SELECTIVE ATTENTION AND LEARNING

JOSHUA SCHWARTZSTEIN*

ABSTRACT. What do we notice and how does this affect what we learn and come to believe? I present a model of an agent who learns to make forecasts on the basis of freely available information, but is selective as to which information he attends because of limited cognitive resources. I model the agent's choice of whether to attend to and encode information along a dimension as a function of his current beliefs about whether such information is predictive, taking as given that he attends to information along other dimensions. If the agent does not attend to and encode some piece of information, it cannot be recalled at a later date. He uses Bayes' rule to update his beliefs given encoded information, but does not attempt to fill in missing information. I show that, as a consequence of selective attention, the agent may persistently fail to recognize important empirical regularities, make biased forecasts, and hold incorrect beliefs about the statistical relationship between variables. In addition, I identify factors that make such errors more likely or persistent. The model sheds light on a set of systematic biases in inference, including the difficulty people have in recognizing relationships that prior theories do not make plausible, and the overattribution of cause to salient event features. The model is applied to help understand the formation and stability of erroneous stereotypes, as well as discrimination based on such stereotypes.

Date: November 9, 2009.

* Department of Economics, Harvard University. e-mail: jschwart@fas.harvard.edu. I am deeply grateful to Drew Fudenberg, Sendhil Mullainathan, and Andrei Shleifer for their generous guidance and encouragement throughout this project, and to Dan Benjamin, John Beshears, Pedro Bordalo, Ryan Bubb, Sylvain Chassang, Ian Dew-Becker, Ignacio Esponda, Nicola Gennaioli, Edward Glaeser, Lawrence Katz, Alex Kaufman, Scott Kominers, David Laibson, Ted O'Donoghue, Giacomo Ponzetto, Jeremy Stein, Rick Townsend, Timothy Vogelsang, Glen Weyl, and Xiaoqi Zhu for extremely helpful comments. This research was supported by the National Institute on Aging, Grant Number T32-AG000186 and by a National Science Foundation Graduate Research Fellowship.

1. INTRODUCTION

We learn to make forecasts through repeated observation. Consider an employer learning to predict worker productivity, a loan officer figuring out how to form expectations about trustworthiness and default, or a student learning which study techniques work best. Learning in this manner often relies on what we remember: characteristics of past workers, details of interactions with small business owners, study practices used for particular tests. Standard economic models of learning ignore memory by assuming that we remember everything. However, there is growing recognition of an obvious fact: memory is imperfect.[1] Memory imperfections do not just stem from limited recall of information stored in memory; not all information will be attended to or encoded in the first place.[2] It is hard or impossible to take note of all the characteristics of a worker, every detail of a face-to-face meeting, each aspect of how we study. Understanding what we attend to has important implications for what we come to believe and how we make forecasts. So what do we notice?

In this paper, I present a formal model which highlights a key feature of what we notice in tasks of judgment and prediction: attention is *selective*. A person engages in selective attention when he narrows his attention to event features currently believed to be informative relative to a prediction task. I draw out the consequences of selective attention in a model of an agent who learns to make forecasts on the basis of freely available information. By analyzing how selective attention affects this learning process, the analysis complements existing studies of "rational inattention" (following Sims 2003), which consider how a flow constraint on the amount of information an agent can process affects his response to signals that have a *known* relationship to a decision-relevant variable (Peng and Xiong 2006, Sims 2006, Mackowiak and Wiederholt 2009).[3]

[1] Schacter (2001) provides an excellent overview of the evidence on memory limitations. Early economic models incorporating memory limitations explored optimal storage of information given limited capacity (e.g., Dow 1991) or analyzed decision problems with exogeneous imperfect recall (e.g., Piccione and Rubinstein 1997). Some of the more recent models, beginning with Mullainathan (2002), have incorporated evidence from psychology, neuroscience, and elsewhere to motivate assumptions regarding what people remember (e.g., Shapiro 2006, Gennaioli and Shleifer 2009).

[2] Schacter (2001, Chapter 2) explores research on the interface between attention and memory. See also Kahneman (1973) for a classic treatment on limited attention and DellaVigna (2007) for a recent survey of field evidence from economics on limited attention.

[3] To take an example, Sims (2006) studies how inattention affects the degree to which someone's consumption will respond to information about her net worth (e.g., the status of her retirement account). Note that there is also a conceptual distinction between the constraint on attention proposed by Sims (2003), which limits how many bits of information an agent can process in a given period, and the constraint implicit in my formulation, which limits the number of different variables an agent can encode (Miller 1956).

Due to selective attention, current beliefs affect which variables are attended to and encoded and, consequently, what is learned. A basic insight of the analysis is that, because of this dependence, the agent may persistently fail to pay attention to an important causal (or predictive) factor and, as a result, will not learn how it is related to the outcome of interest.[4] When we go to a doctor to complain about a persistent headache, we may not be able to answer whether it is particularly strong after eating certain foods, not having suspected a food allergy before.[5] A further insight is that this failure feeds back to create a problem akin to omitted variable bias: by not learning to pay attention to a factor, an individual may persistently misreact to an associated factor. Under the model, whether or not he does misreact, and the extent of his misreaction, depends completely on observable features of the environment: these biases are *systematic*.

For example, suppose a student repeatedly faces the task of predicting whether an individual will act friendly in conversation, $y \in \{0, 1\}$, given information about whether or not the individual is a professor, $x \in \{\text{Prof}, \text{Not Prof}\}$, and whether the conversation will take place at a work or recreational situation, $z \in \{\text{Work}, \text{Play}\}$. His encounters with professors are relatively confined to work situations: letting $g(x, z)$ denote the probability mass function over (x, z), $g(\text{Work}|\text{Prof}) > g(\text{Work}|\text{Not Prof})$. Independent of occupation, every individual is always friendly during recreation but never at work:

$$E[y|\text{Prof}, \text{Play}] = E[y|\text{Not Prof}, \text{Play}] = 1$$

$$E[y|\text{Prof}, \text{Work}] = E[y|\text{Not Prof}, \text{Work}] = 0.$$

Suppose, however, that, as a result of selective attention, the student persistently fails to attend to and encode situational factors. Under the additional assumption that he always notices and is later able to recall someone's occupation and whether she acted friendly, he will mistakenly come to believe that professors are less friendly than non-professors because he tends to encounter

[4] Note the relationship to the literature on bandit problems (e.g., Gittins 1979) and self-confirming equilibrium (e.g., Fudenberg and Levine 1993), which emphasizes that it is possible for individuals to maintain incorrect beliefs about the payoff consequences of actions that have rarely been tried and for these beliefs, in turn, to support suboptimal actions. Aragones et al. (2005) offer a different explanation for why people may not learn empirical relationships that is based on the idea that discovering regularities in existing knowledge is computationally complex.

[5] My model of selective attention is related to Rabin and Schrag's (1999) model of confirmatory bias in that both share the feature that an agent's current beliefs influence how he encodes evidence, with the common implication that first impressions can be important. However, confirmatory bias and selective attention are conceptually quite distinct. An agent who suffers from confirmatory bias has a tendency to *distort* evidence to fit his current beliefs, while an agent who engages in selective attention uses current beliefs to guide encoding. I discuss the relationship between confirmatory bias and selective attention in more detail in Section 3.

them in situations that discourage friendliness. More precisely, if a student persistently fails to attend to and encode the situation, his expectation given $(\text{Occupation}, \text{Situation})$ approaches the empirical frequency of friendliness given just (Occupation) which converges to $E[y|\text{Occupation}]$ by the strong law of large numbers. To illustrate, when the likelihood of encountering different (x, z), $g(x, z)$, is given as in Table 1, the student comes to misforecast

$$\hat{E}[y|\text{Not Prof}, \text{Situation}] = \frac{.4}{.4 + .25} = .62$$
$$\hat{E}[y|\text{Prof}, \text{Situation}] = \frac{.1}{.1 + .25} = .29$$

across situations. He will thus overreact to whether someone is a professor and mistakenly come to believe that professors are less friendly.

	Not Prof	Prof
Work	.25	.25
Play	.4	.1

Table 1: The likelihood that the student interacts with an individual of occupation $x \in \{\text{Not Prof}, \text{Prof}\}$ in situation $z \in \{\text{Work}, \text{Play}\}$

These results and others match many experimentally found biases in inference, such as the difficulty people have in recognizing relationships that prior theories do not make plausible and the overattribution of cause to salient event features. By endogeneizing these biases as a consequence of selective attention, the model illuminates conditions under which we should expect them to persist. In particular, under a version of the model where the agent probabilistically attends to information along a dimension, I relate the speed with which he learns to attend to a predictor variable to features of the joint distribution over observables. Importantly, the same features that contribute to greater bias often make the bias more persistent.

