Channeled Attention and Stable Errors

Tristan Gagnon-Bartsch
FSU

Matthew Rabin
Harvard University

Joshua Schwartzstein[*]
HBS

August 17, 2023

Abstract

We develop a framework for assessing when somebody will eventually notice that she has a misspecified model of the world, premised on the idea that she neglects information that she deems—through the lens of her misconceptions—to be irrelevant. In doing so, we assess the *attentional stability* of both general psychological biases—such as naivete about present bias—and empirical misconceptions—such as false beliefs about consumer demand. We explore which combinations of errors and environments allow an error to persist, versus which errors lead people to *incidentally learn* they have things wrong because even the data they deem relevant tells them that something is amiss. We use the framework to shed light on why fresh eyes are valuable in organizational problems, why people persistently use overly coarse (vs. overly fine) categorizations, why people sometimes recognize their errors in complex environments when they don't in simple environments, and why people recognize errors in others that they don't recognize in themselves.

[*] E-mails: gagnonbartsch@fas.harvard.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
"No theory is kind to us that cheats us of seeing." — Henry James
"Truly nothing is to be expected except for the unexpected." — Alice James

1 Introduction

People's models of the world guide the data they collect, attend to, and analyze. These models then shape people's understanding of the world by channeling their attention to what they deem to be important: per the opening quote by psychologist William James, the experience that matters is not the information people have been exposed to—it is the information they have *attended to* through the lens of their models. But how accurate are those models?

Economists typically ignore this question by assuming people begin with rational expectations, or by imagining those who start from a place of ignorance or bias will eventually form accurate beliefs with rich enough feedback. Someone who suffers from the gambler's fallacy will encounter more streaks than she expects. Someone oblivious to his self-control problems will find himself over-indulging more than he anticipates. Someone who misattributes recovery from ailments to medical treatments rather than the passage of time will see that "treatments" tend to provide no added benefit. If there is a persistent mismatch between a person's understanding of the world and the world itself, it seems reasonable to assume people will notice that something is amiss.

A range of evidence defies this assumption. In finance, investors with large portfolios persistently and systematically make selling decisions that underperform selling assets at random (Akepanidtaworn et al., 2021). In business, firms persistently follow pricing strategies that fail to take into account predictable patterns of consumer demand (DellaVigna and Gentzkow, 2019; Strulov-Shlain, 2022; List et al., 2023) and persistently pursue campaigns that lose money (e.g., internet search advertising for well-known brands, as in Blake et al., 2015). In agriculture, Hanna et al. (2014) show that Indonesian seaweed farmers seem to persistently fail to optimize along a dimension (pod size) they wrongly ignore despite being exposed to rich data from which they could learn if they paid attention. Zambian farmers have "hungry seasons" year after year, even though they could budget their spending to avoid going hungry (Augenblick et al., 2023). And the psychological biases of most interest are not those that reside in the corners of our lives nor in experiments catching people in unfamiliar situations—our retirements are poorer because we are prone to see illusory patterns in financial markets, and lives go astray because of our persistent overoptimism about our self control.

This paper formalizes principles influencing when and how people will notice that their beliefs mismatch reality, generating the following simple explanation for why misconceptions may persist:

we channel our attention to data we deem relevant under our model, whereas it is the data we deem irrelevant that is most likely to contradict our model. The quote from William's novelist brother Henry warns that our theories of the world can point our attention in the wrong direction and block us from seeing information that (unbeknownst to us) is highly relevant. Examples like those above are then not anomalous, but are instead natural consequences of people using models to guide their attention.

Our formal framework builds from a literature pioneered by Sims (2003), which analyzes agents who allocate attention by optimally weighing the expected costs and benefits of attention. But following Schwartzstein (2014), our model of *subjectively* rational inattention departs from this literature in two interrelated ways. First, we repurpose the study of costly attention to address our opening question: when exposed to recurring situations, will patterns she didn't expect eventually alert an agent that her beliefs are systematically wrong? Second, we assume that an agent's perceived benefits from attention—and how she interprets what she notices—are based on her potentially misspecified model. Unlike settings where people both correctly forecast benefits of attention and correctly interpret what they see, the question at hand is whether somebody will notice her error when she is potentially mistaken about both the benefits of attention and the meaning of what she attends to. Because a person's conceptualization of the world channels her attention, harmful errors can persist.

Take the following example: people seem to spend a great deal out of pocket on medical treatments that are seemingly ineffective, such as some popular remedies to help lose weight, recover from a cold, or get better sleep. Why? Consider a patient who is trying to evaluate the cost-effectiveness of a cold medicine she often takes. If she feels uncertain about how quickly she gets better when she takes the medicine but *thinks she knows* she rarely gets better quickly when she *doesn't* take it, she will monitor her speed of recovery when she takes the medicine but not those times when she goes without it. Her underestimation of recovery speed when not taking the medicine means that even when the treatment has no real effect on how quickly she recovers, she will wrongly think the medicine is effective. And this error is what we call attentionally stable: discovering it requires that the person pay careful attention to how quickly she recovers when she's *off* her medication, but she doesn't see the need for attention in that case. The stability of such an error crucially depends on the interaction between the error and channeled attention, since the person would (in our framework) discover any statistically identifiable mistake if she paid full attention.

While our framework suggests that harmful errors as in this example often persist, it also shows how a person *does* sometimes figure out she is wrong through "incidental learning"; that is, when the evidence a person gathers and deems utility-relevant *given her wrong model* nevertheless reveals that something is amiss with her model. Even with channeled attention, the patient from above would wake up to her error if instead of being dogmatic and wrong about how quickly she recovers without medicine, she was initially miscalibrated but sufficiently uncertain along this dimension. One intriguing consequence of this principle is that—in the spirit of an experiment by Esponda et al. (2022)—somebody can be better off when she is initially missing some key information. Her efforts to learn this information may induce her to pay useful attention to feedback. In this way, our framework suggests how fresh eyes may help catch mistakes. We use our framework to identify principles and applications examining which constellations of an agent's misconceptions, goals, and informational environment tend to induce such incidental learning.

Section 2 introduces our formal framework. We focus on errors that can be modeled as an agent who is Bayesian, but with misspecified priors π over the structure of the world.[1] Although not all errors are readily formulated this way, our approach covers a broad array of context-specific empirical misconceptions (as in Barberis et al., 1998's model of investors' misperceptions) and psychological biases (as in Rabin, 2002's model of belief in the law of small numbers). We assume that an agent with such priors will employ a "sufficient attentional strategy" (or "SAS"), where she parses available information in a way that she *perceives* as sufficient for maximizing long-run payoffs. Our framework advances the role that theories play in the act of noticing beyond Schwartzstein (2014): while he highlighted how a person's theory determines whether a given variable is relevant, we further emphasize how the person's theory determines the way she constructs variables in the first place. A manager's beliefs about whether rude employees are less productive team members, for example, depends on how she organizes the complexity of conversations into a variable called "being rude".[2]

The definition of a SAS does not rule out an agent attending to information she believes is useless, but we focus primarily on the implications of employing a "minimal" SAS, whereby a person never attends to more than she finds useful. For instance, if a patient thinks she only recovers quickly from a cold with medicine, then this criterion says she will only attend to the speed of recovery when taking the medicine. Or, if a person is assessing whether her gym membership is worthwhile, then she may find it sufficient to notice the frequency she skips the gym without further analyzing whether the skips are from laziness or from busyness.

To simplify the setup, analysis, and interpretation, we make two strong assumptions about access to information. First, the agent can observe the payoffs generated by each of her potential

[1] As will be seen, it is the support of the agent's model that will matter, not the exact probabilities it assigns.

[2] While such an example illustrates the theory-laden nature of constructing measures of variables we know we want to pay attention to, theories are even more central to selecting variables from an infinity of candidates. Does the manager care about the rudeness of employees, or their helpfulness, or their humorlessness? To take another example, 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.

choices, not just the one chosen. By eliminating the effect of actions on information exposure, our framework focuses solely on the effects of limited attention, rather than limited information that might arise endogenously.[3] Second, we assume that the agent can freely access past information she deems useful at the moment, even if she did not previously take note of that information. This could happen if the agent encodes details in a retrieval-friendly way even if she does not engage with those details at the time, or (more realistically) if the environment is such that relevant data is recorded in an accessible form. While this assumption simplifies the analysis, we show in Appendix A (and in an earlier draft, Gagnon-Bartsch et al., 2018) that similar results on error discovery hold when the retrieval of information is constrained in natural ways. The limited impact of different retrieval assumptions highlights that our results are driven by the agent's perceived (lack of) benefits of attention rather than the costs.[4]

The final component of our framework addresses our motivating question: when will somebody discover his mistakes? We assume a person abandons his model when he notices something that is far more likely under a compelling alternative.[5] In the limit case that we consider, we say that a model π is "attentionally unstable relative to an alternative model" λ if the limit *noticed* data becomes infinitely more likely under λ than under π. In most of our analysis we take λ to be the correct model, but we allow for a more general λ to clarify the crucial role that the availability of alternative explanations plays in our framework. We thus describe an erroneous model π as "attentionally stable" in a context if there exists a SAS where the noticed data is not infinitely less likely under π than under the correct model.[6] Intuitively, our framework says that whether an error is discovered reduces to whether a person is sufficiently surprised by what he *notices* about the world around him, or by his own behavior. For example, a person who believes he has a miracle cure for baldness will eventually notice his unexpected persistence in buying more of the elixir rather than shampoo, implying that his hair isn't growing back. In contrast, a person who believes he has a miracle cure for the common cold may never realize it is ineffective: the amount of time he uses it is never shocking, since the colds always eventually stop. We show how to apply this result to assess the stability of commonly studied biases and empirical misunderstandings.

[3] Our conclusions about noticing mistakes would extend without this assumption so long as the agent either occasionally faces restrictions on her choice set or observes the experiences of other agents.

[4] 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".

[5] Hence, it is not the absolute unlikeliness of outcomes within the person's theory that causes him to abandon it. A person doesn't abandon the view that a coin is fair just from observing any one of the $2^{1,000}$ highly unlikely sequences he will observe; if he sees 1,000 heads in a row, however, he will be drawn to the obvious alternative.

[6] Our formulation may seem knife edge: if the agent put the slimmest of odds rather than *zero* weight on the true model being correct, the agent would in the long run notice his error. Although we think the "zero weight" may well capture the true psychology of fully neglecting an alternative, we also note that our formulation pairs this zero weight with the similarly extreme assumption that the agent pays *full* attention to anything that he thinks *may eventually* prove useful. Having simplified our analysis in this way, we are heuristically capturing the idea of beliefs that are stable in the face of "costly" attention, no matter how small that (non-zero) cost is.

After specifying the formal framework, Section 2 then discusses in greater detail the connection between our assumptions and the motivating evidence from psychology. It also positions the framework relative to recent prominent economic models on limited attention and memory (e.g., Bordalo et al., 2012, 2013, 2020; Woodford, 2012; Gabaix, 2014; Caplin and Dean, 2015; Matějka and McKay, 2015).

Our framework is surely unrealistic in supposing that a person ignores things that fall outside their model of the world—an exploding mushroom cloud on the horizon would be frighteningly interpretable to denizens of the nuclear age, but even earlier generations would notice it with incomprehension. (And, of course, even a person with an infinite supply of elixir may notice he's still bald.) However, psychological evidence reveals the surprising capacity we have for not noticing the unexpected. Famously, when Simons and Chabris (1999) task people with counting the number of passes in a film clip of basketball players, they often fail to notice a faux gorilla walking across the court. They don't see the "gorilla" in front of them because they're channeling their attention to counting passes. Our final opening quote (from the third James sibling) emphasizes that our ubiquitous exposure to highly unlikely things is in fact one of the most predictable features of our lives.

People fail to wake up in our framework because they similarly miss *statistical gorillas*—even when no individual observation is itself a big surprise, they miss a striking mismatch over time between the actual distribution of some outcome and their expected distribution. People can notice instances of their gambling losses, self-indulgence, and drug-free recovery from ailments without recognizing the statistics that would tell them something is amiss with their models. Likewise, a person might be able to answer on a given day whether he forgot to take his pills without tracking how often he forgets. A trader might carefully examine earnings each day, without additionally checking whether the sequence of realized earnings conforms to her assumptions about the stochastic process. A firm leader may not ask questions of the data that would reveal important correlations between control variables and performance indicators if she does not believe them to be relevant. Although such errors are costly, people who don't recognize them miss out on significant monetary or welfare gains because they see no benefits to tracking the statistics.[7]

Section 3 uncovers principles about which errors tend to be stable. Some errors are attentionally stable broadly *irrespective* of the decision context—a feature we call "stable for all preferences". We show that these models are typically ones that neglect relevant outcomes or predictive signals, such as the cold-medicine example where the person ignores the possibility of getting better quickly when not taking a drug. On the other hand, erroneous models that are more likely to

[7] This can be true even when concerned others advise us that something might be important (e.g., NGOs promoting the importance of better agricultural practices). We assume few readers of this article believe that horoscopes are truly informative. You are not carefully reexamining the data now to see if this disbelief is warranted. (Except Capricorns, who tend to be gullible.) But you would deploy this strategy even if you *weren't* right about horoscopes.

be noticed lead the person to attend to the right variables, or make distinctions that are in reality unimportant. Such results shed light on why discussions of coarse and categorical thinking tend to go hand-in-hand (e.g., Mullainathan, 2002b): while people could in principle categorize objects, people, or situations too finely, our results suggest that such schemes are not stable while overly coarse schemes are. We also show examples suggesting that bigger, more costly errors may be more robustly stable than smaller errors in many settings. In the spirit of the medicine example from earlier, when a person wildly versus mildly underestimates the impact of some factor, she is less likely to confront situations where she feels it is worthwhile to pin down an exact understanding of that factor.

Within our framework, waking up to an error depends not only on the nature of that error, but on the particular decision environment. In Section 4, we analyze which types of situations facilitate the stability of errors. We show that environments requiring a person to engage with features of the data across time helps induce her to notice her errors. It also reveals a cost of technologies that allow a person to delegate decisions to others or algorithms: asking others to answer the questions she *thinks* are sufficient may prevent her from recognizing she's asking the wrong questions. These and other examples highlight that it is not simply a matter of situations rendering an error more costly that induces people to notice their errors. When tracking cumulative profits, a family business (e.g., the bagel seller studied by Levitt, 2016) is more likely to discover an error in mispredicting quantity demanded at a given price (and hence in how much to produce at that price) than a potentially more costly error in mispredicting how much people would demand at a *different* price (and hence in how to set prices), since the former error is more likely to be inconsistent with realized profits (and, e.g., the amount of taxes owed).[8]

Section 5 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, we show how our results provide a reason why people sometimes persistently get simple problems wrong and more complicated problems right: even if they wrongly think they know how to answer simpler problems, the process of breaking down a complicated problem into those simpler ones can alert them that they don't actually have all the answers. Third, we combine two central aspects of our framework—that theories guide what we consider to be a variable and that neglectful errors tend to be more stable than overly elaborate ones—in order to consider the role of theory in scientific progress. The former aspect suggests a crucial role for theory; the latter suggests how overly simple theories may lead science to get stuck. Fourth, we consider how people may recognize errors in others even when they don't

[8] Interestingly, Levitt (2016) argues that the seller's suboptimal pricing stems from limited feedback, yet he identifies the seller's mistake from exactly the same data that the seller had. Thus, the issue does not seem to be one of limited data but rather limited attention.

recognize those errors in themselves: correctly thinking we understand ourselves better than others leads us to pay attention to others' behavior in ways we (incorrectly) feel are unnecessary when it comes to our own.

