Channeled Attention and Stable Errors

Tristan Gagnon-Bartsch
Harvard

Matthew Rabin
Harvard

Joshua Schwartzstein*
Harvard Business School

PRELIMINARY

This version: May 25, 2018

Abstract

A common critique of models of mistaken beliefs is that people should recognize their error after observations they thought were unlikely. This paper develops a framework for assessing when a given error is likely to be discovered, in the sense that the error-maker will deem her mistaken theory implausible. The central premise of our approach is that people channel their attention through the lens of their mistaken theory, meaning a person may ignore or discard information her mistaken theory leads her to consider unimportant. We propose solution concepts embedding such channeled attention that predict when a mistaken theory will persist in the long run even with negligible costs of attention, and we use this framework to study the "attentional stability" of common errors and psychological biases. While many costly errors are prone to persist, in some situations a person will recognize her mistakes via "incidental learning": when the data she values given her mistaken theory happen to also tell her how unlikely her theory is. We investigate which combinations of errors, situations, and preferences tend to induce such incidental learning vs. factors that render erroneous beliefs stable. We show, for example, that a person may never realize her self-control problem even when it leads to damaging behavior, and may never notice the correlation in others' advice even when that failure leads her to follow repetitive advice too much. More generally, we show that for every error there exists an environment where the error persists and is costly. Uncertainty about the optimal action paves the way for incidental learning, while being dogmatic creates a barrier.

*E-mails: gagnonbartsch@fas.harvard.edu, matthewrabin@fas.harvard.edu, and jschwartzstein@hbs.edu. For helpful comments, we are grateful to Nava Ashraf, Max Bazerman, Erik Eyster, Nicola Gennaioli, Botond Kőszegi, George Loewenstein, Andrei Shleifer, and seminar participants at BEAM, Boston College, CEU, Cornell, Dartmouth, Harvard, LSE, McGill, Princeton, Stanford, and UC Berkeley.

1 Introduction

A recent explosion of models explore the economic implications of ways that people misunderstand the world. These models span a wide range of mistakes—including simple empirical misconceptions, bad statistical reasoning, faulty social inference, and distorted beliefs about one's personal traits. A common critique of such models is that people should realize their errors after observing events they think are unlikely or impossible. Someone who suffers from the gambler's fallacy, for instance, will see more streaks than she expects. A person who does not think through that others are following the crowd will see an unexpected degree of consensus. And a person who is naive about her self-control problems will indulge herself more often than she anticipates. It seems that error-makers should notice that something is amiss. Won't people get a clue?

This paper develops "solution concepts" to assess when people making specific mistakes in particular situations will eventually notice their errors. These concepts rest on two central criteria. First, we clarify that the absolute unlikeness of an observation wouldn't (and shouldn't) induce a person to abandon an erroneous theory. Rather, a person gets a clue only when an observation is far more likely under a compelling alternative theory that she considers.

Our primary emphasis is a second criterion: building on ideas in Sims (2003), Woodford (2012), Gabaix (2014) and, primarily, Schwartzstein (2014), we examine the learning process of a person who may not notice information that seems to her irrelevant. Attention and memory do not act like cameras that faithfully record all we see. Rather, we attend to and remember a small subset of available information, and we direct this attention toward information we think is task relevant. Our "narrow channel of consciousness" (Chater 2018) is surprisingly effective at blocking out information we're not looking out for. Indeed, while only 20 percent of memory researchers agree with the statement that "people generally notice when something unexpected enters their field of view, even when they're paying attention to something else," 80 percent of the public think this statement is true (Simons and Chabris 2011).[1]

There are vivid and more mundane examples of our failure to notice unexpected events. People tasked with counting the number of passes in a film clip of basketball players often fail to notice a gorilla walking across the court (Simons and Chabris 1999). Indonesian seaweed farmers seem to persistently fail to optimize along a dimension (pod size) they wrongly treat as irrelevant despite being exposed to rich data from which they could learn if they paid attention (Hanna, Mullainathan, and Schwartzstein 2014). And we assess very few of all the possible correlations among variables

[1] Likewise, Simons and Chabris (2011) find that 100 percent of memory researchers disagree with the statement that "human memory works like a video camera, accurately recording the events we see and hear so that we can review and inspect them later" while only 40 percent of the public disagree with it.

that we see day to day, looking for them only when we think we have something useful to learn.[2,3]

We define solution concepts embedding such "channeled attention" that predict when a person's mistaken theory might survive after infinite data, where data is encoded and analyzed *if* and (in most of our analysis) *only if* it is perceived as valuable within that theory. We use this framework to analyze the "attentional stability" of common errors and psychological biases in given situations. In many cases, a person won't discover an error because her misspecified model leads her to ignore the very data necessary to prove that she is wrong.

Our framework jibes with recent theories of rational inattention (e.g., Sims 2003; Gabaix 2014) in the sense that people in our model are more likely to attend to information that they think is valuable. But theories of rational inattention differ by assuming rational expectations: what people think is valuable is, on average, in fact valuable. People who are making errors about the world are likely to misestimate the value of potential information, however. A person overconfident about her self control may see little value in carefully tracking her spending habits. A person who neglects redundancy in advice and treats it as independent does not ask whether it is in fact correlated. Our framework for "subjectively rational inattention" highlights how erroneous priors about the value of information can lead people astray in their attentional strategies.

Attention in our framework is then not only *limited*, it is more generally *misguided*: along with ignoring information she thinks has no value, she may also pay careful attention to information that is, in fact, useless. Consumers may see illusory patterns indicating that ineffective remedies can ward off illness or that branded headache medicines more effectively reduce pain than generics. Investors may pay too much attention to past mutual fund returns to predict future success. Workers may infer too much from the redundant advice of several colleagues when it would be better to follow a single one. In the language of Handel and Schwartzstein (2018), the major stumbling block to processing truly important information in our model is not a "friction" that causes people to view processing as too costly, but rather a "mental gap" that leads people to assign the wrong value to processing.

A person in our model discovers her mistakes only via *incidental learning*; that is, when infor-

[2]We assume few readers of this article (besides Capricorns, who tend to be gullible) believe that the newspaper's daily horoscope holds any information for us. Our skepticism about the goodness of fit between days where we are told "a friend will disappoint you today" and those where a friend actually disappoints us might lead us to ignore the relationship, even if we ruminate on every instance in which a friend has disappointed us and always read our horoscope out of our partner's insistence. (Likewise, those who believe in horoscopes may read them every day without noticing a lack of correlation.) Given our theory that the horoscope is almost surely meaningless, we would never discover if it turned out to be right. If there are people who think we are foolish for ignoring newspaper horoscopes, they may be surprised that we don't believe them, but they shouldn't be surprised that we haven't carefully noticed the correlations.

[3]Research such as Bordalo, Gennaioli, and Shleifer (2012, 2013, 2017) have channeled economists' attention to the intuition that unexpected or surprising events may receive disproportionate weight in decisions. We emphasize that unexpected events we're not looking out for often go unnoticed.

mation she deems useful given her bad theory happens to also tell her how unlikely her theory is. Our framework yields several predictions about what does and does not get people to wake up. First, since people are not on the lookout for their mistakes, errors may persist even when they are very damaging: We show that for every error, there exists an environment where that error persists and is (arbitrarily) costly. Second, a sense of certainty prevents getting a clue: People who think they have nothing to learn need not notice anything; coarse models that ignore predictive signals are more stable than overly elaborate models that assume signals are predictive even when they are not. Third, a person is less likely to get a clue when the environment makes it easier to make decisions without paying detailed attention: Being able to delegate decisions or make them by querying a database creates barriers to incidental learning.

Section 2 provides an extended example about incorrect medical beliefs that introduces some basic concepts of our model. Section 3 then begins formalizing this framework. Although our principles could be applied to a broader class of errors, we focus on "quasi-Bayesian" models: the agent is Bayesian with a misspecified prior π over a set of parameters, where π either assigns positive probability to impossible parameters or zero probability to possible parameters.[4] We impose no further assumptions about which errors people make *ex ante*—they could arise, for instance, from motivated reasoning, cognitive biases, or bad theories of the world. Our goal then is to provide a general framework for assessing when people are prone to discover such an error, taking the misspecified model as a primitive. To this end, we formulate in Section 3 a baseline criterion for when a person deems her misspecified prior "inexplicable"under full attention, meaning that she eventually finds π an implausible explanation for what she observes. Implausibility is assessed relative to an alternative "light-bulb model" λ—a model that people might entertain when doubting their prior conception of the world. More specifically, we say that π is "inexplicable" with respect

[4]Many recent models, spanning a wide range of errors, are either quasi-Bayesian or very close. Examples include Barberis, Shleifer, Vishny (1998) on stock-market misperceptions, Rabin (2002) and Rabin and Vayanos (2010) on the gambler's and hot-hand fallacies, Benjamin, Rabin, and Raymond (2016) on the non-belief in the law of large numbers, and Spiegler (2016) on biases in causal reasoning. Examples of biases about misreading information include Rabin and Schrag (1999) and Fryer, Harms, and Jackson (2018) on confirmation bias and Mullainathan (2002) on naivete about limited memory. Models of coarse or categorical thinking include Mullainathan (2001), Fryer and Jackson (2008), Jehiel (2005), Jehiel and Koessler (2008), Mullainathan, Schwartzstein, and Shleifer (2008), and Eyster and Piccione (2013). Relatedly, models that incorporate errors in reasoning about the informational content of others' behavior include Eyster and Rabin (2005), Esponda (2008), and Madarász (2012). Such errors have been explored in social-learning settings by DeMarzo, Vayanos, Zweibel (2003), Eyster and Rabin (2010), Eyster and Rabin (2014), Bohren (2016), and Gagnon-Bartsch and Rabin (2017). Models that assume false beliefs about others' strategic reasoning or information include Camerer, Ho, and Chong (2004) and Crawford and Iriberri (2007). Misspecified models have also been considered in specific applications, such as firms learning about demand (Kirman 1975; Nyarko 1991) as well as 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 self-control, wherein people believe they will have more self-control in the future than they truly will, falls within this broader category. The model of projection bias in Loewenstein, O'Donoghue and Rabin (2003) likewise posits that somebody may have systematically different beliefs about future tastes as a function of fluctuating contemporaneous tastes.

to λ if observations are "sufficiently" more likely under the light-bulb than her prior. Otherwise, π is explicable with respect to λ. We primarily take the light-bulb model λ to be the true model of the world.

In Section 4, we turn to our primary focus: the role of channeled attention. We assume the agent may *rationally*—from the perspective of her wrong theory—ignore data she considers irrelevant to her payoffs. Our analysis of channeled attention revolves around what we call a "sufficient attentional strategy" (SAS), which specifies what the person notices from the available data. A SAS must satisfy two conditions: First, a person notices any data that she believes is useful for making a decision now or in the future. Second, a person cannot today notice something that happened yesterday unless she noticed it yesterday; that is, one can only remember data she noticed in the first place.[5] Although a SAS does not rule out attention to seemingly useless information, we often focus on "minimal" SAS's where the person notices no more than she finds useful.[6]

To tractably capture the role of a misperceived value of information, we assume that people pay attention to some data if (and often only if) they perceive *any* benefit of attention. This roughly corresponds to limits where attentional costs are vanishingly small (but not-zero) and where people are arbitrarily patient. Our predictions about when people fail to get a clue are therefore conservative: if a misspecified model is stable in our framework, then it would remain stable after incorporating realistic attentional costs.

Relative to Schwartzstein's (2014) stylized setting, where a person channels her attention based on wrongly thinking some variable is unimportant, our "SAS" approach is less restrictive: when following a SAS, the agent recalls a coarsening of the true history each period that she deems sufficient for any future decision. This allows us to apply the framework to a wide range of erroneous models, and captures the important feature that one can pay attention to aspects of variables without noting everything relevant about them. For instance, consider a manager who wants to learn a worker's success rate at a given task and assumes that a worker's performance rate is constant over time. The manager will find it sufficient to notice the worker's historical success frequency and not, for example, any temporal trends in successes and failures. A gym-goer who is naive about his self-control problem finds it sufficient to simply notice whether he wants to skip the gym on any given day without further noticing if this because he is particularly busy or just lazy.

[5]We allow a person to stop noticing and forget information once she deems it no longer useful. We call this *volitional recall* to contrast with more familiar assumptions of imperfect recall that impose exogeneous imperfections in memory. We consider *automatic recall*—where a person forever recalls and notices all data previously noticed—in the appendix.

[6]We show that when the person follows a minimal SAS, she never notices events that her theory deems impossible. Researchers formulating quasi-Bayesian models typically (openly) impose structure such that observables an agent would deem impossible observations cannot happen in order to preserve coherent Bayesian updating within the agent's misspecified model. While tempting to think that such observations would force the agent to reconsider her misspecified model, channeled attention in fact limits the impact of subjectively zero-probability events on the discovery of errors.

We say (roughly) that model π is "attentionally inexplicable" relative to an alternative λ and a given SAS if, over an infinite period of time, the *noticed* data becomes infinitely more likely under λ than π; otherwise, it is "attentionally explicable". When a misspecified model is attentionally explicable relative to the truth, we call the SAS used an "attentionally stable equilibrium" given π. Often there exist attentionally stable equilibria for models that are inexplicable with full attention. This is not simply because inattention leads such data to be lost from memory—indeed, we show that being able to freely look back at all previous data (e.g., it is recorded somewhere) does not lead people to wake up. A person who channels his attention solely on the data his model deems relevant will often fail to notice data that screams his model is wrong.[7]

Before a more general analysis, in Section 5 we examine errors familiar from existing research. We first analyze when and how channeled attention can lead to persistent underestimation of self-control problems.[8] Some researchers in fact suggest that rational learning should correct these problems and Ali (2011) argues that the only form of incorrect beliefs likely to survive is exaggeration of our self-control problems. We argue that this path to sophistication is not a foregone conclusion—overestimation of self control is often part of an attentionally stable equilibrium—which helps explain the empirical reality that people are not fully sophisticated (e.g., Augenblick and Rabin 2018; Fedyk 2017).

We then analyze the implications of channeled attention for a person who neglects the correlated nature of others' advice (as in DeMarzo, Vayanos, and Zwiebel 2003; Eyster and Rabin 2010, 2014; Enke and Zimmermann 2017). Imagine, for instance, a person in a new job who seeks advice from her colleagues and updates about the quality of their advice. If her bad theory is that others' advice is conditionally independent (ignoring that colleagues also talk with each other), then she finds it sufficient to learn about quality simply by comparing recommendations with the eventual outcome. Since this strategy does not record correlation across colleagues, she will fail to notice her mistaken model, and, as a result, persistently overreact to consensus advice. But what if she sometimes cannot observe the outcome associated with her colleagues' recommendations?

[7]There are related models of paradigm shifts and " testability" (e.g., Hong, Stein, and Yu 2007; Ortoleva 2012; Al-Najjar and Shmaya 2014). While studying paradigm shifts has the flavor of analyzing when people have "light-bulb moments", those papers do not study the interaction between light-bulb moments and inattention. The logic behind why channeled attention prevents light-bulb moments is similar to self-confirming equilibrium (Fudenberg and Levine 1993) and why people can maintain incorrect beliefs about options they rarely experiment with in "bandit problems" (Gittins 1979). However, beliefs are consistent with available data in those frameworks—the friction is data collection—while beliefs are only consistent with *encoded* data in ours—the friction is data processing. Finally, we contribute to the growing literature on learning with misspecified models (e.g., Barberis, Shleifer, and Vishny 1998; Rabin 2002; Rabin and Vayanos 2010; Benjamin, Rabin, and Rabin 2015; Spiegler 2016; Esponda and Pouzo 2016; Bohren 2016; Heidhues, Kőszegi, and Strack 2018; Fudenberg, Romanyuk, and Strack 2017) by providing a framework to assess when misspecified models are stable.

[8]A maintained assumption (discussed in greater detail in the conclusion) is that a person's awareness of self-control problems is local to the situation, so he may wake up to those problems in one situation without transferring that knowledge to another.

She then must update about a colleague by "benchmarking" his advice to that of other colleagues she believes are knowledgeable. Any sufficient attentional strategy now requires her to confront the correlation in others' advice. Therefore, with limited feedback, we predict she will discover her mistake—advice is not actually independent. Interestingly, more information makes the error *more* stable.

In Section 6 we turn to general factors that influence an error's stability. Our primary organizing principle behind the discovery of errors is that uncertainty about the optimal action paves the way for incidental learning. We first show that for every error there exists an environment where the error leads to suboptimal decisions, yet is stable. The simplest logic for when people fail to get a clue arises when they feel they have little to learn and, as a result, do not need to pay attention. For instance, if a person is confident that managed funds must outperform index funds, she may blindly invest in managed funds without noticing that her guiding theory is false. We then show that for every erroneous model, there is a decision context where such a logic applies. Hence, waking up to an error is never universal independent of the specific choice problem.

Not waking up, however, can be broadly attentionally stable *irrespective* of a person's preferences or action space. We characterize models that are *preference-independent attentionally explicable* (or PIAE), revealing that these errors are typically "overly coarse"—they ignore relevant outcomes or predictive signals. Errors that are not PIAE, on the other hand, typically make the right distinctions or are "overly fine"—they place importance on truly irrelevant outcomes or signals. A simple intuition underlies this pattern: when a person thinks a variable does not matter, she is certain about how much it matters—not at all. On the other hand, when a person thinks a variable does matter, she may be uncertain about its importance. It is this uncertainty that induces incidental learning. Section 6 concludes by discussing some factors that are likely to induce attentional stability and instability.

Section 7 reiterates and clarifies some of the features of our framework and discusses potential limitations. In particular, our notion of a SAS requires attention to only those data with instrumental value; realistically, there are many things we cannot help but notice even when we think they don't matter. Section 7 concludes by speculating on further applications to delegation, persuasion, and debiasing.

2 Example: Misperceived Benefits of Medical Treatments

This section uses a simple example to illustrate the basic components of our framework. People seem to spend a great deal out of pocket on medical treatments that are seemingly ineffective, such as remedies to help lose weight, recover from a cold, or get better sleep. This section's example reflects a plausible reason why people might exaggerate the efficacy of such treatments:

while people likely notice how quickly they recover with treatments, they may underestimate how quickly they recover *without* treatment and therefore fail to notice this baseline. Miscalibrated beliefs about recovery without treatment along with channeled attention can generate persistent overoptimism about the efficacy of treatments.

Imagine a person learning about the effectiveness of a medical treatment. He accumulates data on the treatment by noticing his own experience and that of acquaintances who might use the treatment. More specifically, each period there is a probability $\nu > 0$ that he must choose whether to use the treatment (e.g., he feels sick on a fraction ν of days). Suppose the treatment has cost $c \in \mathbb{R}_+$ and can potential increase the likelihood of a fast recovery; the event of a fast recovery yields a known benefit $b \in \mathbb{R}_+$. The person's choice to use the treatment thus depends on his current beliefs about the likelihood of fast recovery with treatment, $p_T \in [0,1]$, and without treatment, $p_N \in [0,1]$.[9] In addition to noticing his own recovery results, assume the person can observe (from news media or word-of-mouth) the choices and outcomes of others. This guarantees that an attentive person will have rich enough data to learn parameters p_T and p_N independent of his own behavior: over time, he will be exposed to a large sample of results with and without treatment.[10]

We consider learning when the person is dogmatic about the natural recovery time—that is, he is (perhaps wrongly) sure that $p_N = \hat{p}_N$. Low values of $\hat{p}_N$ capture the belief that treatment is necessary for fast recovery, as when people don't realize the self-limiting nature of colds. The person is uncertain, however, about the probability of fast recovery with treatment. Letting π denote his prior beliefs over (p_N, p_T), we assume π is degenerate on $\hat{p}_N$ and non-degenerate over p_T. We further assume the support of π includes: (1) the true efficacy of treatment, and (2) some p_T that would render treatment optimal in his mind and some where it would not, ensuring that his decisions are non-trivial. That is, letting (p_N^*, p_T^*) denote the true values of (p_N, p_T), the support of π contains values of p_T that are less than and greater than $\hat{p}_N$, as well as $p_T = p_T^*$.

Given these priors, the person continually updates his beliefs about p_T in a Bayesian fashion, but does not update his beliefs at all about the false dogmatic belief that $\hat{p}_N \neq p_N^*$. With the rich data described above, the person would learn the true p_T^*, while maintaining his false belief $\hat{p}_N$.

Here is where a commonly intuited challenge to such models lies: what is the person to make of seeing a proportion of natural recoveries that differs from $\hat{p}_N$? This issue is most stark when $\hat{p}_N = 0$ and $p_N^* > 0$. Hence, the person thinks fast recovery is impossible without treatment, yet he sees it happen. More broadly, for any $\hat{p}_N \neq p_N^*$, the person will witness an extremely unlikely

[9] More precisely, the person derives flow utility $u(x_t, y_t) = b \cdot \mathbf{1}\{r_t = 1\} - c \cdot \mathbf{1}\{x_t = T\}$, where y_t is an indicator for whether the person gets better quickly and $x_t \in \{T, N\}$ is the patient's treatment choice.

