ in every period. Thus, for all h^t, $n^t(h^t)$ perfectly reveals that $\beta_k = \beta$ for all $k \leq t$. Thus,

$$\begin{aligned} \lim_{t\to\infty} \frac{P(n^t(h^t)|\pi)}{P(n^t(h^t)|\pi^*)} &= \lim_{t\to\infty} P\left(\sum_{k=1}^t \mathbf{1}\{\beta_k = \beta\} = t \,\middle|\, \pi\right) \\ &\leq \lim_{t\to\infty} (\bar{q})^t = 0, \end{aligned}$$

where the first equality follows from the fact that $P(n^t(h^t)|\pi^*) = P\left(\sum_{k=1}^t \mathbf{1}\{\beta_k = \beta\} = t \,\middle|\, \pi^*\right) = 1$ when π^* is degenerate on the true parameter, $q^* = 1$.

($\Rightarrow$) We now show that conditions (i) and (ii) are necessary for there to exist no stable attentional strategy given π. If (i) fails to hold, then $\text{supp}(\pi)$ is either a singleton or includes only one value other than zero. First consider the singleton case. This implies a minimal SAS is such that (1) in each period t where the person decides whether to buy a membership, N^t is a singleton; (2) in each period t where the person decides whether to visit the gym, N^t distinguishes only whether or not

$c_t/\beta_t > b$. Since the person does not notice any details of the history beyond her current sentiment about going to the gym, this SAS constitutes a stable attentional strategy given π. More precisely, $\liminf_{t\to\infty} P(c_t/\beta_t > b|\pi)/P(c_t/\beta_t > b|\pi^*) > 0$ and $\liminf_{t\to\infty} P(c_t/\beta_t \leq b|\pi)/P(c_t/\beta_t \leq b|\pi^*) > 0$ since the event $c_t/\beta_t > b$ has positive probability under both models and these probabilities are constant for all t. Now consider the case where $|\text{supp}(\pi)| = 2$ and $0 \in \text{supp}(\pi)$. Since $q^* = 1$, $q = 0$ will be rejected in finite time. At this point, π_t is a singleton and the logic from the previous case applies.

Now suppose that (i) holds but (ii) does not. As above, in each period t where the person decides whether to visit the gym, N^t distinguishes only whether or not $c_t/\beta_t > b$ and thus this data alone cannot render π attentionally unstable. Consider instead periods t where the person decides whether to buy the membership. The only interesting case is where Condition C.1 holds for some $\hat{q} \in \text{supp}(\pi)$ but not all (otherwise the person is subjectively certain about the optimal membership decision). As noted in the text in Section C.2 (details in Footnote 53), each noticing partition of a minimal SAS consists of at most two elements in this case: letting $V(\hat{q})$ denote the right-hand side of (C.1) where $\mathbb{E}_t$ is with respect to π conditional on h^t, the two elements are $n^t_B := \{h^t \in H^t \mid \mathbb{E}_t[V(\hat{q})] \geq m\}$, which contains all histories following which it's optimal to buy the contract, and $n^t_R := \{h^t \in H^t \mid \mathbb{E}_t[V(\hat{q})] < m\}$, which contains all histories following which it's optimal to reject it. Note that at any membership date t, the realized noticed history is either n^t_B or n^t_R. We therefore evaluate the likelihood ratio for the realized cell $n^t(h^t)$. Under $q^* = 1$, this cell has positive probability under π^*, and the preceding argument implies that its probability under π is bounded away from zero. Hence, $\lim_{t\to\infty} P(n^t(h^t)|\pi)/P(n^t(h^t)|\pi^*) > 0$ almost surely under π^*. Thus, the proposed attentional strategy is stable. ■

Proof of Proposition C.3. Consider the setup from Section C.3, and consider a model π that puts probability 1 on $\theta_c = 0$ (independent signals).

Part 1: Suppose there is perfect feedback (i.e., $r_t = \omega_t$ for all t).