Section 6 relates our framework to other approaches for analyzing the stability of erroneous beliefs, including those based on people having incomplete data and motivated reasoning. Section 7 concludes by considering limitations and extensions of our analysis.

2 Framework

This section formalizes our framework. After introducing the environment in 2.1, we develop our notion of how a decision-maker channels his attention in 2.2, and then we present our criteria for assessing whether an attention-channelling decision-maker will discover his errors in 2.3. In 2.4, we discuss related literature and supporting evidence for our central assumptions.

2.1 Environment

Consider a person updating his beliefs over a parameter $\theta \in \Theta$ that influences the distribution of payoff-relevant outcomes. For instance, in the cold-medicine example from above, θ represents the probability of quick recovery with and without medication. More broadly, 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) can observe a realized outcome, or "resolution", $r_t \in R_t$. Let $y_t = (r_t, s_t)$ denote the data generated in round t; we call y_t an "observable". The observable y_t determines payoffs but also reveals information about θ and the optimal action in future periods. 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. The history h^t denotes all the data possibly observed prior to period t's action:

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

Let H^t be the set of all such histories up to period t, and let $H \equiv \cup_{t=1}^{\infty} H^t$. 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 $Y_t \equiv R_t \times S_t$ are finite-valued, each X_t is 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 applications relax these restrictions, they simplify the presentation and proofs of our formal results.

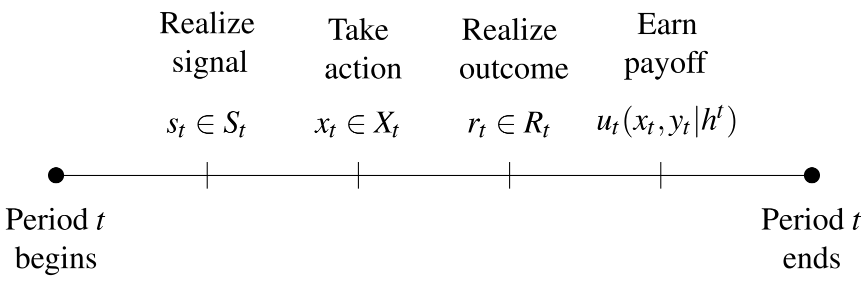

Figure 1: *Timeline of events within period t.*

The observables are correlated with parameter θ. The signal s_t is drawn according to distribution $P_s(s_t|h^t(\neg s_t), \theta)$. The subsequent resolution, r_t, may additionally depend on the current signal and action, and we denote its distribution by $P_r(r_t|x_t, h^t, \theta)$. These two distributions form a joint distribution over observables denoted 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 the agent in our framework trivially fails to learn because he may not entertain the true parameter in his model, this is not the case; it will be clear below that, with full attention, the agent will generally discover his misspecification.[9]

[9] 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

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, we maintain that the person's actions do not affect what he can learn.

Assumption 1 (Maintained assumption)**.** For all $t \in \mathbb{N}$, $h^t \in H^t$, $x_t \in X_t$, $y_t \in Y_t$, and $\theta \in \Theta$: $P(y_t|x_t, h^t(\neg s_t), \theta) = P(y_t|y^t, \theta)$, where $y^t \equiv (y_{t-1}, y_{t-2}, \ldots, y_1)$.

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 $t \in \mathbb{N}$ and $\theta \in \Theta$, the action set X_t is independent of h^t, and for all $x_t \in X_t$, $y_t \in Y_t$, and $h^t \in H^t$, the payoff $u_t(x_t, y_t|h^t)$ is independent of h^t.

Without an incentive for experimentation, 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 $u_t(x_t, y_t)$.

To illustrate an application of our setup, consider the cold-medicine example from the introduction. In each period $t = 1, 2, \ldots$ that a patient experiences a cold, he chooses $x_t \in \{T, N\}$, where T represents taking the medicine and N represents not taking it. The patient may then quickly recover in that period, indicated by $r_t \in \{0, 1\}$ where $r_t = 1$ denotes a quick recovery. The treatment imposes a flow cost $c > 0$, and quick recovery yields a flow benefit $b > 0$: $u_t(x_t, y_t) = b \cdot \mathbf{1}\{r_t = 1\} - c \cdot \mathbf{1}\{x_t = T\}$. Fast recovery occurs with probability $\theta_T \in [0, 1]$ if $x_t = T$ and $\theta_N \in [0, 1]$ otherwise. The person forms beliefs over the parameter $\theta = (\theta_T, \theta_N)$, where $\theta^*_x \equiv \Pr(r_t = 1|x_t = x)$ denotes the true likelihood of fast recovery conditional on action x. To satisfy Assumption 1 (that observations are independent from actions), imagine that the patient also observes the choices and outcomes of others, so he has rich enough data to learn θ_T and θ_N independent of his own behavior.[10] Our introductory example assumes the patient has a misspecified model π over (θ_T, θ_N) such that he is dogmatic that θ_N has a value $\hat{\theta}_N \neq \theta^*_N$ (where the "hat" notation is used to designate a perceived parameter value under π). That is, he thinks he knows how often he quickly recovers when he doesn't take the medicine. We also assume π involves sufficient uncertainty over θ_T so that the patient is uncertain if the medication is worthwhile given $\hat{\theta}_N$ and that $(\hat{\theta}_N, \theta^*_T) \in \text{supp}(\pi)$. The patient therefore has incentive to attend to outcomes in order to learn θ_T. We will see that this is not enough for him to discover his misperception of θ_N.

Madarász (2012, 2021) on information projection and Gagnon-Bartsch (2016) and Gagnon-Bartsch et al. (2021) on taste projection.

[10] To formalize this assumption, suppose the patient observes the recovery (or lack thereof) of an acquaintance who makes the opposite choice as him in each period. The patient observes $y_t = (r_t, r^a_t)$ where $r^a_t \in \{0, 1\}$ indicates the acquaintance's outcome.

2.2 Channeled Attention

People tend to direct their attention to subjectively task-relevant information and away from a mass of other information.[11] 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 the cold-medicine example, the patient sees no purpose in tracking how quickly he recovers on days when he didn't use the medicine; hence, $n^t(h^t)$ need not distinguish outcomes that occurred on those days.

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.

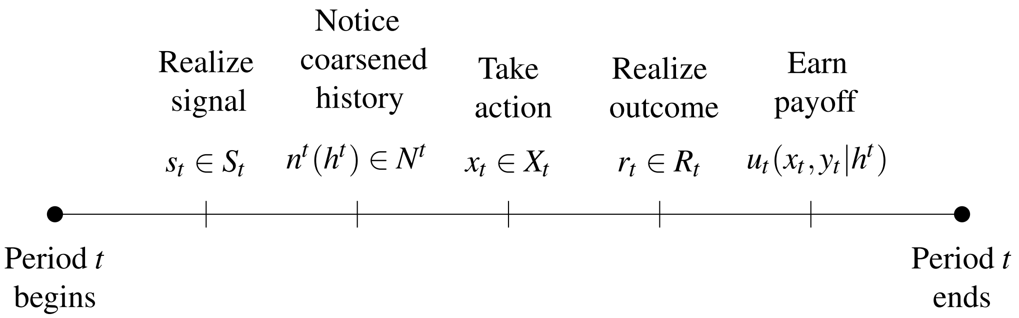

Figure 2: *Timeline of events within period t, including the coarsening of past information.*

A *noticing strategy* $\mathscr{N}$ is the full sequence of the person's noticing partitions, $\mathscr{N} = (N^1, N^2, \ldots)$. 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* $\sigma = (\sigma_1, \sigma_2, \ldots)$, where $\sigma_t : N^t \to \Delta(X_t)$ maps noticed histories to actions. When the

[11] 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 Section 2.4.

person observes noticed history n^t, he uses his theory π and strategy ϕ to assess the probability of each $h^t \in n^t$ and updates accordingly. We assume the person automatically recalls his prior π, attentional strategy ϕ, and current time period t. Hence, the likelihood of n^t given θ and ϕ is $P(n^t|\theta,\phi) = \sum_{h^t \in n^t} P(h^t|\theta,\phi)$, and the resulting posterior over θ is

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

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 piece of 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 information that π deems irrelevant for decisions.

Definition 1. An attentional strategy $\phi = (\mathcal{N},\sigma)$ 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 attentional strategy $\tilde{\phi}$. Under Assumption 2, sufficiency 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 optimized using the precise history. For instance, in the cold-medicine example, the patient follows a SAS that neglects the speed of recovery when not using the medication because he wrongly believes that he already knows this rate with sufficient precision.

Fixing π, there are typically many SASs in a given environment. Our definition of a SAS does not mandate that a person ignores data he deems useless. However, we focus on the "minimal" case where all seemingly extraneous information is ignored. Say that $\tilde{\mathcal{N}}$ *coarsens* $\mathcal{N}$ 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 $(\mathcal{N},\sigma)$ is *minimal* if there does not exist another SAS that can be obtained by coarsening $\mathcal{N}$.

Minimal attentional strategies are perhaps most consistent with our interpretation that the person ignores data due to small costs of attention. However, a minimal SAS also assumes a perhaps implausible ability to ignore data, and we discuss in the conclusion how our framework extends to situations where some data is impossible to ignore.[12]

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

Before considering when a misspecified model is stable under channeled attention, we first address a more basic question: is the model "measurable"? That is, does the person ever confront data he thought was impossible? We say π is *attentionally measurable* with respect to SAS $\phi = (\mathscr{N}, \sigma)$ if all finite noticed histories given ϕ that occur with positive probability under π^* are assigned positive probability under π.

Proposition 1. *If $\phi = (\mathscr{N}, \sigma)$ is a minimal sufficient attentional strategy given π, then π is attentionally measurable with respect to ϕ.*

Proposition 1 shows that for *any* misspecified model π, there exists a sufficient way to filter the data such that the person never notices an outcome he assumed impossible. (See Appendix D for all proofs.) This is true whenever the person follows a minimal SAS. Intuitively, the person sees no benefit to distinguish events he assigns zero probability from those he assigns positive probability. Thus, a minimal noticing strategy will not be fine-tuned to notice when subjectively zero-probability events occur. A person who does not conceive of gaps between his intentions and actions need not notice when his actions deviate from his intentions; a person who thinks he's always in the mood for a run in the morning does not need to recognize when he'd rather run later in the day; a person who thinks that medical doctors all agree on a topic need not appreciate their substantial disagreement; a researcher who dogmatically believes in the rational-expectations hypothesis does not need to notice when people's beliefs clearly cannot be rationalized.[13]

While a person need not notice events he thought were impossible, he may nevertheless come to recognize that there's a much better explanation for the events he *does* notice than his prevailing theory. We next turn to the question of when such incidental learning happens.

[12] There may be multiple minimal SASs for a given π and environment; Proposition 2, 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 make 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.

[13] In the literature on mistaken beliefs, a common fix to prevent people from needing to update on perceived 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 can be seen as providing some justification for another common approach: simply ignoring the issue.

2.3 Light-Bulb Theories and Attentional Stability

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 "light-bulb" model $\lambda \in \Delta(\Theta)$ if the noticed data are infinitely more likely under the light-bulb than the prior. Otherwise, we say π is *attentionally stable* relative to λ.

Definition 3. A model π is *attentionally unstable* with respect to λ and SAS $\phi = (\mathscr{N}, \sigma)$ if the limit inferior as $t \to \infty$ of the ratio $P(n^t|\pi,\phi)/P(n^t|\lambda,\phi)$ equals 0 with positive probability under π^* when the person follows SAS ϕ, 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 notices everything), a model π is unstable if it seems excessively unlikely relative to the light-bulb 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 solely to make optimal decisions given π—happens to make π seem implausible relative to some proposed alternative, then π is attentionally unstable. Put differently, we do not interpret the person as actively noticing aspects of the data for the dedicated purpose of distinguishing π from λ. After all, he sees no reason to question π. Economizing on attention plainly implies a person shouldn't query available evidence for signs of errors they aren't aware they are making.[14]

While our framework allows for any λ, most of our analysis assumes λ is the correct model that we as researchers think the agent ought to entertain. Accordingly, when we assess the attentional stability of π without reference to a particular light-bulb model, we implicitly take $\lambda = \pi^*$. Focusing on $\lambda = \pi^*$ both pins down our analysis and most closely mirrors folk intuition regarding when people should "get a clue".[15]

[14] 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 relative likelihood of π versus 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 a simple example, suppose that a ball is drawn from an urn each day, and the balls from this urn always have the same color. Ex ante, the urn may be one of three types—it may contain purely red, blue, or yellow balls. If the person's erroneous theory π posits that the urn only has two types—purely red or blue—then $P(n^t|\pi)/P(n^t|\lambda)$ may converge to 0 when facing a yellow urn but not when facing a red or blue urn. If there is any chance that the noticed data wakes the person up, we deem π attentionally unstable.

[15] At the same time, taking $\lambda = \pi^*$ gives a potentially misleading impression that if a person discovers his model is wrong, then he necessarily abandons it in favor of the true model. Yet, 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. The dynamics following a "light-bulb moment" where π is deemed unstable—and specifying which model a person

In the interest of presentational parsimony, probabilistic statements and definitions are with respect to the true data-generating process given the realized parameter value, θ^*, that was drawn by nature from π^*. Equivalently, we could imagine π^* as being degenerate on some θ^*.[16] To make the role of channeled attention clear, we will also focus throughout on theories π that are *identifiably wrong*, meaning that π is attentionally unstable with respect to π^* under a full-attention SAS. Hence, with full attention, a person would eventually discover any wrong belief that results in costly errors.[17]

With channeled attention, however, costly errors can persist. As an illustration, return to the cold-medicine example. Under the seemingly sufficient attentional strategy that ignores outcomes while not using the medicine, the patient will eventually act as if he learns the true θ_T^* while maintaining his false belief $\hat{\theta}_N$. Such erroneous beliefs will cause the patient to ultimately use the treatment whenever $\hat{\theta}_N < \theta_T^* - (c/b)$. Furthermore, the patient will not only misperceive the treatment's value, but may also act as if he believes in a false causal relationship. Consider the case where $\theta_N^* = \theta_T^*$, so the medication does not contribute to the patient getting better quickly whatsoever. If $\hat{\theta}_N < \theta_N^*$, the patient acts as if $\theta_T = \theta_N^*$ and thus as if he is more likely to get better fast with treatment. The difference $\theta_T^* - \hat{\theta}_N = \theta_N^* - \hat{\theta}_N > 0$ captures the patient's perceived effect of the medication on getting better quickly, which in truth is zero.

The next proposition shows that checking for attentional stability simplifies to assessing whether a person who notices which of his current actions are optimal will discover his error. The proposition also characterizes a minimal SAS for any given π.

Proposition 2. *For all t, let $X_t^* \subseteq X_t$ denote the set of subjectively optimal actions in period t given model π.*

1. *There exists a minimal SAS given π with the following form: in each period t, the person notices X_t^* and nothing more.*

2. *There exists a minimal SAS given π that is a stable attentional strategy given π if*

adopts after rejecting π—is not part of our formal analysis.

[16] Whether 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.

[17] We discuss some features of attentional stability under full attention in Appendix B.1, and we note that a model π 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 makes the same predictions over observables as θ^*; and (ii) any π that assigns positive probability to $\theta^* \in \text{supp}(\pi^*)$ is attentionally stable with respect to any $\lambda \in \Delta(\Theta)$ and a full-attention SAS.

$\liminf_{t\to\infty} P(X_t^*|\pi,\phi)/P(X_t^*|\lambda,\phi) > 0$ 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 $\liminf_{t\to\infty} P(X_t^*|\pi,\phi)/P(X_t^*|\lambda,\phi) = 0$ with positive probability under π^* when the person follows the SAS ϕ from Part 1.*

Part 1 of Proposition 2 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?" It is then clearly sufficient for the person to ignore everything else, including aspects of actions previously taken that the person deems unnecessary to answer this question. 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 of this result for checking whether there is a stable attentional strategy given π.