[10] This assumption that the person continually receives feedback on both options independent of his action shuts down experimentation concerns that are known to prevent learning even with rational agents. Here, the person's optimal action in each round is simply the myopically-optimal action that maximizes current flow utility. Although in principle our framework can be applied to agents who have experimentation motives, we assume them away here and in most of the paper.

frequency of people recovering quickly without treatment. If he believes $\hat{p}_N = .05$, but sees 800 out of 1,000 patients recover fast without treatment, won't he get a clue?

Our baseline, full-attention concept of explicability (sketched in the introduction and formalized below) reflects this common intuition. A misspecified model π is inexplicable relative to an alternative "light-bulb" model that puts weight on (p_N^*, p_T^*) whenever (p_N^*, p_T^*) is outside the support of π. Thus, if the person questions his initial theory and entertains the idea that $p_N = .80$ rather than .05, our explicability criterion says he will in fact realize this is so.

But if the person is dogmatic about $\hat{p}_N$, why should he attend to how fast people recover when they are not treated? Under a SAS, the person must attend to data from those who used the treatment but can neglect data from those who did not. As a result, he may never learn that $\hat{p}_N$ is wrong. Under a coarsest SAS where the person ignores recovery times of those who abstain from treatment, π is *attentionally* explicable.[11] The person may, as a result, make persistent errors when choosing to use the treatment. There is an attentionally-stable equilibrium where the person neglects recovery times without treatment and ends up best responding to the belief $p_N = \hat{p}_N$. When the misspecified prior underestimates natural recovery rates ($\hat{p}_N < p_N^*$), the equilibrium involves persistent over-treatment. Roughly put, the "self-limiting" nature of certain medical conditions—e.g., colds, pain—is misattributed to treatment: the "treatment effect" is overestimated because of a misconceived baseline recovery rate that remains in place due to the combination of dogmatically incorrect beliefs and channeled attention.[12]

Echoing a broader theme that will emerge later, a reason why the person's bad theory persists here is that his beliefs about natural recovery times are strong enough that he believes he can completely neglect the outcomes that would tell him his theory is wrong. As such, the logic of the example continues to hold even if the person is not dogmatic but thinks that nothing he learns

[11] Such a "coarsest" ("minimal") sufficient attentional strategy has the person keep track of only his posterior beliefs. That is, he enters each round t with a belief π_t over (p_N, p_T) and updates that belief to π_{t+1} based on round t's outcome. He then enters round $t+1$ recalling only π_{t+1}: he does not recall the exact sequence of outcomes that led to π_{t+1}. This is not necessarily the only minimal SAS, and there are examples where π is attentionally inexplicable under other minimal SASs.

[12] Limiting beliefs and actions in attentionally-stable equilibria may be quite biased. In this example, when the person fails to correct his underestimated perception of p_N, perhaps surprisingly he may additionally *over*-estimate p_T. This happens when the person does not always observe others' treatment choices. Specifically, suppose that for each individual whose outcome he observes, he additionally learns that individual's treatment choice with probability $\gamma \in [0,1]$. Additionally, to draw inference about p_T from those with unobserved actions, suppose the person knows that a fraction $q \in [0,1]$ of individuals use the treatment when sick. Under the ASE described above—the person ignores data from those who he knows abstained from treatment—he comes to believe that p_T takes a value that maximizes the likelihood of data he notices. This value of p_T minimizes the distance (in terms of the Kullback-Leibler divergence) between his perceived distribution of recoveries (which assumes $p_N = \hat{p}_N$) and the true distribution (with $p_N = p_N^*$). To see how channeled attention distorts perceptions of p_T, suppose in reality the treatment has no additional benefit, so $p_T = p_N$. For sake of a simple numerical example, assume $p_T = p_N = \frac{1}{2}$, but the person wrongly believes $\hat{p}_N = 0$. Further setting $\gamma = q = \frac{1}{2}$ and $v = 0$, the person comes to believe $\hat{p}_T = \frac{2}{3}$. Interestingly, the extent to which the person overestimates p_T is increasing in the natural recovery rate, p_N^*: the person is more likely to use the treatment when he is more likely to recover naturally.

about p_N will impact his decision regarding treatment. If the person's bad theory instead involves sufficient uncertainty over p_N so that he feels compelled to learn p_N to make optimal decisions, he would incidentally learn his theory is wrong even when the uncertainty is miscalibrated (e.g., π is concentrated on values of p_N less than p_N^*). Below we more systematically explore when people get a clue.

3 Framework With Full Attention

This section formalizes the first ingredients of our framework: light-bulb models and explicability. After describing the learning environment, we present our baseline criteria for assessing whether a decision maker will discover her errors. (We incorporate channeled attention in Section 4.)

3.1 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 medical example above, θ is the vector of recovery rates with and without treatment, (p_T, p_N). Parameter θ may be a feature of the person's surroundings or measure the extent of his biases—like the gambler's fallacy or present bias.

As depicted in Figure 1, each period $t = 1, 2, ...$ is structured as follows: the person (1) receives a signal $s_t \in S_t$ about θ, (2) takes an action $x_t \in X_t$, and (3) can see a realized outcome, or "resolution", $r_t \in R_t$. In the medical example of Section 2, the person chooses a treatment x_t each round that he is sick, and then his recovery time r_t is realized. The person uses vector $y_t = (r_t, s_t)$, which we call an "observable", to learn about θ. At the end of each period t, he earns payoff $u_t(x_t, y_t|h^t)$, which is a function of his action, the observables that period, and (in some cases spelled out below) the history h^t. In addition to being relevant for period-t's payoff, observable y_t may also contain information about the optimal action in future periods.

For simplicity, our general analysis assumes that Θ and each $Y_t \equiv R_t \times S_t$ are finite valued and that each X_t is compact. However, some of our applications drop these assumptions in straightforward ways. We also assume $u_t(x_t, y_t|h^t)$ is continuous and bounded in x_t for all y_t and h^t.

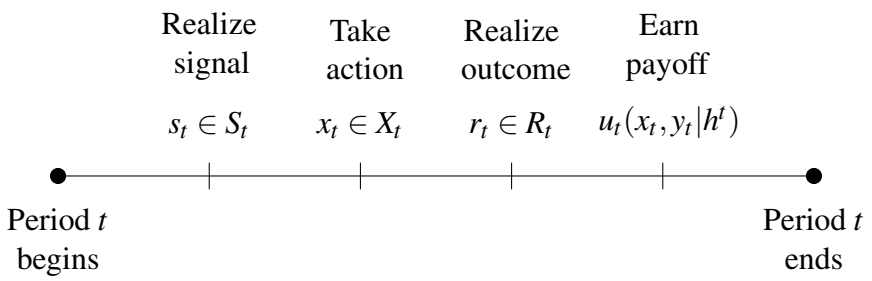
Each round results in the outcome (y_t, x_t). The history through time t,

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

contains all the information possibly observed prior to choosing action x_t. Let H^t be the set of all possible histories h^t up to time t, and let $H \equiv \cup_{t=1}^{\infty} H^t$.

Conditional on h^t and θ, the signal in round t is drawn according to distribution $P_s(s_t|h^t, \theta)$. The

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



resolution in round t depends additionally on the signal and action realized in t, and we denote its distribution by $P_r(r_t|x_t, s_t, h^t, \theta)$. These two distributions form a joint distribution over observations y_t denoted by $P(y_t|x_t, h^t, \theta)$. Finally, we let $\pi^* \in \Delta(\Theta)$ denote the actual probability distribution from which nature draws θ. In sum, the decision environments studied in this paper are described by the tuple $(\Theta, \times_{t=0}^{\infty} X_t, \times_{t=0}^{\infty} Y_t, \times_{t=0}^{\infty} u_t, P, \pi^*)$.

We focus on situations where the person begins with a "misspecified" model of the world. We assume that the person correctly knows the likelihood of outcomes conditional on the parameter, but he may have incorrect beliefs about parameters. The set Θ consists of all parameters that either the person or modeler deem plausible. We define a misspecified model or "theory" as a prior belief over parameters $\pi \in \Delta(\Theta)$. Although misspecified theory could in principle have the same support as the true model, our analysis will make clear that (given our assumption that Θ is finite) the only case of interest is $\text{supp}(\pi) \neq \text{supp}(\pi^*)$. Misspecified theories therefore place positive chance of parameters that never occur or neglects parameters that might in fact occur. Starting with a false model is the person's only mistake: we assume he updates according to Bayes' Rule given his prior π (when possible) and chooses actions that maximize his expected lifetime utility with respect to these updated beliefs.

Unless otherwise noted, we impose the following assumption, which implies that the person knows his actions do not affect observables:

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

Assumption 1 is made to focus on cases where any lack of learning is due to insufficient data or attention, not insufficient experimentation. Unless explicit, we also assume that payoffs in period t are independent of the history. Given the lack of incentive for experimentation, this ensures that myopically optimal actions are in fact long-run optimal:

Assumption 2. For all $t \in \mathbb{N}$ and $\theta \in \Theta$, action set X_t is independent of h^t, and for all $x_t \in X_t$, $y_t \in Y_t$, and $h^t \in H^t$, $u_t(x_t, y_t|h^t)$ is independent of h^t.

Under Assumption 2, we write $u_t(x_t, y_t|h^t)$ simply as $u_t(x_t, y_t)$.

3.2 Illustrative Examples

We often return to the following stylized examples, along with the example in Section 2, to help illustrate our framework.

3.2.1 Stylized Prediction Task

Section 2 describes a situation where a person exaggerates the importance of a variable. Another class of examples involves "predictor neglect," where a person neglects the importance of a variable that actually helps predict payoffs. Examples include seaweed farmers failing to appreciate the importance of pod size (as in Hanna, Mullainathan and Schwartzstein 2014) or small investors failing to appreciate the importance of analyst affiliation when interpreting investment recommendations (as in Malmendier and Shanthikumar 2007).

We will consider a more general form of predictor neglect below in Section 6, but to fix ideas suppose that each round t, a coin is drawn from a jar (with replacement) and the person predicts the likelihood it will land heads. The person then flips the coin and earns payoff $u_t = -(x_t - r_t)^2$, where $x_t \in [0,1]$ denotes the person's prediction and $r_t \in \{0,1\}$ is an indicator for heads.

Suppose the jar consists of two types of coins, A and B, which differ in the likelihood they land heads. Let θ_i denote the probability that coin type $i \in \{A,B\}$ lands heads, so the parameter of interest is $\theta = (\theta_A, \theta_B)$. (In the seaweed example, θ_i would represent the likelihood of high yield for pod-size i.) In a model with predictor neglect, the person dogmatically believes the coins have identical biases: $\theta_A = \theta_B$. Suppose the person can observe the coin's type prior to each flip t to guide his predictions. While these "signals" $s_t \in \{A,B\}$ are useful in predicting r_t when in fact $\theta_A \neq \theta_B$, the dogmatic person finds them useless.

The gambler's fallacy is another misspecification that may plague this prediction task; that is, wrongly thinking past flips are negatively autocorrelated with the current flip. Consider Rabin's (2002) model of the "Law of Small Numbers" where the person updates as if the outcomes from a coin of type $i \in \{A,B\}$ are drawn without replacement from an urn of size $K \in \mathbb{N}$ containing $\lceil \theta_i K \rceil$ heads and $\lfloor (1-\theta_i)K \rfloor$ tails.[13] Given that $K \to \infty$ corresponds to the true model where flips are i.i.d., parameter K measures the extent of the person's misspecification. Hence, the relevant parameter vector is now $\theta = (\theta_A, \theta_B, K)$, and a misspecified model puts positive weight on θ's with $K < \infty$. For instance, if the person dogmatically believes $K = 4$ and $\theta_A = \frac{1}{2}$, then he wrongly thinks the likelihood of two heads in a row from an A coin is $\frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$ instead of $\frac{1}{4}$.

[13] To avoid observations inconsistent with the person's model, Rabin (2002) assumes the urn is "refreshed" every $\lceil \theta_i K \rceil$ periods.

3.2.2 Stylized Choice Task

Now consider a stylized choice task. Suppose the person decides each period whether to take action $x = A$ or $x = B$. Option $x \in \{A, B\}$ yields utility $u_x - c_x$, where $u_x \equiv \alpha v_x$ represents the gross benefit and c_x the cost of option x. Parameter $\alpha > 0$ captures how much the agent cares about these benefits. For instance, if A and B represent two pain medications, then α might capture the agent's sensitivity to pain and hence to differences in drug efficacy. The person is initially uncertain about $\theta = (v_A, v_B)$ but can costlessly observe this pair whenever he wants. Still, the person may fail to gather information on payoffs if he has a bad theory of θ. For instance, as documented by Bronnenberg, Dubé, Gentzkow, and Shapiro (2015), people may choose branded drugs (action A) over less expensive and clinically equivalent generics (action B). Given the price premium for branded drugs, the likely reality is that $u_A^* - c_A > u_B^* - c_B$. Furthermore, readily available information exists (often explicitly printed on the packaging) revealing that the generic is the better option. However, as we spell out below, a dogmatic belief that generics are inferior can prevent a consumer from noticing this information.

3.3 Light-Bulb Theories and Explicability

We now formalize our criterion for assessing when noticed data will lead a person to discover his misspecified theory is false. Roughly, we say that a misspecified model π is *inexplicable* relative to "light-bulb" model $\lambda \in \Delta(\Theta)$ if observables are infinitely more likely under the light-bulb than the prior. Otherwise, we say π is *explicable* relative to λ.

While most of our analysis maintains Assumption 1, so that observables are unaffected by actions, we define explicability more generally. Without Assumption 1, explicability must be designated with respect to the person's behavior. Denote the person's *behavioral strategy* by $\sigma = (\sigma_1, \sigma_2, \dots)$, where each $\sigma_t : H^t \to \Delta(X_t)$. We say that behavioral strategy σ is π-optimal if for all t and $h^t \in H^t$ that occur with positive probability under π and σ, σ_t maximizes expected continuation utility, $U_t \equiv \sum_{\tau=t}^{\infty} \delta^{\tau-t} u_\tau$ for some $\delta \in (0, 1]$.

Definition 1. Given true parameter θ^*, theory $\pi \in \Delta(\Theta)$ is θ^*-*inexplicable* with respect to $\lambda \in \Delta(\Theta)$ and π-optimal behavioral strategy σ if the Bayes Factor $\Pr(h^t|\pi)/\Pr(h^t|\lambda)$ converges to 0 in t with positive probability when the person follows strategy σ. Otherwise, π is θ^*-*explicable* with respect to λ and σ.

To reveal the idea behind this criterion, suppose a Bayesian starts not with prior π but instead puts infinitesimal weight $\varepsilon \approx 0$ on an alternative model λ. That is, his prior is $(1-\varepsilon) \cdot \pi + \varepsilon \cdot \lambda$. Is it possible that the Bayesian's posterior eventually piles up on λ and places less and less weight on π? If so, our formulation says π is inexplicable relative to λ. If not, our formulation says π is

explicable relative to λ. Hence, when $\theta^* \in \text{supp}(\lambda)$, our notion of explicability with full attention essentially amounts to assessing whether long-run Bayesian beliefs settle on the true parameter. While Definition 1 may seem like a convoluted way to present such a familiar concept, we show in Section 4 that it provides an interpretational advantage once we introduce channeled attention. In particular, λ captures possibilities that a person might entertain at moments of reflection and doubt, but these possibilities do not influence what he attends to prior to such reflection.

As highlighted above, our notion of explicability is relative—it depends directly on the alternative theory λ against which π is compared. To say a model π is inexplicable does not simply mean outcomes are unlikely given π: in many models, any given outcome is unlikely.[14] For instance, if a coin is believed to be fair and i.i.d., then a sequence of 200 heads in a row is just as likely as any other sequence of 200 flips, including any with 100 heads and 100 tails. Yet we think the person will doubt his model when he sees the former sequence but not the latter. Thus, in order to rule the fair-coin hypothesis inexplicable, the person must additionally compare it with an alternative model—one where the coin is biased towards heads.

Specifying the alternative model λ is therefore a central feature of our explicability criterion. While we generally allow for any λ, most of our analysis assumes λ is the true theory we as researchers think the agent ought to entertain. Focusing on this case of $\lambda = \pi^*$ both pins down our analysis and most closely mirrors folk intuition regarding when people should get a clue. Although we take λ to be the "true" model, we do not necessarily designate that λ is degenerate on the realized parameter—λ can be probabilistic over the parameters, and thus can represent an array of alternative possibilities against which a person compares his initial theory.[15]

Focusing on the case where $\lambda = \pi^*$ tilts the analysis in favor getting a clue: restricting attention to discrete Θ, if π is explicable with respect to π^*, then π is explicable with respect to *any* light-bulb $\lambda \in \Delta(\Theta)$. 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: if π is inexplicable with respect to π^*, then it is inexplicable 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 inexplicable—and specifying which model a person adopts after rejecting π—are beyond the scope of our analysis.

Throughout the paper, we often simplify the notation in Definition 1 in three ways. First, when

[14] In a related model, Ortoleva (2012) instead evokes the "absolute" likelihood of an outcome. The distinction between the two approaches is obscured by his framework's focus on a finite number of (non-stochastic) states.

[15] Whether λ is degenerate or not becomes relevant for explicability 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 turns out to be biased 0.55 (say), we do not want the person to deem his uncertain-prior model inexplicable 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 inexplicable.

it does not create confusion, we take it as understood that nature draws θ^* from π^* and that probabilistic statements are with respect to the true distribution conditional on θ^*. This allows us to drop θ^* from the presentation. Second, when we say that π is explicable or inexplicable without reference to a particular light-bulb theory, we mean it with respect to $\lambda = \pi^*$. Third, and most substantively, Assumption 1 implies that observations are independent of actions and hence that explicability is independent of σ. Because of this, we typically drop σ from the discussion of explicability.[16]

The remainder of this section presents properties of explicable theories that serve as benchmarks for assessing the impact of channeled attention in Section 4. While these properties hold more generally, we restrict consideration to environments that are "stationary" and have rich enough feedback for a rational agent to learn the true parameter.

Definition 2. The environment is *stationary* if X_t, Y_t, and u_t are independent of t and $P(y_t|x_t,h^t,\theta) = P(y_t|x_t,\theta)$ for all $t \in \{1,2,\ldots\}$, $y_t \in Y_t$, $x_t \in X_t$, $h^t \in H^t$ and $\theta \in \Theta$.

When the environment is stationary, we denote the constant action space, outcome space, and utility function by X, Y, and u, respectively. In such settings, a misspecified model is explicable under full attention if and only if it explains observations as well as the alternative model.[17] 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 $P(\cdot|\theta^*)$.[18] Finally, 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 1. *Suppose Assumptions 1 and 2 hold, the environment is stationary, and $D(\theta^*\|\lambda)$ or $D(\theta^*\|\pi)$ is finite.*

1. *Theory π is θ^*-explicable with respect to λ if $\Delta D(\theta^*\|\lambda,\pi) > 0$ or $\Delta D(\theta^*\|\lambda,\pi) = 0$ for $\theta^* \in \text{supp}(\lambda) \cup \text{supp}(\pi)$.*

2. *Theory π is θ^*-inexplicable with respect to λ if $\Delta D(\theta^*\|\lambda,\pi) < 0$.*

[16] Note that Assumption 1 also tilts the analysis in favor of getting a clue, since it rules out insufficient or misguided experimentation as potential barriers to learning.

[17] As with all of our results, proofs are provided in Appendix B.

[18] The Kullback-Leibler divergence is given by

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

Observation 1, which reflects results known at least since Berk (1966), says that a theory π is explicable with respect to λ if π explains observations better than λ and inexplicable if it does worse. This observation has two immediate implications: (i) theory π is θ^*-explicable with respect to π^* if and only if there exists some $\theta \in \text{supp}(\pi)$ that makes the same predictions over observables, meaning $P(y_t|\theta) = P(y_t|\theta^*)$ for all $y_t \in Y_t$; and (ii) any theory that assigns positive probability to some $\theta^* \in \text{supp}(\pi^*)$ is explicable with respect to any $\lambda \in \Delta(\Theta)$.

Observation 1 additionally suggests that, within environments with rich feedback, any explicable-but-false theory generates no long-run welfare loss. Hence, with full attention, explicable models do not continually generate costly mistakes in the environments we consider.[19] As we show below, this conclusion no longer holds with channeled attention.

4 Channeled Attention and Stable Equilibria

Because our baseline definition of inexplicability in Section 3 assumes full attention, "getting a clue" may rely on the person attending to and remembering things he finds irrelevant. This section instead studies inexplicability when the person selectively notices data based on his misspecified model π. As such, the person may ignore information his model deems irrelevant for payoffs. For instance, in the medical example above, a person who thinks he knows the likelihood of fast recovery without treatment need not attend to how quickly he (or others) recover when untreated. If the person's theory regards as useless precisely those data necessary to discover his error, then he may never do so.

The first part of this section formalizes our model of channeled attention and defines how the person's theory π determines what he pays attention to. We then introduce the concept of attentional explicability and consider when an "attentional strategy" is part of an *attentionally stable equilibrium* given theory π—that is, when following the attentional strategy prevents the person from learning π is false.