Throughout, the model is illustrated with examples concerning the formation of group stereotypes. Like in standard rational statistical discrimination models (Phelps 1972, Arrow 1973), stereotypes are built from experience. However, unlike in rational statistical discrimination models, they may be based on a coarse representation of reality. This implies, for example, that people may persistently misperceive mean differences across groups. There is content to the model because observable features of the environment restrict which misperceptions of mean differences can persist. In other words, while a persistent belief that one group is better than another along some dimension need not reflect an actual difference (conditional on freely available information), it must reflect something. In the context of the earlier example, the student's mistaken belief that professors are relatively unfriendly stems from the fact that his *average* interaction with a professor *is* less pleasant than that with a non-professor.[6]

After presenting the model and results, I illustrate the findings with a variety of other examples. To briefly take one, the model can help make sense of a striking and otherwise puzzling claim made by journalist Michael Lewis in his 2003 book *Moneyball* (Lewis 2003): for decades, people who ran professional baseball teams persistently ignored important components of batter skill (e.g., a batter's ability to get on base by receiving walks) in assessing a player's value to winning games.

Section 2 sets up the formal learning model. An agent learns to predict $y \in \{0, 1\}$ given x and z, where x and z are finite random variables. The agent has a prior belief over mental models specifying whether x (e.g., group identity) and/or z (e.g., situational factors) should be ceteris paribus predictive of y (e.g., whether he will act friendly). Additionally, given a particular mental model, he has prior beliefs over *how* these variables predict y. A standard Bayesian who attends to all details of events eventually learns the true model and makes asymptotically accurate forecasts (Observation 1).

Section 3 introduces selective attention. The likelihood that the agent attends to and encodes z is assumed to be increasing in the current probability he attaches to those mental models which specify z as ceteris paribus predictive of y.[7] In the baseline specification, the agent attends to z if and only if he places sufficient weight on such mental models relative to the fixed degree to which he is cognitively busy.[8] The agent updates his beliefs using Bayes' rule, but, in the spirit of assumptions found in recent work modeling biases in information processing (e.g., Mullainathan 2002, Rabin and Schrag 1999), he is *naive* in the sense that he ignores that a selective failure to attend to z results in a missing data problem that can lead to biased inference. Instead, he uses an update rule which treats a missing value of z as a fixed but distinct non-missing value.

[6] This view of how erroneous group stereotypes may form is consistent with experimental evidence, as detailed in Section 6.

[7] Note the asymmetry between x and z: the agent is assumed to encode x regardless of his beliefs. The interpretation is that some event features, like someone's race, gender, or age, require less top-down attentional effort in order to encode or are particularly *salient* (Fiske 1993). The formal model does not address what makes some features more salient than others.

[8] I do not model optimal cognition, but specify a tractable alternative guided by evidence from psychology. In this manner, my model shares similarities to recent models of costly information acquisition (Gabaix et al. 2006, Gabaix and Laibson 2005), which recognize cognitive limitations, but do not assume that agents optimize given those limitations.

I first establish some basic properties of the learning process. The agent eventually settles on how he mentally represents outcomes: from some period on, he either always encodes z or never encodes z (Proposition 1). The agent is more likely to settle on coarsely representing each period's outcome as $(y, x, \varnothing)$ when he has less of an initial reason to suspect that z is ceteris paribus predictive, or does not devote as much attention to learning to predict y, e.g., he is less motivated or more preoccupied (Proposition 2).

Next, I study limiting forecasts and beliefs given a (settled upon) mental representation. Limiting forecasts must be consistent with the true probability distribution over outcomes as mentally represented (Proposition 3). This implies that there is structure to any limiting biased forecasts: such forecasts can persist only if they are consistent with the true probability distribution over (y, x).[9] The long-run behavior of beliefs over mental models can be described as naively consistent (Proposition 4). When the agent settles on finely representing each period's outcome as (y, x, z), he learns the true model in the sense that he eventually places negligible weight on all mental models other than the true one. However, when the agent settles on coarsely representing each period's outcome as $(y, x, \varnothing)$, then his limiting belief about whether z is ceteris paribus predictive is influenced by his prior since he does not notice variation in z, and, as a consequence of the naivete assumption, his limiting belief about whether x is ceteris paribus predictive is restricted by whether x predicts y *unconditional of* z: if it does, then the agent comes to place full weight on those mental models which specify x as ceteris paribus predictive of y.

Section 4 examines persistent biases that can result from selective attention. First, I show how selective attention can result in the agent effectively suffering from omitted variable bias and persistently over- or underreacting to x depending on features of the joint distribution over (y, x, z) (Proposition 5). Next, interpreting a belief that a variable is ceteris paribus predictive as a belief that it is causally related to the outcome, I show how selective attention can result in misattribution of cause (Corollary 1). In the context of the earlier example, the student becomes convinced that whether someone is a professor influences whether he is friendly and does not just proxy for the situation.

[9] When the agent settles on not encoding z, e.g., situational factors, in my model, then his limiting forecasts will be equivalent to those of a coarse thinker who groups all situational factors together into the same category and applies the same model of inference across members of that category (Mullainathan 2000; Mullainathan, Schwartzstein and Shleifer 2008). Rather than take coarse thinking as given as in much of the previous literature (Eyster and Rabin 2005, Jehiel 2005, Fryer and Jackson 2008, Esponda 2008), I endogeneize it as a potential limiting outcome (or approximate outcome over a reasonable time horizon) of a learning process. As a result, my model has implications regarding which categorizations can persist.

Section 5 develops a result that can be used to study which features of the joint distribution over (y, x, z) make a selective failure to attend to z (and resulting bias) more or less persistent. To do so, Section 5 extends the earlier analysis by assuming there are random fluctuations in the degree to which the agent is cognitively busy in a given period. To make matters as simple as possible, I assume that these fluctuations are such that the likelihood that the agent attends to z varies monotonically and continuously in the intensity of his belief that such processing is decision-relevant. With the continuous attention assumptions, the agent will eventually learn to devote more and more attention to z. The main result of this section (Proposition 7) concerns the speed of convergence: The speed increases in the degree to which the agent finds it difficult to explain what he observes without taking z into account. This is *not* the same as the extent to which the agent misreacts to x by failing to take z into account; the agent may react in a very biased fashion to x but learn very slowly that he should be paying attention to z.

Section 6 presents some illustrative examples. Section 7 considers some basic extensions of the model. Section 8 concludes.

2. Setup and Bayesian Benchmark

2.1. **Setup.** Suppose that an agent is interested in accurately forecasting y given (x, z), where $y \in \{0, 1\}$ is a binary random variable and $x \in \mathsf{X}$ and $z \in \mathsf{Z}$ are finite random variables, which, unless otherwise noted, can each take on at least two values.

- In the earlier example, y represents whether or not an individual will act friendly in conversation, $x \in \{\text{Not Prof}, \text{Prof}\}$ for the individual's occupation, and $z \in \{\text{Work}, \text{Play}\}$ for where the conversation takes place (at work or during recreation).

Each period t, the agent

(1) Observes some draw of (x, z), (x_t, z_t), from fixed distribution $g(x, z)$
(2) Gives his prediction of y, $\hat{y}_t$, to maximize $-(\hat{y}_t - y_t)^2$
(3) Learns the true y_t

The agent knows that, given covariates (x, z), y is independently drawn from a Bernoulli distribution with fixed but *unknown* success probability $\theta_0(x, z)$ each period (i.e., $p_{\boldsymbol{\theta_0}}(y = 1|x, z) = \theta_0(x, z)$). Additionally, he knows the joint distribution $g(x, z)$, which is positive for all (x, z).[10]

[10] The assumption that the agent knows $g(x, z)$ is stronger than necessary. What is important is that he places positive probability on every (x, z) combination and that any learning about $g(x, z)$ is independent of learning about $\boldsymbol{\theta_0}$.

I begin by making an assumption on the (unknown) vector of success probabilities, which makes use of the following definition.

Definition 1. *z is important to predicting y* if and only if there exists x, z, z' such that $\theta_0(x,z) \neq \theta_0(x,z')$. *$x$ is important to predicting y* if and only if there exists x, x', z such that $\theta_0(x,z) \neq \theta_0(x',z)$.

Assumption 1. z is important to predicting y.

I sometimes make the additional assumption that x is not important to predicting y, as in the above example where only situational factors are important to predicting friendliness. Either way, to limit the number of cases considered, I assume that the *unconditional* (of z) success probability depends on x, as in the example where occupation is predictive of friendliness not controlling for situational factors. Formally, defining $p_{\boldsymbol{\theta_0}}(y=1|x) \equiv \sum_{z'} \theta_0(x,z')g(z'|x)$, I make the following assumption.

Assumption 2. $p_{\boldsymbol{\theta_0}}(y=1|x) \neq p_{\boldsymbol{\theta_0}}(y=1|x')$ for some $x, x' \in \mathsf{X}$.