($\Leftarrow$) Suppose π is dogmatic about the precision of one or both information sources. If π is dogmatic about both, then a minimal SAS is such that, for all t, N^t distinguishes only the value of $s_t = (s^1_t, s^2_t)$. Thus, $P(n^t(h^t)|\pi)/P(n^t(h^t)|\pi^*) = P(s_t|\pi)/P(s_t|\pi^*)$ for all t. This ratio is bounded away from zero as $t \to \infty$ since $P(s_t|\pi^*) \geq 1-\gamma$ and $P(s_t|\pi) \geq (1-\gamma)^2$. As such, suppose π is dogmatic about the precision of only one source. WLOG suppose π is dogmatic that source 2 has precision $\tilde{\theta}_2 \in \{0.5, \gamma\}$, and it puts positive probability on both θ^*_1 (the true precision of source 1) and $\tilde{\theta}_1$, which denotes the other possible (yet false) value of θ_1. A minimal SAS is such that, for all $t \geq 2$, N^t will (i) notice $s_t = (s^1_t, s^2_t)$; (ii) ignore all past signals from source 2; and (iii) distinguish the value of

$W_t^1 := \sum_{k=1}^{t-1} \mathbf{1}\{s_k^1 = \omega_k\}$, which is sufficient for updating beliefs about θ_1. Thus:

$$\begin{aligned}
\frac{P(n^t(h^t)|\pi)}{P(n^t(h^t)|\pi^*)} &= \frac{P\big(W_t^1\big|\pi\big)}{P\big(W_t^1\big|\theta_1^*\big)}\frac{P\big(s_t\big|W_t^1,\pi\big)}{P\big(s_t\big|\pi^*\big)} \\
&= \left(\sum_{\theta_1\in\{\tilde{\theta}_1,\theta_1^*\}} \pi(\theta_1)\frac{P\big(W_t^1\big|\theta_1\big)}{P\big(W_t^1\big|\theta_1^*\big)}\right)\frac{P\big(s_t\big|W_t^1,\pi\big)}{P\big(s_t\big|\pi^*\big)} \\
&= \left(\pi(\theta_1^*)+\pi(\tilde{\theta}_1)\left(\frac{\tilde{\theta}_1}{\theta_1^*}\right)^{W_t^1}\left(\frac{1-\tilde{\theta}_1}{1-\theta_1^*}\right)^{t-1-W_t^1}\right)\frac{P\big(s_t\big|W_t^1,\pi\big)}{P\big(s_t\big|\pi^*\big)}. \qquad \text{(D.27)}
\end{aligned}$$

Since in truth $W_t^1 \sim \text{Binomial}(t-1,\theta_1^*)$, $\lim_{t\to\infty}\left(\frac{\tilde{\theta}_1}{\theta_1^*}\right)^{W_t^1}\left(\frac{1-\tilde{\theta}_1}{1-\theta_1^*}\right)^{t-1-W_t^1} = 0$.[54] Hence, $\liminf_{t\to\infty} P(n^t(h^t)|\pi)/P(n^t(h^t)|\pi^*) > 0$ since $P\big(s_t\big|W_t^1,\pi\big)/P\big(s_t\big|\pi^*\big)$ is bounded away from 0 as $t\to\infty$.

($\Rightarrow$) Suppose π admits a stable attentional strategy. Toward a contradiction, suppose that π is not dogmatic about either information source, and define W_t^2 analogously to W_t^1. Therefore, a minimal SAS is such that, for all $t\geq 2$, N^t must distinguish the value of W_t^1, W_t^2, and s_t. (Note that the count of "successes" is a minimal sufficient statistic for updating about the parameter governing a binomial distribution.) Thus, following the logic of (D.27) above, we have

$$\begin{aligned}
\frac{P(n^t(h^t)|\pi)}{P(n^t(h^t)|\pi^*)} &= \frac{P\big(W_t^1,W_t^2\big|\pi\big)}{P\big(W_t^1,W_t^2\big|\pi^*\big)}\frac{P\big(s_t\big|W_t^1,W_t^2,\pi\big)}{P\big(s_t\big|\pi^*\big)} \\
&= \frac{P\big(W_t^1\big|\pi\big)P\big(W_t^2\big|\pi\big)}{P\big(W_t^1\big|\theta_1^*\big)}\frac{P\big(s_t\big|W_t^1,W_t^2,\pi\big)}{P\big(s_t\big|\pi^*\big)} \\
&= P\big(W_t^2\big|\pi\big)\left(\pi(\theta_1^*)+\pi(\tilde{\theta}_1)\left(\frac{\tilde{\theta}}{\theta^*}\right)^{W_t^1}\left(\frac{1-\tilde{\theta}_1}{1-\theta_1^*}\right)^{t-1-W_t^1}\right)\frac{P\big(s_t\big|W_t^1,W_t^2,\pi\big)}{P\big(s_t\big|\pi^*\big)}.
\end{aligned}$$

As shown above, the middle term of the expression above (in parentheses) converges to $\pi(\theta_1^*)$ and $P\big(s_t\big|W_t^1,W_t^2,\pi\big)/P\big(s_t\big|\pi^*\big)$ is finite. Furthermore, $\lim_{t\to\infty} P\big(W_t^2\big|\pi\big) = 0$ given that for each $\theta_2 \in \text{supp}(\pi)$, the person presumes $W_t^2 \sim \text{Binomial}(t-1,\theta_2)$. Thus, $\lim_{t\to\infty} P(n^t(h^t)|\pi)/P(n^t(h^t)|\pi^* = 0$, which contradicts π admitting a stable attentional strategy.