This result is helpful in assessing the stability of commonly studied biases in environments often explored in the literature. For instance, consider an investor who wrongly thinks that innovations to an asset's earnings are autocorrelated as in Barberis et al. (1998), when in reality earnings follow a random walk. As we show in Appendix C.1, this bias is stable when the investor simply decides whether to buy or sell the asset in each period. Intuitively, any isolated buy-sell decision does not entirely reveal the path of earnings and thus does not force the investor to confront her misconception of the earnings process. For another example, consider naivete about a self-control problem as in O'Donoghue and Rabin (1999, 2001) or Eliaz and Spiegler (2006). Such naivete can be attentionally stable when the person's sole decision each day is whether to take an action with an immediate cost and delayed benefit, such as going to the gym. Reflecting Proposition 2, 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 herself "do I want to go to the gym today?"—she doesn't ask the additional question "if not, why?" This limited data is insufficient to reveal the severity of her self-control problem, as we detail in Appendix C.2.

Proposition 2 is also useful for examining comparative statics on the stability of a bias across environments, or across biases in a given environment. To illustrate the latter, consider a small family business that makes pricing and production decisions using a misspecified model of demand. For concreteness, imagine a bagel seller as in Levitt (2016) who decides how many bagels to produce and how to price them. Suppose the seller both cares about maximizing current profits and about accurately reporting cumulative profits (e.g., for tax purposes). In each period, the seller chooses a price from $\{p_1, p_2\}$ with $p_1 < p_2$. Let $\theta^*(p_k)$ denote the true likelihood that an individual consumer buys a bagel at price p_k while the seller believes this value is $\hat{\theta}(p_k)$. Further, suppose price p_1 is optimal under $(\hat{\theta}(p_1), \hat{\theta}(p_2))$.

The seller will discover errors that lead them to mispredict demand at their model-optimal price p_1, but may not discover errors that lead them to mispredict the truly-optimal price. More formally, if the seller is wrong about $\theta(p_1)$, their reported profits will eventually look much more consistent with $\theta^*(p_1)$ than with $\hat{\theta}(p_1)$, which will lead them to discover their mistake. On the other hand, if the seller is correct about $\theta(p_1)$ but incorrectly believes $\hat{\theta}(p_2) < \theta^*(p_2)$, then they need not discover this mistake: in the long run, anticipated cumulative profits will equal actual cumulative profits even if the seller persistently sets p_1 when p_2 is in fact optimal.[18] This pattern reflects Levitt (2016)'s evidence that the (MIT-trained economist-turned) bagel seller consistently set suboptimal prices despite setting reasonable quantities—and that the seller did this despite Levitt (2016) being able to uncover evidence of suboptimal pricing using the seller's *own* data. We will return to this example below by considering which environments could induce the seller to notice their error.

2.4 Discussion of Our Assumptions and Motivating Evidence

Motivating Evidence. 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. Some research in this tradition also illustrates our assumption that people can parse data in a way that ignores distinctions they deem irrelevant, even when they must notice and *act* on those details. For instance, Johansson et al. (2005) had experimental participants choose which of two faces was more attractive, but found a considerable fraction defended choices as recorded by the experimenters without noticing that those recorded choices did not match their previously stated preferences.

Beyond simply showing that conspicuous stimuli may go unnoticed, the psychological research also 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), and others. The notion of sparsity following Gabaix (2014) has likewise elaborated on the calculus of optimal attention.[19] And Payzan-LeNestour and Woodford (2020) experimentally demonstrate that a form of "outlier blindness",

[18] This is the case even when the seller could access the quantities they would have sold at price p_2 (e.g., the seller sometimes varies prices in response to changes in the cost of ingredients or sees signals of demand for rival sellers with different prices), since the seller sees no need to attend to this data.

[19] 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.

whereby observers do not distinguish between improbable extreme values along a dimension, may be consistent with rational information processing. Although we (crucially) differ in allowing attention to be guided by misspecified models, our framework is in this tradition of goal-directed attention.

Other Features of Attention. One limitation of this approach is that it abstracts from some other realistic components of attention. Much of the psychological research identifies aspects of stimuli, such as visual features, that tend to induce or escape attention, but various strands have emphasized very different themes. A substantial literature emphasizes the notion of "attentional capture", whereby certain stimuli tend to capture attention independent of the instrumental value of noticing them.[20]

Recent economic models have drawn on these insights. Bordalo et al. (2012, 2013, 2020) build on themes explored in the attentional-capture research to formalize ways in which surprising features of a context may receive disproportionate weight in decisions. Continuing this idea that stimulus-driven attention plays a role in economic choices, Li and Camerer (2021) show that machine-learned algorithms trained to predict which features of visual scenes attract attention also predict choices in experimental games.

Our focus on how theories channel our attention clarifies how it's simultaneously true that surprising features of the data often grab our attention, while unexpected features do not. The word "surprise" embeds the act of noticing. Of the infinity of unexpected things that happen to us at every moment, it is only those things that our theories focus on that typically attract our attention. 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.

All this said, it is worth emphasizing that what fails to wake people up are almost always *statistical* gorillas that accrue over time. Even if an aspect of the environment automatically drives someone's attention in the moment, that individual observation is typically not enough to alert him that his theory is wrong.

Interaction Between Attention and Memory. Without further restrictions on the noticing strategy,

[20] See Bordalo et al. (2022) for a review targeted to economists.

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. A neglectful patient could, for example, notice whether the empirical frequency of feeling better quickly absent medication exceeds some threshold without ever noticing the exact empirical frequency. However, at some level the exact frequency must be tracked in order to determine whether it exceeds a threshold. 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).[21] As we discuss in that section, 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. We also think this is a reasonable first pass in situations where a person doesn't need to rely solely on memory to guide decisions (e.g., information is stored in a database or can be written down).[22]

3 Which Errors Tend to be Stable?

This section examines which errors tend to be attentionally stable both within and across environments. Our formal results reflect a relatively simple intuition: errors that involve being wrong and certain are more stable than errors that involve being wrong yet uncertain. In the former case, a person finds it sufficient to notice the answer to fewer questions, providing less scope for incidental learning. We begin by formalizing and probing the limits of this intuition.

Revisiting the bagels example from Section 2.3, suppose the seller now thinks the state of demand may be either high (H) or low (L). For example, the seller may be uncertain about the degree of competition in the market. The seller thinks the likelihood of a sale at price p_k in state $d \in \{H,L\}$ is $\hat{\theta}_d(p_k)$ and $(\hat{\theta}_d(p_1), \hat{\theta}_d(p_2))$ are such that the seller thinks p_1 and p_2 are optimal in the low and high states, respectively. Suppose the seller is right about $\hat{\theta}_L(p_1)$ and $\hat{\theta}_H(p_2)$ but is wrong about $\hat{\theta}_L(p_2)$—in fact, the true values, $\theta^*(p_k)$, are such that the optimal price is always p_2.

If the seller needs to learn whether demand is high or low through sales history, then her incorrect

[21] Recent work in economics studies implications of memory-based models of imperfect recall (Mullainathan, 2002a; Bernheim and Thomadsen, 2005; Bodoh-Creed, 2019; Enke et al., 2019; Kőszegi et al., 2021) and the interaction between attention and memory (Bordalo et al., 2014, 2020).

[22] 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 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.

model is attentionally unstable, since carefully tracking the frequency of sales at both prices in order to determine the optimal price will reveal data inconsistent with her model. On the other hand, if the seller were informed about the state of demand *ex ante*, then she thinks she knows the optimal price, and thus she need not attend to profits at alternative prices. Moreover, the observed profits will be consistent with the seller's model despite charging the suboptimal price. The seller will therefore fail to discover her misspecification about demand at alternative prices. Because uncertainty promotes incidental learning, gleaning answers to questions from experience may provide more opportunities to wake up than knowing some of those answers ahead of time.

The following result generalizes this example and clarifies when uncertainty promotes or hinders the discovery of errors.

Proposition 3. *Fix any environment* $\Gamma \equiv (\Theta, \times_{t=1}^{\infty} X_t, \times_{t=1}^{\infty} Y_t, \times_{t=1}^{\infty} u_t, P, \pi^*)$. *Compare two models* π *and* $\tilde{\pi}$, *where* $\tilde{\pi}$ *deems a set of parameters* $\widetilde{\Theta} \subset \text{supp}(\pi)$ *entertained by* π *are not possible: model* $\tilde{\pi}$ *satisfies* $\text{supp}(\tilde{\pi}) = \text{supp}(\pi) \setminus \widetilde{\Theta}$. *Define*

$$\Theta^{reject}(\Gamma, \phi) = \{\theta \mid \Pr(\theta | n^t, \pi) \to 0 \textit{ almost surely given } \pi^* \textit{under SAS } \phi\}.$$

1. *Suppose* $\widetilde{\Theta} \subseteq \Theta^{reject}(\Gamma, \phi^{min})$, *where* ϕ^{min} *is a minimal SAS given* π *that is a stable attentional strategy if one exists, and is some minimal SAS otherwise. If there exists a stable attentional strategy given* π, *then there exists a stable attentional strategy given* $\tilde{\pi}$. *The converse, however, does not generally hold.*
2. *When* $\widetilde{\Theta} \not\subseteq \Theta^{reject}(\Gamma, \phi^{min})$, *there may exist a stable attentional strategy given* π, *but not given* $\tilde{\pi}$.

Part 1 of this proposition generalizes the example above, and shows that incidental learning is hindered when somebody knows ex ante the answer to a question *that she would otherwise figure out herself.* Even under the wrong model and a coarsely-attentive SAS, the seller in the example could correctly learn over time whether the demand is high or low.[23] But knowing this information upfront prevents her from waking up.

On the other hand, Part 2 of the proposition clarifies that providing a person with information ex ante that contradicts what they would otherwise conclude (given their misspecification and SAS) promotes waking up. To illustrate with an extension of the previous example, an erroneous belief that the likelihood of a sale may fluctuate over time could be attentionally stable, since it provides an explanation for why the observed frequencies of sales are inconsistent with $\hat{\theta}_d(p_k)$ in the long run. However, if the seller were informed ex ante that these frequencies are fixed, then her model goes back to being attentionally unstable.

[23] This would be the case, for instance, if $(\hat{\theta}_L(p_1), \hat{\theta}_L(p_2))$ better explain (in the Kullback-Leibler sense) the true distribution of outcomes in state L than $(\hat{\theta}_H(p_1), \hat{\theta}_H(p_2))$.

So Part 1 of Proposition 3 says that truthfully informing a person of something they would've concluded on their own can hinder the discovery of an error, whereas Part 2 says that truthfully informing them of something that contradicts what they would've concluded can help with this discovery.

These results match the findings of Esponda et al. (2022), 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 instead 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.

We now turn to showing how the principle from above—that it is better to be wrong and uncertain than wrong and certain—provides guidance in assessing which errors tend to be stable *across* environments. We first focus on environments that meet Assumption 2 and are stationary in the following sense.

Definition 4. 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.

Given Assumption 1, we can separate the environment into two components: 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.

Definition 5. Restricting attention to stationary environments where Assumption 2 holds and fixing the outcome environment, a model π is *stable for all preferences* if for any choice environment (action space X and payoffs $u : X \times Y \to \mathbb{R}$) there exists a stable attentional strategy given π.

Lemma B.1 in Appendix B.2 characterizes when a model π is stable for all preferences. It roughly says that a theory is stable for all preferences if and only if the person would not be alerted to her errors when she pays attention to all information helpful for updating her beliefs about θ under π.[24] The idea is that there are some preferences for which the person finds it necessary to attend to such information, but none (in stationary environments) where the person finds it necessary to attend to more.

[24] That is, she pays attention to what is called a "minimal sufficient statistic" under π in the language of probability theory (e.g., Lehmann and Casella, 1998).

Lemma B.1 helps categorize models that are or are not stable across choice environments. We discuss the categorization in intuitive terms here. Maintaining the structure of our previous cold-medicine example, we fix ideas throughout the illustration of these results by considering examples where a person holds beliefs over the likelihood θ_x that she recovers from a condition (a cold, baldness, cancer, etc.) with treatment ($x = T$) or without ($x = N$). In this context, our results could be viewed as drawing out the limits of gullibility when it comes to medical beliefs.

The following three classes of models *are* stable for all preferences in stationary environments (see Appendix B.2 for corresponding formal results):

1. Dogmatic errors where π is degenerate. For instance, suppose the person has false but dogmatic beliefs about the impact of medical treatments and is confident about which action to take. Then the minimal sufficient statistic distinguishes no data, and the person will not be alerted to her error.

2. "Censored" models that ignore possible outcomes. An example is the cold-medicine example from the introduction with $\hat{\theta}_N = 0$—the person thinks she will only recover quickly if she uses the treatment. Accordingly, the person's attentional strategy does not distinguish how quickly she improves when she does not use the treatment, since she thinks she knows the outcome in those instances.

3. Models that neglect predictive signals and thus treat some truly predictive signals as uninformative. For example, the person may fail to recognize that the efficacy of some medical treatments depends on the situation, such as by neglecting that antibiotics are necessary for only some types of bacterial infections.[25]

In contrast, the following three classes of models *are not* stable for all preferences, meaning they are prone to incidental learning in some stationary environments.

1. Uncertain models that correctly specify the set of outcomes but incorrectly specify their probabilities.[26] As we saw above, if a person's only error is mistakenly believing that fast recovery is impossible absent treatment (e.g., $\hat{\theta}_N = 0$ when in reality $\theta_N^* = 1/2$), then his model is stable for all preferences. On the other hand, if the person's only error is the

[25] Other examples of predictor neglect include farmers failing to appreciate the importance of pod size (Hanna et al., 2014); small investors failing to appreciate the importance of analyst affiliation when interpreting investment recommendations (Malmendier and Shanthikumar, 2007); investors failing to appreciate that the way a manager chooses to report current earnings predicts future earnings (Teoh et al., 1998a,b); or physicians using overly simple models to predict the value of testing for heart attacks (Mullainathan and Obermeyer, 2022).

[26] Specifically, if no two observations within a period lead to the same beliefs over parameters (defined in Appendix B.2 as the *Varying Likelihood Ratio Property*), then the person believes that separately noticing every outcome would aid in learning θ. When the agent has incentives to learn θ, he will thus notice that his model is miscalibrated.

seemingly "smaller" one where he is uncertain whether $\theta_N = 1/8$ or $\theta_N = 1/4$, then his model is no longer stable for all preferences.

2. "Overly elaborate" (and non-dogmatic) models that anticipate too wide a range of outcomes. These models represent a counterpoint to "censored" models. A person's model is overly elaborate, for example, if he believes in miracle cures for chronic conditions that cannot be cured, putting weight on the impossible outcome of curing the incurable. An illustration would be a person who thinks he has a chance of de-aging 10 years each time he applies a "miracle oil" and is uncertain of that chance (i.e., $\theta_T^* = 0$ in reality, but the person's model π puts weight only on $\theta_T > 0$). Such models are prone to discovery when the person has an incentive to track the frequency of various outcomes, since the person will eventually notice that an impossible outcome fails to materialize.

3. "Over-fit" (and non-dogmatic) models that assume the set of predictive signals is wider than it truly is. These models represent a counterpoint to those that neglect predictive signals. Such models are unstable for some preferences when there is uncertainty about how useful the signals are (e.g., a person thinks the miracle oil might systematically work better for some people than others). In these contexts, the person would attend to the sequence of signals and resolutions, which would ultimately prove his model false.