4.1 Channeled Attention

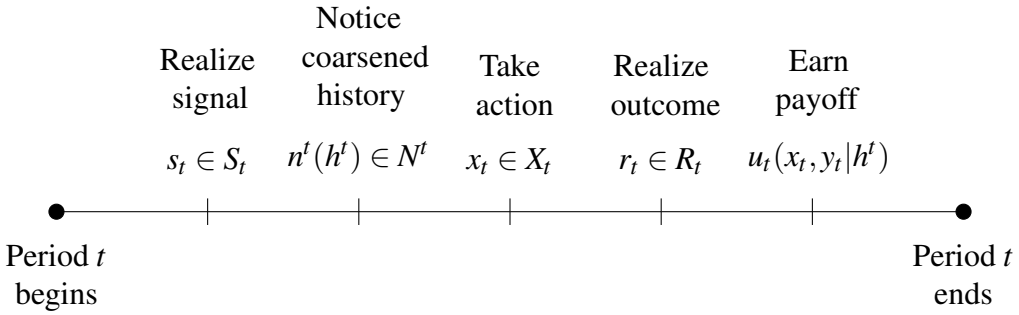
To model channeled attention, we assume the person notices and remembers coarse "signals" about the history rather than its exact value. For each $t = 1, 2, \ldots$, let N^t denote a partition of the set of histories, H^t. Following $h^t \in H^t$, the person recalls only the element of N^t containing h^t, denoted by $n^t(h^t)$. We call $n^t(h^t)$ the *noticed history*. In the coin example, suppose the person ignores

[19] That said, explicable-but-false models may reduce welfare in richer environments. If the person grows convinced of wrong parameters in one setting, he may naturally use them to make predictions in other environments where they do worse than the true model.

which type of coins are flipped and attends only to whether they land heads or tails. Then $n^t(h^t)$ contains all histories with the same sequence of heads and tails as h^t.

Figure 2, below, revisits the timeline of an individual period (Figure 1), and now includes the timing of information coarsening that happens each round: prior to taking action x_t, the person summarizes all realized data into a "sufficient statistic" $n^t(h^t)$ which he uses to guide x_t.

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



A *noticing strategy* $\mathscr{N}$ is the full sequence of the person's "noticing partitions": $\mathscr{N} = (N^1, N^2, \dots)$. This strategy specifies for each point in time what the person has noticed conditional on the true history. While we specify below how $\mathscr{N}$ depends on both the person's preferences and model π, we first consider restrictions on the noticing strategy.

With channeled attention, it becomes crucial to specify what a person recalls about data that he either did not notice in the first place or has subsequently forgotten. We limit our focus to noticing strategies that are *memory consistent*: roughly speaking, once data has been ignored or forgotten, it cannot be later noticed.[20]

Definition 3. A noticing strategy $\mathscr{N}$ is *memory consistent* 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 does not say that the person notices all information he previously encoded. Indeed, a primary focus of our analysis will concern the possibility that people do not notice information once they deem it no longer useful.

That said, there are situations where it seems plausible that data would be top of mind even when a person no longer finds it useful—for instance, immediately following an action based on a particular piece of data. To handle such scenarios, the proofs in Appendix B formally consider how our

[20]Memory consistency also rules out convergence complications that arise when the person's noticed history expands and contracts over time. While useful for such technical issues, this assumption does rule out some well-known aspects of memory. For example, an environmental cue today may elicit associative memories which were not top of mind yesterday (formalized in economics by, e.g., Mullainathan 2002 and Bordalo, Gennaioli, and Shleifer 2017).

results extend in the limiting situation of *automatic recall* (as opposed to *volitional recall* assumed in the text) where a person notices anything he previously noticed. (See Definition B. 1 for a formal definition.) For example, if a consumer considered a product's price when deciding whether to buy it, then automatic recall says that price is always top of mind. Although automatic recall is an extreme assumption, many of our results on when an error is attentionally stable continue to hold even if we make it.[21]

We assume the person is aware that he filters information. More formally, define an attentional strategy as a pair $\phi = (\mathscr{N}, \sigma)$ such that (i) noticing strategy $\mathscr{N}$ is memory consistent and (ii) for each period t, behavioral strategy $\sigma_t : N^t \to \Delta(X_t)$ maps noticed histories to actions. If the person observes noticed history n^t, he uses his theory π and knowledge of ϕ to weight the probability of each $h^t \in n^t$ and updates accordingly. Specifically, the likelihood of n^t given θ conditional on ϕ is $\Pr(n^t|\theta,\phi) = \sum_{h^t \in n^t} P(h^t|\theta,\phi)$ and the resulting posterior over θ is

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

Our analysis implicitly assumes that the person recalls her prior π, her strategy ϕ, and the time period $t \in \mathbb{N}$.[22]

4.2 Sufficient Attentional Strategies

Since we assume attentional costs are negligible, an optimizing decision maker ignores some piece of data only when he perceives it as useless for guiding future decisions. Accordingly, we say his attentional strategy is "sufficient" with respect to his theory π if he filters out only information that π deems irrelevant for decisions.

Definition 4. An attentional strategy $\phi = (\mathscr{N}, \sigma)$ is a *sufficient attentional strategy (SAS) given* π if, under π, the person expects to do no worse by following ϕ than he would by following any other attentional strategy $\tilde{\phi}$. Under Assumptions 1 and 2, sufficiency amounts to

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

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

Loosely, under a sufficient attentional strategy (SAS), the person believes that his expected pay-

[21] We conjecture that all our results carry over to a less extreme (and probably more realistic) form of automatic recall where the decision maker can freely revisit any previously noticed data if he so chooses.

[22] Although we assume throughout the paper that the agent is aware of his channeled attention, we do not think natural forms of naivete (e.g., as described in Schwartzstein, 2014) would impact our analysis given our focus on long-run questions.

off is independent of whether he optimizes using the coarsened history or the precise history. For instance, attending solely to flip outcomes in the coin-flip example is part of a SAS under the theory in which a coin's bias is independent of type. Additionally, given that we define a SAS without reference to t, we implicitly assume it is dynamically consistent. However, it would be straightforward to extend our framework to handle dynamic inconsistencies.[23]

Our definition of a SAS does not mandate that a person ignores data he deems useless. However, as a benchmark, we sometimes consider the "minimal" case where all seemingly extraneous information is ignored. Say that $\tilde{\mathscr{N}}$ *is coarser than* $\mathscr{N}$ if, for all t, $\tilde{N}^t$ is coarser than N^t and at least one of these coarsenings is strict.

Definition 5. Given π, a SAS $(\mathscr{N}, \sigma)$ is *minimal* if there does not exist another SAS that can be obtained by coarsening $\mathscr{N}$.

Minimal attentional strategies are 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 approach could accommodate situations where some data is impossible to ignore.

There can be multiple minimal sufficient attentional strategies, and the particular way a person filters out data can dictate which long-run beliefs she comes to hold. For example, consider a person who thinks that all doctors make the same recommendation given a fixed set of symptoms. Under a minimal SAS, the person follows a single doctor's advice and ignores the rest. If there is in fact heterogeneity across doctors, the patient's (seemingly inconsequential) choice of who to follow will determine his long-run beliefs.

4.3 Attentional Stability and Measurability

We are interested in when mistaken models are stable in the sense that observations noticed while following a sufficient attentional strategy are "explicable". The following extends our definition of explicability to channeled attention.

Definition 6. Given true parameter θ^*, theory π is θ^*-*attentionally inexplicable* with respect to λ and SAS $\phi = (\mathscr{N}, \sigma)$ if the Bayes Factor $\Pr(n^t|\pi)/\Pr(n^t|\lambda)$ converges to 0 in t with positive probability when the person follows SAS ϕ. Otherwise, π is θ^*- *attentionally explicable* with respect to λ and ϕ. When the latter case holds relative to $\lambda = \pi^*$, we call ϕ an *attentionally stable equilibrium (ASE)* given π.

[23] To give an example of a dynamically inconsistent SAS, consider a person who has self-control problems but is naive about them. He is at a restaurant and orders tiramisu, thinking this will be the last time he orders dessert. Because he thinks he will go on a life-long diet starting tomorrow, he does not pay attention to the tiramisu's quality. But when tomorrow comes, he again wants to order dessert and wishes he could recall how much he liked the tiramisu.

As in the medical example where the inattentive person failed to learn about natural recovery rates, channeled attention limits when "Eureka moments" will happen. With full attention, model π is inexplicable if it seems excessively unlikely relative to the light-bulb model when all available data is actively used to assess the relative likelihood of π versus λ. Our interpretation of inexplicability with channeled attention differs in the following way: we ask when the seemingly-relevant data under π will alert the person about his model's misspecification. That is, we do not interpret the agent as *actively* collecting data for the dedicated purpose of distinguishing π from λ. After all, our agent sees no reason to question π. In this interpretation, potential "Eureka moments" happen *after* the person attends to the data. If at that point the attended-to data—which was noticed solely to make optimal decisions given π—happens to make π seem implausible relative to some proposed alternative, then we say π is attentionally inexplicable. Returning to the medical example, even when natural recovery rates are equal to those with treatment, the theory that treatment is superior will be explicable when the person follows a SAS that ignores natural recovery rates.

The following proposition further emphasizes that attentional stability is not about the limited *availability* of data per se that may arise from channeled attention, but rather a failure to notice the right features in the data. Perhaps surprisingly, any attentionally stable error will remain stable if we grant the agent access to a complete archive of past outcomes. That is, an agent who can always go back and look at all past data (e.g., it is written down somewhere) still need not get a clue.

Proposition 1. *Consider any environment* $\Gamma \equiv (\Theta, \times_{t=0}^{\infty} X_t, \times_{t=0}^{\infty} Y_t, \times_{t=0}^{\infty} u_t, P, \pi^*)$ *that satisfies Assumptions 1 and 2, and consider a modified environment identical to* Γ *aside from allowing the person access to the full history each period: in the modified environment, the person receives signals* $\tilde{s}_t = (s_t, h^t)$ *for all* $t = 1, 2, \ldots$*, where* s_t *follows the signal structure of* Γ *and* h^t *is the history up to period* t*. If a model* π *is attentionally stable in the original environment* Γ*, then it is attentionally stable in the modified environment.*

To provide intuition, suppose ϕ is an ASE given π in an environment where the history is not exogenously archived. 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 he believes would improve his decision beyond the data he gathered following ϕ. If this is the case, then ϕ was not sufficient in the first place: any sufficient attentional strategy extracts all seemingly useful data from the history.

Another fact helps strengthen the result above:

Proposition 2. *Consider any environment* $\Gamma \equiv (\Theta, \times_{t=0}^{\infty} X_t, \times_{t=0}^{\infty} Y_t, \times_{t=0}^{\infty} u_t, P, \pi^*)$ *that satisfies Assumptions 1 and 2, and consider the modification of this environment described in Proposition 1. There exists a minimal SAS* ϕ *in the modified environment where in each period* t*, the person notices the optimal action* $x_t^* \in X_t$ *given* h^t *and* π *and nothing more.*

Given the SAS in Proposition 2, attentional explicability boils down to the existence of $\theta \in \text{supp}(\pi)$ such that

$$\frac{P(x_t^*|\theta)}{P(x_t^*|\theta^*)} \tag{2}$$

is bounded away from zero, where x_t^* is the optimal action in period t conditional on h^t and π. Thus, so long as the optimal action with full attention does not grow excessively unlikely under π relative to θ^*, ϕ is an attentionally stable equilibrium when h^t is always available. While Proposition 1 shows that continual access to h^t makes the discovery of errors weakly less likely (relative to environments without it), Proposition 2 suggests that often this access makes the discovery *strictly* less likely.

In other words, increasing the availability of historical data decreases the propensity to discover errors. Access to the history limits the data an agent must notice over time. Indeed, when h^t is available each round, a minimal SAS requires the person to simply query the history each period asking "what action should I take today?" It is then sufficient to notice this recommended action and nothing more (including actions previously taken). However, knowing solely today's optimal action is typically not sufficient to reveal a model's misspecification: one must additionally notice details of the outcomes leading to that action.

To illustrate using an example we return to later, suppose a manager must assign a worker to either an important or unimportant task each period. The optimal assignment depends on the worker's success rate, θ, which is unknown ex ante (e.g., the manager should assign the important task only if the worker's estimated success rate exceeds 50%). Must the manager discover when his prior puts zero weight on the true success rate, θ^*, for example if he is overly pessimistic about the worker? If the worker's outcomes are recorded in a database so that the manager notices solely the optimal action each period, then the answer is no: the optimal assignment provides only a rough sense of how often the worker has been successful (e.g., whether she's been successful more than 50% percent of the time). For the manager to discover that $\theta^* \notin \pi$, he must further notice the worker's precise success rate over time. But when h^t is always available, the manager has no incentive to engage with this seemingly superfluous data.

These incentives change when h^t is not always available . If the manager can no longer query a database to derive the optimal action, then he must notice and remember the worker's success rate himself. In this case, tracking the history to update about θ could incidentally lead the manager to discover his mistake—eventually noticing a success rate inconsistent with any $\theta \in \text{supp}(\pi)$.

This highlights a sense in which incidental learning comes from a discrepancy in the data necessary to make an optimal decision and that necessary to precisely learn parameters. Providing access to data that pinpoints the current optimal action (e.g., by writing things down or delegating tasks) allows the decision maker to bypass other details required for belief updating and hence

limits the scope for incidental learning.[24]

Similar to the logic above, when following a minimal SAS, the person will never confront data that he thought was *impossible*—his model will be "measurable". This will be the case even when the person does not automatically have access to h^t each period (e.g., some data is not naturally recorded in a database). 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 3. *If $\phi = (\mathscr{N}, \sigma)$ is a minimal sufficient attentional strategy given π, then π is attentionally measurable with respect to ϕ.*

Proposition 3 demonstrates 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. In particular, this is true when the person follows a minimal SAS. The idea behind the result is that the person sees no benefit to distinguish events he assigns zero probability from those he assigns positive probability. Hence, he need not notice when subjectively zero probability events occur.

There are three basic points to take away from this result. First, surprising events do not necessarily lead a person to get a clue because he may not be on the lookout for such events: while nothing is more surprising than an event he thought impossible, these are precisely the events he assumes are not worth looking out for. Second, whether a person is on the lookout depends on his theory, and thus surprising events that a person's theory anticipates are more likely to lead him to get a clue. Third, since an immeasurable event would surely lead a person with full attention to get a clue, channeled attention modifies key comparative statics predicted by full attention.

While a person's theory will be attentionally measurable if he follows a minimal SAS, it need not be attentionally explicable. In the next section, we examine some applications that demonstrate how attentional stability depends on the environment. We then discuss broader principles of attentional stability in the following section.

5 Applying the Framework to Some Well-Known Errors

5.1 Self Control

This application explores how channeled attention can lead a person to persistently underestimate her self-control problems. While some researchers (e.g., Ali 2011) argue that rational learning

[24] Of course, our analysis abstracts from realistic attentional and memory constraints and the associated benefits of not having to rely on one's memory to access data. Our analysis also abstracts from what happens following light-bulb moments—that is, what a person comes to believe after concluding that π is false. Such dynamics will depend heavily on whether a person is able to revisit data he did not notice originally.

should correct such overestimation of self control, we demonstrate that the path to sophistication in fact faces many obstacles.

To study scenarios where a person chooses to take an action with immediate cost and delayed benefit, we frame the analysis around the daily decision of whether to visit the gym. If the person goes on day t, then she pays an immediate effort (or opportunity) cost equal to c_t and earns benefit $b > 0$ in the future. If she doesn't go, she incurs no cost or benefit. For now, we assume there is no monetary fee to attend; for example, the gym is in her apartment building.

Following Laibson (1997) and O'Donoghue and Rabin (1999, 2001), we consider a (β, δ) discounter with $\delta = 1$: in each period t, the person discounts any future costs or benefits by a factor $\beta \in [0, 1)$. While the person may acknowledge that she discounts future utility, we assume she underestimates the extent of her time inconsistency. The person's priors π^β over β are concentrated on the interval $(\beta, 1]$: she is overoptimistic about her self control. An important special case is where the person dogmatically believes in some $\hat{\beta} \in (\beta, 1]$.

We additionally allow for uncertainty over the distribution of effort costs. For instance, the person could be uncertain how busy she will be. Suppose costs c_t are i.i.d. draws from $\mathrm{U}[0, \bar{c}]$, where $\bar{c}$ is potentially unknown. The person's model is thus $\pi = (\pi^\beta, \pi^{\bar{c}})$, where $\pi^{\bar{c}}$ is her prior over $\bar{c}$. We assume $\mathrm{supp}(\pi^{\bar{c}})$ includes the true value of $\bar{c}$ and that $\bar{c} > b$ to avoid trivial cases.

The person thinks prospectively that she should visit the gym on day t if and only if $b > c_t$, but she in fact does so if and only if $\beta b > c_t$. As it aids our exposition, we re-write this last inequality as $b > c_t + \tau$, where $\tau \equiv (1 - \beta)b$ is the person's "temptation" to avoid the gym. Naivete implies that the person misperceives τ. In particular, if she dogmatically believes in $\hat{\beta} > \beta$, then she forecasts that she will visit the gym whenever $b > c_t + \hat{\tau}$, where $\hat{\tau} \equiv (1 - \hat{\beta})b < \tau$. We treat temptation τ as observable and assume that the person would notice that $\tau \neq \hat{\tau}$ for her behavior to differ from her forecast. Hence, with full attention, π is inexplicable: through her daily assessment of whether the gym seems worthwhile, the person notices that $\tau \neq \hat{\tau}$.

With channeled attention, however, π is part of an attentionally stable equilibrium. Recall that the person decides to visit the gym on day t if and only if $b > c_t + \tau$. Hence, the decision only requires her to notice whether the *sum* $[c_t + \tau]$—her total disinclination to visit the gym that day—exceeds b. She need not separately notice the values of c_t and τ.[25] Furthermore, because she does not think there is anything payoff relevant to learn, she need not remember past values of $[c_k + \tau]$ or her past gym-going. So, when following a SAS, the person need not realize that τ is greater

[25] When π^β is degenerate on some $\hat{\beta} > \beta$ then there is an additional SAS in which the person's false sense of self control is "self confirming": she behaves exactly *as if* her present-bias discount factor were in fact $\hat{\beta}$. Under this SAS, the person attends only to whether c_t exceeds $b - \hat{\tau}$ each round, and, in effect, attends to c_t instead of $[c_t + \tau]$. She thus visits the gym if her *predicted* over-all desire to avoid the gym that day, $[c_t + \hat{\tau}]$, falls short of the benefit. Since this behavior matches the optimal plan of a sophisticated agent with discount factor $\hat{\beta}$, this SAS is also an attentionally stable equilibrium.

than she thought possible.

Proposition 4. *Consider an environment in which a person decides each day whether to go to the gym and let* $\pi = (\pi^{\beta}, \pi^{\bar{c}})$*, where* $\beta < \min\left[\text{supp}(\pi^{\beta})\right]$.

1. *With full attention,* π *is inexplicable relative to* π^*.

2. π *is part of an* attentionally *stable equilibrium in which the person visits the gym on any day* t *if and only if* $\beta b > c_t$.

The logic behind Proposition 4 is straightforward: if a person has no incentive to track her own behavior, then she need not recognize the extent of her self-control problem. Interestingly, we show next that this result continues to hold even the person does have incentives to track her behavior—for example, if she must decide whether it is worthwhile to pay for a gym membership.

Suppose now that at the start of each period, the person decides whether to buy a gym membership at up-front cost $m > 0$. If she buys it, she can visit the gym that period at no (pecuniary) cost. We assume the person can observe her effort cost (c_t) and her temptation to avoid the gym (τ) regardless of whether she purchases the membership, so having it does not help her learn. Thus, her optimal choice each period is the myopically optimal decision of whether to buy the membership. For a given point belief $(\hat{\beta}, \hat{c})$, the person desires the membership if [26]

$$m < \mathbb{E}\left[b - c | \hat{\beta} b > c\right] \Pr(\hat{\beta} b > c) \;\Leftrightarrow\; m < \frac{1}{\hat{c}} \int_0^{\hat{\beta} b} (b - c) dc \Leftrightarrow m < \hat{\beta} b^2 \left(\frac{2 - \hat{\beta}}{2\hat{c}}\right). \tag{3}$$

That is, she buys the membership if she thinks its cost is lower than the option value of being able to use the gym at zero pecuniary cost. This option value is higher when perceived effort costs are low (small $\hat{c}$) and perceived self control is high (large $\hat{\beta}$).

It is straightforward that the person may persistently "pay not to go to the gym" in an attentionally stable equilibrium: when inequality (3) holds for all $(\hat{\beta}, \hat{c}) \in \text{supp}(\pi)$, the person thinks membership is worthwhile no matter what, and thus sees no need to track her behavior. The analysis is then similar to Proposition 4. (And it's also similar if inequality (3) is reversed for all $(\hat{\beta}, \hat{c})$ in

[26] More generally, suppose the person is offered two contracts: (1) a "pay-per-visit" contract, consisting of a visit fee v and no up-front membership fee, and (2) a "membership" contract consisting of no visit fee but an up-front membership fee m. The case we consider above assumes v is so large that the person only considers the membership contract. For a given $v < \infty$ and $(\hat{\beta}, \hat{c})$, the person chooses the membership over the pay-per-visit contract if

$$m < \mathbb{E}\left[b - c | \hat{\beta} b > c\right] \Pr(\hat{\beta} b > c) - \mathbb{E}\left[b - c - v | \hat{\beta} b > c + v\right] \Pr(\hat{\beta} b > c + v)$$
$$\Leftrightarrow m < \frac{1}{\hat{c}} \left(\int_0^{\hat{\beta} b} (b - c) dc - \int_0^{\hat{\beta} b - v} (b - c - v) dc \right).$$

When $\hat{\beta} b - v > 0$, then the last inequality reduces to $m < \frac{v(2b - v)}{2\hat{c}}$.

the support of π.) So, if the person is certain she has sufficient self control to make the membership worthwhile, then she need not notice that she goes too little to justify the membership.