Since the agent does not know $\boldsymbol{\theta_0} = (\theta_0(x',z'))_{x' \in \mathsf{X}, z' \in \mathsf{Z}}$, he estimates it from the data using a hierarchical prior $\mu(\boldsymbol{\theta})$, which is now described.[11] He entertains and places positive probability on each of four different models of the world, $M \in \{M_{X,Z}, M_{\neg X,Z}, M_{X,\neg Z}, M_{\neg X,\neg Z}\} \equiv \mathcal{M}$. These models correspond to whether x and/or z are important to predicting y and each is associated with a prior distribution $\mu^{i,j}(\boldsymbol{\theta})$ ($i \in \{X, \neg X\}, j \in \{Z, \neg Z\}$) over vectors of success probabilities. The vector of success probabilities $\boldsymbol{\theta} = (\theta(x',z'))_{x' \in \mathsf{X}, z' \in \hat{\mathsf{Z}}}$ has dimension $|\mathsf{X}| \times |\hat{\mathsf{Z}}|$, where $\hat{\mathsf{Z}} \supset \mathsf{Z}$. The importance of defining $\hat{\mathsf{Z}}$ will be clear later on when describing selectively attentive forecasts, but, briefly, it will denote the set of ways in which a selectively attentive agent can encode z.

Under $M_{\neg X,\neg Z}$, the success probability $\theta(x,z)$ (e.g., the probability that an individual is friendly) depends on neither x nor z (neither occupation nor the situation):

$$\mu^{\neg X,\neg Z}(\{\boldsymbol{\theta} : \theta(x,z) = \theta(x',z') \equiv \theta \text{ for all } x, x', z, z'\}) = 1,$$

[11] This prior is similar to the one used by Diaconis and Freedman (1993) in studying the consistency properties of non-parametric binary regression. The prior is called hierarchical because it captures several levels of uncertainty: uncertainty about the correct model of the world and uncertainty about the underlying vector of success probabilities given a model of the world.

Models	Parameters	Interpretation
$M_{\neg X,\neg Z}$	θ	Neither x nor z predicts y
$M_{X,\neg Z}$	$(\theta(x'))_{x'\in \mathsf{X}}$	Only x predicts y
$M_{\neg X,Z}$	$(\theta(z'))_{z'\in \hat{\mathsf{Z}}}$	Only z predicts y
$M_{X,Z}$	$(\theta(x',z'))_{(x',z')\in \mathsf{X}\times\hat{\mathsf{Z}}}$	Both x and z predict y

Table 2: Set of Mental Models

so $M_{\neg X,\neg Z}$ is a one parameter model. Under $M_{X,\neg Z}$, $\theta(x,z)$ depends only on x (occupation):

$$\mu^{X,\neg Z}(\{\boldsymbol{\theta} : \theta(x,z)=\theta(x,z')\equiv\theta(x) \text{ for all } x,z,z'\})=1,$$

so $M_{X,\neg Z}$ is a $|\mathsf{X}|$ parameter model. Under $M_{\neg X,Z}$, $\theta(x,z)$ depends only on z (the situation)

$$\mu^{\neg X,Z}(\{\boldsymbol{\theta} : \theta(x,z)=\theta(x',z)\equiv\theta(z) \text{ for all } x,x',z\})=1,$$

so $M_{\neg X,Z}$ is a $|\hat{\mathsf{Z}}|$ parameter model. Finally, under $M_{X,Z}$, $\theta(x,z)$ depends on both x and z (on both occupation and the situation) so it is a $|\mathsf{X}|\times|\hat{\mathsf{Z}}|$ parameter model; i.e., $\mu^{X,Z}(\boldsymbol{\theta})$ places weight on those vectors for which $\theta(x,z)\neq\theta(x,z')$ and $\theta(x',z'')\neq\theta(x'',z'')$ for some x,x',x'',z,z',z''. Table 2 summarizes the four different models. All effective parameters under $M_{i,j}$ are taken as independent with respect to $\mu^{i,j}$ and distributed according to common density, $\psi(\cdot)$.[12] I make a technical assumption on the density ψ which guarantees that a standard Bayesian will have correct beliefs in the limit (Diaconis and Freedman 1990, Fudenberg and Levine 2006).

Assumption 3. The density ψ is *non-doctrinaire*: It is continuous and strictly positive at all interior points.

Denote the prior probability placed on model $M_{i,j}$ by $\pi_{i,j}$ and assume the following

$$\pi_{X,Z}=\pi_X\pi_Z$$
$$\pi_{X,\neg Z}=\pi_X(1-\pi_Z)$$
$$\pi_{\neg X,Z}=(1-\pi_X)\pi_Z$$
$$\pi_{\neg X,\neg Z}=(1-\pi_X)(1-\pi_Z)$$

[12] I provide an alternative, more explicit, description of the agent's prior in Appendix A.1.

for some $\pi_X, \pi_Z \in (0,1]$. π_X is interpreted as the subjective prior probability that x is important to predicting y (e.g., that occupation is important to predicting friendliness); π_Z is interpreted as the subjective prior probability that z is important to predicting y (e.g., that situational factors are important to predicting friendliness).

2.2. **Standard Bayesian.** Denote the history through period t by

$$h^t = ((y_{t-1}, x_{t-1}, z_{t-1}), (y_{t-2}, x_{t-2}, z_{t-2}), \ldots, (y_1, x_1, z_1)).$$

The probability of such a history, given the underlying data generating process, is derived from the probability distribution over infinite horizon histories $h^\infty \in H^\infty$ as generated by $\boldsymbol{\theta_0}$ together with g. I denote this distribution by $P_{\boldsymbol{\theta_0}}$.[13]

The agent's prior, together with g, generates a joint distribution over $\Theta, \mathcal{M}$, and H, where Θ is the set of all possible values of $\boldsymbol{\theta_0}$, $\mathcal{M}$ is the set of possible models, and H is the set of all possible histories. Denote this distribution by $\Pr(\cdot)$.[14] The (standard) Bayesian's beliefs are derived from $\Pr(\cdot)$. His period-t forecast of y given x and z equals

$$E[y|x,z,h^t] = E[\theta(x,z)|h^t] = \sum_{i,j} \pi^t_{i,j} E[\theta(x,z)|h^t, M_{i,j}] \tag{1}$$

$$\xrightarrow{a.s.} \pi^t_{X,Z}\bar{y}_t(x,z) + \pi^t_{X,\neg Z}\bar{y}_t(x) + \pi^t_{\neg X,Z}\bar{y}_t(z) + \pi^t_{\neg X,\neg Z}\bar{y}_t, \tag{2}$$

where

- $\bar{y}_t(x,z)$ equals the empirical frequency of $y = 1$ given (x,z), $\bar{y}_t(x)$ equals the empirical frequency of $y = 1$ collapsed across z, $\bar{y}_t(z)$ equals the empirical frequency of $y = 1$ collapsed across x, and $\bar{y}_t$ denotes the empirical frequency of $y = 1$ collapsed across both x and z.

[13] $P_{\boldsymbol{\theta_0}}$ is defined by setting

$$P_{\boldsymbol{\theta_0}}(E(h^t)) = \prod_{\tau=1}^{t-1} \theta(x_\tau, z_\tau)^{y_\tau}(1-\theta(x_\tau,z_\tau))^{1-y_\tau} g(x_\tau, z_\tau)$$

at each event $E(h^t) = \{\widetilde{h}^\infty : \widetilde{h}^t = h^t\}$.

[14] For any $\widetilde{\Theta} \subset \Theta, M \in \mathcal{M}, h^t \in H$

$$\Pr(h^t, \widetilde{\Theta}, M) = \pi_M \int_{\widetilde{\Theta}} \rho(h^t|\boldsymbol{\theta})\mu^M(d\boldsymbol{\theta})$$

where

$$\rho(h^t|\boldsymbol{\theta}) = \prod_{\tau=1}^{t-1} \theta(x_\tau, z_\tau)^{y_\tau}(1-\theta(x_\tau,z_\tau))^{1-y_\tau} g(x_\tau, z_\tau)$$

- $\pi^t_{i,j} \equiv \Pr(M_{i,j}|h^t)$ equals the posterior probability placed on model $M_{i,j}$.
- Convergence is uniform across histories where (x, z) is encountered infinitely often as a result of the non-doctrinaire assumption (Diaconis and Freedman 1990).[15]

Equation (1) says that the period-t likelihood the Bayesian attaches to $y = 1$ given x and z is asymptotically a weighted average of (i) the empirical frequency of $y = 1$ given (x, z) (e.g., the empirical frequency of the individual being friendly given both occupation and situational factors), (ii) the empirical frequency of $y = 1$ given (x) (e.g., the empirical frequency of the individual being friendly only given occupation), the empirical frequency of $y = 1$ given (z) (e.g., the empirical frequency of the individual being friendly only given situational factors), and the unconditional empirical frequency of $y = 1$ (e.g., the unconditional empirical frequency of the individual being friendly).