This collection of results above can, in the context of medical beliefs, be seen as revealing the limits to gullibility under channeled attention by classifying errors that a person would indeed notice given the appropriate circumstances. While a person may continually believe some medicine is necessary to recover quickly from a temporary ailment—as in the "censored model" case—they are more apt to discover false claims of a miracle cure for a chronic condition—as in the "overly elaborate" case.

More generally, 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*. Put in terms of the questions people ask themselves, failing to ask a question is more stable than asking the question but entertaining the wrong set of answers. In the previous examples, a person may fail to wake up when he simply doesn't ask whether his colds go away quickly without medicine, but he may discover a false miracle cure when he continually asks how effective it has been, only to discover an answer that wasn't anticipated—it hasn't worked at all.

This categorization of errors may help clarify researchers' often implicit assumption that "thinking-through-categories" and "thinking coarsely" go together. Indeed, researchers often interchangeably talk about coarse and categorical thinking (see, e.g., Mullainathan, 2002b; Mullainathan et al., 2008). While people in principle could categorize social groups, situations, or objects too finely, our results suggest they are more likely to eventually wake up to such errors relative to those where they categorize too coarsely. This may be why we rarely see such "overly fine" categorizations.

We now relax the stationarity restriction and ask which errors are attentionally stable across history-dependent action spaces and utility functions.

Definition 6. Fixing the outcome environment, a model π is *strongly stable for all preferences* 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.

While any model that is strongly stable for all preferences also satisfies Definition 5, the converse is not true.

Proposition 4. *There are outcome environments where there exist identifiably wrong models π that are strongly stable for all preferences. Any such model must assign zero probability to some finite history h^t that occurs with positive probability under π^*.*

Proposition 4 implies that models that anticipate the correct set of outcomes are *less* universally stable than models that anticipate the wrong set of outcomes. The idea is that in order for an error to be universally stable it must hold up in any environment where the person has an incentive to track all patterns over time that she deems possible. If she's correct about which patterns are possible, she discovers any statistically identifiable mistake with such an incentive.

"Bigger" errors may then be more robustly stable than "smaller" errors, since bigger errors engender less overlap between the questions a person seeks to ask and those she should be asking. As an illustration, consider a worker who is able to expense work travel, but needs to submit receipts in order to do so. If she thinks that she always remembers to submit them immediately when she returns from a trip, then she doesn't need to notice that she sometimes forgets. On the other hand, if she recognizes that she sometimes forgets but underestimates how often, then there's an environment where she'll discover she forgets more often than expected. For example, if she knows she sometimes forgets, she might write "submit travel receipts" on her to-do list and periodically review that list to see what still needs to be done. She will then be shocked to find this item still on her list a year later—she's more forgetful than she thought possible.

4 What Features of Situations Tend to Make Errors Stable?

We now explore features of the environment that influence an error's stability. We separately analyze the effects of the person's objectives and of the information structure.

4.1 How A Person's Goals Influence Stability

We first establish that there is no tight link between whether a person wakes up to an error and the cost of the error. For any outcome environment and erroneous model, there exists *some* choice environment in which that model is stable. For instance, if u_t is independent of outcomes, then the person has no incentive to attend to data and his model is thus attentionally 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. Before stating these results, we first define our notion of a costly error more formally.

Definition 7. Consider a model π and SAS $\phi = (\mathcal{N}, \sigma)$ given π. Let $\bar{u}_t(\phi|h^t) \equiv \mathbb{E}_{(\theta^*,\phi)}[u_t(x,y)|h^t]$ denote the expected utility in period t given the true parameter θ^*, history h^t, and SAS ϕ. Let ϕ^* be any SAS given π^*. The SAS ϕ is *costless* if $|\bar{u}_t(\phi|h^t) - \bar{u}_t(\phi^*|h^t)| \to 0$ almost surely given 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 5. *Consider an outcome environment $(\times_{t=1}^{\infty} Y_t, \Theta, P, \pi^*)$ and a model π.*

1. *There exists a choice environment $(\times_{t=1}^{\infty} X_t, \times_{t=1}^{\infty} u_t)$ and a corresponding SAS ϕ such that ϕ is a costly stable attentional strategy given π.*

2. *If π is attentionally measurable under a full-attention SAS, then there exists a choice environment $(\times_{t=1}^{\infty} X_t, \times_{t=1}^{\infty} u_t)$ where (i) every SAS given π is costless, yet (ii) there exists no stable attentional strategy given π.*

Part 1 can be proven by construction in a way that also provides the underlying intuition. Given the outcome environment $(\times_{t=1}^{\infty} Y_t, \Theta, P, \pi^*)$, we can construct a choice environment with a fixed binary action space, $X = \{H, L\}$, such that H is optimal for any $\theta \in \text{supp}(\pi)$ but L is optimal for parameter θ^*. The proof considers a utility function with two properties: (i) the person incurs a big penalty if he ever switches actions—so he's effectively choosing between "always H" and "always L"—and (ii) "always H" yields a higher payoff in any period where π provides a better fit of the empirical distribution than does π^*. When the person believes θ^* is impossible, he thinks there is nothing payoff-relevant to learn because he is confident H is optimal in the first period and, because of the switching penalty, he thinks he should never revise his action. While the proof constructs a context where the person is ex ante dogmatic about the optimal action, many of our examples show that costly errors can remain attentionally stable even with active learning about which action to take.

Moreover, the costly attentionally stable strategies identified in Proposition 5 can be arbitrarily costly. A person fails to discover a costly mistake only when he wrongly deems valuable data entirely useless and ignores it. Yet, once deemed useless, the true value of this data—which determines the harm from the person's mistake—has no influence on his decision to ignore it. That is, no matter how great the true benefit of some data, the decision-maker may continually ignore this data if his *perceived* benefit is sufficiently small.[27]

Proposition 5 provides two messages. First, for many mistakes, it is situation dependent whether a person wakes up to them. 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. This may shed light, for instance, on why most of us have experience using a commitment device in *some* situation to counteract our self-control or memory problems, despite making few commitments overall (Laibson, 2015): a person may recognize her limited self control or limited prospective memory in one situation yet remain naive in another.[28]

Second, whether the person wakes up is not intimately tied to the cost of the mistake.[29] This is contrary to intuitions from the rational-inattention literature, where inattention is more likely to enable small mistakes than big ones. So, if the cost of the mistake doesn't encourage waking up, what does?

The basic answer, suggested by Proposition 2, is whether the person's objective encourages him to track and discern more features of the data. We formalize this idea and provide a simple result in Appendix B.3. As an illustration, a patient who does not appreciate how often he forgets to take his pills is more likely to notice he's wrong when his physician provides him with incentives to keep *and review* a diary of how often he takes his pills. Or, following the logic of the chronic-illness example, a person who intends to sign up for TSA PreCheck but keeps procrastinating or forgetting will be forced to recognize the degree to which his intentions deviate from his intentions the 100th time he's unable to skip the long security line, which serves as a quite direct summary statistic that his past behavior involved never signing up. In such examples, the person does not recognize that carefully tracking what he's seen will help him learn about the payoff-relevant parameter—such tracking therefore enables incidental learning.

[27] To see the intuition behind Part 2, consider a choice environment where in each period the person earns a payoff of 1 if he correctly repeats back the entire history and earns -1 otherwise. If π is attentionally measurable under a full-attention SAS (equivalently, π^* is absolutely continuous with respect to π), then π assigns zero probability only to events that are truly impossible under the true model. Hence, given π, the person has incentive to actively notice h^t in each period. In this case, any error is costless: if the person knows h^t, he will always take the optimal action despite holding a misspecified model. Furthermore, knowledge of h^t forces the person to wake up to any unstable error.

[28] 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).

[29] This result mirrors recent experimental findings by Enke et al. (2023) showing that very high stakes don't significantly reduce widely documented cognitive biases (e.g., base-rate neglect and anchoring) in the settings they study.

4.2 How the Information Structure Influences Stability

The information structure also influences attentional stability. In particular, 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 would respond discontinuously to an increase in price from, say, \$14.99 to \$15.00. The firm, however, fails to appreciate this discontinuity and employs a pricing algorithm that estimates 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—this 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 their error.

The following proposition summarizes and formalizes such intuitions.

Proposition 6. *Consider any environment $\Gamma \equiv (\Theta, \times_{t=1}^{\infty} X_t, \times_{t=1}^{\infty} Y_t, \times_{t=1}^{\infty} u_t, P, \pi^*)$ and a model π.*

1. *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.*

2. *Consider a modified environment identical to Γ aside from allowing the person to observe the full history each period: in the modified environment, the person receives signals $\tilde{s}_t = (s_t, h^t(\neg s_t))$ for all $t \in \mathbb{N}$, where s_t follows the signal structure of Γ and $h^t(\neg s_t)$ is the history prior to period t. There exists a stable attentional strategy given π in the modified environment if and only if there exists a stable attentional strategy given π in the original environment.*

Part 1 of Proposition 6 was discussed above. The logic of the result is illustrated by the following whimsical twist on "a needle in a haystack". If you can ask an algorithm whether the needle you're searching for is in a haystack, then you need not notice the gorilla lying in wait below the surface of the hay. If you instead have to search for it yourself, then you will inevitably confront the gorilla. The benefits of delegation or using algorithms—reducing noise and effort in decision-making—are obvious, 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.

Part 2 of Proposition 6 highlights that attentional stability is not about the limited *availability* of data per se, but rather a failure to notice the right features of the data. Any attentionally stable error will remain stable if we grant the person continual access to a complete archive of past outcomes (e.g., outcomes are recorded somewhere). That is, a person who can always revisit any past data—if he so chooses—still need not get a clue.[30]

Relaxing the assumption of perfect memory clarifies that giving a person access to the full history each period, if anything, *hinders* a person from recognizing an error. Tracking data by itself is not enough to learn when we are asking the wrong questions—we need to have an incentive to ask and answer questions that we don't feel are instrumental for our decisions. A forgetful patient does not wake up to her forgetfulness just by keeping a diary of when she takes her pills—she does so by additionally reviewing the diary. In Appendix A, we define refinements on SASs that incorporate memory imperfections and thus introduce interactions between attention and memory. We then establish that giving a person access to a recording of the full history each period impedes getting a clue for a person sophisticated about her memory imperfections. If a person thinks she should preemptively notice information today because it *might* be useful tomorrow, then she will end up noticing more today than if she had access to a recording of the full history. That is, a SAS with memory imperfections remains a SAS without them, but not vice-versa.

The results above reflect the most basic implication of our framework: a person is alerted to her mistakes not by the statistical unlikeliness of all the data in front of her, but rather by how surprising she finds the data she notices. This means that factors that are often intuited as promoting learning—increasing the stakes, decreasing the cost of information gathering, simplifying the choice, etc.—may not help in our framework (and may even backfire) depending on how they influence the person's perceived uncertainty about the optimal action.

5 Further Applications and Principles

5.1 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

[30] To provide intuition, suppose ϕ is a StAS given π in an environment where the agent does not have continual access to an archive of past outcomes. If π is not attentionally stable when the person *can* access the history prior to any decision, then there must exist information in the history that she believes would improve her decision beyond the data she gathered under ϕ. But this contradicts the assumption that ϕ is sufficient in the first place.

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's model is π. By contrast, the outsider is unsure if the appropriate model of the world is π or some alternative, π'. The outsider thus entertains both possibilities, resulting in model $\pi'' = (1-\alpha)\pi + \alpha\pi'$ for some $\alpha \in (0,1)$. Suppose that π is misspecified, so both the insider and outsider initially suffer from the same mistake. When will this mistake be stable for the insider yet recognized by the outsider?

The interesting case here is when, under full attention, $\lim_{t\to\infty} \pi''_t(\theta) = 0$ for all $\theta \in \text{supp}(\pi') \setminus \text{supp}(\pi)$. In this case, the outsider's broader model still does not encapsulate the true model of the world, but instead tacks on an alternative π' that the insider already knows is false. This is the key case of interest because if π' *did* include the true model, then the outsider would trivially learn correctly. Therefore, 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 superior information can prevent the insider from noticing as much as the outsider, and hence the outsider may discover that π is misspecified while the insider does not. This intuition clearly reflects that from Proposition 3 above, and can be formalized as a corollary.

Corollary 1. *Consider $\pi'' = (1-\alpha)\pi + \alpha\pi'$ for $\alpha \in (0,1)$. Suppose that under full attention $\lim_{t\to\infty} \pi''_t(\theta) = 0$ with probability one for all $\theta \in \text{supp}(\pi') \setminus \text{supp}(\pi)$. Then a stable attentional strategy given π'' is a stable attentional strategy given π, but the converse is not necessarily true.*

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 to be true that wakes her up to the insider's mistake.

To illustrate, consider again the small-business example from Section 3. Suppose the seller (the insider) seeks advice from a consultant (the outsider). As above, the seller holds miscalibrated beliefs, correctly thinking demand is low (e.g., because of high competition in the industry) but wrongly thinking they should set a low versus high price conditional on demand being low (e.g., because they are miscalibrated about the slope of demand). The consultant shares these miscalibrated beliefs but, because she is uncertain about the industry, she is uncertain about whether demand is low or high and thus which price to set. The consultant will carefully attend to sales data over time to assess demand. In doing so, she will confront a history that is inconsistent with her miscalibrated beliefs and incidentally discover that these beliefs are wrong. The seller, however, need not notice this because they think they already know the demand at various prices. The act of rediscovering something the seller already knows to be true (demand is low) allows the consultant, as the outsider, to notice errors that the more knowledgeable seller, as the insider, would otherwise fail to recognize.

This result demonstrates how, 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.[31]

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. 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. (2022) shows that encouraging entrepreneurs to think in this way improves the accuracy of their forecasts about the revenue growth of their companies.

5.2 The Complexity Error-Recognition Disconnect

Why do we sometimes get simple problems wrong—such as the famous bat-and-ball problem (Frederick, 2005)—and more complex problems (eventually) right—such as more abstract math problems? Our framework illustrates a force at play: if we think we know how to answer a simple problem, we don't see the need to notice information that might reveal we're wrong. On the other hand, even if we (wrongly) think we know how to answer simple problems, the process of trying to map a complex problem into these simple problems can alert us that we do not, in fact, have all the answers.[32]

[31] 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). Corollary 1 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.

[32] Indeed, Rabin (2013b) argues that many important errors are "astray" errors where the right answer is not very hard to see but the wrong answer is enticing.

To illustrate the logic, consider a prediction problem where, in each period, the person reports a prediction $x_t \in [0,1]$ about a binary outcome $r_t \in \{0,1\}$ and earns a payoff of $u_t = -(x_t - r_t)^2$. The probability that the outcome equals 1 in any period is $\theta_s \in [0,1]$, which depends on the "situation" $s \in S \subset [0,1]$. Suppose for all $s \in S$, $\theta_s \neq \theta_{s'}$ if $s \neq s'$, both in truth and under the person's theory π. However, the person is misspecified about the value of θ_s for each s, thinking it is $\hat{\theta}_s \neq \theta_s$. If the person knows the situation ahead of making her prediction, then there's a stable attentional strategy involving her repeatedly reporting $x = \hat{\theta}_s$ and noticing nothing else. On the other hand, if she does not know the situation and instead needs to learn it, then she'd notice that the true probability that $r = 1$ is not $\hat{\theta}_s$ for any $s \in S$.[33] The reason is that any SAS requires the person to track the number of times that $r = 1$ to discern the situation and relevant value of $\hat{\theta}_s$. This is enough to incidentally alert her that empirical frequency of $r = 1$ does not match $\hat{\theta}_s$ for any s. This is an example where the person thinks she knows the right answer in any situation. If she knows the situation, she won't wake up to being wrong about the answer. However, if she doesn't know the situation, then she could wake up to not having the right answer as a byproduct of trying to figure out which situation she's in.