Part 2: Suppose there is no feedback (i.e., $r_t = \varnothing$ for all t).

($\Leftarrow$) Suppose π is dogmatic about the precision of both information sources. (Below we deal with

[54] To see this, write $\left(\frac{\tilde{\theta}_1}{\theta_1^*}\right)^{W_t^1}\left(\frac{1-\tilde{\theta}_1}{1-\theta_1^*}\right)^{t-1-W_t^1}$ as $L(t,W_t^1,|\theta_1^*,\tilde{\theta}_1)^{t-1}$, where $L(t,W_t^1,|\theta_1^*,\tilde{\theta}_1) := \left(\frac{\tilde{\theta}_1}{\theta_1^*}\right)^{W_t^1/(t-1)}\left(\frac{1-\tilde{\theta}_1}{1-\theta_1^*}\right)^{(t-1-W_t^1)/(t-1)}$. Note that $\lim_{t\to\infty} L(t,W_t^1|\theta_1^*,\tilde{\theta}_1) = \left(\frac{\tilde{\theta}_1}{\theta_1^*}\right)^{\theta_1^*}\left(\frac{1-\tilde{\theta}_1}{1-\theta_1^*}\right)^{1-\theta_1^*} := \bar{L}(\tilde{\theta}_1|\theta_1^*)$. Furthermore, $\arg\max_{\theta_1}\bar{L}(\theta_1|\theta_1^*) = \theta_1^*$ and $\bar{L}(\theta_1^*|\theta_1^*) = 1$. Thus, $\bar{L}(\tilde{\theta}_1|\theta_1^*) < 1$ since $\tilde{\theta}_1 \neq \theta_1^*$, and hence $\lim_{t\to\infty} L(t,W_t^1,|\theta_1^*,\tilde{\theta}_1)^{t-1} = \lim_{t\to\infty}\bar{L}(\tilde{\theta}_1|\theta^*)^{t-1} = 0$.

the additional case where π is dogmatic that one information source is uninformative.) As in the similar case above with feedback, a minimal SAS is such that, for all t, N^t distinguishes only the value of $s_t = (s_t^1, s_t^2)$. Thus, $\lim_{t\to\infty} P(n^t(h^t)|\pi)/P(n^t(h^t)|\pi^*) = \lim_{t\to\infty} P(s_t|\pi)/P(s_t|\pi^*) > 0$.

($\Rightarrow$) Suppose π admits a stable attentional strategy. Toward a contradiction, assume the person is uncertain about at least one of the information sources and is not dogmatic that either source is uninformative. As we show below, the person must track the number of periods in which the two sources agree. Let $a_t := \mathbf{1}\{s_t^1 = s_t^2\}$ be an indicator for agreement in period t and let $Q_t := \sum_{k=1}^t a_k$ count the number of agreements through round t.[55] Assuming the person notices Q_t (and s_t) each period, then π is attentionally unstable:

$$\begin{aligned}\frac{P(n^t(h^t)|\pi)}{P(n^t(h^t)|\pi^*)} &= \frac{P\big(Q_t\big|\pi\big)}{P\big(Q_t\big|\pi^*\big)}\frac{P\big(s_t\big|Q_t,\pi\big)}{P\big(s_t\big|\pi^*\big)} \\ &= P\big(Q_t\big|\pi\big)\frac{P\big(s_t\big|Q_t,\pi\big)}{P\big(s_t\big|\pi^*\big)},\end{aligned} \tag{D.28}$$

where the second equality follows from the fact that Q_t is deterministic under θ^* given that signals are in truth perfectly correlated. Furthermore, $\lim_{t\to\infty} P\big(Q_t = t\big|\pi\big) = 0$ given that π treats s_t^1 and s_t^2 as independent for all t. Thus (D.28) converges a.s. to 0 given that $P\big(s_t\big|Q_t,\pi\big)/P\big(s_t\big|\pi^*\big)$ is bounded from above.