Definition 2. The agent *learns the true model* if

(1) Whenever x (in addition to z) is important to predicting y, $\pi^t_{X,Z} \to 1$

(2) Whenever x (unlike z) is *un*important to predicting y, $\pi^t_{\neg X,Z} \to 1$

Observation 1. *Suppose the agent is a standard Bayesian. Then*

(1) $E[y|x, z, h^t] \to E_{\boldsymbol{\theta}_0}[y|x, z]$ *for all* (x, z)*, almost surely with respect to* $P_{\boldsymbol{\theta}_0}$.

(2) *The agent learns the true model, almost surely with respect to* $P_{\boldsymbol{\theta}_0}$.

Proof. Unless otherwise noted, proofs can be found in Appendix B. ■

According to Observation 1 the Bayesian with access to the full history h^t at each date makes asymptotically accurate forecasts. In addition, he learns the true model. In particular, whenever x is unimportant to predicting y his posterior eventually places negligible weight on all models other than $M_{\neg X,Z}$. This latter result may be seen as a consequence of the fact that Bayesian model selection procedures tend not to overfit (see, e.g., Kass and Raftery 1995). In the context of the earlier example, the standard Bayesian will learn that knowledge of situational factors but not whether someone is a professor helps predict friendliness and, over time, will come arbitrarily close to correctly predicting that an individual is always friendly during recreation but never at work.

[15] When $\psi(\theta) \sim \mathrm{U}[0, 1]$, then, for any t, h^t, (2) is an accurate approximation of the agent's period-t forecast to order $\frac{1}{N(x,z)}$, where $N(x, z)$ equals the number of times (x, z) has appeared along history h^t.

3. Selective attention

An implicit assumption underlying the standard Bayesian approach is that the agent perfectly encodes (y_k, x_k, z_k) for all $k < t$. But, if the individual is "cognitively busy" (Gilbert et al. 1988) in a given period k, he may not attend to and encode all components of (y_k, x_k, z_k) because of selective attention (Fiske and Taylor 2008). Specifically, there is much experimental evidence that, under stress, individuals narrow their attention to stimuli perceived to be important in performing a given task (e.g., Mack and Rock 1998, von Hippel et al. 1993).[16] Consequently, at later date t, the agent may only have access to a coarse mental representation of history h^t, denoted by $\hat{h}^t$.

To place structure on $\hat{h}^t$, I make several assumptions. First, I take as given that both y and x are always encoded: selective attention operates only on z. To model selective attention, I assume that the likelihood that the agent attends to and encodes z is increasing in the current probability he attaches to such processing being decision-relevant. Formally, his mental representation of the history is

$$\hat{h}^t = ((y_{t-1}, x_{t-1}, \hat{z}_{t-1}), (y_{t-2}, x_{t-2}, \hat{z}_{t-2}), \ldots, (y_1, x_1, \hat{z}_1)) \tag{3}$$

where

$$\hat{z}_k = \begin{cases} z_k & \text{if } e_k = 1 \text{ (the agent encodes } z_k\text{)} \\ \varnothing & \text{if } e_k = 0 \text{ (the agent does not encode } z_k\text{)} \end{cases} \tag{4}$$

and

$$e_k = \begin{cases} 1 & \text{if } \hat{\pi}_Z^k > b_k \\ 0 & \text{if } \hat{\pi}_Z^k \leq b_k \end{cases} \tag{5}$$

$e_k \in \{0, 1\}$ stands for whether or not the agent encodes z in period k, $0 \leq b_k \leq 1$ captures the degree to which the agent is cognitively busy in period k, and $\hat{\pi}_Z^k$ denotes the probability that the agent attaches to z being important to predicting y in period k. I assume that b_k is a random variable which is independent of (x_k, z_k) and independently drawn from a fixed and known distribution

[16] To take one example, Mack and Rock (1998) describe results from a research paradigm developed by Mack, Rock, and colleagues. In a typical task, participants are asked to judge the relative lengths of two briefly displayed lines that bisect to form a cross. On the fourth trial, an unexpected small object is displayed at the same time as the cross. After that trial, participants are asked whether they observed anything other than the cross. Around 25 percent of participants show 'inattentional blindness'. In the fifth and the sixth trial again only the cross appears. In the seventh, an unexpected object again appears. This time, however, almost all participants notice the object.

across periods. If b_k is distributed according to a degenerate distribution with full weight on some $b \in [0,1]$, I write $b_k \equiv b$ (with some abuse of notation).

When $b_k \equiv 1$ (the agent is always extremely busy), (5) tells us that he never encodes z_k; when $b_k \equiv 0$ (the agent is never busy at all), he always encodes z_k. For most of the paper, I assume that $b_k \equiv b$ for some $b \in (0,1)$ so the agent is always somewhat busy, and, as a result, encodes z if and only if he believes sufficiently strongly that it aids in predicting y. In Section 5 I consider the case where there are random, momentary, fluctuations in the degree to which the agent is cognitively busy in a given period; i.e., b_k is drawn according to a non-degenerate distribution. In this case, the likelihood that the agent attends to z varies more continuously in the intensity of his belief that z is important to predicting y.

For later reference, (4) and (5) (together with the agent's prior as well as an assumption about how b_k is distributed) implicitly define an *encoding rule* $\xi : \mathsf{Z} \times \hat{H} \rightarrow \Delta(\mathsf{Z} \cup \{\varnothing\})$ for the agent, where $\hat{H}$ denotes the set of all possible recalled histories and $\xi(z, \hat{h}^k)[\hat{z}']$ equals the probability (prior to b_k being drawn) that $\hat{z}_k = \hat{z}' \in \mathsf{Z} \cup \{\varnothing\}$ given z and $\hat{h}^k$. In other words, the encoding rule specifies how the agent encodes z given any history.[17]

To derive forecasts and beliefs given coarse history, $\hat{h}^t$, I need to specify how the agent treats missing values of z. I assume that he is naive and ignores any memory imperfections that result from selective attention when drawing inferences. I model this by assuming that the agent's prior treats missing and non-missing information the exact same: it treats $\varnothing$ as if it were a fixed but distinct non-missing value. Before stating the formal assumption, recall that the agent's prior μ is on $[0,1]^{|\mathsf{X}| \times |\hat{\mathsf{Z}}|}$ ($\hat{\mathsf{Z}} \supset \mathsf{Z}$) and that all effective parameters under $M_{i,j}$ are taken as independent with respect to $\mu^{i,j}$. For example, subjective uncertainty regarding $\theta(z')$ and $\theta(z'')$ is independent with respect to $\mu^{\neg X, Z}$ for any $z' \neq z''$ with both z' and z'' in $\hat{\mathsf{Z}}$.

Assumption 4. The agent is *naive* in performing statistical inference: $\hat{\mathsf{Z}} = \mathsf{Z} \cup \{\varnothing\}$.

It is easiest to understand this assumption by comparing the naive agent with the more familiar sophisticated agent. In constrast to the naive agent, a sophisticated agent's prior only needs to be

[17] ξ, $\boldsymbol{\theta_0}$, and g generate a measure $P_{\boldsymbol{\theta_0},\xi}$ over $\hat{H}^\infty$, where $\hat{H}^\infty$ denotes the set of all infinite-horizon recalled histories. In particular, $P_{\boldsymbol{\theta_0},\xi}$ is defined by setting

$$P_{\boldsymbol{\theta_0},\xi}(E(\hat{h}^t)) = \prod_{\tau=1}^{t-1} \sum_{z'} \theta_0(x_\tau, z')^{y_\tau}(1-\theta_0(x_\tau, z'))^{1-y_\tau} g(x_\tau, z') \xi(z', \hat{h}^\tau)[\hat{z}_\tau]$$

at each event $E(\hat{h}^t) = \{\hat{h}'^\infty : \hat{h}'^t = \hat{h}^t\}$. All remaining statements regarding almost sure convergence are with respect to this measure.

over $[0,1]^{|\mathsf{X}|\times|\mathsf{Z}|}$ since he takes advantage of the structural relationship relating the success probability following missing versus non-missing values of z. Thus, whereas the naive agent treats missing and non-missing values of z the exact same for purposes of inference, the sophisticated agent treats missing information differently than non-missing information: He attempts the difficult task of inferring what missing data could have been when updating his beliefs.[18]

I maintain the naivete assumption in what follows because it seems to be more realistic than the assumption that people are sophisticated.[19] It also is in the spirit of assumptions found in recent work modeling biases in information processing (e.g., Mullainathan 2002, Rabin and Schrag 1999). I will highlight which arguments and results rely on this assumption as they arise.