There are other natural forms of complexity that similarly lead a person to recognize her errors, for example reducing the quality of feedback. Consider the tendency to neglect correlations in others' opinions (as in DeMarzo et al., 2003; Eyster and Rabin, 2010, 2014; Enke and Zimmermann, 2017). Appendix C.3 fleshes out an example of a newly hired employee who seeks advice from her colleagues. Over time, this person updates her beliefs about the quality of advice that each colleague provides. However, she wrongly treats their advice as conditionally independent, ignoring that colleagues also talk with each other. Whether the person wakes up to this error depends on the amount of feedback she receives. If she always observes whether advice was useful or not ex post, then she can learn about an individual colleague simply by comparing his recommendations with the realized outcomes. Since this strategy does not depend on correlations across colleagues' advice, her mistaken model is attentionally stable under a minimal SAS; consequently, she may persistently overreact to consensus advice. In contrast, if she does not always observe whether advice was useful or not, then any SAS requires her to notice the correlation across others' advice: in the absence of feedback, the efficient way to update about the quality of a colleague is to "benchmark" his advice against that of other colleagues believed to be knowledgeable. Thus, with limited feedback, she will notice that colleagues are simply echoing one another and hence discover her mistake.

Such examples illustrate how complexity can help one discover that they're looking at a problem the wrong way. In both examples, complexity improves long-run decision making by increasing engagement with feedback, which matches recent experimental findings (Esponda et al., 2022).

[33] That is, suppose the situation s is such that $\theta_s \notin \{\hat{\theta}_s : s \in S\}$.

5.3 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.[34]

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, e.g., 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, perhaps in part because of channeled attention. Indeed, as discussed in Section 3, theories that omit factors are robustly attentionally stable, while theories that include too many factors are not.

We see something similar with the rise of "sufficient statistics" approaches in public finance (e.g., Chetty, 2009; Mullainathan et al., 2012) and "portable modeling" in applied (e.g., behavioral) economics (Rabin, 2013a). 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 important 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 ignore data on beliefs that defy rational expectations (e.g., Greenwood and Shleifer, 2014).[35]

[34] 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.

[35] 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

5.4 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 naive about their prospective memory, their self control, and the heuristics and biases they display. 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, 2021). Consider the following scenario. Biddy (she) is thinking about working with Addy (he) on a project. She and Addy both think that Addy is 90% productive 50% of working hours and is 10% productive (doomscrolling on Twitter) 50% of working hours. In deciding whether to work with Addy, Biddy wants to monitor his work habits to figure out whether his productive hours align with hers' (she's an afternoon person). He doesn't feel the need to monitor his own work habits because he believes that he knows when he's productive—in the later hours. Biddy ends up agreeing with Addy that he's more productive later in the day but, by feeling the need to understand when he's relatively more productive, she incidentally learns something else: he's much less productive overall than either of them thought possible—he's always doomscrolling![36]

The logic of this example 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 things they don't, which enables incidental learning. 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. In the previous example, Θ^1 captures whether Addy is more productive in the morning or afternoon and Θ^2 captures how productive Addy is at these times. A person is dogmatic about his own θ^1, while an observer is unsure of this value. Corollary 1 may then apply (as it does in the Addy example), suggesting the observer is more likely to wake up to a misspecification than the person himself.

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, espe-

on those consequences, it can seem like such models assume agents are willfully ignoring useful information.

[36] Formally, suppose θ is two-dimensional, where the first dimension captures whether Addy is more productive in the morning or afternoon, and the second dimension captures how productive Addy is in more- and less-productive hours. Addy is dogmatic he is more productive in the afternoon and that he completes tasks 90% of the time in productive hours and 10% of the time in unproductive hours. In reality, he only completes tasks 20% of the time in productive hours. Biddy's prior is the same as Addy's, except she's unsure whether Addy is more productive in the afternoon, placing prior probability ψ on Addy being more productive in the afternoon and probability $(1-\psi)$ on him being more productive in the morning. We assume in the example that Biddy has an incentive to accurately track her posterior probability that Addy is more productive in the afternoon. Her posterior belief (together with the time period) will reveal over time that Addy is less productive than either of them thought possible.

cially when this failure concerns their own behavior that they think they understand.[37]

6 Related Literature on the Stability of Erroneous Beliefs

There are of course many reasons beyond channeled attention why people may not recognize that their models are wrong. The plainest and most pervasive 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, as in self-confirming equilibrium (Fudenberg and Levine, 1993) and "bandit problems" (Gittins, 1979). Or it could more broadly be because the available data doesn't allow people to reliably distinguish between models (Ba, 2022). 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 has instead been on the many scenarios where people have sufficient data to discover their mistakes yet fail to do so, such as those where fresh eyes catch mistakes, where people recognize that others make an error but don't notice themselves making that same error, and where people continue using overly-coarse models when the data would allow them to estimate further parameters.

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).[38] 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

[37] Recent experimental work in economics demonstrates how people may fail to recognize biases in themselves (e.g., their own forgetfulness), despite experience and feedback that could correct those biases (Bronchetti et al., 2023).

[38] Some of these statistical errors, such as base-rate neglect (Bodoh-Creed, 2019) or the non-belief in the law of large numbers (Benjamin et al., 2016) even predict cases where 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 or reinforce, for instance, an optimistic view of their abilities (Bénabou and Tirole, 2002; Gottlieb, 2019; Möbius et al., 2022).

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).[39] 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 are motivated to not acknowledge certain conclusions despite knowing they are true. Relative to this literature, we shed light on scenarios where people repeatedly make very costly mistakes in the face of data that reveal those mistakes.

Overall, these alternative factors surely contribute to the persistence of some erroneous beliefs, yet, by neutralizing them, we demonstrate how channeled attention itself enables the persistence of errors. We speak 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.

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]

As a byproduct, this focus also stacks the deck *in favor* of a person noticing his errors, as does

[39] 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.

[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., 2022), 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.

our study of repetitive choice contexts that provide sufficient data to identify the true model. In this sense, our attentional 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.

An additional limitation inheres in our ambition to provide a sharp framework that we and others can broadly apply: our approach 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 they would notice because it is hard not to notice. One could embed assumptions about what captures attention as a primitive, and our general framework leaves such assumptions to be imposed on a case-by-case basis.[42] However, it is worth noting again that any form of attentional capture that influences what *momentarily* draws attention without systematically influencing what comes to mind from memory will typically not overturn our results on when people discover their errors.

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 are exposed to the correct alternative and data supporting it, it may not be immediately compelling because they don't think they need to pay attention.[43] Our theory also highlights the importance of providing data that is relevant *within* that 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," Mimeo 2021.

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

[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] 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.

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 2023.

Ba, Cuimin, "Robust Model Misspecification and Paradigm Shifts," Mimeo 2022.

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., 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.

Bernheim, B. Douglas and Raphael Thomadsen, "Memory and Anticipation," *The Economic Journal*, 2005, *115* (503), 271–304.

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.

Blake, Thomas, Chris Nosko, and Steven Tadelis, "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment," *Econometrica*, 2015, *83* (1), 155–174.

Bodoh-Creed, Aaron L., "Mood, Memory, and the Evaluation of Asset Prices," *Review of Finance*, 2019, *24* (1), 227 – 262.

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*, 2022, *Forthcoming.*

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,” Mimeo 2014.

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

_ **,** _ **, and** _ , “Salience,” *Annual Review of Economics*, 2022, *Forthcoming.*

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 F. Stehr, Justin R. Sydnor, and Dmitry Taubinsky, “How are Preferences For Commitment Revealed?,” Mimeo 2021.

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.

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.

_ **, Frederik Schwerter, and Florian Zimmermann**, "Associative Memory and Belief Formation," Mimeo 2019.

_ **, 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.

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

_ **, Emanuel Vespa, and Sevgi Yuksel**, "Mental Models and Learning: The Case of Base-Rate Neglect," Mimeo 2022.

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

Eyster, Erik 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 Naivete: Beliefs About Self-Control," *Available at SSRN 2727499*, 2021.

Frederick, Shane, "Cognitive Reflection and Decision Making," *Journal of Economic perspectives*, 2005, *19* (4), 25–42.

_ **, 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.

Gagnon-Bartsch, Tristan, "Taste Projection in Models of Social Learning," Mimeo 2016.

_ **and Benjamin Bushong**, "Learning with Misattribution of Reference Dependence," Mimeo 2022.

_ **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.

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.

Johansson, Petter, Lars Hall, Sverker Sikström, and Andreas Olsson, "Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task," *Science*, 2005, *310*, 116 – 119.

Kőszegi, Botond, George Loewenstein, and Takeshi Murooka, "Fragile Self-Esteem," *Review of Economic Studies*, 2021, *Forthcoming.*

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.

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.

__ , "Why Don't Present-Biased Agents Make Commitments?," *American Economic Review Papers and Proceedings*, 2015, *105* (5), 267–72.

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

Levitt, Steven D., "Bagels and donuts for sale: A case study in profit maximization," *Research in Economics*, 2016, *70* (4), 518–535.

Li, Xiaomin and Colin Camerer, "Predictable Effects of Bottom-up Visual Salience in Experimental Decisions and Games," Mimeo 2021.

List, John A, Ian Muir, Devin Pope, and Gregory Sun, "Left-Digit Bias at Lyft," *The Review of Economic Studies*, 02 2023.

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

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': Efficient Coding Generates an Inability to Represent Extreme Values," Mimeo 2020.

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.

_ , "Incorporating Limited Rationality into Economics," *Journal of Economic Literature*, 2013, *51* (2), 528–543.

_ 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.

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.

Teoh, Siew Hong, Ivo Welch, and T. J. Wong, "Earnings Management and the Long-Run Market Performance of Initial Public Offerings," *The Journal of Finance*, 1998, *53* (6), 1935–1974.

_ , _ , and _ , "Earnings management and the Underperformance of Seasoned Equity Offerings," *Journal of Financial Economics*, 1998, *50* (1), 63–99.

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," 2022.

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 refines noticing strategies, if there exists a stable attentional strategy under memory consistency, then there exists a stable attentional strategy without requiring memory consistency.[44]

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 he currently believes the worker's success rate, θ, exceeds 50%. Suppose the manager is overly

[44] 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.

pessimistic about the worker and the support of his misspecified model has an upper bound strictly below the true rate, θ^*. 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 2). The optimal assignment in a single period provides but a rough sense of how often the worker has succeeded (e.g., only whether this frequency exceeds 50%). 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.[45] 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.

Definition A.2. A noticing strategy $\mathscr{N}$ satisfies *automatic recall (AR)* if for all $t \in \mathbb{N}$ and $h^t \in H^t$,

[45] For instance, assigning the high task reveals only that the employee has delivered good performance in at least 50% of the previous periods, 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.

$\tilde{h}^t \notin n^t(h^t)$ 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 2.3; see Footnote 17). These serve as benchmarks for assessing the impact of channeled attention. For simplicity, we consider environments that are stationary (Definition 4). Given our focus on models that are identifiably wrong, a misspecified model is stable under full attention in such settings if and only if it explains observations as well as the alternative model.

More precisely, Observation B.1 below shows that a theory π is attentionally stable with respect to λ and a full-attention SAS if π explains observations better than λ (in the Kullback-Leibler sense) and attentionally unstable if it does worse.[46] This observation suggests that, within environments with rich feedback (i.e., π is identifiably wrong), any false theory that is stable under a full-attention SAS necessarily generates no long-run welfare loss. Hence, with full attention, stable models do not continually generate costly mistakes in the environments we consider.

Given Assumption 1, let $D(\theta^* \| \lambda) \equiv \min_{\theta \in \text{supp}(\lambda)} D(\theta^* \| \theta)$, where $D(\theta^* \| \theta)$ is the Kullback-Leibler Divergence of $P(\cdot|\theta)$ from $P(\cdot|\theta^*)$, and define $\Delta D(\theta^* \| \lambda, \pi) \equiv D(\theta^* \| \lambda) - D(\theta^* \| \pi)$ as 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.*

[46] The Kullback-Leibler divergence is given by

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

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 $\theta^* \in \operatorname{supp}(\lambda) \cup \operatorname{supp}(\pi)$.*

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

B.2 Stability for All Preferences in Stationary Environments

In this section, we formalize the results discussed in Section 3 pertaining to stability for all preferences in stationary environments. Accordingly, we maintain Assumption 2 and stationarity (Definition 4) throughout this section. As a prelude to results below that focus on specific classes of models, we first provide a more general characterization of when a model π is stable for all preferences, 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 $Y(\pi) \equiv \cup_{\theta\in\operatorname{supp}(\pi)} Y(\theta)$. Concepts from probability theory regarding minimal sufficient statistics (e.g., Lehmann and Casella, 1998) help us analyze when models are stable for all preferences. Let

$$m_\pi(y) \equiv \left\{ y' \in Y(\pi) \mid \exists g(y',y) > 0 \text{ s.t. } \forall\theta \in \operatorname{supp}(\pi), P(y'|\theta) = P(y|\theta)g(y',y) \right\}. \qquad \text{(B.2)}$$

In words, the partition defined by $m_\pi(y)$ 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' as she would after noticing y. We analogously define minimal sufficient statistics over histories $y^t = (y_{t-1}, y_{t-2}, \ldots, y_1)$, which we denote by $m_\pi(y^t)$.[47] To ensure that the "minimal sufficient statistic" partition defined by (B.2) is fixed over time, we assume that $P(y|\theta) \in (0,1)$ for all $(y,\theta) \in Y(\pi) \times \operatorname{supp}(\pi)$.[48]

With these concepts in hand, the following result characterizes when a theory is stable for all preferences.

[47] For an example of $m_\pi(y)$ and $m_\pi(y^t)$, suppose $Y = \{0,1\}$ and $\theta \in [0,1]$ represents the probability that an employee delivers a successful performance (i.e., $y = 1$) on any given day. Suppose a manager'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

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

is independent of θ if and only if $\tilde{k} = k$.

[48] In principle, $m_{\pi_t}(y)$ could vary in t if $P(y|\theta) \in \{0,1\}$ for some $(y,\theta) \in Y(\pi) \times \operatorname{supp}(\pi)$ since the support of π_t may differ from that of π.

Lemma B.1. *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 for all preferences if and only if there exists* $\theta \in \text{supp}(\pi)$ *such that with probability one*

$$\liminf_{t\to\infty} \frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)} > 0. \tag{B.3}$$

The idea behind this result is straightforward: in stationary environments, it is always sufficient—and sometimes necessary—for the person to attend to all information helpful in updating beliefs about θ. Therefore, if noticing $m_\pi(y^t)$ does not force the person to wake up, then there is *no* choice environment that will. On the other hand, if noticing $m_\pi(y^t)$ does force the person to wake up, then we can find a choice environment where noticing this data is necessary under any SAS. The proof (presented in Appendix D.2) also shows how this result extends straightforwardly under the natural restrictions on memory introduced above in Appendix A (i.e., under memory consistency and automatic recall from Definitions A.1 and A.2).

The remainder of this section formalizes the enumerated results discussed in Section 3. The following propositions follow from Lemma B.1. Thus, we maintain the assumptions of Lemma B.1 throughout this subsection. For intuitive descriptions of the following results, see the main text; see Appendix D.2 for proofs.

As in in Section 3, we first consider classes of models that are stable for all preferences (Definition 5).

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 Lemma B.1 hold. If* π *exhibits a dogmatic error, then* π *is stable for all preferences.*

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 Lemma B.1 hold. If* π *is censored, then* π *is stable for all preferences.*

   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 for all preferences under automatic recall (see Definition A.2). Thus, $Y(\pi) \subset Y(\theta^*)$ alone is enough to imply that π is stable for all preferences.