To complete the proof, we show that a SAS must notice (Q_t, s_t) in each period t. Let $\pi_t(\theta_1,\theta_2)$ denote beliefs over $(\theta_1,\theta_2) \in \{0.5,\gamma\}^2$ conditional on h^t and note that the optimal action given these beliefs is

$$x_t^* = \sum_{(\theta_1,\theta_2)} \pi_t(\theta_1,\theta_2)P(\omega_t = 1|s_t,\theta_1,\theta_2). \tag{D.29}$$

We consider what minimal data the person finds sufficient to form the beliefs $\pi_t(\theta_1,\theta_2)$ that determine (D.29). First consider how the person (who believes $\theta_c = 0$) updates these beliefs upon observing a single period: given beliefs π_{t-1} and signal s_t, Bayes' rule implies

$$\pi_t(\theta_1,\theta_2|s_t) = \frac{P(s_t|\theta_1,\theta_2)\pi_{t-1}(\theta_1,\theta_2)}{\sum_{(\tilde\theta_1,\tilde\theta_2)} P(s_t|\tilde\theta_1,\tilde\theta_2)\pi_{t-1}(\tilde\theta_1,\tilde\theta_2)}. \tag{D.30}$$

Examining $P(s_t|\theta_1,\theta_2)$ for various values of s_t reveals that beliefs update identically when signals agree, i.e. $s_t \in \{(0,0),(1,1)\}$, and update identically when signals disagree, i.e. $s_t \in \{(1,0),(0,1)\}$. To see this, note that $P(s_t = (1,1)|\theta_1,\theta_2) = \frac{1}{2}P(s_t = (1,1)|\theta_1,\theta_2,\omega_t = 1) + \frac{1}{2}P(s_t = (1,1)|\theta_1,\theta_2,\omega_t = 0) = \frac{1}{2}[1+2\theta_1\theta_2-\theta_1-\theta_2]$. Similar calculations show that $P(s_t = (0,0)|\theta_1,\theta_2) = P(s_t = (1,1)|\theta_1,\theta_2) = \frac{1}{2}[1+2\theta_1\theta_2-\theta_1-\theta_2]$ and $P(s_t = (1,0)|\theta_1,\theta_2) = P(s_t = (0,1)|\theta_1,\theta_2) = \frac{1}{2}[\theta_1+\theta_2-2\theta_1\theta_2]$. This

[55] Note that a_t is based solely on signals and not resolutions. This means that a_t is observed at the start of period t rather than the end, and hence a_t is given by h^t.

implies that $a_t = \mathbf{1}\{s_t^1 = s_t^2\}$ is sufficient for updating beliefs following a single arbitrary period.

Now consider updating the prior π based on h^t. We will show that $Q_t = \sum_{k=1}^{t} a_k$ is sufficient for h^t to form updated beliefs $\pi_t(\theta_1, \theta_2)$. The calculations above reveal that for all parameter combinations $(\theta_1, \theta_2) \neq (\gamma, \gamma)$, $P(a_t = 1|\theta_1, \theta_2) = 1/2$. In contrast, $P(a_t = 1|\gamma, \gamma) = 1 - 2\gamma(1-\gamma)$. Thus,

$$\pi_t(\gamma, \gamma|h^t) = \frac{[1-2\gamma(1-\gamma)]^{Q_t}[2\gamma(1-\gamma)]^{t-Q_t}\pi(\gamma,\gamma)}{[1-2\gamma(1-\gamma)]^{Q_t}[2\gamma(1-\gamma)]^{t-Q_t}\pi(\gamma,\gamma) + [0.5]^t(1-\pi(\gamma,\gamma))}, \tag{D.31}$$

but for any $(\theta_1, \theta_2) \neq (\gamma, \gamma)$,

$$\pi_t(\theta_1, \theta_2|h^t) = \frac{[0.5]^t\pi(\theta_1,\theta_2)}{[1-2\gamma(1-\gamma)]^{Q_t}[2\gamma(1-\gamma)]^{t-Q_t}\pi(\gamma,\gamma) + [0.5]^t(1-\pi(\gamma,\gamma))}. \tag{D.32}$$

From Equations (D.31) and (D.32), it is clear that Q_t is sufficient for h^t to update beliefs. Furthermore, so long as $\pi_t(\gamma, \gamma)$ is non-degenerate, Q_t is also necessary. But as shown in (D.28), noticing Q_t each period renders π attentionally unstable, which is a contradiction.

Finally, consider the case where the person is dogmatic that one of the information sources is uninformative. This implies that $\pi_t(\gamma, \gamma) = 0$ for all t and hence Equation (D.32) implies that beliefs about all other combinations of (θ_1, θ_2) are independent of h^t. Therefore a minimal SAS is such that, for all t, N^t distinguishes only the value of s_t. Thus, π is attentionally stable in this case. ■