While an individual treats $\varnothing$ as a fixed but distinct non-missing value when drawing inferences, I assume that he is otherwise sophisticated in the sense that he "knows" the conditional likelihood of not encoding z given his encoding rule: His beliefs are derived from $\Pr_\xi(\cdot)$, which is the joint distribution over $\Theta, \mathcal{M}$, and $\hat{H}$ as generated by his prior together with g and ξ.[20] The important feature of an individual being assumed to have such "knowledge" is that, whenever his encoding rule dictates not encoding z_t with positive probability, he places positive probability on the event that he will not encode z_t: He never conditions on (subjectively) zero probability events. While

[18] It may also be helpful to compare the "likelihood functions" applied by naive and sophisticated agents, as implicit in the specification of their priors. For every $\widetilde{\Theta} \subset \Theta, M \in \mathcal{M}, \hat{h}^t \in \hat{H}$, the naive agent applies "likelihood function"

$$\Pr(\hat{h}^t|\widetilde{\Theta}, M) \propto \frac{\int_{\widetilde{\Theta}} \prod_{\tau=1}^{t-1} p_{\boldsymbol{\theta}}(y_\tau|x_\tau, \hat{z}_\tau)\mu^M(d\boldsymbol{\theta})}{\int_{\widetilde{\Theta}} \mu^M(d\boldsymbol{\theta})}, \tag{6}$$

where $p_{\boldsymbol{\theta}}(y=1|x,\hat{z}) = \theta(x,\hat{z})$ for all $(x,\hat{z}) \in \mathsf{X} \times \hat{\mathsf{Z}}$. On the other hand, for every $\widetilde{\Theta} \subset \Theta, M \in \mathcal{M}$, and $\hat{h}^t \in \hat{H}$, the sophisticated agent applies "likelihood function"

$$\Pr^{\mathrm{S}}(\hat{h}^t|\widetilde{\Theta}, M) \propto \frac{\int \prod_{\tau \in \mathcal{E}(t)} p_{\boldsymbol{\theta}}(y_\tau|x_\tau, z_\tau) \prod_{\tau \notin \mathcal{E}(t)} p_{\boldsymbol{\theta}}(y_\tau|x_\tau)\mu^M(d\boldsymbol{\theta})}{\int_{\widetilde{\Theta}} \mu^M(d\boldsymbol{\theta})}, \tag{6S}$$

where $\mathcal{E}(t) = \{k < t : \hat{z}_k \neq \varnothing\}$ equals the set of periods $k < t$ in which the agent encodes z and $p_{\boldsymbol{\theta}}(y=1|x) = \sum_{z' \in \mathsf{Z}} \theta(x,z')g(z'|x)$ equals the unconditional (of z) success probability under $\boldsymbol{\theta}$ as a consequence of Bayes' rule.

[19] See Mullainathan (2002) for evidence that people do not seem to correct for memory limitations when making inferences.

[20] In detail, for any $\widetilde{\Theta} \subset \Theta, M \in \mathcal{M}, \hat{h}^t \in \hat{H}$

$$\Pr_\xi(\hat{h}^t, \widetilde{\Theta}, M) = \pi_M \int_{\widetilde{\Theta}} \rho_\xi(\hat{h}^t|\boldsymbol{\theta})\mu^M(d\boldsymbol{\theta})$$

where

$$\begin{aligned} \rho_\xi(\hat{h}^t|\boldsymbol{\theta}) &= \prod_{\tau=1}^{t-1} \theta(x_\tau, \hat{z}_\tau)^{y_\tau}(1-\theta(x_\tau,\hat{z}_\tau))^{1-y_\tau} g_\xi(x_\tau, \hat{z}_\tau|\hat{h}^\tau) \\ g_\xi(x, \hat{z}|\hat{h}^t) &= \sum_{z'} g(x,z')\xi(z', \hat{h}^t)[\hat{z}]. \end{aligned}$$

there are many other ways to specify the agent's beliefs such that they fulfill this (technical) condition, I make this assumption in order to highlight which departures from the standard Bayesian model drive my results.[21]

Discussion of assumptions. It is worth discussing the assumptions underlying (3)-(5) in a bit more detail. First, note the asymmetry between x and z: the agent is assumed to encode x regardless of his beliefs. This assumption can be thought of as capturing in a simple (albeit extreme) way the idea that information along certain dimensions is more readily encoded than information along others, across many prediction tasks. For example, there is much evidence that people instantly attend to and categorize others on the basis of age, gender, and race (Fiske 1993).[22] While what makes some event features more automatically encoded than others lies outside the scope of the formal analysis, it is reasonable to expect that event features which are useful to making predictions and arriving at utility maximizing decisions in many contexts are likely to attract attention, even when they may not be useful in the context under consideration. For example, gender may be salient in economic interactions because considering gender is useful in social interactions. Consistent with this idea of a spillover effect, the amount of effort required to process and encode information along a stimulus dimension decreases with practice (Bargh and Thein 1985).

Second, note that, since $\varnothing \notin \mathsf{Z}$, individuals do not fill in missing details of events and remember distorted versions but instead represent missing information differently than they would a specific value of z (similar to in Mullainathan 2002). For example, if an individual does not encode situational factors he knows that he cannot remember whether a given conversation took place during work or recreation. It may be helpful to think of the individual as representing events at coarser or finer levels, depending on what he encodes. If he encodes the situation, he represents the event as (Friendliness, Occupation, Work) or (Friendliness, Occupation, Play). If he does not, he represents the event as (Friendliness, Occupation, Real-World Interaction).

Finally, the formalization of selective attention (Equation (5)) has the simplifying feature that whether the agent encodes z depends on his period-k belief about whether it is predictive but not

[21] For example, I could instead assume that the agent believes that he fails to encode z with independent probability f each period. In other words, he believes that the joint distribution over $(x_t, \hat{z}_t)$ equals

$$\hat{g}_t(x, \hat{z}) = \begin{cases} (1-f)g(x, \hat{z}) & \text{for all } x, \hat{z} \neq \varnothing \\ fg(x) & \text{for all } x, \hat{z} = \varnothing \end{cases}$$

for each t.

[22] Researchers have identified "preconcious" or "preattentive" processes that result in some event features being more automatically processed and encoded than others (see, e.g., Bargh 1992 for a review).

his assessment of by how much. I conjecture that my qualitative results for the discrete attention case would continue to hold if I was to relax this assumption. Intuitively, the only real change would be that the agent could not persistently encode z if z is not *sufficiently* predictive, expanding the circumstances under which the agent's limiting forecasts and beliefs would be biased.

3.1. **Beliefs and forecasts.** The probability that the selectively attentive agent assigns to model $M_{i,j}$ in period t is given by

$$\hat{\pi}^t_{i,j} = \Pr_\xi(M_{i,j}|\hat{h}^t).$$

As a result, the probability he assigns to z being important to predicting y is

$$\hat{\pi}^t_Z = \Pr_\xi(M_{\neg X,Z}|\hat{h}^t) + \Pr_\xi(M_{X,Z}|\hat{h}^t)$$

and the probability he assigns to x being important to predicting y is

$$\hat{\pi}^t_X = \Pr_\xi(M_{X,\neg Z}|\hat{h}^t) + \Pr_\xi(M_{X,Z}|\hat{h}^t).$$

His period-t forecast of y given x and z is[23]

$$\hat{E}[y|x,z,\hat{h}^t] = E_\xi[\theta(x,\hat{z})|\hat{h}^t], \tag{7}$$

which converges to

$$\hat{\pi}^t_{X,Z}\bar{y}_t(x,\hat{z}) + \hat{\pi}^t_{X,\neg Z}\bar{y}_t(x) + \hat{\pi}^t_{\neg X,Z}\bar{y}_t(\hat{z}) + \hat{\pi}^t_{\neg X,\neg Z}\bar{y}_t \tag{8}$$

uniformly across those mentally represented histories where $(x,\hat{z})$ appears infinitely often.[24]

Equation (8) says that the period-t likelihood the selectively attentive agent attaches to $y = 1$ given x and z approaches a weighted average of (i) the empirical frequency of $y = 1$ given $(x,\hat{z})$ (e.g., the empirical frequency of the individual being friendly given both occupation and the mental representation of situational factors), (ii) the empirical frequency of $y = 1$ given (x) (e.g., the empirical frequency of the individual being friendly only given occupation), the empirical frequency of $y = 1$ given $(\hat{z})$ (e.g., the empirical frequency of the individual being friendly only

[23] I discuss the agent's period-t forecast in greater detail in Appendix A.

[24] When $\psi(\theta) \sim \mathrm{U}[0,1]$, then, for all $t, \hat{h}^t$, (8) is an accurate approximation of the agent's period-t forecast to order $\frac{1}{N(x,\hat{z})}$, where $N(x,\hat{z})$ equals the number of times $(x,\hat{z})$ has appeared along history $\hat{h}^t$.

given situational factors as mentally represented), and the unconditional empirical frequency of $y = 1$ (e.g., the unconditional empirical frequency of the individual being friendly).