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 $s_t \equiv (s_t^1, \ldots, s_t^K)$. 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 Lemma B.1 hold. If π exhibits predictor neglect and there exists some $\theta \in \text{supp}(\pi)$ such that $P(r_t|s_t^1, \ldots, s^J, \theta) = P(r_t|s_t^1, \ldots, s^J, \theta^*)$ for all possible $(r, s^1, \ldots, s^J)$ under θ^*, then π is stable for all preferences.*

   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 for all preferences.

Next, we consider classes of models that are not stable for all preferences (i.e., they are unstable for some preferences).

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 $y \in Y(\pi)$.

[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.

Proposition B.4. *Suppose the assumptions underlying Lemma B.1 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 for all preferences.*

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 Lemma B.1 hold. If π is overly elaborate, then π is not stable for all preferences.*

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 $s_t \equiv (s_t^1, \ldots, s_t^K)$. Model π is *over-fit* if it has the following properties:

   (a) 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 $(s^{J+1}, \ldots, s^K) \neq (\tilde{s}^{J+1}, \ldots, \tilde{s}^K)$, $P(r|s, \theta^*) = P(r|\tilde{s}, \theta^*)$.

   (b) For all $\theta \in \text{supp}(\pi)$, $P(r_t|s_t, \theta)$ depends on both $(s_t^1, \ldots, s_t^J)$ and $(s_t^{J+1}, \ldots, s_t^K)$. That is, for all $s, \tilde{s} \in S$ such that $s \neq \tilde{s}$, $P(r|s, \theta) \neq P(r|\tilde{s}, \theta)$.

   (c) 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 $(r, s), (r, \tilde{s}) \in Y(\pi)$.

   To summarize, over-fit models are certain that some useless signals help predict outcomes (properties 1 and 2), yet exhibit some uncertainty about the extent to which they help (property 3).

   Proposition B.6. *Suppose the assumptions underlying Lemma B.1 hold. If π is over-fit, then π is not stable for all preferences.*

   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.

B.3 Choice Environments that Create Incentives to Track Statistics

Definition B.6. Fix a model π and an outcome environment $(\times_{t=1}^{\infty} Y_t, \Theta, P, \pi^*)$. Say that a choice environment $(\times_{t=1}^{\infty} X_t, \times_{t=1}^{\infty} u_t)$ *creates incentives for a person to keep track of what she's seen and not just what she's learned* when only keeping track of the minimal sufficient statistic for updating beliefs is not a SAS. Otherwise, say that the choice environment *creates incentives for a person to only keep track of what she's learned.*

A simple example of an environment that creates incentives for the person to track what she's seen is one that requires her to accurately report back h^t or statistics of h^t.

Corollary B.1. *Consider a model π and any outcome environment $(\times_{t=1}^{\infty} Y_t, \Theta, P, \pi^*)$. If π is stable for all preferences but not strongly so (i.e., it satisfies Definition 5 but not Definition 6), then (i) there exists a stable attentional strategy in every choice environment that creates incentives for a person to only keep track of what she's learned, but (ii) there is no stable attentional strategy in some choice environment that creates incentives for a person to keep track of what she's seen and not just what she's learned.*

Incentives for a person to track what she's seen and not just what she's learned may induce the discovery of a model that is stable for all stationary preferences (Definition 5): while such a model is stable across all stationary environments that meet Assumption 2, it need not be stable more broadly.

C Additional Applications

This section provides details for some of the applications discussed in main text: empirical misconceptions of asset patterns (Section 2.3), self control (Section 2.3), and redundancy neglect (Section 5.2).

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 2.3. 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.

More formally, 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 (BSV, 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 $V_t \equiv \sum_{k=0}^{\infty} M_{t+k}/(1+\rho)^k$ is the discounted value of all future earnings and $\mathbb{E}_t$ denotes expectations under π 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 $\mathbb{E}_t[V_t] \geq p_t$ 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 $\mathbb{E}_t[M_t]$.

[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 $(y_{t-1}, \ldots, y_1)$.

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 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 2, 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: $n^t_B = \{h^t \in H^t \mid \mathbb{E}_t[V_t] \geq p_t\}$ and $n^t_S = \{h^t \in H^t \mid \mathbb{E}_t[V_t] < p_t\}$. 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 2.3.

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 investigate when a person might come to realize that her self-control problem is more severe than she deemed plausible. As with all errors, 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 $1 - q$. 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 $\hat{q} \in (0, 1)$.

[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.

We consider a naive person with prior π over q such that $1 \notin \text{supp}(\pi)$.[52] 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 $m < T\big\{(1-\hat{q}) \cdot \mathbb{E}[b-c|b>c]\Pr(b>c) + \hat{q} \cdot \mathbb{E}[b-c|\beta b > c]\Pr(\beta b > c)\big\}$, or, equivalently, if

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

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

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

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., 2021).

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 gym less often than she thought possible. By Proposition 2, 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 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]

[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 $\sum_{k=1}^{t} \mathbf{1}\{\beta_k = 1\} = 0$, which becomes increasingly improbable under π.

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). In each period, the agent encounters a problem that has a solution dependent on a binary state that is i.i.d. over time. For instance, the state may be the optimal way to resolve a problem at work, and new problems crop up over time. Denote the state in period t by $\omega_t \in \{0,1\}$. Suppose ω_t is i.i.d. across periods with $P(\omega_t = 1) = 1/2$. In each period t, the agent receives signals about ω_t from 2 information sources, denoted $s_t = (s_t^1, s_t^2) \in \{0,1\}^2$. These signals, for instance, may be colleagues' advice on how to resolve a problem. Let $\theta_i \equiv P(s_t^i = \omega_t | \omega_t)$ denote the precision of signals 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., $\theta_i = \gamma$). 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 her subjective probability that $\omega_t = 1$ conditional on s_t.

Suppose the two information sources are in fact perfectly correlated: $s_t^1 = s_t^2$ for all t. For example, one colleague has good intuition (or access to good information), while the second colleague has none and simply mimics the first. We explore the stability of a misspecified model where the agent treats these two information sources as independent, thereby thinking she receives two informative signals each period instead of one. We also describe some 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 $\theta_c = 0$.

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 her information sources. For instance, a salesperson may encounter new clients each period and must decide how to pitch her product. She takes into account client-specific advice from her colleagues, but she may also need to learn which of them gives good advice. Or, for another example, a new professor teaches different lectures each day and learns over time which colleagues give good advice on how to lead those classes.

The second feature concerns the feedback available to the agent. 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 worker can readily assess whether her colleagues' advice was 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, a teacher only gets noisy feedback about whether classes go well or poorly—some of this uncertainty is resolved only when she receives course evaluations. 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.*

This result 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 her erroneous "independent-signals" model admits a stable attentional strategy if and only if she knows the precision of one or both information sources. If she is certain about one of the sources, then she can ignore feedback about it. Under this SAS there is no way to notice the correlation between sources. However, if she is uncertain about the precision of both, then tracking how often each gives good advice will cause her 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 she is dogmatic about the precision of both information sources. If she uncertain about the precision of at least one information source, then—unlike the case with perfect feedback—she will discover her error. This is because she 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 the worker is dogmatic that Colleague 1 has high precision yet is uncertain about the precision of Colleague 2. In this case, even without feedback on whether s_t^1 is correct, the mere event of agreement between colleagues would be good news about the quality of Colleague 2's advice.

D Proofs

D.1 Proofs of Results in the Main Text

Proof of Proposition 1. Consider a minimal SAS $(\mathscr{N}, \sigma)$ given π. Toward a contradiction, suppose that π is not attentionally measurable with respect to $(\mathscr{N}, \sigma)$. This implies that there exists a sample path h^t, $t \geq 2$ that occurs with positive probability under θ^* with the following property: there exists a finite $\tilde{t} \leq t$ such that $P(n^{\tilde{t}}(h^{\tilde{t}})|\theta) = 0$ for all $\theta \in \text{supp}(\pi)$, where $h^{\tilde{t}}$ is the history up to time $\tilde{t}$ consistent with h^t. Let τ be the smallest such $\tilde{t}$. Consider a modified noticing strategy $\widehat{\mathscr{N}} = (\widehat{N}^1, \widehat{N}^2, \dots)$ derived from $\mathscr{N}$ in the following way. First, let $\widehat{N}^k = N^k$ for all $k < \tau$ and all $k > \tau$. Second, since N^τ is a finite partition, enumerate its elements arbitrarily by $N^\tau = \{n_1^\tau, \dots, n_J^\tau\}$ for some $J \geq 1$. By the assumption above, there exits some element $n^\tau \in N^\tau$ such that $P(n^\tau|\theta) = 0\, \forall \theta \in \text{supp}(\pi)$. Since the enumeration of N^τ is arbitrary, label this element by n_J^τ. There must, however, exist some $n_i^\tau \in N^\tau$, $i \neq J$, such that $P(n_i^\tau|\theta) > 0$ for some $\theta \in \text{supp}(\pi)$. Let $\widehat{N}^\tau$ consist of $J-1$ elements, $\widehat{N}^\tau = \{\hat{n}_1^\tau, \dots, \hat{n}_{J-1}^\tau\}$, such that $\hat{n}_k^\tau = n_k^\tau$ if $k \neq i, J$ and $\hat{n}_i^\tau = n_i^\tau \cup n_J^\tau$. That is, $\widehat{N}^\tau$ is a coarsening of N^τ where the (subjectively) zero-probability cell, n_J^τ, is merged with a positive-probability cell, n_i^τ. The noticing strategy $\widehat{\mathscr{N}}$ is thus coarser than $\mathscr{N}$ and the attentional strategy $(\widehat{\mathscr{N}}, \sigma)$ is also sufficient given π, since altering how a person behaves in subjectively zero-probability situations does not impact his expected payoffs. Hence, $(\mathscr{N}, \sigma)$ is not minimal, yielding a contradiction. ■

Proof of Proposition 2. 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, let $X^*(h^t) \subseteq X_t$ denote the set of actions to which $\sigma_F(h^t)$ assigns positive probability. Define $\mathscr{N}$ such that for all t, $n^t(h^t) = \left\{\tilde{h}^t \in H^t \mid X^*(\tilde{h}^t) = X^*(h^t)\right\}$. Let σ be a behavioral strategy where, for all t, $\sigma_t(n^t(h^t))$ places probability one on $X^*(h^t)$; $\phi = (\mathscr{N}, \sigma)$ 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 set of optimal actions under π, and this set is distinct from the one prescribed by all $h^t \in \hat{n}^t$. Thus, $\tilde{\mathscr{N}}$ 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 3. Part 1. Suppose $\widetilde{\Theta} \subseteq \Theta^{\text{reject}}(\Gamma, \phi^{\min})$, where $\phi^{\min}$ is a minimal SAS. Suppose $\phi^{\min}$ is a stable attentional strategy given π, and hence $\liminf_{t\to\infty} P(n^t|\pi)/P(n^t|\pi^*) > 0$, where n^t are noticed histories under $\phi^{\min}$. Because $\text{supp}(\tilde{\pi}) = \text{supp}(\pi) \setminus \widetilde{\Theta}$ and because $\Theta^{\text{reject}}(\Gamma, \phi^{\min}) = \{\theta | \Pr(\theta|n^t) \to 0$ almost surely given π^*under SAS $\phi^{\min}\}$, it follows that long-run beliefs under $\phi^{\min}$ assign positive probability to all $\theta \in \text{supp}(\pi) \setminus \Theta^{\text{reject}}(\Gamma, \phi^{\min})$ when either starting with prior π or prior $\tilde{\pi}$. Thus, under model $\tilde{\pi}$, we have $\liminf_{t\to\infty} P(n^t|\tilde{\pi})/P(n^t|\pi^*) > 0$ given that, under model π, we have $\liminf_{t\to\infty} P(n^t|\pi)/P(n^t|\pi^*) > 0$. Furthermore, $\phi^{\min}$ is still a SAS given $\tilde{\pi}$ (albeit not necessarily minimal). It therefore follows that $\phi^{\min}$ is a stable attentional strategy given $\tilde{\pi}$. Turning to the converse, the text provides an example demonstrating that there can exist a stable attentional strategy given $\tilde{\pi}$ but not one given π.

Part 2. Now suppose $\widetilde{\Theta} \not\subseteq \Theta^{\text{reject}}(\Gamma, \phi^{\min})$. In contrast to the case above, long-run beliefs under $\phi^{\min}$ assign positive probability to all $\theta \in \text{supp}(\pi) \setminus \Theta^{\text{reject}}(\Gamma, \phi^{\min})$ when starting with prior π but not when starting with prior $\tilde{\pi}$. Thus, it is not necessarily so that $\liminf_{t\to\infty} P(n^t|\tilde{\pi})/P(n^t|\pi^*) > 0$ under $\phi^{\min}$. The text provides an example where there exists a stable attentional strategy given π but not one given $\tilde{\pi}$. ■

Proof of Proposition 4. To construct an outcome environment in which there exists an identifiably wrong model π that is strongly stable for all preferences, consider repeated flips of a coin where the outcome in each period is whether it lands heads (denoted by $y_t = 1$) or tails ($y_t = 0$). Suppose $y_t \in Y = \{0,1\}$ are i.i.d. with $\theta \in \Theta = [0,1]$ denoting the probability of $y_t = 1$, and let the true parameter be $\theta^* \in (0,1)$. Consider π that places probability one on $\theta = 1$. Then π is strongly stable for all preferences because the agent believes the setting is deterministic and hence finds it sufficient to ignore all data in any choice environment. To prove the second statement of the proposition, we prove its contrapositive. Suppose π places positive probability on all finite histories that occur with positive probability under π^*. Then π is not strongly stable for all preferences since π is attentionally unstable in environments where, in each period, the person is incentivized to accurately report the complete history up to that point. To be concrete, suppose she gets a strictly positive payoff for accurately reporting h^t and 0 otherwise. In this environment, π is attentionally unstable with respect to every SAS: since π places positive probability on every truly possible history, every SAS must be equivalent to the full attention SAS. Since π is attentionally unstable with respect to a full-attention SAS (because π is identifiably wrong), it then follows that π is attentionally unstable with respect to every SAS in this environment. ■

Proof of Proposition 5. We begin by stating and proving a simple lemma.

Lemma D.1. *Suppose Assumptions 1 and 2 hold. If with probability 1 under θ^* there exists some $\tilde{t} \in \mathbb{N}$ such that, for all $t > \tilde{t}$, the optimal action given π_t is independent of $\theta \in supp(\pi_t)$, then there exists a stable attentional strategy $(\mathscr{N}, \sigma)$ given π whether or not it is costly.*

Proof of Lemma D.1: Suppose that with probability 1 under θ^* there exists some $\tilde{t}$ such that for all $t > \tilde{t}$ the optimal action given π_t is independent of $\theta \in \text{supp}(\pi_t)$. Then there exists a SAS $(\mathscr{N}, \sigma)$ given π such that after $\tilde{t}$: (i) the noticed history $n^t(h^t)$ discards all information except possibly aspects of the current signal s_t, and (ii) the noticing strategy $\mathscr{N}$ lumps together any signal that is impossible under π_t with a signal that is possible under π_t. Under such a SAS, $P(n^t|\pi)/P(n^t|\lambda)$ is bounded away from 0 because the length of the noticed history in any $t > \tilde{t}$ is finite and essentially bounded by $\tilde{t}$. Hence, conditional on θ^*, a stable attentional strategy given π exists, completing the proof of Lemma D.1.