A perhaps more interesting case emerges when the person is so uncertain about $(\hat{\beta},\hat{c})$ that she is initially unsure whether the membership is worthwhile. That is, inequality (3) holds for some $(\hat{\beta},\hat{c}) \in \text{supp}(\pi)$ but not all. In this case, any SAS requires the person to track statistics either on realized values of (c,τ) or on how often she visits the gym. As it happens, even in this case it is attentionally stable for her to remain overconfident about her self control. To illustrate in a simple way, assume that π^{β} is degenerate on $\hat{\beta} > \beta$ and $\pi^{\bar{c}}$ concentrates on $\{c^L, c^H\}$, where $c^L < c^H$. Suppose the person forecasts that she we would want the membership if $\bar{c} = c^L$ but not if $\bar{c} = c^H$:

$$\hat{\beta} b^2 \left(\frac{2-\hat{\beta}}{2c^H} \right) < m < \hat{\beta} b^2 \left(\frac{2-\hat{\beta}}{2c^L} \right).$$

To determine whether the membership is worthwhile, any SAS requires the person to register enough data to differentiate between c^L and c^H. To do this, the person need only track whether c ever exceeds c^L. There are at least two ways she can do this while still noticing whether c_t is low enough to visit the gym on day t:

1. *Track non-temptation costs.* In any period where she purchases a membership—and thus has not yet determined that $\bar{c} = c^H$—she attends to whether $b > c_t + \tau$ (to determine if she wants to visit the gym that day) and, if not, whether $c_t > c^L$ (to update her beliefs over $\bar{c}$). If she does not purchase a gym membership and has not yet determined that $\bar{c} = c^H$, then she just attends to whether $c_t > c^L$.

2. *Track overall costs.* In any period where she purchases a membership, she attends to whether $b > c_t + \tau$ and, if not, whether $c_t + \tau > c^L + \hat{\tau}$. If she does not purchase a gym membership and has not yet determined that $\bar{c} = c^H$, then she just attends to whether $c_t + \tau > c^L + \hat{\tau}$.

Under either SAS above, the person remains overoptimistic about her self control. However, they lead to different behavior when $\bar{c} = c^L$. Following the first SAS, the person will discover costs are low and, by not learning that $\hat{\beta} < \beta$, she may mistakenly buy gym memberships in perpetuity. Following the second SAS, by contrast, the person reconciles large $c + \tau$ while maintaining her belief in $\hat{\tau} < \tau$ by concluding that $\bar{c} = c^H$. Hence, she will turn down memberships from some period on despite remaining overconfident about her self control.

Generalizing the example above, π is part of an attentionally stable equilibrium even when $\pi^{\bar{c}}$ and π^{β} are diffuse.

Proposition 5. *Consider an environment in which a person decides at the start of each period whether to buy a gym membership and, if she does, decides later that period whether to go to*

the gym. Let $\pi = (\pi^{\beta}, \pi^{\bar{c}})$, where subjective uncertainty over $\hat{\beta}$ and $\hat{c}$ is independent and $\beta < \min\left[\text{supp}(\pi^{\beta})\right]$. Model π is part of an attentionally stable equilibrium under which the person visits the gym on any day t if and only if she buys a membership and $\beta b > c_t$.

In an attentionally stable equilibrium, the person may end up correctly predicting her behavior without realizing it is suboptimal. To illustrate, consider the example above where the person follows the second SAS and suppose $c^H = \frac{\hat{\beta}}{\beta} c^L$. If in truth $\bar{c} = c^L$, then the person explains her behavior by thinking $\bar{c} = c^H$ and declines the membership from some period on when

$$m > \hat{\beta} b^2 \left(\frac{2 - \hat{\beta}}{2c^H} \right) = \beta b^2 \left(\frac{2 - \hat{\beta}}{2\bar{c}} \right).$$

But this is the wrong inequality for her to use: the $\hat{\beta}$ inside the parentheses should be $\beta < \hat{\beta}$. By not learning the extent of her self-control problem, the person neglects the commitment value of the membership and may mistakenly "learn" to give up on her membership when it would be in her interest to keep it.

5.2 Neglecting Correlations

This section considers an agent who neglects the correlated nature of others' advice (as in DeMarzo, Vayanos, and Zwiebel 2003; Eyster and Rabin 2010; or Enke and Zimmermann 2017). Each period, the agent encounters a problem that has a solution dependent on a binary state that fluctuates from period to period. 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 $\omega_t \in \{A, B\}$, where $q_A \in (0,1)$ denotes the prior probability that $\omega_t = A$. This prior probability captures a person's intuition about the optimal action. Each period t, the person receives signals from K information sources, denoted $s_t = (s_t^1, \ldots, s_t^K) \in \{A, B\}^t$, that are potentially informative about ω_t. These signals, for instance, may be colleagues' advice on how to resolve a problem.

We examine situations where the agent learns how to use these signals for his decisions. For example, a salesperson may encounter new clients each period and must decide how aggressive to be. He takes into account customer-specific advice from his colleagues while initially being unsure who gives good advice. Or a new professor teaches different lectures each day and learns over time which colleagues give good advice on how to lead those classes.

We allow for the possibility that the agent does not always receive feedback about whether an information source made a good or bad recommendation. (E.g., colleagues give advice about how to teach a class, but it is sometimes hard to tell how the class went.) The outcome r_t each round is such that the person receives feedback ($r_t = \omega_t$) with probability $\rho \in [0,1]$ and does not (denoted

$r_t = \varnothing$) with probability $1-\rho$. Hence, ρ is the frequency of feedback on the quality of the person's information sources.

The person's objective is to take an action $x_t \in \{A,B\}$ that matches the state: $u_t = 1$ if both $x_t = \omega_t$ and $r_t = \omega_t$, and $u_t = 0$ otherwise. For example, x_t is how the person resolves the problem he faces on day t—this clearly pays off if he receives feedback that x_t was the appropriate response.

The agent is initially uncertain about the quality of his information sources. We explore the stability of a misspecified model that treats his information sources as independent. We let $\theta^k \equiv \Pr(s_t^k = \omega_t | \omega_t)$ denote the precision of signals from source k. The overall parameter governing the environment is $\theta = (\theta^1, \ldots, \theta^K, \theta^{K+1}) \in (.5,1)^K \times \{0,1\}$, where the final element $\theta^{K+1} \in \{0,1\}$ parameterizes the correlation in information sources. We let $\theta^{K+1} = 0$ denote the case where the information sources are independent (conditional on ω_t) and $\theta^{K+1} = 1$ denote the specific correlation structure introduced below, which for sake of exposition focuses on the case of two sources of advice.

To fix ideas, consider a new employee who receives advice from two colleagues $i \in \{1,2\}$ each period. In truth, s_t^1 and s_t^2 are not independent conditional on ω_t. Colleague 1 always communicates her independent private information $s_t^1 \in \{A,B\}$ with both the new employee and Colleague 2. Colleague 2, however, simply repeats Colleague 1's information to the new employee unless she has perfect private information about ω_t, which we suppose happens with probability $\iota \in [0,1]$. That is, fixing ω_t, $s_t^2 = s_t^1$ with probability $1-\iota$ and $s_t^2 = \omega_t$ with probability ι.

Correlation neglect (being dogmatic that $\theta^{K+1} = 0$) can have sharp behavioral consequences. For example, when $\theta^1 = .6, \iota = .3$, and $q_A = .7$, then the employee goes against her intuition (given her prior that A is optimal) only when both colleagues agree that she should take action B. This is clearly suboptimal given the information structure, since in fact $(s^1 = A, s^2 = B)$ reveals for sure that $\omega = B$, while $\omega = A$ is actually more likely following $(s^1 = B, s^2 = B)$. The employee *should* go against her intuition of taking action A only when her colleagues disagree about which action is better, not when they agree that B is better: disagreement reveals that Colleague 2 has strong private information.

Will the new employee get a clue that his colleagues' advice is correlated and avoid the mistake above? With volitional recall, the answers are perhaps surprising: roughly, the person maintains his false model if and only if he receives *perfect* feedback about his colleagues' advice.

Proposition 6. *Consider any misspecified model π that is non-doctrainaire about θ^1 and θ^2 but puts probability 1 on $\theta^3 = 0$ (independent signals). π is part of an attentionally stable equilibrium if and only if $\rho = 1$ and $\iota > 0$.*

Since the agent wants to learn θ^1 and θ^2, he must track his colleagues' advice.[27] But the agent

[27]While non-doctrinaire priors violate our earlier assumption of a finite parameter space, they are not necessary

need not track the precise ordered history, so he may ignore some data on how often the two colleagues agree. The extent of this ignorance depends on the rate of feedback, ρ. With perfect feedback ($\rho = 1$), a minimal SAS records only the number of times each colleague gives correct advice. Since the employee need not notice the inexplicable rate at which the two colleagues agree (i.e., the frequency with which $s_t^1 = s_t^2$), he can persist in believing their advice is independent. With limited feedback ($\rho < 1$), however, a minimal SAS must additionally record how often the two colleagues agree in periods without feedback. Although the employee never learns whether his colleagues were right or wrong in such periods, their signals are still useful for updating. To see why, consider the extreme case where the employee is confident that Colleague 1 typically delivers high-quality advice: even without feedback on whether s_t^1 is correct, the mere fact that Colleague 2 agrees with Colleague 1 would be good news about the quality of Colleague 2. As this logic extends to less-extreme cases, $\rho < 1$ implies that the agent learns about θ^1 and θ^2 in part by *comparing* the advice of the two colleagues. Such benchmarking requires the agent to notice a rate of agreement inconsistent with independent signals, leading him to incidentally discover that signals are in fact correlated.

In this environment, richer information (on the quality of advice) *prevents* the person from getting a clue. The next section considers in more generality which combinations of errors and environments create barriers to getting a clue.

6 When Are Errors Attentionally Stable?

We now explore general factors that influence an error's stability. We partially characterize which combinations of environments and erroneous models generate stable yet *costly* mistakes. The analysis revolves around three central questions: (1) Which errors are likely to be discovered independent of the environment? (2) Which errors are *un*likely to be discovered independent of the environment? (3) When the stability of an error is not universal, what factors lead it to be discovered?

Throughout this section, we will illustrate our results with the following simple example. Consider a manager who must assign an employee a task each period. There are two tasks $x \in \{H, L\}$, "high" and "low", which differ in their importance to the manager. The manager assigns tasks

and are assumed only for sake of exposition. As the proof makes clear, we need only ensure that the agent has incentive to learn (θ^1, θ^2), which can be achieved with a finite parameter space meeting the following conditions. Consider a model π with supports over θ^1 and θ^2 that share common minimum and maximum values, denoted by $\underline{\theta}$ and $\overline{\theta}$, respectively, and define the sets $Q_A \equiv \left[\frac{\underline{\theta}^2}{(1-\underline{\theta})^2}, \frac{\overline{\theta}^2}{(1-\overline{\theta})^2}\right]$, $Q_B \equiv \left[\frac{(1-\overline{\theta})^2}{\overline{\theta}^2}, \frac{(1-\underline{\theta})^2}{\underline{\theta}^2}\right]$, $Q_M \equiv \left[\frac{\underline{\theta}(1-\overline{\theta})}{\overline{\theta}(1-\underline{\theta})}, \frac{\overline{\theta}(1-\underline{\theta})}{\underline{\theta}(1-\overline{\theta})}\right]$, and $\mathscr{Q} \equiv Q_A \cup Q_B \cup Q_M$. The result of Proposition 6 holds so long as π yields a joint distribution over (θ^1, θ^2) such that (a) the supports of θ^1 and θ^2 both contain at least two elements and the true values of θ^1 and $\theta^2 = \iota + (1-\iota)\theta^1$; (b) $(1-q_A)/q_A \in \mathscr{Q}$; and (c) uncertainty about θ^1 and θ^2 is independent under π.

based on his beliefs about the employee's ability, $\theta \in [0,1]$. We assume the employee's output $y_t \in \{0,1\}$ is i.i.d. conditional on θ with $P(y_t = 1|\theta) = \theta$. That is, $y_t = 1$ denotes a "successful" job in round t and θ is the employee's success rate. Suppressing the utility function determining this choice, the manager assigns the important task in round t (i.e., $x_t = H$) if and only if $\mathbb{E}_t[\theta] \geq k$, where $k \in [0,1]$ is some cutoff, and $\mathbb{E}_t$ is the manager's expectation of competence based on updated beliefs entering round t, π_t.[28] The results in this section will shed light, for example, on when the manager is likely to discover that π—his view of the employee's ability or the factors that predict this ability—is wrong.

6.1 All Errors are Sometimes Stable and Costly

All errors are sometimes stable and costly. Some errors, however, will be discovered across a wide range of environments. Consider settings satisfying Assumption 1, which can be written in terms of two components: the *outcome environment*, (Y, Θ, P, π^*), describing possible distributions over outcomes, and the *choice environment*, (X,u), describing the action space and utility function. Trivially, for every outcome environment and erroneous model, there exists *some* choice environment in which the model is stable:if u is independent of outcomes, the person has no incentive to attend to data.

More interestingly, for every error there is always a choice environment in which that error is both stable *and costly*, meaning that limiting behavior under a SAS given π is suboptimal relative to limiting behavior under a SAS given π^*. The following definition more formally spells out what we mean by costly.

Definition 7. Consider a misspecified model π and SAS $\phi = (\mathscr{N}, \sigma)$ given π. Let $\bar{u}_t(\phi|h^t) \equiv \mathbb{E}_{(\theta^*,\phi)}[u_t(x,y)|h^t]$ denote 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 SAS ϕ is not costless it is *costly*. When ϕ is both costly and an attentionally stable equilibrium, it is a *costly attentionally stable equilibrium*.

As a step towards showing that all errors can be stable and costly in some environment, we first demonstrate the stability of "dogmatic models" that provide a sense of certainty about the optimal action.

[28] For an example of payoffs that generate this decision rule, consider payoffs for $x \in \{H,L\}$ given by $\theta b_x - (1-\theta)\psi_x$ for $b_H \geq b_L$ and $\psi_H \geq \psi_L$. That is, the "high" importance task is associated with a greater net benefit of getting things right. The manager then wants to assign task H if and only if

$$\theta \geq \frac{\psi_H - \psi_L}{b_H - b_L + \psi_H - \psi_L} \equiv k.$$

Lemma 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 an attentionally stable equilibrium $(\mathscr{N}, \sigma)$ given π whether or not it is costly.*

With channeled attention, a sense of certainty—even if it is misguided—hinders the discovery of errors: if the person eventually believes there is no further data that would change his action, then he sees no benefit from paying attention. In terms of our example, if the manager is certain that the employee's ability θ is sufficiently large or small then he is confident about which action to take. This logic extends to situations where the person is initially uncertain but believes outcomes will follow some identifiable pattern. For instance, a manager who thinks an employee's performance is constant across situations may judge the worker based on a single observation and, by neglecting subsequent performance, never learn that his theory of steady performance is perhaps wrong.

Lemma 1 is useful in establishing the main result of this sub-section: given any model π that does not entertain the true parameter and *any* outcome environment (Y, Θ, P, π^*) that meets some simplifying regularity conditions, there exists a choice environment (X, u) such that π is attentionally explicable and costly.

Proposition 7. *Consider any outcome environment (Y, Θ, P, π^*) such that $P(\cdot|\theta) \neq P(\cdot|\theta')$ if $\theta \neq \theta'$. For every π with $\theta^* \notin \text{supp}(\pi)$, there exists a choice environment (X, u) and a corresponding SAS ϕ such that ϕ is a costly attentionally stable equilibrium given π.*

Here is some intuition. Fixing the outcome environment (Y, Θ, P, π^*), we can construct a choice environment with binary actions, $X = \{H, L\}$, 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 particular properties: (1) the person incurs a big penalty if he ever switches actions—so he's effectively choosing between "always H" and "always L"—and (2) "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. Hence, invoking Lemma 1, there is an attentionally stable equilibrium given π, and it is clearly costly. While the proof constructs a context where the person is *ex ante* dogmatic about the optimal action, many of our examples show that errors can remain attentionally stable even with active updating about which action to take.

What is more (and in a sense obvious given our framework), the costly attentionally stable equilibria identified in Proposition 7 can be *arbitrarily costly* to the decision maker. A person fails to discover a costly mistake only when he wrongly deems valuable data entirely useless and ignores it. Once deemed useless, the true value of this data—which determines the scale of the person's mistake—does not influence the decision to ignore it. That is, no matter how great the

true benefit of some data relative to the attentional cost of processing it, the decision maker may continually ignore this data if his *perceived* benefit is sufficiently small.

Proposition 7 should be read as saying that the attentional stability of an error must be assessed with reference to the choice environment, not that all errors tend to be stable. Indeed, some errors are attentionally *un*stable across "large" classes of choice environments. In particular, there are models π that are never both stable and costly when the choice environment is *stationary*. Consider, for instance, the employee example above assuming the manager earns a stationary utility $u(x_t, y_t)$ that depends solely on the current task assignment and outcome. If the manager wrongly believes the employee's success rate θ has support $\{.25, .75\}$ when in reality $\theta^* = .5$ (i.e., the manager believes the employee is either typically successful or unsuccessful, when in fact her success is a coin flip), then this error is stable only when it is costless—any (wrongheaded) attempt to distinguish $\theta = .25$ from $\theta = .75$ must alert the manager about the intermediate value $\theta^* = .5$.[29]

6.2 Some Errors are Almost Always Stable

While the discussion above illustrates that the choice environment influences attentional stability, some errors are part of an attentionally stable equilibrium *broadly irrespective* of the decision maker's preferences and action space. These include, for instance, errors that say some truly important distinctions or variables do not matter. In this section, we characterize errors that are stable across choice environments meeting stationarity and Assumptions 1 and 2.

Definition 8. In a stationary environment that meets Assumptions 1 and 2, theory π is *preference-independent attentionally explicable (PIAE)* given true parameter θ^* if for any action space X and $u : X \times Y \to \mathbb{R}$, there exists a θ^*-attentionally stable equilibrium given π.

Below, we discuss specific errors that are PIAE. To reach that point, however, we first provide a more general characterization of when a model π satisfies PIAE, 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 $Y(\pi) \equiv \cup_{\theta \in \text{supp}(\pi)} Y(\theta)$. Concepts from probability theory regarding minimal sufficient statistics (e.g., Lehmann and Casella 1998) help us analyze when models are PIAE. Let

$$m_\pi(y) = \left\{ y' \in Y(\pi) | P(y'|\theta) = P(y|\theta) h(y', y) \; \forall \; \theta \in \text{supp}(\pi) \text{ and some } h(y', y) > 0 \right\}. \quad (4)$$

[29] The reason is that, for the error to be stable, the manager must think that he does not need to attend to the frequency of success to determine his optimal decision (otherwise he'd incidentally learn that $\theta = .5$). So for the error to be stable, the manager must believe that one of the tasks, say $x = H$, is weakly optimal irrespective of $\theta = .25$ or $\theta = .75$. Optimality of $x = H$ given $\theta = .25$ implies $.25u(H,1) + .75u(H,0) \geq .25u(L,1) + .75u(L,0)$ and given $\theta = .75$ it implies $.75u(H,1) + .25u(H,0) \geq .75u(L,1) + .25u(L,0)$. Normalizing $u(L,0) = 0$, this means $u(H,1) + u(H,0) \geq u(L,1)$. But in order for this error to additionally be *costly*, L must be strictly optimal given the true parameter $\theta^* = .5$, which implies $u(H,1) + u(H,0) < u(L,1)$, a contradiction.

Intuitively, $m_\pi(y)$ 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. To accommodate volitional recall, we analogously define minimal sufficient statistics over histories $y^t = (y_{t-1}, y_{t-2}, \ldots, y_1)$, which we denote by $m_\pi(y^t)$.[30] Note that $m_{\pi_t}(y)$ could, in principle, vary in t: if $P(y|\theta) = 0$ or 1 for some $(y,\theta) \in Y(\pi) \times \text{supp}(\pi)$, then the support of π_t may differ from that of π. However, to reduce the number of cases, we assume $P(y|\theta) \in (0,1)$ for all $(y,\theta) \in Y(\pi) \times \text{supp}(\pi)$, which implies that $m_{\pi_t}(y)$ is constant in t.

With these concepts in hand, the following result characterizes when theories are PIAE.

Lemma 2. *Assume Assumptions 1 and 2 hold, and that the environment is stationary. Further suppose* $P(y|\theta) \in (0,1) \; \forall \; (y,\theta) \in Y(\pi) \times \text{supp}(\pi)$. *The theory* π *is PIAE given* θ^* *if and only if there exists* $\theta \in \text{supp}(\pi)$ *such that with probability 1 the ratio*

$$\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)} \tag{5}$$

is bounded away from zero.