Observation 2. *Suppose the selectively attentive agent is never at all cognitively busy ($b_k \equiv 0$). Then, each period, his forecasts coincide with the Bayesian's: $\hat{E}[y|x, z, \hat{h}^t] = E[y|x, z, h^t]$ for all x, z, h^t, t.*

Proof. Follows directly from definitions. ■

Observation 2 shows that the selective attention model nests the Bayesian one as a special case.

3.2. **Stable mental representations.** I now establish some basic properties of the selective attention learning process for the discrete attention case. First, I show that the agent eventually settles on how he mentally represents events, or, equivalently, on whether he encodes or does not encode z.

Definition 3. The agent *settles on encoding* z if there exists some $\tilde{t}$ such that $e_k = 1$ for all $k \geq \tilde{t}$. The agent *settles on not encoding* z if there exists some $\tilde{t}$ such that $e_k = 0$ for all $k \geq \tilde{t}$.

Proposition 1. *Assuming $b_k \equiv b$ for a constant $b \in [0, 1]$, the agent settles on encoding or not encoding z almost surely.*

The intuition behind Proposition 1 is the following. Suppose that, with positive probability, the agent does not settle on encoding or not encoding z and condition on the event that he does not settle on encoding or not encoding z. Then the agent must encode z infinitely often (otherwise he settles on not encoding z). As a result, he learns that z is important to predicting y almost surely and will eventually always encode z, a contradiction.

Proposition 1 implies that the selective attention learning process is well behaved in the sense that, with probability one, it does not generate unrealistic cycling, where the agent goes from believing that he should encode z, to believing that he should not encode z, back to believing that he should encode z, etc. This implies that to characterize potential long-run outcomes of the learning process, it is enough to study the potential long-run outcomes when the agent does or does not settle on encoding z. Before doing so, I identify factors that influence whether or not the agent settles on encoding z.

Proposition 2. *Suppose $b_k \equiv b$ for a constant $b \in (0, 1)$. Then*

(1) *As* $\pi_Z \to 1$ *the probability that the agent settles on encoding* z *tends towards* 1*. As* $\pi_Z \to 0$ *the probability that the agent settles on not encoding* z *tends towards* 1*.*

(2) *As* $b \to 0$ *the probability that the agent settles on encoding* z *tends towards* 1*. As* $b \to 1$ *the probability that the agent settles on not encoding* z *tends towards* 1*.*

The intuition behind Proposition 2 is the following. As $\pi_Z \to 1$ or $b \to 0$, the "likelihood ratio"

$$(9) \qquad \Lambda(\hat{h}^t) = \frac{\Pr_\xi(\hat{h}^t | z \text{ important})}{\Pr_\xi(\hat{h}^t | z \text{ unimportant})} = \frac{\Pr_\xi(\hat{h}^t | M_{X,Z})\pi_X + \Pr_\xi(\hat{h}^t | M_{\neg X,Z})(1-\pi_X)}{\Pr_\xi(\hat{h}^t | M_{X,\neg Z})\pi_X + \Pr_\xi(\hat{h}^t | M_{\neg X,\neg Z})(1-\pi_X)}$$

would have to get smaller and smaller to bring $\hat{\pi}_Z^t$ below b. But the probability that $\Lambda(\hat{h}^t)$ never drops below some cutoff λ tends towards one as λ approaches zero. In the other direction, as $\pi_Z \to 0$ or $b \to 1$, $\pi_Z < b$ and the agent starts off not encoding z. In this case, the agent never updates his belief about whether z is important to predicting y and settles on not encoding z since, by treating $\varnothing$ as he would a distinct non-missing value of z (the naivete assumption), he forms beliefs as if there has been no underlying variation in z and, consequently, believes that he does not have access to any data relevant to the determination of whether z is important to predicting y. Note that this argument relies on the naivete assumption: If the agent is sophisticated then a greater degree of variation in y conditional on x may provide a subjective signal that there is an underlying unobserved variable (z) that influences the success probability.

Proposition 2 highlights that, unlike with a standard Bayesian, whether the selectively attentive agent ever detects the relationship between z and y and learns to properly incorporate information about z in making predictions depends on the degree to which he initially favors models that include z as a causal or predictive factor. This is consistent with evidence presented by Nisbett and Ross (1980, Chapter 5). As they note, the likelihood that a relationship is detected is increasing in the extent to which prior "theories" put such a relationship on the radar screen. One example they provide is that "few insomniacs are aware of how much more difficult their sleep is made by an overheated room, by the presence of an odd smell, by having smoked a cigarette, or by having engaged in physical exercise or intense mental concentration just before retiring" (Nisbett and Ross 1980, page 110).[25]

[25] Interestingly, the tendency to more readily detect relationships in the data which prior "theories" make plausible may not be confined to humans:

> If a rat is allowed to eat a new-tasting food and then many hours later is made ill ... it will avoid the new food thereafter ... If the animal is made ill several hours after eating a food of *familiar taste but unfamiliar shape*, it does not show subsequent avoidance of the new-shaped food. Conversely,

Proposition 2 also illustrates how the degree to which an agent is cognitively busy (the level of b) when learning to predict an output influences the relationships he detects and, as demonstrated later, the conclusions he draws. This relates to experimental findings that the degree of cognitive load or time pressure influences learning, as does the agent's level of motivation (Fiske and Taylor 2008, Nisbett and Ross 1980). To take one example, Gilbert et al. (1988) had experimental participants watch seven clips of a visibly anxious woman discussing various topics without the audio on. Half of the participants were told that some of the topics were "anxiety-provoking" (e.g., sexual fantasies). The other half were told that all of the topics were rather mundane (e.g., world travel). Additionally, half of the participants were placed under cognitive load while watching the clips. After watching the clips, participants were asked to predict how anxious the woman would feel in various hypothetical situations (e.g., when asked to give an impromptu presentation in a seminar). Participants who were not under cognitive load were sensitive to the topics manipulation - those in the anxious topics condition predicted less future anxiety than did those in the mundane topics condition. In contrast, participants under cognitive load at the time of encoding did not use the situational-constraint information.

3.3. **Long-run forecasts given a mental representation.** Recall that Proposition 1 implies that to characterize potential long-run outcomes of the learning process, it is enough to study the potential long-run outcomes when the agent does or does not settle on encoding z. In this subsection, I characterize the potential long-run forecasts. In the next, I characterize the potential long-run beliefs over mental models.

Proposition 3. *Suppose that $b_k \equiv b$ for a constant $b \in [0, 1]$.*

(1) *If the agent settles on encoding z, then, for each (x, z), $\hat{E}[y|x, z, \hat{h}^t]$ converges to $E_{\boldsymbol{\theta_0}}[y|x, z]$ almost surely.*

(2) *If the agent settles on not encoding z, then, for each (x, z), $\hat{E}[y|x, z, \hat{h}^t]$ converges to $E_{\boldsymbol{\theta_0}}[y|x]$ almost surely.*

if the animal eats food of a new shape and then is shocked immediately afterward, it will learn to avoid eating food of that shape even though it will *not* learn to avoid eating food having a *new taste* that is followed immediately by electric shock. The rat thus may be desecribed as possessing two "theories" useful in its ecology: (1) Distinctive gustatory cues, when followed by delayed gastric distress, should be considered suspect. (2) Distinctive spacial cues, when followed by immediate somatic pain, should be considered suspect. (Nisbett and Ross 1980, page 105)

The intuition behind Proposition 3 is the following. If the agent settles on encoding z, then, from some period on, he finely represents each period's outcome as (y, x, z). On the other hand, if he settles on not encoding z, then, from some period on, he coarsely represents each period's outcome as $(y, x, \varnothing)$ (this is coarser because $\varnothing$ is fixed). Either way, his asymptotic forecasts will be consistent with the true probability distribution over outcomes as mentally represented (his effective observations).

Together with Proposition 1, Proposition 3 implies that forecasts converge and there is structure to any limiting biased forecasts: Such forecasts can persist only if they are consistent with the true probability distribution over (y, x). Returning to the earlier example, incorrectly predicting professors to almost never be friendly cannot persist since such a forecast is inconsistent with any coarse representation of outcomes. On the other hand, incorrectly forecasting professors to only be friendly around 30 percent of the time during recreation can persist because such a prediction is consistent with actual outcomes as averaged across work and recreation.

Note how the predictions of my model are sharper than those of general theories of hypothesis maintenence, like confirmatory bias. The logic of confirmatory bias - i.e., the tendency of individuals to misinterpret new information as supporting previously held hypotheses (Rabin and Schrag 1999) - does not by itself pin down which incorrect beliefs we can expect to persist. For example, if an individual begins with a belief that professors are almost never friendly, then, because of confirmatory bias, he may selectively scrutinize and discount evidence to the contrary (e.g., examples of kind acts on the part of professors) and become more and more convinced in this incorrect hypothesis. However, under my model of selective attention, such an incorrect belief cannot persist because evidence is filtered at the level of mental models of *which* factors influence an outcome and not at the level of hypotheses about *how* those factors influence an outcome. As a result, the selectively attentive agent can only become more and more convinced of hypotheses that are consistent with some coarse representation over outcomes, no matter his initial beliefs.[26]

3.4. **Long-run beliefs given a mental representation.** In this subsection, I consider the agent's long-run beliefs over mental models.