Proof of Part 1. Consider a stationary and binary action space $X = \{0,1\}$. Let $y^{t+1} = (y_t, \dots, y_1) \in \times_{k=1}^{t} Y_k$ denote the sequence of realized outcomes through period t. For each $t \in \mathbb{N}$, consider the history-dependent utility function u_t defined as follows:

$$u_t(x_t, y_t | h^t) = \begin{cases} \frac{\max_{\theta \in \text{supp}(\pi)} P(y^{t+1}|\theta)}{\max_{\theta \in \text{supp}(\pi^*)} P(y^{t+1}|\theta)} & \text{if} \quad \sum_{k=1}^{t} x_k = 0, \\ \frac{\max_{\theta \in \text{supp}(\pi^*)} P(y^{t+1}|\theta)}{\max_{\theta \in \text{supp}(\pi)} P(y^{t+1}|\theta)} & \text{if} \quad \sum_{k=1}^{t} x_k = t, \\ -1 & \text{if} \quad \sum_{k=1}^{t} x_k \notin \{0, t\}. \end{cases} \tag{D.1}$$

Given that π is unstable under full attention (since we assume π is identifiably wrong), the true parameter θ^* is such that $\theta^* \notin \text{supp}(\pi)$, and conditional on θ^*, $\lim_{t \to \infty} P(y^{t+1}|\theta)/P(y^{t+1}|\theta^*) = 0$ for all $\theta \in \text{supp}(\pi)$. Thus, according to model π, it is optimal to choose $x_t = 0$ for all t, and strictly so whenever $\text{supp}(\pi)$ is not a subset of $\text{supp}(\pi^*)$. As such, there exists a minimal SAS in which the person chooses $x_t = 0$ for all t and ignores all feedback (i.e., $N^t = \{H^t\}$ for all t). This SAS yields a stable attentional strategy given π (by Lemma D.1), and it is costly given that, under π^*, it is actually optimal to choose $x_t = 1$ for all t.

Proof of Part 2. Suppose π^* is absolutely continuous with respect to π (i.e., π is attentionally measurable under a full-attention SAS). Consider the choice environment where for all $t = 1, 2, \dots$, we have $X_t = H^t$ and $u_t(x_t|h^t) = 1$ if $x_t = h^t$ and $u_t(x_t|h^t) = -1$ otherwise. That is, the agent has incentive to accurately report the full history in each period.

Consider an arbitrary minimal SAS $\phi = (\mathscr{N}, \sigma)$ given π. Absolute continuity in this environment implies that noticing strategy $\mathscr{N} = (N^1, N^2, \dots)$ must distinguish all histories that arise with positive probability under θ^* since any such histories happen with positive probability under π as well.[54] Since the minimal SAS ϕ must precisely distinguish the true history h^t each period (i.e., $\tilde{h}^t \notin n^t(h^t)$ for all h^t and $\tilde{h}^t$ that occur with positive probability under π), the behavioral strategy $\sigma_t : N^t \to X_t$ is such that $x_t = h^t$ with probability 1. Hence, the person acts optimally each period and the SAS ϕ is thus costless.

Finally, we can show that π is attentionally unstable under the arbitrary minimal SAS ϕ. For all $\theta \in \pi$, $P(n^t(h^t)|\theta) = \sum_{\tilde{h}^t \in n^t(h^t)} P(\tilde{h}^t|\theta) = P(h^t|\theta)$, since sufficiency of ϕ implies that for any $\theta \in \text{supp}(\pi)$, $\tilde{h}^t \in n^t(h^t)$ where $\tilde{h}^t \neq h^t$ only if $\tilde{h}^t$ is assigned probability zero under θ. Thus, the

[54] That is, for all t, if h^t and $\tilde{h}^t \neq h^t$ happen with positive probability under θ^* (and hence under π), then $\tilde{h}^t \notin n^t(h^t)$. To see why, consider any period t and $h^t, \tilde{h}^t \in H^t$ such that $h^t \neq \tilde{h}^t$, $P(h^t|\theta^*) > 0$, and $P(\tilde{h}^t|\theta^*) > 0$. Toward a contradiction, suppose $\tilde{h}^t \in n^t(h^t)$. Sufficiency then requires $\max_{x \in X_t} \mathbb{E}_{(\pi,\sigma)}[u_t(x|h^t)|n^t(h^t)] = \max_{x \in X_t} \mathbb{E}_{(\pi,\sigma)}[u_t(x|h^t)|h^t]$ for all $h^t \in H^t$ that occur with positive probability under (π, σ). Note that this condition fails if both h^t and $\tilde{h}^t \in n^t(h^t)$ are assigned positive probability under π: in this case, $\max_{x \in X_t} \mathbb{E}_{(\pi,\sigma)}[u_t(x|h^t)|n^t(h^t)] < \max_{x \in X_t} \mathbb{E}_{(\pi,\sigma)}[u_t(x|h^t)|h^t]$. Thus, no cell of any N^t can contain more than one history assigned positive probability under π. Absolutely continuity further implies that for all t and any history $h^t \in H^t$ that can occur under θ^*, there exists $\theta \in \text{supp}(\pi)$ such that h^t has positive probability under θ. Hence, $P(h^t|\pi) > 0$ and $P(\tilde{h}^t|\pi) > 0$, and thus $\tilde{h}^t \in n^t(h^t)$ would imply a contradiction to our sufficiency assumption.

relevant Bayes' Factor for assessing attentional stability under ϕ is equivalent to the one for assessing stability under full attention: for all $h^t \in H^t$, $P(n^t(h^t)|\theta)/P(n^t(h^t)|\theta^*) = P(h^t|\theta)/P(h^t|\theta^*)$. Since π is attentionally unstable under full attention (because we assume π is identifiably wrong), $P(h^t|\theta)/P(h^t|\theta^*)$ converges to zero in t with positive probability and therefore $P(n^t(h^t)|\theta)/P(n^t(h^t)|\theta^*)$ does so as well. Thus, π is attentionally unstable under the arbitrary minimal SAS. ■

Proof of Proposition 6. Part 1. In the original environment, let $\phi = (\mathcal{N}, \sigma)$ denote the SAS derived in the proof of Proposition 2 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 $\tilde{\phi} = (\tilde{\mathcal{N}}, \tilde{\sigma})$ where $\tilde{\mathcal{N}}$ 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.

Part 2. Starting with the "if" direction, let $\phi = (\mathcal{N}, \sigma)$ be a stable attentional strategy given π in the original environment. In the original environment, the history each period looks like $h^t = (s_t, y_{t-1}, x_{t-1}, y_{t-2}, x_{t-2}, \ldots, y_1, x_1)$, while in the modified environment, it instead looks like

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

Let $\tilde{H}^t$ be the set of modified histories up to period t, and let $\tilde{h}^t_1$ denote the first component of $\tilde{h}^t$, which (for illustration) above is h^t. Now construct a noticing strategy for the modified environment, $\tilde{\mathcal{N}}$, such that

$$\tilde{n}^t(\tilde{h}^t) = \left\{\hat{h}^t \in \tilde{H}^t | \hat{h}^t_1 \in n^t(\tilde{h}^t_1)\right\} \; \forall\, t, \, \tilde{h}^t \in \tilde{H}^t.$$

Since $n^t(\tilde{h}^t_1) = n^t(h^t)$, the person notices the same data under $\tilde{\mathcal{N}}$ that she does under $\mathcal{N}$. Now derive $\tilde{\sigma}$ from σ in the obvious way and let $\tilde{\phi} = (\tilde{\mathcal{N}}, \tilde{\sigma})$. Note that $\tilde{\phi}$ leads to the same noticed information, beliefs, and behavior as ϕ. Since ϕ is a SAS given π, so is $\tilde{\phi}$. Moreover, since ϕ is a StAS given π in the original environment, $\tilde{\phi}$ is a StAS given π in the modified environment. The other direction is analogous and omitted. ■

Proof of Corollary 1. The result follows immediately from Proposition 3. ■

D.2 Proofs of Supplemental Results

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

Lemma D.2. *Assume Assumptions 1 and 2 hold, and that the environment is stationary. Enumerate the stationary outcome space Y arbitrarily by $Y = \{y_1, \dots, y_N\}$. Suppose the true parameter is θ^*, and for any $\theta, \theta' \in \Theta$ define*

$$\bar{Z}(\theta, \theta'|\theta^*) \equiv \prod_{n=1}^{N} \left(\frac{P(y_n|\theta)}{P(y_n|\theta')} \right)^{P(y_n|\theta^*)}. \tag{D.2}$$

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

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

where

$$Z_t \equiv \prod_{n=1}^{N} \left(\frac{P(y_n|\theta)}{P(y_n|\theta')} \right)^{k_n(y^t)/(t-1)}. \tag{D.4}$$

If $\bar{Z}(\theta, \theta'|\theta^*) < 1$, fix any $\widetilde{Z} \in \left(\bar{Z}(\theta, \theta'|\theta^*), 1\right)$. 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 $\limsup_{t\to\infty} (Z_t)^{t-1} \leq \lim_{t\to\infty} \left(\tilde{Z}\right)^{t-1} = 0$ 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.2 and D is the KL divergence defined in Equation B.1. Hence, the three conditions of Lemma D.2 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 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 $\tilde{\Theta} \subseteq \Theta$, $P(h^t|\tilde{\Theta}) = \sum_{\theta' \in \tilde{\Theta}} P(h^t|\theta')\pi(\theta'|\tilde{\Theta})$. We can then expand $P(h^t|\pi)$ as follows:

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

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

$$\frac{P(h^t|\pi)}{P(h^t|\lambda)} = \frac{P(h^t|\Theta_\pi^{\min}) \cdot \left[\pi(\Theta_\pi^{\min}) + \frac{P(h^t|\text{supp}(\pi)\setminus\Theta_\pi^{\min})}{P(h^t|\Theta_\pi^{\min})} \cdot (1 - \pi(\Theta_\pi^{\min}))\right]}{P(h^t|\Theta_\lambda^{\min}) \cdot \left[\pi(\Theta_\lambda^{\min}) + \frac{P(h^t|\text{supp}(\lambda)\setminus\Theta_\lambda^{\min})}{P(h^t|\Theta_\lambda^{\min})} \cdot (1 - \pi(\Theta_\lambda^{\min}))\right]}. \tag{D.5}$$

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.2 that $\frac{P(h^t|\theta)}{P(h^t|\theta')} \xrightarrow{\text{a.s.}} 0$, implying that $\frac{P(h^t|\text{supp}(\pi)\setminus\Theta_\pi^{\min})}{P(h^t|\Theta_\pi^{\min})} \xrightarrow{\text{a.s.}} 0$. Thus, (D.5) implies that

$$\frac{P(h^t|\pi)}{P(h^t|\lambda)} \xrightarrow{\text{a.s.}} \frac{P(h^t|\Theta_\pi^{\min})}{P(h^t|\Theta_\lambda^{\min})} \cdot \frac{\pi(\Theta_\pi^{\min})}{\pi(\Theta_\lambda^{\min})}. \tag{D.6}$$

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.2 that $\frac{P(h^t|\theta)}{P(h^t|\theta')} \xrightarrow{\text{a.s.}} \infty$, and thus $\frac{P(h^t|\Theta_\pi^{\min})}{P(h^t|\Theta_\lambda^{\min})} \xrightarrow{\text{a.s.}} \infty$. Therefore, D.6 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.2 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.2 implies that $\frac{P(h^t|\Theta_\pi^{\min})}{P(h^t|\Theta_\lambda^{\min})} \xrightarrow{\text{a.s.}} 0$, and thus π is attentionally unstable with respect to λ and a full-attention SAS. ■

Proof of Lemma B.1. We begin by proving a similar result under memory consistency (MC) and automatic recall (AR), as defined in Definitions A.1 and A.2.

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 for all preferences under memory consistency and automatic recall if and only if there exists $\theta \in \text{supp}(\pi)$*

such that

$$P(m_\pi(y)|\theta) \geq P(m_\pi(y)|\theta^*)\ \forall y \in Y(\pi). \tag{D.7}$$

Proof of Lemma D.3. ($\Leftarrow$) 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, $Y(\theta^*)$:

1. Suppose $Y(\pi) = Y(\theta^*)$. This implies that condition (D.7) 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 $k_n(y^t) \equiv \sum_{\tau=1}^{t-1} \mathbf{1}\{y_\tau \in m_\pi^n\}$ denote the count of entries in y^t that are in m_π^n. Then for any $\theta \in \text{supp}(\pi)$,

$$\frac{P(n^t(y^t)|\theta)}{P(n^t(y^t)|\theta^*)} = \frac{\prod_{n=1}^N P(m_\pi^n|\theta)^{k_n(y^t)}}{\prod_{n=1}^N P(m_\pi^n|\theta^*)^{k_n(y^t)}} = \left(\frac{\prod_{n=1}^N P(m_\pi^n|\theta)^{k_n(y^t)/(t-1)}}{\prod_{n=1}^N P(m_\pi^n|\theta^*)^{k_n(y^t)/(t-1)}}\right)^{t-1}. \tag{D.8}$$

   This likelihood ratio is identical to the one considered in Lemma D.2 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.8) 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,

$$n^t(y^t) = \{\tilde{y}^t \in Y(\pi)^t \mid m_\pi(\tilde{y}_\tau) = m_\pi(y_\tau)\ \forall \tau = 1,\ldots,t-1\}. \tag{D.9}$$

   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.

2. Suppose $Y(\pi) \neq Y(\theta^*)$. Thus, condition (D.7) 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 $Y^0 \equiv Y(\theta^*) \setminus Y(\pi)$, is non-empty. Let $P^0 \equiv \sum_{y \in Y^0} P(y|\theta^*)$. 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 $P\left(\tilde{m}_\pi^{(n,t)}|\theta\right) = P\left(\tilde{m}_\pi^{(n,t)}|\theta^*\right)$ $\forall n = 1, \ldots, N$ 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

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

In other words, 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 $t \in \mathbb{N}$, $P\left(\tilde{m}_\pi^{(n,t)}|\theta^*\right) = P(m_\pi^n|\theta^*) + P^0 \cdot P(z_t = n) = P(m_\pi^n|\theta) = P\left(\tilde{m}_\pi^{(n,t)}|\theta\right)$, 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.