The idea behind this result is simple: It is always sufficient and sometimes necessary for the person to attend to any and all information helpful in updating beliefs about θ. Lemma 2 can be used to identify specific classes of models that are stable across choice environments. One immediate corollary of Lemma 2 is that dogmatic errors (e.g., about ability or self-control problems) are PIAE: when π is degenerate, then $m_\pi(y) = Y(\pi)$ and $m_\pi(y^t) = Y(\pi)^{t-1}$—that is, no data is distinguished, and hence (5) necessarily holds.

We can use Lemma 2 to further categorize when classes of models are PIAE. Below, we discuss this categorization in intuitive terms, and relegate formal definitions and results to Appendix A.

Roughly, the following two classes of models are PIAE:

1. *"Censored" models that ignore possible outcomes:* A censored model neglects some possible outcomes, but predicts correct marginals over the subset of outcomes it entertains. Imagine the manager example with three outcomes: the employee's performance may be successful, mediocre, or poor, so $Y = \{-1, 0, 1\}$. A manager with a censored model believes, for instance, that for all $\theta \in \text{supp}(\pi)$, the probability of $P(y = -1|\theta) = 0$—he thinks

[30] For an example of $m_\pi(y)$ and $m_\pi(y^t)$, consider the manager's theory about an employee's success rate, $\theta \in [0,1]$. He places some weight on the rate being θ' and some on $\theta'' \neq \theta'$. In this case, $m_\pi(0) = \{0\}$ and $m_\pi(1) = \{1\}$, since $P(1|\theta)/P(0|\theta) = \theta/(1-\theta)$ depends on θ. Also, letting $k(y^t)$ denote the number of successes in y^t, $m_\pi(y^t) = \{\tilde{y}^t | k(\tilde{y}^t) = k(y^t)\}$ 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 $\tilde{k} = k$.

the employee is good enough to avoid poor performance—yet π is consistent with the noticed frequency of mediocre and successful performances. In this case, the manager's attentional strategy could simply coarsen outcomes based on "successful" or "unsuccessful" without further distinguishing poor from mediocre performance. So long as there exists a $\theta \in \text{supp}(\pi)$ such that $P(y=0|\theta) = P(y=0|\theta^*) + P(y=-1|\theta^*)$, then the manager's observations are entirely consistent with his theory. Generally, such "censored" models are PIAE: the agent can ignore those outcomes he deems impossible while being able to explain the relative frequencies of those outcomes he anticipates.

2. *Models that neglect predictive signals:* These models treat a subset of signals as independent of outcomes. For instance, an investor might, as in Malmendier and Shanthikumar (2007), neglect the importance of analyst affiliation in interpreting investment advice. Alternatively, consider the managerial example where the manager additionally receives a signal $s \in \{h,l\}$ each period that helps predict the employee's productivity. In truth, $P(y_t = 1|s_t = h,\theta) = \theta + \varepsilon$ and $P(y=1|s=l,\theta) = \theta - \varepsilon$ for some $\varepsilon < \min\{\theta, 1-\theta\}$; the manager, however, thinks $P(y_t = 1|s_t,\theta) = \theta$ regardless of s_t. Such neglectful models are PIAE so long as they are otherwise well calibrated (i.e., the manager can explain the frequency of signals).

We now describe three classes of models that are *not* PIAE, meaning they are prone to incidental learning. These include models that are miscalibrated but emphasize the right distinctions and those that emphasize truly unimportant distinctions or variables.

1. *Uncertain models that correctly specify the set of outcomes but incorrectly specify their probabilities:* When the person is unsure of his optimal action, he will collect the data his model deems necessary to learn. The amount of uncertainty he perceives—which is determined by uncertainty in π and the sensitivity of $P(\cdot|\theta)$ to θ—thus impacts the scope of attention and, ultimately, the stability of his model. Specifically, if no two observations lead to the same beliefs over parameters (what in the appendix we call the *Varying Likelihood Ratio Property*, or VLRP), then the person finds it necessary to separately notice every outcome in order to learn θ. When the agent has incentive to learn θ, he will thus notice that his model is miscalibrated. For example, if the manager is mistakenly convinced that the employee has a certain ability $\theta \neq \theta^*$, then his mistake is PIAE. However, a mistaken theory that θ has support $\{.25,.75\}$ is not PIAE (as discussed above).

2. *"Overly elaborate" models that anticipate too wide a range of outcomes:* These models can be viewed as a counterpoint to "censored" models. Consider again the extended managerial example with three outcomes, $Y = \{-1,0,1\}$. The manager's model is overly elaborate if $P(y=-1|\theta) > 0$ for all $\theta \in \text{supp}(\pi)$ when in fact $P(y=-1|\theta^*) = 0$; that is, the manager

wrongly thinks the employee will sometimes be dismal when in reality her performance always exceeds this overly-pessimistic lower bound. Such models are prone to discovery, so long as the person has an incentive to track the frequency of the different outcomes. In this case, the person will eventually notice that an impossible outcome fails to materialize.

3. *"Over-fit" models that assume the set of predictive signals is wider than it truly is:* These models can be viewed as a counterpoint to those that neglect predictive signals. To illustrate, consider again the employee example with signals, but suppose now $P(y_t = 1|s_t, \theta) = \theta$ for either $s_t \in \{h, l\}$, meaning that the signals are truly useless. A manager with an over-fit model treats these useless signals as informative: he believes, for instance, that $P(y_t = 1|s_t, \theta) = \theta + \varepsilon$ when $s_t = h$ and $P(y_t = 1|s_t, \theta) = \theta - \varepsilon$ when $s_t = l$. Such models are *not* PIAE when there is uncertainty about how useful the signals are (e.g., the manager wants to determine which employee characteristics predict success). In these contexts the person would record the sequence of both signals and resolutions, which would ultimately prove his initial theory false.

The following table summarizes the categorizations above:

	Not PIAE	PIAE
Censored		✓
Predictor Neglect		✓
Miscalibrated with VLRP	✓	
Overly Elaborate	✓	
Over-Fit	✓	

Overall, these results demonstrate a sense in which "overly coarse" models—those that ignore relevant outcomes or predictive signals—have a greater tendency to be stable than "overly fine" models that place importance on truly irrelevant outcomes or signals.

A simple intuition underlies this pattern: when a person thinks a variable does not matter, she is certain about how much it matters—she thinks it does not matter at all. Conversely, when a person thinks a variable does matter, she is often initially uncertain about its importance. It is this uncertainty that enables incidental learning.

6.3 What Factors Influence the Discovery of Errors?

We've seen that some errors are broadly attentionally stable independent of the choice environment yet others are stable only in particular environments. In the latter case, which features of the choice environment promote the discovery of errors?

Again, the main organizing principle behind our results on waking up is the following: uncertainty about the optimal action paves the way for incidental learning, while being dogmatic creates a barrier. Thus, factors that are often intuited as promoting learning—increasing the stakes, decreasing the cost of information gathering, simplifying the choice, etc.—may not do so in our framework (and may even backfire) depending on how they influence the person's perceived uncertainty about the optimal action.

We illustrate this idea through a binary choice example, where we characterize the impact of raising the stakes on incidental learning. Consider a stochastic variant of the stylized choice task from Section 3.2.2: the payoff from option $x \in \{A, B\}$ each round is random and (v_A, v_B) represents the average benefit of each option. Specifically, choice $x_t \in \{A, B\}$ yields utility $\alpha y_t^x - c_x$, where y_t^x are i.i.d. with $\Pr(y_t^x = 1) = v_x$. The person initially has non-degenerate yet misspecified priors π about parameters (v_A, v_B), but he can costlessly observe feedback (y_t^A, y_t^B) each period. Hence, the person must keep track of outcomes if he cares to learn (v_A, v_B).

For example, each period a manager must assign one of two workers to the high-importance task and the other to the low importance task. Here, α parameterizes the stakes of the task assignment (the gravity of assigning the better worker to the more important task) and c parameterizes how much the manager has to compensate the worker to engage in the high-importance task. The manager may have to compensate one worker more than the other, $c^A \neq c^B$, when, for instance, one has better credentials.[31]

What the person attends to and ultimately chooses will depend on both the "stakes" α and his prior π over (v_A, v_B). Let $\underline{v}_x$ and $\bar{v}_x$ denote the minimum and maximum value of v_x in the support of π, respectively.

Definition 9. Let $x, x' \in X$. Option x *is perceived to dominate* x' in terms of gross benefit if $\underline{v}_x > \bar{v}_{x'}$.

If, for instance, A is perceived to dominate B, then the support of the marginal over v_A, $\text{supp}(\pi_A)$, lies strictly above that of v_B: the lowest conceivable expected benefit of A exceeds the highest such benefit of B. Put differently, if the two options cost the same, the person believes that A is

[31] In the notation of Footnote 28, choosing worker $x = A$ for the high-importance task and B for the low-importance task yields utility

$$\theta^A b_H - (1 - \theta^A)\psi_H - c_A + \theta^B b_L - (1 - \theta^B) b_L,$$

while choosing $x = B$ for the high-importance task and A for the low-importance task yields utility

$$\theta^B b_H - (1 - \theta^B)\psi_H - c_B + \theta^A b_L - (1 - \theta^A) b_L.$$

The manager is better off choosing $x = A$ for the high-importance task whenever

$$\theta^A \cdot [b_H - b_L + \psi_H - \psi_L] - c_A > \theta^B \cdot [b_H - b_L + \psi_H - \psi_L] - c_B.$$

So, in this example, $v_x = \theta^x$ and $\alpha = [b_H - b_L + \psi_H - \psi_L]$.

(on average) certainly the right choice. The case where neither option is perceived to dominate the other is equivalent to assuming the supports of v_A and v_B overlap; that is, $\text{co}(\text{supp}(\pi_A)) \cap \text{co}(\text{supp}(\pi_B))$ has positive measure. In this case, even if the two options cost the same, the person is still uncertain about the right choice.

We now derive long-run beliefs and behavior for the case where option A costs more than B: $c_A > c_B \geq 0$. Whether the person discovers his error will depend on the stakes and on whether option A is perceived to dominate option B.

Proposition 8. *Consider the stylized choice task above and define the constants* $\underline{\alpha} \equiv \frac{c_A - c_B}{\overline{v}_A - \underline{v}_B}$ *and* $\overline{\alpha} \equiv \frac{c_A - c_B}{\underline{v}_A - \overline{v}_B}$.

1. *Suppose that neither option is perceived to dominate the other. In this case, increasing* α *promotes incidental learning:*

   (a) *If* $\alpha < \underline{\alpha}$*, then any minimal SAS is such that for all* $t \in \mathbb{N}$*, the agent ignores* (y_t^A, y_t^B) *and chooses* $x_t = B$*. Beliefs remain at prior* π*.*

   (b) *If* $\alpha > \underline{\alpha}$*, then any minimal SAS is such that for all* $t \in \mathbb{N}$*, the person must attend to* (y_t^A, y_t^B)*. If* $(v_A^*, v_B^*) \in \text{supp}(\pi)$*, then beliefs concentrate on* (v_A^*, v_B^*)*. Otherwise, there is no attentionally stable equilibrium.*

2. *Suppose A is perceived to dominate B. In this case, increasing* α *has a non-monotonic impact on incidental learning—increasing* α *to intermediate levels promotes incidental learning, while increasing* α *to high levels creates a barrier to incidental learning:*

   (a) *If* $\alpha < \underline{\alpha}$*, then the minimal SAS is such that for all* $t \in \mathbb{N}$*, the person ignores* (y_t^A, y_t^B) *and chooses* $x_t = B$*. Beliefs remain at prior* π*.*

   (b) *If* $\alpha \in (\underline{\alpha}, \overline{\alpha})$*, then any minimal SAS is such that for all* $t \in \mathbb{N}$*, the agent must attend to* (y_t^A, y_t^B)*. If* $(v_A^*, v_B^*) \in \text{supp}(\pi)$*, then beliefs concentrate on* (v_A^*, v_B^*)*. Otherwise, there is no attentionally stable equilibrium.*

   (c) *If* $\alpha > \overline{\alpha}$*, then any minimal SAS is such that for all* $t \in \mathbb{N}$*, the agent ignores* (y_t^A, y_t^B) *and chooses* $x_t = A$*. Beliefs remain at prior* π*.*

3. *Suppose B is perceived to dominate A. Then* α *has no impact on incidental learning. For all* $\alpha > 0$*, the minimal SAS is such that for all* $t \in \mathbb{N}$*, the agent ignores* (y_t^A, y_t^B) *and chooses* $x_t = B$*. Beliefs remain at prior* π*.*

Comparing the first two cases reveals an interesting non-monotonic relationship between stakes and attention. In case 1, where the person is *ex ante* uncertain which option has the higher benefit

(e.g., choosing between two workers with similar credentials), his attention is monotonic in the extent to which he cares about this benefit, α. With a higher α (e.g., more important tasks), he is more apt to attend to (v_A, v_B) to guide his decisions. By contrast, in case 2 where the person is (subjectively) certain which option has the higher benefit (e.g., choosing between two workers where one has better credentials than the other), his attention is non-monotonic. For intermediate α, the person keeps track of outcomes to determine if the supposedly high-benefit option (A) is worth the higher cost. This data would lead the person to "incidentally" learn that B is optimal and hence that his prior theory was false. However, for high values of α, the person is convinced that the benefit of A is worthwhile no matter its exact value, and thus ignores feedback and wrongly sticks with A.[32] This result contrasts with the intuition that the more one cares about a dimension, the more likely she is to correctly learn about it.

7 Discussion of Potential Limitations and Further Applications

This paper develops a framework for assessing when people "get a clue", emphasizing a fundamental barrier to correcting errors: it often requires that people pay attention to, and remember, data they deem irrelevant. We use this framework to partially characterize when people are more or less likely to discover their mistakes, and to investigate the stability of common erroneous beliefs and psychological biases.

7.1 Potential Limitations

Many of our assumptions about the environment stack the deck in *favor of getting a clue*: (1) vanishingly small costs of attention, (2) repetitive choice contexts that provide sufficient data to identify the true model, and (3) sufficient patience on behalf of the decision maker to record data that may be useful at any point in the future. Our attentional stability criterion can in some sense be thought of as a "stress test": if a person does not discover his erroneous beliefs even in repetitive environments where attention is cheap, they are unlikely to do so in other settings.

However, we simultaneously focus on erroneous models that put zero weight on the true parameter. This, of course, *impedes* getting a clue. We take this focus to accommodate our assumption that attentional costs are negligible, making our results independent of details regarding the attentional cost function: if the person puts positive weight on the true parameter θ^* and faces no attentional costs, he will always learn θ^* whenever he has incentive to do so. But with non-negligible costs of

[32] The logic above suggests, for example, that overoptimistic models of an investment's returns may go uncorrected by investors with low enough risk aversion that they think investing is necessarily optimal. However, they would be discovered and corrected by more discerning, risk-averse investors.

attention, the person need not learn θ^* even if he assigns it positive weight, so long as that weight is small relative to the cost of attention.

Our framework abstracts away from factors beyond channeled attention that could hinder learning. We assume the agent processes information in a fully Bayesian way given his priors. Hence, psychological biases such as confirmation bias or motivated misreading of information do not play a role in our analysis. Such factors surely contribute to the persistence of some erroneous beliefs, yet, by neutralizing them, our framework demonstrates the extent to which channeled attention itself could prevent the discovery of errors.

An additional limitation inheres in our ambition to provide a sharp framework that we and others can broadly apply: our approach ignores non-instrumental factors in determining what draws attention. Especially when we focus on minimal attentional strategies, our framework permits agents to ignore non-instrumental data that they might, in reality, obviously have to notice. While one could embed assumptions about what a person automatically notices as a primitive, our general framework leaves such assumptions to be imposed on an application-by-application basis.

One final feature of our framework—the extent to which agents correct their world view following a "light-bulb moment"—warrants further discussion. Again, our framework is meant to be applied once a misspecified model is provided as a primitive, and we rely on empirical literature in both psychology and economics to suggest these primitives. However, this literature is often silent on whether a particular error is "local" or "global"; that is, if a person must correct an error in one context if he corrects it in another. It is unclear, for example, if a person naive about his self-control problem has varying degrees of naivete depending on the context (e.g., spending vs. exercise) or instead applies the same misspecified model of his self control across all inter-temporal decision problems. In practice, to apply the model consistently (and in a way we think is broadly realistic) we assume that awareness of errors is local. Under this assumption, a person may discover his mistake in one context, but continue to make that same mistake in other contexts. Such a "local learning" assumption can limit or facilitate waking up. Of course, our framework accommodates alternative assumptions on the scope of learning.

7.2 Further Applications

One natural application of our analysis is to shed light on hidden costs of delegation. Suppose a manager must decide each period whether to assign a worker to an easy, moderate, or elite task. The manager thinks for sure the worker should be assigned the easy or moderate task when in fact he should be assigned the elite task. Will the manager get a clue? Not necessarily if she delegates this assignment task to an outside agent: in this case, following the logic of Proposition 2, the manager will never notice that the worker performs better than expected since she does not

track the worker's performance herself. If additionally the manager faces costs and benefits from delegation—e.g., the agent can acquire more precise signals about the worker but requires a fixed payment to do so—then the manager is likely to delegate those tasks with higher stakes. Thus, increasing the stakes is likely to (i) increase the propensity that a manager delegates the task, (ii) increase the accuracy of decisions within the misspecified model, π, and (iii) reduce the likelihood that the manager discovers π is wrong. Delegation improves decisions within a paradigm, but prevents getting a clue when paradigms are incorrect.

Our framework could also shed light on types of effective persuasion—and on how policymakers and researchers might go about debiasing those who are making errors. If the goal is to provide data to convince a person that his model π is wrong, our theory highlights the importance of providing data that is relevant *within* that model. For instance, making it easier to track gym attendance for a person who is naive about his self-control problems may not help him recognize his mistake if he is convinced he goes enough to justify a membership. While many researchers find that debiasing people is particularly difficult (e.g., Soll, Milkman, and Payne 2014), 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

AUGENBLICK, N. AND M. RABIN (2018): "An Experiment on Time Preference and Misprediction in Unpleasant Tasks." *Review of Economic Studies*, forthcoming.

ALI, S.N. (2011): "Learning Self Control." *Quarterly Journal of Economics*, 126(2), 857–893.

AL-NAJJAR, N. AND E. SHMAYA (2015): "Uncertainty and Disagreement in Equilibrium Models." *Journal of Political Economy*, 123, 778–808.

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

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

BENJAMIN, D., M. RABIN, AND C. RAYMOND (2016): "A Model of Non-Belief in the Law of Large Numbers." *Journal of the European Economic Association*, 14(2), 515–544.

BERK, R. (1966): "Limiting Behavior of Posterior Distributions when the Model is Incorrect." *Annals of Mathematical Statistics*, 37(1), 51–58.

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

BORDALO, P., N. GENNAIOLI, AND A. SHLEIFER (2012): "Salience Theory of Choice Under Risk." *The Quarterly Journal of Economics*, 127(3), 1243–1285.

BORDALO, P., N. GENNAIOLI, AND A. SHLEIFER (2013): "Salience and Consumer Choice." *Journal of Political Economy*, 121(5), 803–843.

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

BROADBENT, D. (1958): *Perception and Communication*, Pergamon Press.

BRONNENBERG, B., J. DUBÉ, M. GENTZKOW, AND J. SHAPIRO (2015): "Do Pharmacists Buy Bayer? Informed Shoppers and the Brand Premium." *Quarterly Journal of Economics*, 130(4), 1669–1726.

CAMERER, C., T. HO, AND J. CHONG (2004): "A Cognitive Hierarchy Model of Games." *Quarterly Journal of Economics*, 119(3), 861–898.

CHATER, N. (2018): *The Mind is Flat*, Yale University Press.

CRAWFORD, V. AND N. IRIBERRI (2007): "Level-K Auctions: Can a Non-equilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?" *Econometrica*, 75(6), 1721–1770.

DELLAVIGNA, S., AND U. MALMENDIER (2006): "Paying Not to Go to the Gym." *American Economic Review*, 96, 694–719.

DEMARZO, D. VAYANOS, AND J. ZWIEBEL (2003): "Persuasion Bias, Social Influence, and Uni-Dimensional Opinions." *Quarterly Journal of Economics*, 118, 909–968.

ENKE, B. AND F. ZIMMERMANN (2017): "Correlation Neglect in Belief Formation." *Review of Economic Studies*, forthcoming.

ESPONDA, I. (2008): "Behavioral Equilibrium in Economies with Adverse Selection." *American Economic Review*, 98(4), 1269–1291.

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

EVANS, G. AND S. HONKAPOHJA (2001): *Learning and Expectations in Macroeconomics*, Princeton University Press.

EYSTER, E. AND PICCIONE, M. (2013): "An Approach to Asset Pricing Under Incomplete and Diverse Perceptions." *Econometrica*, 81, 1483–1506.

EYSTER, E. AND M. RABIN (2005): "Cursed Equilibrium." *Econometrica*, 73(5), 1623–1672.