[26] Another way to think of the distinction between Rabin and Schrag's (1999) model of confirmatory bias and my model of selective attention is the following. Their model highlights a general mechanism that helps understand why all sorts of erroneous first impressions can persist or become more strongly held in the face of contradictory or ambiguous data; mine helps explain the persistence of a *specific* erroneous first impression (i.e., that a causal factor is unimportant to prediction) and how this may be responsible for the persistence of a set of systematic biases.

Proposition 4. *Suppose that $b_k \equiv b$ for a constant $b \in [0, 1]$.*

(1) *If the agent settles on encoding z, then he learns the true model almost surely.*

(2) *If the agent settles on not encoding z, then $\hat{\pi}_X^t \overset{a.s.}{\to} 1$ and, for large t, $\hat{\pi}_Z^t \leq b$.*

The first part of Proposition 4 says that when the agent settles on encoding z, then, like the standard Bayesian, he learns the true model.[27] The second part says that when the agent settles on not encoding z, then, almost surely, he eventually places negligible weight on models where x is unimportant to predicting y because the *unconditional* success probability depends on x (recall Assumption 2). On the other hand, the limiting behavior of $\hat{\pi}_Z^t$ is largely unrestricted because he effectively does not observe any variation in z. Interestingly, although the agent "knows" that he sometimes cannot recall z and does not have access to all data, he still becomes convinced that x is important to predicting y. This is because, by *treating* $\varnothing$ as a non-missing value of z (the naivete assumption), he believes he has access to all *relevant* data necessary to determine whether x is important to prediction. Put differently, the agent can identify $\theta_0(x, \varnothing) - \theta_0(x', \varnothing)$ for all x, x', which he considers the same as being able to identify $\theta_0(x, z') - \theta_0(x', z')$ for all x, x' and any $z' \neq \varnothing$.[28]

4. Persistent biases

The results from Section 3 establish that the selectively attentive agent may fail to learn to pay attention to an important causal (or predictive) factor and contrast such an agent's long-run forecasts and beliefs with the standard Bayesian's. In this Section, I explore how a failure to learn to pay attention to a variable creates a problem akin to omitted variable bias, where the agent will persistently and systematically misreact to an associated factor and may mistakenly attribute cause to it as well.

4.1. Misreaction.

In the long run, how will the selectively attentive agent misreact to x when he fails to learn to attend to z? To study this question, it is useful to specialize to the case where x is

[27] A bit more precisely, Proposition 4.1 should be read as saying the following: Suppose that the agent settles on encoding z with positive probability under $P_{\boldsymbol{\theta}_0,\boldsymbol{\xi}}$. Then, conditional on the event that the agent settles on encoding z, he learns the true model almost surely. Proposition 4.2 can similarly be made more precise.

[28] This result can be interpreted as saying that the agent sometimes acts as if he believes that correlation implies cause. This belief has been ranked as "probably among the two or three most serious and common errors of human reasoning" (Gould 1996, page 272).

a binary random variable and $\mathsf{X} = \{0, 1\}$. Define

$$\begin{aligned}\mathcal{R}_x(z') &= E_{\boldsymbol{\theta_0}}[y|x = 1, z'] - E_{\boldsymbol{\theta_0}}[y|x = 0, z'] \\ \mathcal{R}_x &= E_z[\mathcal{R}_x(z)|x = 1] \\ \phi &= \mathrm{Cov}_z(E_{\boldsymbol{\theta_0}}[y|x = 0, z], g(x = 1|z)),\end{aligned}$$

where

- $\mathcal{R}_x(z')$ is the *standard Bayesian's limiting reaction to x conditional on $z = z'$*: It equals the gap between the true conditional expectation of y given $(x, z) = (1, z')$ and that given $(x, z) = (0, z')$.
- $\mathcal{R}_x$ is the *standard Bayesian's average limiting reaction to x* : It equals the expected gap between the true conditional expectation of y given $(x, z) = (1, z')$ and that given $(x, z) = (0, z')$, where the expectation is taken over z' conditional on $x = 1$.[29]
- ϕ is the covariance between the likelihood that $y = 1$ given $(x, z) = (0, z')$ and the likelihood that $x = 1$ given z'. $\phi > 0$ means that z which are associated with $x = 1$ are also associated with $y = 1$; $\phi < 0$ means that z which are associated with $x = 1$ are also associated with $y = 0$. The magnitude $|\phi|$ measures the degree to which variation in z induces a relationship between the expected value of y and the likelihood that $x = 1$.

Additionally, let $\hat{E}[y|x, z] \equiv \lim_{t\to\infty} \hat{E}[y|x, z, \hat{h}^t]$ denote the selectively attentive agent's limiting forecast given (x, z), which almost surely exists by Propositions 1 and 3.

Proposition 5. *Suppose $b_k \equiv b$, the agent settles on not encoding z, and* $\mathsf{X} = \{0, 1\}$. *Then*

(10) $$\hat{\mathcal{R}}_x(z') \equiv \hat{E}[y|x = 1, z'] - \hat{E}[y|x = 0, z'] = \mathcal{R}_x + \frac{\phi}{\mathrm{Var}(x)}$$

almost surely for all z'.

Proposition 5 says that when the agent settles on not encoding z, his limiting reaction to x conditional on $z = z'$, $\hat{\mathcal{R}}_x(z')$, differs from the standard Bayesian's, $\mathcal{R}_x(z')$, in two key ways corresponding to the two terms on the right hand side of (10). When $\phi = 0$, the agent's limiting reaction reduces to the first term, $\mathcal{R}_x$: By persistently failing to encode z, the agent's limiting

[29] $\mathcal{R}_x$ is formally equivalent to what is referred to as the population average treatment effect for the treated in the statistical literature on treatment effects, where $x = 1$ corresponds to a treatment and $x = 0$ to a control.

conditional reaction equals the standard Bayesian's limiting *average* reaction. Thinking of z as a situation, this is one of the distortions exploited in Mullainathan, Schwartzstein, and Shleifer (2008): By grouping distinct situations together in forming beliefs, an agent transfers the informational content of data across situations. For example, the agent may react to a piece of information which is uninformative in a given situation, z, because it is informative in another situation, z'.

When $\phi \neq 0$, the agent's limiting conditional reaction differs from the standard Bayesian's limiting average reaction in an amount and direction determined by ϕ, which can be thought of as the magnitude and direction of omitted variable bias. A non-zero ϕ creates the possibility that, by settling on not encoding z, an agent will conclude a relationship between y and x that (weakly) reverses the true relationship conditional on *any* z' (e.g., that non-professors are always more likely to be friendly than professors when, in reality, they are equally likely conditional on the situation).

Definition 4. Suppose $\hat{\mathcal{R}}_x(z')$ and $\mathcal{R}_x(z')$ have the same sign. Then the agent *overreacts to x at z'* if $|\hat{\mathcal{R}}_x(z')| > |\mathcal{R}_x(z')|$ and *underreacts to x at z'* if $|\hat{\mathcal{R}}_x(z')| < |\mathcal{R}_x(z')|$. He *overreacts to x* if he overreacts to x at all $z' \in \mathsf{Z}$ and *underreacts to x* if he underreacts to x at all $z' \in \mathsf{Z}$.

It is easy to see from Proposition 5 that a selectively attentive agent who fails to learn to pay attention to z can either over- or underreact to x at z', depending on features of the joint distribution over (y, x, z). It is useful to consider factors that influence whether the selectively attentive agent will persistently over- or underreact to x at z' for two special cases: when $\phi = 0$ and when $\mathcal{R}_x(z') = \mathcal{R}_x$ for all $z' \in \mathsf{Z}$.[30]

Special case 1: $\phi = 0$. Consider first the case where $\phi = 0$. From Equation (10), the agent's limiting reaction to x then equals $\hat{\mathcal{R}}_x(z') = \mathcal{R}_x$. To apply the definition of over- or underreaction, suppose $\mathcal{R}_x(z')$ and $\mathcal{R}_x$ have the same sign, say positive. Making this additional assumption, the agent will persistently overreact to x at z' if

$$\mathcal{R}_x > \mathcal{R}_x(z') \tag{11}$$

[30] Note that my definition of over- or underreaction only applies when $\hat{\mathcal{R}}_x(z')$ and $\mathcal{R}_x(z')$ have the same sign. This is because it is difficult to label the phenomenon where the agent mistakenly reacts positively (negatively) to x when the true conditional relationship is negative (positive) as either over- or underreaction. Such a phenomenon is sometimes referred to as Simpson's paradox or association reversal in the statistics literature (Samuels 1993).

and will underreact to x at z' if

$$\mathcal{R}_x < \mathcal{R}_x(z'). \tag{12}$$

To interpret conditions (11) and (12), suppose that $\mathcal{R}_x(z) \geq 0$ for all $z \in \mathsf{Z}$, so we can view $\mathcal{R}_x(z)$ as a measure of the degree to which x is informative given z. Then (11) says that whenever x is less than average informative at z', the agent will overreact to x at z'. Similarly, (12) says that whenever x is more than average informative at z', the agent will underreact to x at z'. This is the sort of over- and underreaction emphasized in the literature on coarse thinking (e.g., Mullainathan 2000, Mullainathan et al. 2008). For example, someone might overreact to past performance information in forecasting the quality of mutual fund managers, z', because such information tends to be more informative in assessing the quality of other professionals (e.g., doctors or lawyers); i.e., other z (Mullainathan et al. 2008).