($\Rightarrow$) Suppose π is stable for all preferences under MC and AR. Enumerate $Y(\pi) = \{y_1, \ldots, y_N\}$ and consider the action space $X = [0,1]^N$ along with utility function $u(x,y) = -\sum_{n=1}^N (x_n - \mathbf{1}\{y = y_n\})^2$. We first show that under (X,u), any SAS requires that the person notices at least the information contained in $m_\pi(y_t)$ each period since this is a minimal sufficient statistic (see, e.g., Lehmann and Casella, 1998). To establish this, we show that the person's optimal action after noticing $y \in m_\pi$ differs from the optimal action following any $y' \in m'_\pi$ where $m'_\pi \neq m_\pi$. The optimal action under (X,u) is such that $x_n = \sum_{\theta \in \text{supp}(\pi)} P(y_n|\theta)\pi_t(\theta)$. First, if $m_\pi(y) = Y(\pi)$ $\forall y$, then we are trivially done. If there exists $y \in Y(\pi)$ such that $m_\pi(y) \neq Y(\pi)$, then it suffices to show the following: $y' \notin m_\pi(y) \Rightarrow \sum_{\theta \in \text{supp}(\pi)} P(y|\theta)\pi(\theta|y) \neq \sum_{\theta \in \text{supp}(\pi)} P(y|\theta)\pi(\theta|y')$, where $\pi(\theta|y)$ is the posterior probability of θ following outcome y given prior $\pi(\theta)$. Note that $\sum_{\theta \in \text{supp}(\pi)} P(y|\theta)\pi(\theta|y) \neq$

$\sum_{\theta \in \text{supp}(\pi)} P(y|\theta)\pi(\theta|y') \Leftrightarrow \sum_{\theta \in \text{supp}(\pi)} P(y|\theta)[\pi(\theta|y) - \pi(\theta|y')] \neq 0$. Further,

$$\begin{aligned}
\sum_{\theta \in \text{supp}(\pi)} P(y|\theta)[\pi(\theta|y) - \pi(\theta|y')] &= \sum_{\theta \in \text{supp}(\pi)} P(y|\theta) \left[\frac{P(y|\theta)\pi(\theta)}{P(y)} - \frac{P(y'|\theta)\pi(\theta)}{P(y')} \right] \\
&\propto \sum_{\theta \in \text{supp}(\pi)} \pi(\theta) P(y|\theta) \left[P(y|\theta)P(y') - P(y'|\theta)P(y) \right] \\
&= \sum_{\theta \in \text{supp}(\pi)} \pi(\theta) \left[P(y|\theta)^2 P(y') - P(y|\theta)P(y'|\theta)P(y) \right] \\
&= \mathbb{E}_\theta \left[P(y|\theta)^2 P(y') - P(y|\theta)P(y'|\theta)P(y) \right] \\
&> \mathbb{E}_\theta \left[P(y)^2 P(y') - P(y'|\theta)P(y)^2 \right] = 0,
\end{aligned}$$

where the strict inequality follows from Jensen's inequality given that $y' \notin m_\pi(y)$ and hence $P(y'|\theta)/P(y|\theta)$ depends on θ. Thus, any π that is stable for all preferences (under MC and AR) must be part of a stable attentional strategy with a SAS that distinguishes $m_\pi(y_t)$ each round (as in Equation D.9).

To finally establish that condition (D.7) must hold, we proceed by contradiction: suppose condition (D.7) does not hold, so for any $\theta \in \text{supp}(\pi)$, there exists $y \in Y(\pi)$ such that $P(m_\pi(y)|\theta) < P(m_\pi(y)|\theta^*)$. Under a SAS where the person records each instance of $m_\pi(y)$, the predicted distribution over noticed outcomes for each t and $\theta \in \text{supp}(\pi)$ will differ from the true distribution in the limit. As such, the KL distance between these distributions is positive and $P(n^t|\theta)/P(n^t|\theta^*) \xrightarrow{\text{a.s.}} 0$ by Remark D.1. Thus π is not stable for all preferences (under MC and AR), yielding a contradiction.

Completing the Proof of Lemma B.1. Unlike Lemma D.3 above, Lemma B.1 does not assume memory consistency and automatic recall. Thus, we now extend the proof above by dropping these two assumptions.

($\Leftarrow$) For any (X,u), the person believes it is sufficient to notice $m_\pi(y^t)$ and 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 $m_\pi(y^t)$ and s_t constitutes the noticing strategy for some SAS. 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 for all preferences.

($\Rightarrow$) Suppose π is stable for all preferences. The proof follows along the same lines as the analogous result assuming MC and AR, above. First, we show that there exist (X,u) under which any SAS requires the person to notice $m_\pi(y^t)$ (and s_t) for all t. As above, consider $X = [0,1]^N$ along with utility function $u(x,y) = -\sum_{n=1}^N (x_n - \mathbf{1}\{y = y_n\})^2$. Analogous to the proof under MC and AR, the person's optimal action after noticing $\tilde{y}^t \in m_\pi(y^t)$ differs from the optimal action following any $\tilde{y}^t \notin m_\pi(y^t)$ and hence the person must distinguish any $m_\pi(y^t)$ from $m_\pi(\tilde{y}^t) \neq m_\pi(y^t)$. Given this result, if π is stable for all preferences, then by definition there exists $\theta \in \text{supp}(\pi)$ and a SAS such that $\liminf_{t\to\infty} P(n^t|\theta)/P(n^t|\theta^*) > 0$ with probability 1 under θ^*, which in turn implies $\liminf_{t\to\infty} P(m_\pi(y^t)|\theta)/P(m_\pi(y^t)|\theta^*) > 0$ with probability 1 under θ^* since n^t must contain at least as much information as $m_\pi(y^t)$. ■

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 Lemma B.1 that π is stable for all preferences. ■

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 preferences (with memory consistency and automatic recall) from the proof of Lemma B.1 holds (Condition D.7). If π is stable for all preferences with memory consistency and automatic recall, then it is stable for all preferences more generally. ■

Proof of Proposition B.3. Assume π exhibits predictor neglect and 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^*. Any SAS must distinguish $y = (r,s^1,\ldots,s^K)$ from $\tilde{y} = (\tilde{r},\tilde{s}^1,\ldots,\tilde{s}^K)$ only if $(r,s^1,\ldots,s^J) \neq (\tilde{r},\tilde{s}^1,\ldots,\tilde{s}^J)$. Let N be the number of distinct values of $(r,s^1,\ldots,s^J)$ under π. Then for each $n = 1,\ldots,N$, $m_\pi(y^t)$ must record the count $k_n(y^t)$ of outcomes y_τ, $\tau < t$, such that $y_\tau \in m^n_\pi$. Then $\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)}$ is identical to (D.11) from the proof of Proposition B.5, below. It then follows from that proof that π is stable for all preferences if $P(m^n_\pi|\theta^*) = P(m^n_\pi|\theta)$ 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 $P(m^n_\pi|\theta^*) = P(m^n_\pi|\theta)$ for all $n = 1,\ldots,N$, so π is stable for all preferences. ■

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 $k_n(y^t)$). Then

$$\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)} = \frac{\prod_{n=1}^N P(y_n|\theta)^{k_n(y^t)}}{\prod_{n=1}^N P(y_n|\theta^*)^{k_n(y^t)}} = \left(\frac{\prod_{n=1}^N P(y_n|\theta)^{k_n(y^t)/(t-1)}}{\prod_{n=1}^N P(y_n|\theta^*)^{k_n(y^t)/(t-1)}}\right)^{t-1}. \tag{D.10}$$

Hence, Lemma D.2 along with (D.10) 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 for all preferences. ■

Proof of Proposition B.5. Let m_π be the partition of $Y(\pi)$ defined in (B.2). 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 $k_n(y^t)$). Then

$$\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)} = \frac{\prod_{n=1}^N P(m_\pi^n|\theta)^{k_n(y^t)}}{\prod_{n=1}^N P(m_\pi^n|\theta^*)^{k_n(y^t)}} = \left(\frac{\prod_{n=1}^N P(m_\pi^n|\theta)^{k_n(y^t)/(t-1)}}{\prod_{n=1}^N P(m_\pi^n|\theta^*)^{k_n(y^t)/(t-1)}}\right)^{t-1}. \tag{D.11}$$

The likelihood ratio (D.11) is effectively identical to the one considered in the proof of Lemma B.1 (Equation D.8). 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.2 and Remark D.1 that ratio (D.11) converges to 0 a.s. and π is therefore not stable for all preferences. ■

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 (B.2) 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.11) 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 $P_m(\cdot|\theta) = P_m(\cdot|\theta^*)$. We now show that the previous equality is violated for any $\theta \in \text{supp}(\pi)$ when π is over-fit: since π is over-fit, there exists $s, \tilde{s} \in S$ 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 $(r,s),(r,\tilde{s}) \in Y(\pi)$. For all $\theta \in \text{supp}(\pi)$, $P(r|s,\theta) \neq P(r|\tilde{s},\theta)$, but $P(r|s,\theta^*) = P(r|\tilde{s},\theta^*)$. This implies that, for each $\theta \in \text{supp}(\pi)$, one of the following inequalities must hold: $P(r|s,\theta) \neq P(r|s,\theta^*)$ or $P(r|\tilde{s},\theta) \neq P(r|\tilde{s},\theta^*)$. Consider an arbitrary $\theta \in \text{supp}(\pi)$, and suppose WLOG that the first or the two previous inequalities holds: $P(r|s,\theta) \neq P(r|s,\theta^*)$. Since $P(s)$ is independent of the parameter value (by assumption), the previous inequality implies that $P((r,s)|\theta) \neq P((r,s)|\theta^*)$, and therefore $P_m(m_\pi((r,s))|\theta) \neq P_m(m_\pi((r,s))|\theta^*)$, violating the above equality necessary for π to be stable for all preferences. ■

Proof of Corollary B.1. Part (i) is immediate from Lemma B.1's characterization of models that are stable for all preferences. Part (ii) follows from the definition of not being strongly stable for all preferences. ■

Proof of Proposition C.1. For the analyst who does not need to provide predictions, a minimal SAS contains at most two elements, $n_B^t = \{h^t \in H^t \mid \mathbb{E}_t[V_t] \geq p_t\}$ and $n_S^t = \{h^t \in H^t \mid \mathbb{E}_t[V_t] < p_t\}$. Thus, the analyst must distinguish only whether $\mathbb{E}_t[V_t]$ is high enough to justify buying at the current price. The analyst who does need to provide predictions, on the other hand, needs to additionally notice $\mathbb{E}_t[M_{t+1}]$. 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 $\bar{q} = \max \operatorname{supp}(\pi)$. 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 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 his 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 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 $n_B^t \equiv \{h^t \in H^t \mid \mathbb{E}_t[V(\hat{q})] \geq m\}$, which contains all histories following which it's optimal to buy the contract, and $n_R^t \equiv \{h^t \in H^t \mid \mathbb{E}_t[V(\hat{q})] < m\}$, which contains all histories following which it's optimal to reject it. Note that $\lim_{t\to\infty} P(n_B^t|\pi)/P(n_B^t|\pi^*) > 0$ and $\lim_{t\to\infty} P(n_R^t|\pi)/P(n_R^t|\pi^*) > 0$, since $P(n_x^t|\pi)$ is bounded away from 0 as $t \to \infty$ given the assumption that Condition C.1 holds for some $\hat{q} \in \text{supp}(\pi)$ but not all. ■

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 $s_t = (s_t^1, s_t^2)$. 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 $s_t = (s_t^1, s_t^2)$; (ii) ignore all past signals from source 2; and (iii) distinguish the value of $W_t^1 \equiv \sum_{k=1}^{t-1} \mathbf{1}\{s_k^1 = \omega_k\}$, which is sufficient for updating beliefs about θ_1. Thus:

$$\begin{aligned}
\frac{P(n^t(h^t)|\pi)}{P(n^t(h^t)|\pi^*)} &= \frac{P\big(W_t^1\big|\pi\big)}{P\big(W_t^1\big|\theta_1^*\big)} \frac{P\big(s_t\big|W_t^1,\pi\big)}{P\big(s_t\big|\pi^*\big)} \\
&= \left(\sum_{\theta_1 \in \{\tilde{\theta}_1, \theta_1^*\}} \pi(\theta_1) \frac{P\big(W_t^1\big|\theta_1\big)}{P\big(W_t^1\big|\theta_1^*\big)} \right) \frac{P\big(s_t\big|W_t^1,\pi\big)}{P\big(s_t\big|\pi^*\big)} \\
&= \left(\pi(\theta_1^*) + \pi(\tilde{\theta}_1) \left(\frac{\tilde{\theta}_1}{\theta_1^*}\right)^{W_t^1} \left(\frac{1-\tilde{\theta}_1}{1-\theta_1^*}\right)^{t-1-W_t^1} \right) \frac{P\big(s_t\big|W_t^1,\pi\big)}{P\big(s_t\big|\pi^*\big)}. \qquad \text{(D.12)}
\end{aligned}$$

Since in truth $W_t^1 \sim \text{Binomial}(t-1, \theta_1^*)$, $\lim_{t\to\infty} \left(\frac{\tilde{\theta}_1}{\theta_1^*}\right)^{W_t^1} \left(\frac{1-\tilde{\theta}_1}{1-\theta_1^*}\right)^{t-1-W_t^1} = 0$.[55] Hence,

[55] 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 $L(t, W_t, |\theta_1^*, \tilde{\theta}_1)^{t-1}$, where $L(t, W_t, |\theta_1^*, \tilde{\theta}_1) \equiv$

$\liminf_{t\to\infty} P(n^t(h^t)|\pi)/P(n^t(h^t)|\pi^*) > 0$ 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. (To see this, note that the count of "successes" is a minimal sufficient statistic for updating about the parameter governing a binomial distribution; see Footnote 47.) Thus, following the logic of (D.12) 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 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 $a_t \equiv \mathbf{1}\{s_t^1 = s_t^2\}$ be an indicator for agreement in period t and let $Q_t \equiv \sum_{k=1}^t a_k$ count the number of agreements through round t.[56] Assuming the person notices Q_t (and s_t) each

$\left(\frac{\tilde\theta}{\theta^*}\right)^{W_t^1/(t-1)}\left(\frac{1-\tilde\theta_1}{1-\theta_1^*}\right)^{(t-1-W_t^1)/(t-1)}$. Note that $\lim_{t\to\infty} L(t,W_t|\theta_1^*,\tilde\theta_1) = \left(\frac{\tilde\theta}{\theta^*}\right)^{\theta_1^*}\left(\frac{1-\tilde\theta_1}{1-\theta_1^*}\right)^{1-\theta_1^*} \equiv \bar L(\tilde\theta_1|\theta_1^*)$. 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 $\lim_{t\to\infty} L(t,W_t,|\theta_1^*,\tilde\theta_1)^{t-1} = \lim_{t\to\infty} \bar L(\tilde\theta_1|\theta^*)^{t-1} = 0$.

[56] 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.

period, then π is attentionally unstable:

$$\begin{aligned}\frac{P(n^t(h^t)|\pi)}{P(n^t(h^t)|\pi^*)} &= \frac{P\big(Q_t\big|\pi\big)}{P\big(Q_t\big|\pi^*\big)}\frac{P\big(s_t\big|Q_t,\pi\big)}{P\big(s_t\big|\pi^*\big)} \\ &= P\big(Q_t\big|\pi\big)\frac{P\big(s_t\big|Q_t,\pi\big)}{P\big(s_t\big|\pi^*\big)}, \end{aligned} \tag{D.13}$$

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.13) 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

$$x_t^* = \sum_{(\theta_1,\theta_2)} \pi_t(\theta_1,\theta_2)P(\omega_t = 1|s_t,\theta_1,\theta_2). \tag{D.14}$$

We consider what minimal data the person finds sufficient to form the beliefs $\pi_t(\theta_1,\theta_2)$ that determine (D.14). 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

$$\pi_t(\theta_1,\theta_2|s_t) = \frac{P(s_t|\theta_1,\theta_2)\pi_{t-1}(\theta_1,\theta_2)}{\sum_{(\tilde{\theta}_1,\tilde{\theta}_2)} P(s_t|\tilde{\theta}_1,\tilde{\theta}_2)\pi_{t-1}(\tilde{\theta}_1,\tilde{\theta}_2)}. \tag{D.15}$$

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 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,

$$\pi_t(\gamma,\gamma|h^t) = \frac{[1-2\gamma(1-\gamma)]^{Q_t}[2\gamma(1-\gamma)]^{t-Q_t}\pi(\gamma,\gamma)}{[1-2\gamma(1-\gamma)]^{Q_t}[2\gamma(1-\gamma)]^{t-Q_t}\pi(\gamma,\gamma) + [0.5]^t(1-\pi(\gamma,\gamma))}, \tag{D.16}$$

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

$$\pi_t(\theta_1, \theta_2 | h^t) = \frac{[0.5]^t \pi(\theta_1, \theta_2)}{[1 - 2\gamma(1-\gamma)]^{Q_t} [2\gamma(1-\gamma)]^{t-Q_t} \pi(\gamma, \gamma) + [0.5]^t (1 - \pi(\gamma, \gamma))}. \tag{D.17}$$

From Equations (D.16) and (D.17), 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.13), 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.17) 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. ■