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

EYSTER, E. AND M. RABIN (2014): "Extensive Imitation is Irrational and Harmful." *Quarterly Journal of Economics*, 129(4), 1861–1898.

FEDYK, A. (2017): "Asymmetric Naivete: Beliefs About Self-Control." *Working Paper*.

FRYER, R. AND M. JACKSON (2008): "A Categorical Model of Cognition and Biased Decision-Making." *B. E. Journal of Theoretical Economics*, 8, 1–44.

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

FUDENBERG, D., AND D. LEVINE (1993): "Self-Confirming Equilibrium." *Econometrica*, 61, 523–545.

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

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

GAGNON-BARTSCH, T., AND M. RABIN (2017): "Naive Social Learning, Mislearning, and Unlearning." *Working Paper*.

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

HANDEL, B., AND J. SCHWARTZSTEIN (2018) "Frictions or Mental Gaps: What's Behind the Information We (Don't) Use and When Do We Care?" *Journal of Economic Perspectives*, 32(1): 155–178.

HANNA, R., S. MULLAINATHAN, AND J. SCHWARTZSTEIN (2014): "Learning Through Noticing: Theory and Evidence from a Field Experiment." *Quarterly Journal of Economics*, 129(3), 1311–1353.

HEIDHUES, P., B. KŐSZEGI, AND P. STRACK (2018): "Unrealistic Expectations and Misguided Learning." *Econometrica*, forthcoming.

HONG, H., J. STEIN, AND J. YU (2007): "Simple Forecasts and Paradigm Shifts." *Journal of Finance*, 62, 1207–1242.

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

JEHIEL, P. AND F. KOESSLER (2008): "Revisiting games of incomplete information with analogy-based expectations." *Games and Economic Behavior*, 62(2), 533–557.

KIRMAN, A. (1975): "Learning by Firms About Demand Conditions." In R. Day and T. Groves (Eds), *Adaptive Economic Models*, Academic Press, 137–156.

KŐSZEGI, B. AND A. SZEIDL (2014): "A model of focusing in economic choice." *The Quarterly Journal of Economics*, 128(1), 53–104.

LAIBSON, D. (1997): "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics*, 112(2): 443–477.

LEHMANN, E., AND G. CASELLA (1998): *Theory of Point Estimation*, Springer New York.

LOEWENSTEIN, G., T. O'DONOGHUE, AND M. RABIN (2003): "Projection Bias in Predicting Future Utility." *Quarterly Journal of Economics*, 118(4), 1209–1248.

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

MALMENDIER, U. AND D. SHANTHIKUMAR (2007): "Are Small Investors Naive About Incentives?" *Journal of Financial Economics*, 85, 457–489.

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

MULLAINATHAN, S., J. SCHWARTZSTEIN, AND A. SHLEIFER (2008): "Coarse Thinking and Persuasion." *Quarterly Journal of Economics*, 123, 577–619.

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

O'DONOGHUE, T. AND M. RABIN (1999) "Doing It Now or Later." *American Economic Review*, 89(1), 103–124.

O'DONOGHUE, T. AND M. RABIN (2001) "Choice and Procrastination." *Quarterly Journal of Economics*, 116(1), 121–160.

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

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

RABIN, M. AND J. SCHRAG (1999): "First Impressions Matter: A Model of Confirmatory Bias." *Quarterly Journal of Economics*, 114(1): 37–82.

RABIN, M. AND D. VAYANOS (2010): "The Gambler's and Hot-Hand Fallacies: Theory and Applications." *Review of Economic Studies*, 77: 730–778.

SARGENT, T. (1993): *Bounded Rationality in Macroeconomics*, Oxford University Press.

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

SIMONS, D. AND C. CHABRIS (1999): "Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events." *Perception*, 28, 1059–1074.

SIMONS, D. AND C. CHABRIS (2011): "What People Believe about How Memory Works: A Representative Survey of the U.S. Population." *PLoS ONE*, 6(8), e22757.

SIMS, C. (2003): "Implications of Rational Inattention." *Journal of Monetary Economics*, 50, 665–690.

SOLL, J., K. MILKMAN, AND J. PAYNE (2015). *A User's Guide to Debiasing*. In The Wiley Blackwell Handbook of Judgment and Decision Making (eds G. Keren and G. Wu).

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

WOODFORD, M. (2012): "Inattentive Valuation and Reference-dependent Choice." *Working Paper*.

A Preference-Independent Attentional Explicability

The following propositions are essentially corollaries of Lemma 2 and formalize results described in Section 6.2 on when models are PIAE. See the main text for intuitive descriptions of these results and Appendix B for proofs.

Following the presentation in the main text, we first consider the two classes of models that are PIAE.

1. *"Censored" models that ignore possible outcomes:* We formally define censored models as follows:

   Definition A. 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 A.1. *Suppose the assumptions underlying Lemma 2 hold. If π is censored, then π is PIAE.*

2. *Models that neglect predictive signals:* Consider environments 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, the person may be uncertain about how s predicts r, but is certain about the frequency of s. 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 A.2. *Suppose the assumptions underlying Lemma 2 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 PIAE.*

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

We next consider the three classes of models that are not PIAE.

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 A. 2. 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 $\theta=\theta'$.[33]

   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, each $m_\pi(y)$ is a singleton.

   Proposition A.3. *Suppose the assumptions underlying Lemma 2 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 PIAE.*

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

   Definition A. 3. Model π is *overly elaborate* if $Y(\pi)\supset Y(\theta^*)$ and there exists $y\in Y(\pi)$ such that $m_\pi(y)\cap Y(\theta^*)=\varnothing$.

   Proposition A.4. *Suppose the assumptions underlying Lemma 2 hold. If π is overly elaborate, then π is not PIAE.*

3. *"Over-fit" models that assume the set of predictive signals is wider than it truly is:* Consider the environment introduced before Proposition A.2, above: 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:

[33] 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. Of course, VLRP holds for any family of distributions that satisfy (strict) MLRP.

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

(b) For all $\theta \in \text{supp}(\pi)$, $P(r_t|s_t,\theta)$ depends on both $(s_t^1,\dots,s_t^J)$ and $(s_t^{J+1},\dots,s_t^K)$. 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 $(s^{J+1},\dots,s^K)$ are useful for updating under model π. That is, there exist $s,\tilde{s} \in S$ such that $(s^1,\dots,s^J) = (\tilde{s}^1,\dots,\tilde{s}^J)$, $(s^{J+1},\dots,s^K) \neq (\tilde{s}^{J+1},\dots,\tilde{s}^K)$, 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 A.5. *Suppose the assumptions underlying Lemma 2 hold. If π is over-fit, then π is not PIAE.*

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

B Proofs (Preliminary)

Some proofs consider how results would change or extend with automatic recall.

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

Automatic recall requires the person to distinguish the continuations of any two histories that were previously distinguished.

The following lemma will be useful in establishing explicability in many of the proofs to follow.

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

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

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

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

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

Proof. For any $y^t \in Y^t$, let $k_n(y^t)$ be the count of outcomes y_τ from y^t such that $y_\tau = y_n$. Then

$$\frac{P(y^t|\theta)}{P(y^t|\theta^*)} = \frac{\prod_{n=1}^N P(y_n|\theta)^{k_n(y^t)}}{\prod_{n=1}^N P(y_n|\theta^*)^{k_n(y^t)}} = \left(\frac{\prod_{n=1}^N P(y_n|\theta)^{k_n(y^t)/t}}{\prod_{n=1}^N P(y_n|\theta^*)^{k_n(y^t)/t}}\right)^t = (Z_t)^t, \tag{B.2}$$

where

$$Z_t \equiv \prod_{n=1}^N \left(\frac{P(y_n|\theta)}{P(y_n|\theta^*)}\right)^{k_n(y^t)/t}. \tag{B.3}$$

If $\bar{Z}(\theta,\theta'|\theta^*) < 1$, then there exists $\widetilde{Z} \in \left(\bar{Z}(\theta,\theta'|\theta^*), 1\right)$ such that

$$\left(\frac{Z_t}{\widetilde{Z}}\right)^t \xrightarrow{\text{a.s.}} 0 \tag{B.4}$$

by the strong law of large numbers. Property (B.4) implies that $(Z_t)^t \xrightarrow{\text{a.s.}} 0$ since $\widetilde{Z}^t \xrightarrow{\text{a.s.}} 0$. Similarly, if $\bar{Z}(\theta,\theta'|\theta^*) > 1$, then there exists $\widetilde{Z} \in \left(1, \bar{Z}(\theta,\theta'|\theta^*)\right)$ such that

$$\left(\frac{Z_t}{\widetilde{Z}}\right)^t \xrightarrow{\text{a.s.}} \infty \tag{B.5}$$

by the strong law of large numbers. Property (B.5) implies that $(Z_t)^t \xrightarrow{\text{a.s.}} \infty$ since $\widetilde{Z}^t \xrightarrow{\text{a.s.}} \infty$. 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 1. Note that $\ln(\bar{Z}(\theta,\theta'|\theta^*)) = D(\theta^*\|\theta') - D(\theta^*\|\theta)$, where D is the KL divergence defined in Equation 1. Hence, the three conditions of Lemma B. 1 are equivalent to (1) $D(\theta^*\|\theta') < D(\theta^*\|\theta)$, (2) $D(\theta^*\|\theta') > D(\theta^*\|\theta)$, and (3) $D(\theta^*\|\theta') = D(\theta^*\|\theta)$ and θ or θ' equals θ^*.

Proof of Observation 1

Proof. Supposing $D(\theta^*\|\lambda)$ and $D(\theta^*\|\pi)$ are finite (the case where one is infinite is obvious), $\Pr(h^t|\pi) > 0$ and $\Pr(h^t|\lambda) > 0$ for all h^t in the support of $P(\cdot|\theta^*)$. Now letting $\Theta_\pi^{\min} = \arg\min_{\tilde{\theta}\in\text{supp}(\pi)} D(\theta^*\|\tilde{\theta})$ and $\Theta_\lambda^{\min} = \arg\min_{\tilde{\theta}\in\text{supp}(\lambda)} D(\theta^*\|\tilde{\theta})$, expand

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

Similarly expand $\Pr(h^t|\lambda)$.

As a result,

$$\frac{\Pr(h^t|\pi)}{\Pr(h^t|\lambda)} = \frac{\Pr(h^t|\Theta_\pi^{\min}) \cdot \left[\pi(\Theta_\pi^{\min}) + \frac{\Pr(h^t|\text{supp}(\pi)\setminus\Theta_\pi^{\min})}{\Pr(h^t|\Theta_\pi^{\min})} \cdot (1-\pi(\Theta_\pi^{\min}))\right]}{\Pr(h^t|\Theta_\lambda^{\min}) \cdot \left[\pi(\Theta_\lambda^{\min}) + \frac{\Pr(h^t|\text{supp}(\lambda)\setminus\Theta_\lambda^{\min})}{\Pr(h^t|\Theta_\lambda^{\min})} \cdot (1-\pi(\Theta_\lambda^{\min}))\right]}$$
$$\xrightarrow{\text{a.s.}} \frac{\Pr(h^t|\Theta_\pi^{\min})}{\Pr(h^t|\Theta_\lambda^{\min})} \cdot \frac{\pi(\Theta_\pi^{\min})}{\pi(\Theta_\lambda^{\min})}$$

by successive applications of Lemma B. 1 (abusing $\xrightarrow{\text{a.s.}}$ notation slightly). The result then follows from further successive applications of Lemma B. 1.

■

Proof of Proposition 1

Proof. Let $\phi = (\mathscr{N}, \sigma)$ be an attentionally stable equilibrium given π in the original environment. In the original environment, the history each period looks like

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

In the modified environment, the history each period instead looks like

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

Let $\tilde{h}_1^t$ be the first component of $\tilde{h}^t$, which (for illustration) above is h^t. Now, let $\tilde{\mathscr{N}}$ be defined by

$$\tilde{n}^t(\tilde{h}^t) = \left\{\hat{h}^t \in \tilde{H}^t | \hat{h}_1^t \in n^t(\tilde{h}_1^t)\right\} \ \forall\, t,\ \tilde{h}^t \in \tilde{H}^t.$$

So the person notices the same things under $\tilde{\mathscr{N}}$ that she does under $\mathscr{N}$. Now derive $\tilde{\sigma}$ from σ in the obvious way and let $\tilde{\phi} = (\tilde{\mathscr{N}}, \tilde{\sigma})$.

Note that $\tilde{\phi}$ leads to the same behavior, beliefs, and effective information sets as ϕ. Since ϕ is a SAS, so is $\tilde{\phi}$. And since ϕ is an ASE given π in the original environment, $\tilde{\phi}$ is an ASE given π in the modified environment.

■

Proof of Proposition 2

Proof. Let $\mathscr{N}^{FA}$ be the noticing strategy where a person notices everything (i.e., each n^{tFA} is a singleton) and σ^{FA} be a pure behavioral strategy such that $\phi^{FA} = (\mathscr{N}^{FA}, \sigma^{FA})$ is a SAS. (FA is shorthand here for "full attention".) For each h^t, let $x(h^t)$ equal the x that $\sigma^{FA}(h^t)$ places probability 1 on. To simplify the presentation, suppose that for every t and $\tilde{x} \in X_t$ there exists an h^t that occurs with positive probability according to π under which $x(h^t) = \tilde{x}$.

For each t and $x \in X_t$, let

$$n^t_{\tilde{x}} = \left\{ h^t \in H^t | x(h^t) = \tilde{x} \right\}$$

and define $\mathscr{N}$ by

$$n^t(h^t) = n^t_{x(h^t)} \; \forall \, t, h^t.$$

Under $\mathscr{N}$, the person comes up with a recommendation for herself of what action to take and notices only this recommendation.

It is clear that it is a best response for the person to follow the noticed recommendation and he cannot do better by noticing any more than the recommendation. That is, $\phi = (\mathscr{N}, \sigma)$ is a SAS where $\sigma(n^t_x)$ always places probability 1 on x.

■

Proof of Proposition 3

Proof. 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, $\widehat{N}^k = N^k$ for all $k < \tau$. Second, since N^τ is a finite partition, enumerate its elements arbitrarily by $N^\tau = \{n^\tau_1, \dots, n^\tau_J\}$ for some $J \geq 1$. By assumption, there exits some element $n^\tau \in N^t$ 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^τ_J. There must, however, exist some $n^\tau_i \in N^\tau$, $i \neq J$, such that $P(n^\tau_i|\theta) > 0$ for some $\theta \in \text{supp}(\pi)$. Let $\widehat{N}^\tau$ consist of $J-1$ elements, $\widehat{N}^\tau = \{\hat{n}^\tau_1, \dots, \hat{n}^\tau_{J-1}\}$, such that $\hat{n}^\tau_k = n^\tau_k$ if $k \neq i, J$ and $\hat{n}^\tau_i = n^\tau_i \cup n^\tau_J$. That is, $\widehat{N}^\tau$ is a coarsening of N^τ where the zero probability cell n^τ_J is merged with a positive probability cell, n^τ_i. For periods beyond τ, the partitions $\widehat{N}^{\tau+k}$ are derived from $N^{\tau+k}$ in order to maintain memory consistency given the coarsening in period 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, a contradiction.

■

Proof of Proposition 4

Proof. In text. ■

Proof of Proposition 5

Proof. Let $\mathscr{B} = \text{supp}(\pi^{\beta})$ and let $\mathscr{B}^{\text{marg}} = \{\hat{\beta} \in \mathscr{B} | \exists \hat{c} \in \text{supp}(\pi^{\bar{c}})$ where (3) holds and another $\hat{c} \in \text{supp}(\pi^{\bar{c}})$ where (3) is reversed$\}$. Let $\mathscr{B}^{\text{infra}} = \mathscr{B} \setminus \mathscr{B}^{\text{marg}}$. Order the members of $\mathscr{B}$ as $\hat{\beta}_1 < \hat{\beta}_2 < \cdots < \hat{\beta}_{|\mathscr{B}|}$ and the members of $\text{supp}(\pi^{\bar{c}})$ as $\hat{c}_1 < \hat{c}_2 < \cdots < \hat{c}_{|\text{supp}(\pi^{\bar{c}})|}$. As argued in the text, the non-trivial case is where inequality (3) holds for some $(\hat{\beta}, \hat{c})$ in the support of π and is reversed for another. So the interesting case is where $\mathscr{B}^{\text{marg}} \neq \emptyset$.

We will construct an ASE using a SAS along the lines of the second one described in the text surrounding the statement of this proposition—where the person tracks overall costs.

In the first period, the person chooses whether to purchase a membership in a (myopically) subjectively optimal manner. She then notices if $\beta < \hat{\beta}_2, \beta \in [\hat{\beta}_2, \hat{\beta}_3), \beta \in [\hat{\beta}_3, \hat{\beta}_4), \ldots, \beta \geq \hat{\beta}_{|\mathscr{B}|}$. Given her subjective uncertainty, she then believes she knows β for sure. Let $\tilde{\beta}$ denote this belief and $\tilde{\tau} = (1 - \tilde{\beta})b$.

If $\tilde{\beta} \in \mathscr{B}^{\text{infra}}$, then the person does not feel she needs to track costs since she thinks she knows whether a membership is worth it. She does not attend to anything if she does not purchase a membership. She attends to whether $b > c_t + \tau$ if she purchases a membership (to determine if she wants to visit the gym that day). She does not remember anything from previous periods except for $\tilde{\beta}$.

If $\tilde{\beta} \in \mathscr{B}^{\text{marg}}$, then the person feels she needs to track costs. Let $c^L(\tilde{\beta}) \equiv \max\{\hat{c} \in \text{supp}(\pi^{\bar{c}}) |$ inequality (3) holds given $(\tilde{\beta}, \hat{c})\}$. In any period where she purchases a membership (and has not yet determined that $\bar{c} > c^L(\tilde{\beta})$), she can attend to whether $b > c_t + \tau$ (to determine if she wants to visit the gym that day) and, if not, place $c_t + \tau$ in one of the intervals $[\hat{c}_1 + \tilde{\tau}, \hat{c}_2 + \tilde{\tau}), [\hat{c}_2 + \tilde{\tau}, \hat{c}_3 + \tilde{\tau}), \ldots, [\hat{c}_{|\text{supp}(\pi^{\bar{c}})-1|} + \tilde{\tau}, \hat{c}_{|\text{supp}(\pi^{\bar{c}})|})$ or note whether it lays outside these intervals (to update her beliefs over $\bar{c}$). If she does not purchase a gym membership and has not yet determined that $\bar{c} > c^L(\tilde{\beta})$, then she can just attend to which of the above intervals $c_t + \tau$ lies in (to update her beliefs over $\bar{c}$). From previous periods, she remembers $\tilde{\beta}$ and her beliefs over $\bar{c}$.

The constructed SAS is an ASE. ■

Proof of Proposition 6

Proof. We first show that non-doctrinaire priors (i.e., they are given by continuous density functions that are non-zero on interior points $(.5, 1)$) imply that a SAS requires the agent to notice data

informative about (θ^1, θ^2). To do so, we in fact show that a condition weaker than non-doctrinaire is sufficient.

Consider a model π with supports over θ^1 and θ^2 that share common minimum and maximum values, denoted by $\underline{\theta}$ and $\overline{\theta}$, respectively, and define the sets $Q_A \equiv \left[\frac{\underline{\theta}^2}{(1-\underline{\theta})^2}, \frac{\overline{\theta}^2}{(1-\overline{\theta})^2}\right]$, $Q_B \equiv \left[\frac{(1-\overline{\theta})^2}{\overline{\theta}^2}, \frac{(1-\underline{\theta})^2}{\underline{\theta}^2}\right]$, $Q_M \equiv \left[\frac{\underline{\theta}(1-\overline{\theta})}{\overline{\theta}(1-\underline{\theta})}, \frac{\overline{\theta}(1-\underline{\theta})}{\underline{\theta}(1-\overline{\theta})}\right]$, and $\mathscr{Q} \equiv Q_A \cup Q_B \cup Q_M$. We show that if $\bar{q} \equiv (1 - q_A)/q_A \in \mathscr{Q}$, then the agent must update her beliefs over each θ^i, $i \in \{1,2\}$. To see this, note that the agent must record information about (θ^1, θ^2) whenever there exist (θ^1, θ^2) and $(\tilde{\theta}^1, \tilde{\theta}^2)$ both in $\text{supp}(\pi)$ and a signal realization (s^1, s^2) such that, given (s^1, s^2), $x = A$ is optimal under (θ^1, θ^2) and $x = B$ is optimal under $(\tilde{\theta}^1, \tilde{\theta}^2)$. Additionally, conditional on (s^1, s^2), the agent takes A under (θ^1, θ^2) iff $\Pr(\omega = A | s^1, s^2, \theta^1, \theta^2) \geq \Pr(\omega = B | s^1, s^2, \theta^1, \theta^2)$ iff

$$\frac{\Pr(s^1, s^2 | \omega = A, \theta^1, \theta^2)}{\Pr(s^1, s^2 | \omega = B, \theta^1, \theta^2)} \geq \bar{q}. \tag{B.6}$$

We now derive values of $\bar{q}$ such that the optimal strategy given (s^1, s^2) varies across parameters (θ^1, θ^2) in $\text{supp}(\pi)$. Starting with $(s^1, s^2) = (A, A)$, the strategy varies in (θ^1, θ^2) if A is optimal under the parameters most supportive of action A following signal (A, A) and B is optimal under the parameters least supportive of A. From (B.6), this is equivalent to

$$\bar{q} \in \left[\frac{\underline{\theta}^2}{(1-\underline{\theta})^2}, \frac{\overline{\theta}^2}{(1-\overline{\theta})^2}\right] = Q_A. \tag{B.7}$$

Turning to $(s^1, s^2) = (B, B)$, the optimal strategy given (B, B) varies in (θ^1, θ^2) if

$$\bar{q} \in \left[\frac{(1-\overline{\theta})^2}{\overline{\theta}^2}, \frac{(1-\underline{\theta})^2}{\underline{\theta}^2}\right] = Q_B. \tag{B.8}$$

Finally, for a mixed signal $(s^1, s^2) \in \{(A, B), (B, A)\}$, the optimal strategy given such a signal varies in (θ^1, θ^2) if

$$\bar{q} \in \left[\frac{\underline{\theta}(1-\overline{\theta})}{\overline{\theta}(1-\underline{\theta})}, \frac{\overline{\theta}(1-\underline{\theta})}{\underline{\theta}(1-\overline{\theta})}\right] = Q_M. \tag{B.9}$$

Thus, if $\bar{q} \in \mathscr{Q} = Q_A \cup Q_B \cup Q_M$, then there necessarily exist (θ^1, θ^2) and $(\tilde{\theta}^1, \tilde{\theta}^2)$ in $\text{supp}(\pi)$ and signal realization (s^1, s^2) such that, given (s^1, s^2), A is optimal under (θ^1, θ^2) and B is optimal under $(\tilde{\theta}^1, \tilde{\theta}^2)$. Non-doctrinaire priors imply $\underline{\theta} = .5$ and $\overline{\theta} = 1$, which in turn implies $\mathscr{Q} = [0, \infty)$ and hence guarantees $\bar{q} \in \mathscr{Q}$.

We now consider when π is part of an attentionally stable equilibrium. For notational ease, we focus on the case where $q_A = 1/2$ (note that $q_A = 1/2 \Rightarrow \bar{q} \in \mathscr{Q}$ for any $\underline{\theta}, \overline{\theta} \in (.5, 1)$ such that $\underline{\theta} < \overline{\theta}$). The proof for $q_A \neq 1/2$ is similar.

We first consider the case of automatic recall (Definition B. 1). Since a SAS requires the person to continually update beliefs over θ (shown above), the agent must track whether each $i \in \{1,2\}$ was correct when $r_t = \omega_t$ and whether the two parties agree when $r_t = \varnothing$. Thus the various sufficient statistics $m_\pi((s^1,s^2,r))$ are:

$$\begin{aligned}
m_\pi((A,A,A)) &= \{(A,A,A),(B,B,B)\},\\
m_\pi((A,B,A)) &= \{(A,B,A),(B,A,B)\},\\
m_\pi((B,A,A)) &= \{(B,A,A),(A,B,B)\},\\
m_\pi((B,B,A)) &= \{(B,B,A),(A,A,B)\},\\
m_\pi((A,B,\varnothing)) &= \{(A,B,\varnothing),(B,A,\varnothing)\},\\
m_\pi((A,A,\varnothing)) &= \{(A,A,\varnothing),(B,B,\varnothing)\}.
\end{aligned}$$

Under automatic recall, a minimal SAS must distinguish $m_\pi(y)$ each period. Enumerate the elements of $m_\pi(\cdot)$ above by $\{m_\pi^1,\ldots,m_\pi^N\}$. For any $y^t \in Y(\pi)^t$, let $k_n(y^t)$ be the count of outcomes y_τ from y^t such that $y_\tau \in m_\pi^n$. Then for any $\theta \in \text{supp}(\pi)$,

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

Likelihood ratio (B.10) is identical to the one considered in Lemma B. 1 aside from the fact that we take the "noticed" outcome space in this case to be $\{m_\pi^1,\ldots,m_\pi^N\}$ rather than $Y(\pi)$. As such, we can immediately invoke Lemma B. 1: $\lim_{t\to\infty} P(m_\pi(y^t)|\theta)/P(m_\pi(y^t)|\theta^*) > 0$ with probability 1 given θ^* iff $\bar{Z}(\theta,\theta^*|\theta^*) \geq 1$, where $\bar{Z}(\theta,\theta^*|\theta^*) \equiv \prod_{n=1}^N \left(\frac{P(m_\pi^n|\theta)}{P(m_\pi^n|\theta^*)}\right)^{P(m_\pi^n|\theta^*)}$. From Remark 1, $\bar{Z}(\theta,\theta^*|\theta^*) \geq 1$ iff $D(\theta^*\|\theta) \leq 0$, 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. By Gibb's inequality, $D(\theta^*\|\theta) \leq 0 \Leftrightarrow P_m(\cdot|\theta) = P_m(\cdot|\theta^*)$. Note, however, that $P(m_\pi(A,A,A)|\theta) = \theta^1\cdot\theta^2 = \theta^1((1-\iota)\theta^1+\iota)$ and $P(m_\pi(A,A,A)|\theta^*) = (1-\iota)\theta^1+\iota\theta^1 = \theta^1$. Thus, if θ^* involves $\theta^1 < 1$, then $P(m_\pi(A,A,A)|\theta) < P(m_\pi(A,A,A)|\theta^*)$ and hence $\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)}$ converges to 0 a.s. under θ^*, implying that π is not θ^*-attentionally explicable with respect to the minimal SAS that distinguishes $m_\pi(y)$ each round.

We now consider volitional recall. To assess attentional explicability, we first derive the minimal SAS following any history. For any $y^t \in Y(\pi)^t$, let $f(y^t)$ be the number of rounds with feedback (i.e., $r \neq \varnothing$) and let $k^i(y^t)$ be the number of rounds with feedback in which s^i, $i \in \{1,2\}$, is correct. Let $a(y^t)$ be the number of rounds without feedback in which $s^1 = s^2$. Now consider two distinct histories, y^{t+1} and $\tilde{y}^{t+1}$. Let $f = f(y^{t+1})$, $\tilde{f} = f(\tilde{y}^{t+1})$, $k^i = k^i(y^{t+1})$, $\tilde{k}^i = k^i(\tilde{y}^{t+1})$, $j = j(y^{t+1})$,

and $\tilde{j} = j(\tilde{y}^{t+1})$. Then

$$\frac{\Pr(\tilde{y}^{t+1}|\theta)}{\Pr(y^{t+1}|\theta)} = \frac{\left[\binom{t}{\tilde{f}}\rho^{\tilde{f}}(1-\rho)^{t-\tilde{f}}\right]\left[\binom{\tilde{f}}{\tilde{k}^1}(\theta^1)^{\tilde{k}^1}(1-\theta^1)^{\tilde{f}-\tilde{k}^1}\right]\left[\binom{\tilde{f}}{\tilde{k}^2}(\theta^2)^{\tilde{k}^2}(1-\theta^2)^{\tilde{f}-\tilde{k}^2}\right]}{\left[\binom{t}{f}\rho^{f}(1-\rho)^{t-f}\right]\left[\binom{f}{k^1}(\theta^1)^{k^1}(1-\theta^1)^{f-k^1}\right]\left[\binom{f}{k^2}(\theta^2)^{k^2}(1-\theta^2)^{f-k^2}\right]}\Psi(\tilde{y}^{t+1}, y^{t+1}|\theta),$$

where

$$\Psi(\tilde{y}^{t+1}, y^{t+1}|\theta) = \frac{\Pr(\tilde{j} \text{ agreements in } t-\tilde{f}|\theta)}{\Pr(j \text{ agreements in } t-f|\theta)} = \frac{\binom{t-\tilde{f}}{\tilde{j}}(1-\theta^1-\theta^2+2\theta^1\theta^2)^{\tilde{j}}(\theta^1+\theta^2-2\theta^1\theta^2)^{t-\tilde{f}-\tilde{j}}}{\binom{t-f}{j}(1-\theta^1-\theta^2+2\theta^1\theta^2)^{j}(\theta^1+\theta^2-2\theta^1\theta^2)^{t-f-j}}.$$

Now define $C(\tilde{y}^{t+1}, y^{t+1})$ as the components of $\Pr(\tilde{y}^{t+1}|\theta)/\Pr(y^{t+1}|\theta)$ that are independent of θ:

$$C(\tilde{y}^{t+1}, y^{t+1}) \equiv \frac{\left[\binom{t}{\tilde{f}}\rho^{\tilde{f}}(1-\rho)^{t-\tilde{f}}\right]\left[\binom{\tilde{f}}{\tilde{k}^1}\right]\left[\binom{\tilde{f}}{\tilde{k}^2}\right]\left[\binom{t-\tilde{f}}{\tilde{j}}\right]}{\left[\binom{t}{f}\rho^{f}(1-\rho)^{t-f}\right]\left[\binom{f}{k^1}\right]\left[\binom{f}{k^2}\right]\left[\binom{t-f}{j}\right]}.$$

Hence,

$$\frac{\Pr(\tilde{y}^{t+1}|\theta)}{\Pr(y^{t+1}|\theta)} = C(\tilde{y}^{t+1}, y^{t+1})\left[(\theta^1)^{\tilde{k}^1-k^1}(1-\theta^1)^{(\tilde{f}-\tilde{k}^1)-(f-k^1)}\right]\left[(\theta^2)^{\tilde{k}^2-k^2}(1-\theta^2)^{(\tilde{f}-\tilde{k}^2)-(f-k^2)}\right]$$
$$\times\left[(1-\theta^1-\theta^2+2\theta^1\theta^2)^{\tilde{j}-j}(\theta^1+\theta^2-2\theta^1\theta^2)^{f+j-\tilde{f}-\tilde{j}}\right] \quad \text{(B.11)}$$

Equation B.11 is independent of θ iff each exponent is zero. In turn, this requires $k^1 = \tilde{k}^1$, $f = \tilde{f}$, and $k^2 = \tilde{k}^2$, $j = \tilde{j}$. This implies:

$$m_\pi(y^{t+1}) = \left\{\tilde{y}^{t+1} \in Y^{t+1}(\pi) \mid f(\tilde{y}^{t+1}) = f(y^{t+1}),\ k^1(\tilde{y}^{t+1}) = k^1(y^{t+1}),\ k^2(\tilde{y}^{t+1}) = z^2(y^{t+1}),\ j(\tilde{y}^{t+1}) = f(y^{t+1})\right\}.$$

Thus, any SAS must record summary statistics $f(y^{t+1})$, $k^1(y^{t+1})$, $k^2(y^{t+1})$, and $j(y^{t+1})$.

Given such a SAS, we now analyze whether π is θ^*-attentionally explicable; i.e., whether $\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)} = \frac{P(f,k^1,k^2,j|\theta)}{P(f,k^1,k^2,j|\theta^*)}$ converges to 0 in t (variables f, k^i, and j are implicitly functions of t, but we suppress reference to t to avoid notational clutter). Note that $\frac{P(f,k^1,k^2,j|\theta)}{P(f,k^1,k^2,j|\theta^*)}$ is equal to

$$\frac{\binom{t}{f}\rho^{f}(1-\rho)^{1-f}\cdot\binom{f}{k^1}(\theta^1)^{k^1}(1-\theta^1)^{f-k^1}\cdot\binom{f}{k^2}(\theta^2)^{k^2}(1-\theta^2)^{f-k^2}\Pr(j|f,\theta)}{\binom{t}{f}\rho^{f}(1-\rho)^{1-f}\cdot\binom{f}{k^1}(\theta^1)^{k^1}(1-\theta^1)^{f-k^1}\cdot\Pr(k^2|k^1,f,\theta^*)\Pr(j|f,\theta^*)}$$
$$= R_1(k^1,k^2,f)\cdot R_2(j,f), \quad \text{(B.12)}$$

where

$$R_1(k^1,k^2,f) \equiv \frac{\binom{f}{k^2}(\theta^2)^{k^2}(1-\theta^2)^{f-k^2}}{\Pr(k^2|k^1,f,\theta^*)} \quad \text{and} \quad R_2(j,f) \equiv \frac{\Pr(j|f,\theta)}{\Pr(j|f,\theta^*)}.$$

In turn, we derive the limit of R_1 and R_2 as $t \to \infty$. Beginning with R_1, following t rounds

$$R_1 = \frac{\binom{f}{k^2}(\theta^2)^{k^2}(1-\theta^2)^{f-k^2}}{\Pr(k^2|k^1,f,\theta^*)} = \frac{\binom{f}{k^2}(\theta^2)^{k^2}(1-\theta^2)^{f-k^2}}{\binom{f-k^1}{k^2-k^1}(\iota)^{k^2-k^1}(1-\iota)^{f-k^2}} = Z_t^a \cdot Z_t^b, \tag{B.13}$$

where $Z_t^a = \binom{f}{k^2}/\binom{f-k^1}{k^2-k^1}$ and $Z_t^b = \frac{(\theta^2)^{k^2}(1-\theta^2)^{f-k^2}}{(\iota)^{k^2-k^1}(1-\iota)^{f-k^2}}$. First note that as $t \to \infty$, $k^i/t \to \rho\theta^i$ and $f/t \to \rho$. We will make use of the fact that $\sqrt{2\pi t}\left(\frac{t}{e}\right)^t/t! \to 1$ as $t \to \infty$. As such, the fraction $Z_t^a = \frac{f!}{(f-k^1)!}\frac{(k^2-k^1)!}{k^2!}$ can be written as

$$Z_t^a = W_t \cdot \frac{\sqrt{2\pi f}\left(\frac{f}{e}\right)^f}{\sqrt{2\pi(f-k^1)}\left(\frac{f-k^1}{e}\right)^{f-k^1}} \frac{\sqrt{2\pi(k^2-k^1)}\left(\frac{k^2-k^1}{e}\right)^{k^2-k^1}}{\sqrt{2\pi k^2}\left(\frac{k^2}{e}\right)^{k^2}} \tag{B.14}$$

where

$$W_t \equiv \frac{f!}{(f-k^1)!}\frac{(k^2-k^1)!}{k^2!}\left[\frac{\sqrt{2\pi f}\left(\frac{f}{e}\right)^f}{\sqrt{2\pi(f-k^1)}\left(\frac{f-k^1}{e}\right)^{f-k^1}} \frac{\sqrt{2\pi(k^2-k^1)}\left(\frac{k^2-k^1}{e}\right)^{k^2-k^1}}{\sqrt{2\pi k^2}\left(\frac{k^2}{e}\right)^{k^2}}\right]^{-1} \tag{B.15}$$

is such that $W_t \to 1$ given the fact above. Thus, we can write (B.14) as $Z_t^a = W_t \cdot G_t \cdot (H_t)^t$, where

$$G_t = \sqrt{\frac{\frac{f}{t}\left(\frac{k^2-k^1}{t}\right)}{\left(\frac{f-k^1}{t}\right)\frac{k^2}{t}}} \tag{B.16}$$

with $G_t \to \sqrt{\frac{\theta^2-\theta^1}{(1-\theta^1)\theta^2}}$ and

$$H_t = \frac{\left(\frac{f}{t}\right)^{\left(\frac{f}{t}\right)}\left(\frac{k^2-k^1}{t}\right)^{\left(\frac{k^2-k^1}{t}\right)}}{\left(\frac{f-k^1}{t}\right)^{\left(\frac{f-k^1}{t}\right)}\left(\frac{k^2}{t}\right)^{\left(\frac{k^2}{t}\right)}} \tag{B.17}$$

with

$$H_t \to \bar{H} \equiv \left(\frac{(\theta^2-\theta^1)^{(\theta^2-\theta^1)}}{(1-\theta^1)^{(1-\theta^1)}(\theta^2)^{\theta^2}}\right)^{\rho}. \tag{B.18}$$

Similarly,

$$Z_t^b = \frac{(\theta^2)^{k^2}(1-\theta^2)^{f-k^2}}{(\iota)^{k^2-k^1}(1-\iota)^{f-k^2}} = (I_t)^t, \tag{B.19}$$

where

$$I_t = \frac{(\theta^2)^{\frac{k^2}{t}}(1-\theta^2)^{\frac{f-k^2}{t}}}{(\iota)^{\frac{k^2-k^1}{t}}(1-\iota)^{\frac{f-k^2}{t}}} \tag{B.20}$$

with

$$I_t \to \bar{I} = \left(\frac{(\theta^2)^{\theta^2}(1-\theta^2)^{1-\theta^2}}{(\iota)^{\theta^2-\theta^1}(1-\iota)^{1-\theta^2}}\right)^{\rho}. \tag{B.21}$$

Thus $R_1 = W_t \cdot G_t \cdot (H_t \cdot I_t)^t$, where

$$\lim_{t\to\infty} H_t \cdot I_t = (\bar{H}\cdot\bar{I})^{\rho} = \left(\frac{(\theta^2-\theta^1)^{\theta^2-\theta^1}}{(1-\theta^1)^{1-\theta^1}(\theta^2)^{\theta^2}}\frac{(\theta^2)^{\theta^2}(1-\theta^2)^{1-\theta^2}}{(\iota)^{\theta^2-\theta^1}(1-\iota)^{1-\theta^2}}\right)^{\rho} = 1, \tag{B.22}$$

which follows from the fact that $\theta^2-\theta^1 = \iota(1-\theta^1)$ and $1-\theta^2 = (1-\iota)(1-\theta^1)$. Hence, an argument analogous to Lemma B. 1 implies that $\lim_{t\to\infty}(H_t\cdot I_t)^t$ is positive and finite. Finally, because $\lim_{t\to\infty} W_t\cdot G_t$ is positive and finite, $\lim_{t\to\infty} R_1$ must be as well.

Since R_1 converges to a positive finite value, the limiting behavior of $\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)}$ is determined by $\lim_{t\to\infty} R_2$. Note that under θ, the probability of $s^1 = s^2$ in any given period is $\theta^1\theta^2 + (1-\theta^1)(1-\theta^2) = 1+2\theta^1\theta^2-\theta^1-\theta^2$. Under θ^*, this probability is $(1-\iota)+\iota\theta^1 = 1-\iota(1-\theta^1)$. Thus, following t rounds,

$$R_2 = \frac{\Pr(j|f,\theta)}{\Pr(j|f,\theta^*)} = \frac{\binom{t-f}{j}(1-\theta^1-\theta^2+2\theta^1\theta^2)^j(\theta^1+\theta^2-2\theta^1\theta^2)^{t-f-j}}{\binom{t-f}{j}(1-\iota(1-\theta^1))^j(\iota(1-\theta^1))^{t-f-j}}. \tag{B.23}$$

As $t\to\infty$, $j/t \to (1-\rho)(1-\iota(1-\theta^1))$. Hence, we can write $R_2 = (Z_t)^t$ where

$$Z_t = \frac{(1-\theta^1-\theta^2+2\theta^1\theta^2)^{j/t}(\theta^1+\theta^2-2\theta^1\theta^2)^{(t-f-j)/t}}{(1-\iota(1-\theta^1))^{j/t}(\iota(1-\theta^1))^{(t-f-j)/t}} \tag{B.24}$$

with

$$\lim_{t\to\infty} Z_t \equiv \bar{Z} = \left[\left(\frac{1-(1-\theta^1)(\iota+(2-\iota)\theta^1)}{1-(1-\theta^1)\iota}\right)^{1-(1-\theta^1)\iota}\left(\frac{\iota+(2-\iota)\theta^1}{\iota}\right)^{(1-\theta^1)\iota}\right]^{(1-\rho)}. \tag{B.25}$$

First consider the case of $\rho = 1$. In this case, $\bar{Z} = 1$ and thus (by arguments analogous to Lemma B. 1), $R_2 = (Z_t)^t$ is strictly positive and finite as $t\to\infty$. Hence, there exists $\theta\in\pi$ such that the

likelihood ratio $\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)}$ remains positive as $t \to \infty$, implying that π is θ^*-attentionally explicable.

Now consider $\rho < 1$. Let $q = 1-(1-\theta^1)(\iota+(2-\iota)\theta^1)$ and $q = 1-(1-\theta^1)\iota$. Then $\bar{Z} < 1$ (and hence R_2 tends to zero) iff

$$\left(\frac{q}{p}\right)^p \left(\frac{1-q}{1-p}\right)^{1-p} < 1 \Leftrightarrow p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) > 0. \tag{B.26}$$

Since $p\log\left(\frac{p}{q}\right)+(1-p)\log\left(\frac{1-p}{1-q}\right)$ is the KL divergence from distribution $(q,1-q)$ to distribution $(p,1-p)$, Gibb's inequality implies that this value is strictly positive iff $p \neq q$. Since $p \neq q$ for all θ^1 and ι, condition (B.26) holds, meaning that π is not θ^*-attentionally explicable whenever $\rho < 1$.

■

Proof of Lemma 1

Proof. 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 $(\mathcal{N}, \sigma)$ given π under which: with probability 1 there exists some period of time $\tilde{t}$ after which (1) the noticed history $n^t(h^t)$ discards all information from $\tilde{t}$ on except possibly aspects of the current signal s_t, and (2) the noticing strategy $\mathcal{N}$ lumps together any signal that is impossible under π_t with a signal that is possible under π_t. Under such a SAS, $\Pr(n^t|\pi)/\Pr(n^t|\lambda)$ is bounded away from 0 because the effective history length is essentially bounded at $\tilde{t}$. Hence, an attentionally stable equilibrium given θ^* exists.

If, in addition, there exists a $\tilde{t}' \geq \tilde{t}$ such that the optimal action given π_t is independent of $s_t \in S_t$ for all $t > \tilde{t}'$, then there exists a SAS satisfying automatic recall where the person ignores all information after $\tilde{t}'$. Under this SAS, $\Pr(n^t|\pi)/\Pr(n^t|\lambda)$ is positive and constant in t beyond $\tilde{t}'$. Hence, an attentionally stable equilibrium given θ^* satisfying automatic recall exists.

■

Proof of Proposition 7

Proof. Suppose $\theta^* \notin \text{supp}(\pi)$ and consider a stationary and binary action space $X = \{0,1\}$. Because $\theta^* \notin \text{supp}(\pi)$, $P(\cdot|\theta) \neq P(\cdot|\theta^*)$ for all $\theta \in \text{supp}(\pi)$. Let $y^{t+1} = (y_1,\ldots,y_t) \in \times_{k=1}^t Y_k$ denote the sequence of realized outcomes through period t. For each $t \in \mathbb{N}$, consider a history-dependent utility function u_t defined as follows:

$$u_t(x_t, y_t; h^t) = \begin{cases} \frac{\max_{\theta \in \text{supp}(\pi)} P(y^{t+1}|\theta)}{\max_{\tilde{\theta} \in \text{supp}(\pi^*)} P(y^{t+1}|\tilde{\theta})} & \text{if} \quad \sum_{k=1}^{t} x_k = 0 \\ \frac{\max_{\tilde{\theta} \in \text{supp}(\pi^*)} P(y^{t+1}|\tilde{\theta})}{\max_{\theta \in \text{supp}(\pi)} P(y^{t+1}|\theta)} & \text{if} \quad \sum_{k=1}^{t} x_k = t \\ -1 & \text{if} \quad \sum_{k=1}^{t} x_k \notin \{0, t\}. \end{cases} \tag{B.27}$$

If $\theta^* \notin \pi$, then 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. This SAS yields an attentionally stable equilibrium given π (by Lemma 1), and it is costly given that, under π^*, it is in fact optimal to choose $x_t = 1$ for all t.

■

Proof of Lemma 2

Proof. *Part 1*: We first prove a variant of the claim under automatic recall (AR) (Definition B. 1): Further suppose S is a singleton and $P(y|\theta) \in (0,1) \; \forall \; (y, \theta) \in Y(\pi) \times \text{supp}(\pi)$. Then the theory π is PIAE (with AR) given θ^* if and only if there exists $\theta \in \text{supp}(\pi)$ such that

$$P(m_\pi(y)|\theta) \geq P(m_\pi(y)|\theta^*) \; \forall \; y \in Y(\pi). \tag{B.28}$$

($\Leftarrow$) 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 straightforwardly that condition (B.28) 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., recording $m_\pi(y_t)$ each period) is part of an attentionally stable equilibrium. To see this, consider any history of outcomes $y^t \in Y(\pi)^{t-1}$ and corresponding noticed history $n^t = (m_\pi(y_{t-1}), \ldots, m_\pi(y_1))$. The model π is θ^*-attentionally explicable if for some $\theta \in \text{supp}(\pi)$, $\lim_{t \to \infty} P(n^t|\theta)/P(n^t|\theta^*) > 0$ 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 $y^t \in Y(\pi)^{t-1}$, let $k_n(y^t)$ be the count of outcomes y_τ in y^t such that $y_\tau \in m_\pi^n$. Then for any $\theta \in \text{supp}(\pi)$,

$$\frac{P(n^t(y^t)|\theta)}{P(n^t(y^t)|\theta^*)} = \frac{\prod_{n=1}^{N} P(m_\pi^n|\theta)^{k_n(y^t)}}{\prod_{n=1}^{N} P(m_\pi^n|\theta^*)^{k_n(y^t)}} = \left(\frac{\prod_{n=1}^{N} P(m_\pi^n|\theta)^{k_n(y^t)/t}}{\prod_{n=1}^{N} P(m_\pi^n|\theta^*)^{k_n(y^t)/t}} \right)^t. \tag{B.29}$$

This likelihood ratio is identical to the one considered in Lemma B. 1 except the "noticed" outcome space in this case to be $\{m_\pi^1, \ldots, m_\pi^N\}$ rather than $Y(\pi)$. Hence (as described in the

proof of Proposition 6), Lemma B. 1 and Remark 1 imply $\lim_{t\to\infty} P(n^t|\theta)/P(n^t|\theta^*) > 0$ with probability 1 given θ^* iff $D(\theta^*\|\theta) \leq 0$, 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. By Gibb's inequality, $D(\theta^*\|\theta) \leq 0 \Leftrightarrow P_m(\cdot|\theta) = P_m(\cdot|\theta^*)$. Thus, there exists $\theta \in \text{supp}(\pi)$ such that $\lim_{t\to\infty} P(n^t|\theta)/P(n^t|\theta^*) > 0$ with probability 1 given θ^*, implying that π is part of an attentionally stable equilibrium irrespective of (X,u).

2. Suppose $Y(\pi) \neq Y(\theta^*)$. As such, condition (B.28) 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 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 of outcomes in $Y(\theta^*)$ but outside $Y(\pi)$, defined as $Y^0 \equiv [Y(\theta^*) \cup Y(\pi)] \setminus Y(\pi)$, 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 $P\left(\tilde{m}_\pi^{(n,t)}\middle|\theta\right) = P\left(\tilde{m}_\pi^{(n,t)}\middle|\theta^*\right)\ \forall n = 1,\ldots,N$ and $\forall t \in \mathbb{N}$. Suppose the person merges $y \in Y^0$ with elements of a partition over $Y(\pi)$ according to a randomizing device governed by discrete i.i.d. random variables z_t with support $\mathscr{Z} = \{1,\ldots,N\}$ and mass function $\Pr(z_t = n) = [P(m_\pi^n|\theta) - P(m_\pi^n|\theta^*)]/P^0$. We augment the observation space to $Y \times \mathscr{Z}$. 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 $\Pr(z_t = n)$, it follows that for all $n \in \{1,\ldots,N\}$ and all $t \in \mathbb{N}$, $P\left(\tilde{m}_\pi^{(n,t)}\middle|\theta^*\right) = P(m_\pi^n|\theta^*) + P^0 \cdot \Pr(z_t = n) = P(m_\pi^n|\theta) = P\left(\tilde{m}_\pi^{(n,t)}\middle|\theta\right)$, where the last equality follows from the fact that the only realizations of y_t included in $\tilde{m}_\pi^{(n,t)}$ beyond those in m_π^n have probability zero under θ. Given that the distribution of noticed outcomes under $\tilde{m}_\pi(\cdot)$ is equivalent for both θ and θ^*, the proof concludes along the same lines as the case above with $Y(\pi) = Y(\theta^*)$ aside from the simple difference that the m_π^n's above are replaced with the $\tilde{m}_\pi^{(n,t)}$'s.

($\Rightarrow$) Suppose π is PIAE (with AR) given θ^*. 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, for example, 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 $x_n = \sum_{\theta \in \text{supp}(\pi)} P(y_n|\theta)\pi_t(\theta)$. 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,

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

where the inequality follows from Jensen's inequality. Thus, any π that is PIAE (with AR) must be part of an ASE involving a SAS that records $m_\pi(y_t)$ each round.

To finally establish that condition (B.28) must hold, we proceed by contradiction: suppose condition (B.28) 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 $\Pr(n^t|\theta)/\Pr(n^t|\theta^*) \xrightarrow{\text{a.s.}} 0$ by Remark 1. Thus π is not PIAE (with AR), a contradiction.

Part 2: We now prove the claim without automatic recall.

($\Leftarrow$) For any (X,u), the person believes it is sufficient to notice $m_\pi(y^t)$ each period since this is sufficient for updating beliefs about θ (given S is a singleton). Hence, for any (X,u), noticing $m_\pi(y^t)$ constitutes the noticing strategy for some SAS. By assumption, there exists $\theta \in \text{supp}(\pi)$ such that $\lim_{t\to\infty} P(m_\pi(y^t)|\theta)/P(m_\pi(y^t)|\theta^*) > 0$ with probability 1 under θ^*. Thus, under the noticing strategy described above, $\lim_{t\to\infty} P(n^t|\theta)/P(n^t|\theta^*) = \lim_{t\to\infty} P(m_\pi(y^t)|\theta)/P(m_\pi(y^t)|\theta^*) > 0$ (with probability 1 under θ^*), which implies that π is part of an attentionally stable equilibrium given θ^*. Finally, since (X,u) was arbitrary, π is PIAE given θ^*.

($\Rightarrow$) Suppose π is PIAE given θ^*. The proof follows along the same lines as the analogous result assuming automatic recall, above. First, we show that there exist (X,u) under which any SAS requires the person to notice $m_\pi(y^t)$ 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 with automatic recall, 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, PIAE by definition implies that there exists $\theta \in \text{supp}(\pi)$ and a SAS such that $\lim_{t\to\infty} P(n^t|\theta)/P(n^t|\theta^*) > 0$ with probability 1 under θ^*, which in turn implies $\lim_{t\to\infty} P(m_\pi(y^t)|\theta)/P(m_\pi(y^t)|\theta^*) > 0$ with probability 1 under θ^* since n^t must contain at least as much information as $m_\pi(y^t)$.

■

Proof of Proposition 8

Proof. The proof is organized as follows: we first describe sufficient statistics for h^t based on the perceived support of (v_A, v_B). We then show that in scenarios where the person deems it necessary to attend to outcomes, π is attentionally inexplicable whenever $(v_A^*, v_B^*) \notin \text{supp}(\pi)$. Finally, we apply these results to each case considered in the proposition.

For any noticing strategy $\mathcal{N}$, the person chooses A in round $t \Leftrightarrow \alpha\mathbb{E}[y_t^A - y_t^B|n^t(h^t), \pi] > c^A - c^B$. For $\mathcal{N}$ to be sufficient, we require that for all t, $\alpha\mathbb{E}[y_t^A - y_t^B|n^t(h^t), \pi] > c^A - c^B \Leftrightarrow \alpha\mathbb{E}[y_t^A - y_t^B|h^t, \pi] > c^A - c^B$—that is, the person predicts that his behavior would be identical regardless of whether he notices h^t or $n^t(h^t)$.

Since $\mathbb{E}[y_t^A - y_t^B|h^t, \pi] \in [\underline{v}_A - \overline{v}_B, \overline{v}_A - \underline{v}_B]$, the person finds it useless to attend to h^t whenever they are initially certain about the optimal action. This corresponds to either:

$$\alpha(\underline{v}_A - \overline{v}_B) > c_A - c_B \tag{B.30}$$

or

$$\alpha(\overline{v}_A - \underline{v}_B) < c_A - c_B, \tag{B.31}$$

and we assume that $\underline{v}_A - \overline{v}_B \neq c_A - c_B$ and $\overline{v}_A - \underline{v}_B \neq c_A - c_B$ in order to rule out non-generic cases of indifference. When either (B.30) or (B.31) hold, the minimal SAS is such that for all $t \geq 1$, N^t does not distinguish any h^t. Under this minimal SAS, π is trivially attentionally explicable.

We now describe the minimal SAS when neither (B.30) nor (B.31) hold—i.e., when the person is not initially certain of the optimal action. Note that $\mathbb{E}[y_t^x|h^t, \pi] = \sum_{v\in\text{supp}(\pi_x)} \Pr(v|h^t)v_x$. For all t, let $s_x(h^t) \equiv \sum_{k=1}^{t-1} \mathbf{1}\{y_k^x = 1\}$, so

$$\Pr(v_x|h^t) = \frac{\binom{t-1}{s_x(h^t)} v_x^{s_x(h^t)}(1-v_x)^{t-1-s_x(h^t)}}{\sum_{\tilde{v}_x\in\text{supp}(\pi_x)} \binom{t-1}{s_x(h^t)} \tilde{v}_x^{s_x(h^t)}(1-\tilde{v}_x)^{t-1-s_x(h^t)}} = \frac{v_x^{s_x(h^t)}(1-v_x)^{t-1-s_x(h^t)}}{\sum_{\tilde{v}_x\in\text{supp}(\pi_x)} \tilde{v}_x^{s_x(h^t)}(1-\tilde{v}_x)^{t-1-s_x(h^t)}}. \tag{B.32}$$

Thus, if the person deems updating useful in round t (i.e., under π there exist distinct histories that would lead to different optimal actions), then tracking $s_x(t)$ for each $x \in \{A, B\}$ is sufficient for h^t.

Furthermore, if neither (B.30) nor (B.31) hold, then a minimal SAS must track each $s_x(h^t)$ for all t. To see this, note that $\alpha(\overline{v}_A - \underline{v}_B) > c_A - c_B$ and $\alpha(\underline{v}_A - \overline{v}_B) < c_A - c_B$ implies the existence of a

value $\bar{c} \in (\underline{v}_A - \bar{v}_B, \bar{v}_A - \underline{v}_B)$ such that $\alpha \mathbb{E}[y_t^A - y_t^B | h^t, \pi] > c^A - c^B \iff \mathbb{E}[y_t^A - y_t^B | h^t, \pi] > \bar{c}$. For any h^t such that $\mathbb{E}[y_t^A - y_t^B | h^t, \pi] > \bar{c}$, there exists a continuation history $h^t + k$ with positive probability under π such that $\mathbb{E}[y_t^A - y_t^B | h^{t+k}, \pi] < \bar{c}$. For instance, consider a continuation where $y_{t+j}^A = 0$ and $y_t^B = 1$ for all $j = 1, \ldots, k$ for a sufficiently large finite value k. (A similar logic implies that if $\mathbb{E}[y_t^A - y_t^B | h^t, \pi] < \bar{c}$, there is positive probability that $\mathbb{E}[y_t^A - y_t^B | h^{t+k}, \pi] > \bar{c}$ for sufficiently large k.) As such, there is no finite time at which the agent can stop attending to outcomes from that point on, and therefore a minimal SAS is comprised of noticing partitions N^t with cells characterized by

$$n^t(h^t) = \{\tilde{h}^t | s_x(\tilde{h}^t) = s_x(h^t) \; \forall x \in X\}. \tag{B.33}$$

Finally, if N^t is given by (B.33) for all t, then (following the proof of Proposition A.3) π is attentionally explicable iff $P(\cdot|(v_A, v_B)) = P(\cdot|(v_A^*, v_B^*))$ for some $(v_A, v_B) \in \text{supp}(\pi)$ where $P(\cdot|(v_A, v_B))$ is the probability distribution over (y_t^A, y_t^B). Note that $P(\cdot|(v_A, v_B)) = P(\cdot|(v_A^*, v_B^*))$ iff $v_x = v_x^*$ for each $x = A, B$, and therefore π is attentionally explicable given a SAS that tracks each $s_x(h^t)$ iff $(v_A^*, v_B^*) \in \text{supp}(\pi)$.

To complete the proof, we assess under which conditions both (B.30) and (B.31) fail to hold. As shown above, it is under these conditions (and only these conditions) that there is no ASE under the minimal SAS when $(v_A^*, v_B^*) \notin \text{supp}(\pi)$.

Part 1. Suppose that neither option is perceived to dominate the other. This implies that $\underline{v}_A - \bar{v}_B < 0$ and $\bar{v}_A - \underline{v}_B > 0$.

Case a: If $\alpha \leq \underline{\alpha}$, then $\alpha(\bar{v}_A - \underline{v}_B) \leq \underline{\alpha}(\bar{v}_A - \underline{v}_B) = c_A - c_B$, which implies that Condition (B.31) holds and therefore π is attentionally explicable under the minimal SAS.

Case b: If $\alpha > \underline{\alpha}$, then $\alpha(\bar{v}_A - \underline{v}_B) > \underline{\alpha}(\bar{v}_A - \underline{v}_B) = c_A - c_B$, which implies that Condition (B.31) fails to hold. Furthermore, $\underline{v}_A - \bar{v}_B < 0$ implies that Condition (B.30) fails. Therefore there is no ASE under the minimal SAS when $(v_A^*, v_B^*) \notin \text{supp}(\pi)$.

Part 2. Suppose A is perceived to dominate B. Thus, $\underline{v}_A - \bar{v}_B > 0$.

Case a: If $\alpha \leq \underline{\alpha}$, then $\alpha(\bar{v}_A - \underline{v}_B) \leq \underline{\alpha}(\bar{v}_A - \underline{v}_B) = c_A - c_B$, which implies that Condition (B.31) holds and therefore π is attentionally explicable under the minimal SAS.

Case b: If $\alpha > \underline{\alpha}$, then $\alpha(\bar{v}_A - \underline{v}_B) > \underline{\alpha}(\bar{v}_A - \underline{v}_B) = c_A - c_B$, which implies that Condition (B.31) fails to hold. Furthermore, if $\alpha \leq \overline{\alpha}$, then $\alpha(\underline{v}_A - \bar{v}_B) \leq \overline{\alpha}(\underline{v}_A - \bar{v}_B) = c_A - c_B$, which implies that Condition (B.30) fails to hold. Therefore π is attentionally explicable under the minimal SAS when $(v_A^*, v_B^*) \notin \text{supp}(\pi)$.

Case c: If $\alpha > \overline{\alpha}$, then $\alpha(\underline{v}_A - \bar{v}_B) > \overline{\alpha}(\underline{v}_A - \bar{v}_B) = c_A - c_B$, which implies that Condition (B.30) holds and therefore π is attentionally explicable under the minimal SAS.

Part 3. Suppose B is perceived to dominate A. This implies that $\bar{v}_A - \underline{v}_B < 0$. Thus, Condition (B.31) holds and π is attentionally explicable under the minimal SAS.

■

Proof of Proposition A.1

Proof. 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 PIAE (with AR) from the proof of Lemma 2 holds (Condition B.28). If π is PIAE with AR, then it is PIAE more generally. ■

Proof of Proposition A.2

Proof. 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 π. Any SAS must distinguish the $y = (r,s^1,\ldots,s^K)$ 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 $y_\tau \in m_\pi^n$. Then $\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)}$ is identical to (B.35) from the proof of Corollary A.4. As such, π is PIAE if $P(m_\pi^n|\theta^*) = P(m_\pi^n|\theta)$ for all $n = 1,\ldots,N$. The fact that $P(r|s^1,\ldots,s^J,\theta) = P(r|s^1,\ldots,s^J,\theta^*)$ for all possible $(r,s^1,\ldots,s^J)$ under π along with the definition of m_π implies $P(m_\pi^n|\theta^*) = P(m_\pi^n|\theta)$ for all $n = 1,\ldots,N$, so π is PIAE. ■

Proof of Proposition A.3

Proof. 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 $y^t \in Y(\pi)^{t-1}$ 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

$$\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)} = \frac{\prod_{n=1}^N P(y_n|\theta)^{k_n(y^t)}}{\prod_{n=1}^N P(y_n|\theta^*)^{k_n(y^t)}} = \left(\frac{\prod_{n=1}^N P(y_n|\theta)^{k_n(y^t)/t}}{\prod_{n=1}^N P(y_n|\theta^*)^{k_n(y^t)/t}}\right)^t. \tag{B.34}$$

Hence, Lemma B. 1 along with (B.34) 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 PIAE.

■

Proof of Proposition A.4

Proof. Let m_π be the partition of $Y(\pi)$ defined in (4). Since this partition is unique and finite, enumerate its elements as $\{m_\pi^1, \ldots, m_\pi^N\}$. For any $y^t \in Y(\pi)^{t-1}$, $m_\pi(y^t)$ 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

$$\frac{P(m_\pi(y^t)|\theta)}{P(m_\pi(y^t)|\theta^*)} = \frac{\prod_{n=1}^N P(m_\pi^n|\theta)^{k_n(y^t)}}{\prod_{n=1}^N P(m_\pi^n|\theta^*)^{k_n(y^t)}} = \left(\frac{\prod_{n=1}^N P(m_\pi^n|\theta)^{k_n(y^t)/t}}{\prod_{n=1}^N P(m_\pi^n|\theta^*)^{k_n(y^t)/t}} \right)^t. \tag{B.35}$$

Likelihood ratio (B.35) is identical to the one considered in in Part 1 of Lemma 2 (Equation B.29). 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 $D(\theta^*\|\theta) = 0 \Leftrightarrow P_m(\cdot|\theta) = P_m(\cdot|\theta^*)$, ratio (B.35) converges to 0 a.s. and π is therefore not PIAE. ■

Proof of Proposition A.5

Proof. Following the setup of the proof of Proposition A.4, let m_π be the partition of $Y(\pi)$ defined in (4) and enumerate its elements as $\{m_\pi^1, \ldots, m_\pi^N\}$. Again following the proof of Proposition A.4, any $y^t \in Y(\pi)^{t-1}$, $m_\pi(y^t)$ 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 B.35) 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 (by assumption), the previous inequality implies that $P((r,s)|\theta) \neq P((r,s)|\theta^*)$, and therefore $P_m(m_\pi((r,s))|\theta) \neq P_m(m_\pi((r,s))|\theta^*)$.

■