Special case 2: $\mathcal{R}_x(z') \equiv \mathcal{R}_x$. Now consider the case where $\mathcal{R}_x(z') = \mathcal{R}_x$ for all $z' \in \mathsf{Z}$. From Equation (10), the agent's limiting reaction to x then equals $\hat{\mathcal{R}}_x(z') = \mathcal{R}_x(z') + \phi/\operatorname{Var}(x)$. To apply the definition of over- or underreaction, suppose that $\hat{\mathcal{R}}_x(z')$ and $\mathcal{R}_x$ have the same sign, say positive. Making this additional assumption, it immediately follows that the agent overreacts to x at all z' when $\phi > 0$, and underreacts to x at all z' when $\phi < 0$. The agent will overreact to x at z' when z which are associated with $x = 1$ are also associated with $y = 1$, but will underreact to x at z' when z which are associated with $x = 1$ are negatively associated with $y = 1$. The intuition for this sort of over- and underreaction is familiar from the econometric literature on omitted variable bias.

To take an example, someone who persistently fails to take situational factors, z, into account may overreact to the identity of an organization's leader, x, in predicting whether or not an organizational activity (e.g., coordination among workers) will be successful if higher quality leaders also tend to be "lucky" and placed in more favorable situations (e.g., tend to manage smaller sized groups) than others. Alternatively, he could underreact to the identity of a leader if higher quality leaders tend to be "unfortunate" and placed in less favorable situations than others. In the extreme case where there is no actual variation in quality among leaders, there must be overreaction, creating "the illusion of leadership" (Weber et al. 2001).

4.2. **Misattribution of cause.** The results on misreaction concern long-run forecasts. A related question is to ask what the selectively attentive agent comes to believe about the causal relationship between variables, interpreting $\hat{\pi}_X^t$ (resp. $\hat{\pi}_Z^t$) as the intensity of the agent's belief that x (resp. z) is causally related to y. Proposition 4 established that, in the limit, the agent will attribute cause to a factor whenever he reacts to it. Corollary 1 emphasizes an implication of this result, namely that the selectively attentive agent will attribute cause to a factor even when it only proxies for selectively unattended to predictors.

Corollary 1. *Suppose the conditions of Proposition 5 hold and, additionally, x is unimportant to predicting y. Then, so long as $\phi \neq 0$,*

(1) $|\hat{\mathcal{R}}_x(z')| = \frac{|\phi|}{\mathrm{Var}(x)} \neq 0$ *almost surely for all z': The agent overreacts to x and the extent of overreaction is increasing in $\frac{|\phi|}{\mathrm{Var}(x)}$.*

(2) $\hat{\pi}_X^t \overset{a.s.}{\to} 1$*: The agent becomes certain that x is important to predicting y even though it is not.*

Proof. By the assumption that x is unimportant to predicting y, $\mathcal{R}_x(z') = 0$ for all z' so $\mathcal{R}_x = 0$. Then, by Proposition 5,

$$\hat{\mathcal{R}}_x(z') = E_{\boldsymbol{\theta_0}}[y|x=1] - E_{\boldsymbol{\theta_0}}[y|x=0] \tag{13}$$

$$= \frac{\phi}{\mathrm{Var}(x)}, \tag{14}$$

which establishes the first part of the Proposition. Additionally, $E_{\boldsymbol{\theta_0}}[y|x=1] - E_{\boldsymbol{\theta_0}}[y|x=0] \neq 0$ whenever $\phi \neq 0$ (by (14)) and the second part of the Proposition then follows from Proposition 4. ■

Corollary 1 considers the situation where x is completely unimportant to prediction and the selectively attentive agent settles on not encoding z. The first part says that, as a result of the possibility that the selectively attentive agent will settle on not encoding z, he may come to overreact to x; i.e., to salient event features.[31] The degree to which the agent overreacts depends on the extent to which there is a tendency for z's that are associated with $x = 1$ to have relatively high (or low) corresponding success probabilities. Weakening this tendency will mitigate overreaction.

[31] Whenever $\mathsf{X} = \{0, 1\}$ and x is unimportant to predicting y, Proposition 5 establishes that Assumption 2 holds if and only if $\phi \neq 0$, so it is technically redundant to include this condition in the statement of Proposition 1; it is included for clarity.

The second part of Corollary 1 says that, as a result of the possibility that the selectively attentive agent will settle on not encoding z, he may eventually become certain that x is ceteris paribus predictive of y even when it is not. This is true whenever z is associated with both x and y and the agent effectively suffers from omitted variable bias. Again, in this case, the agent mistakenly comes to view x as more than a proxy for selectively unattended to predictors.

These results relate to experimental findings that individuals attribute more of a causal role to information that is the focus of attention and to salient information more generally (Fiske and Taylor 2008, Chapter 3; also see Nisbett and Ross 1980, Chapter 6). To take an example, Taylor and Fiske (1975, Experiment 2) had participants watch a videotape of two people interacting in conversation. In the most relevant experimental condition, a third of the participants were instructed to pay particular attention to one of the conversationalists, a third were instructed to pay particular attention to the other, and the final third were told only to observe the conversation (i.e., they were not instructed to attend to anything in particular). Later, participants rated the extent to which each conversationalist determined the kind of information exchanged, set the tone of the conversation, and caused the partner to behave as he did. An aggregate score served as the dependent measure. The interaction between instructions and conversationalist was highly significant: Participants were more likely to see the conversationalist they attended to as causal in the interaction.[32]

Friendliness and occupation example continued. Return to the earlier example, but generalize it a bit and assume that, independent of whether an individual is a professor, he is friendly with probability p^H during recreation and with probability $p^L < p^H$ at work. In addition, assume that $g(\text{Occupation}, \text{Situation})$ is uniformly positive but otherwise place no initial restrictions on this joint distribution.

If the student settles on attending to situational factors then Proposition 3 says that he will eventually stop reacting to whether an individual is a professor ($\mathcal{R}_{Occup}(\text{Work}) = \mathcal{R}_{Occup}(\text{Play}) = 0$), and, by Proposition 4, will learn to place full weight on mental models which do not include whether an individual is a professor among factors influencing friendliness. On the other hand, if the student settles on not attending to situational factors then Corollary 1 says that his limiting reaction to whether the individual is a professor equals

$$\hat{R}_{Occup}(\text{Situation}) = \hat{E}[y|\text{Prof}, \text{Situation}] - \hat{E}[y|\text{Not Prof}, \text{Situation}] = \frac{\phi}{g(\text{Prof})(1 - g(\text{Prof}))}.$$

[32] Participants also retained more information about the conversationalist they attended to.

A simple calculation gives us that $\phi = (p^H - p^L)(g(\text{Prof})g(\text{Work}) - g(\text{Prof}, \text{Work}))$, so the student's limiting reaction is

$$(15) \qquad \hat{R}_{Occup}(\text{Situation}) = \frac{(p^H - p^L)(g(\text{Work}) - g(\text{Work}|\text{Prof}))}{(1 - g(\text{Prof}))}.$$

From (15), the student will react to whether an individual is a professor in the limit whenever occupation and situational factors are associated in the sense that $g(\text{Work}|\text{Prof}) \neq g(\text{Work}|\text{Not Prof})$ and, in particular, will predict professors to be less friendly than others when $g(\text{Work}|\text{Prof}) > g(\text{Work}|\text{Not Prof})$.

In addition, whenever $g(\text{Work}|\text{Prof}) \neq g(\text{Work}|\text{Not Prof})$, Corollary 1 tells us that the student will become certain that whether an individual is a professor is a ceteris paribus predictor of friendliness even though he "knows" that he sometimes does not attend to situational factors. Again, the reason is that, by the naivete assumption, he treats the mentally represented history as if it were complete. In particular, he mistakenly treats observed variation in (Friendliness, Occupation|Real-World Interaction) as being equally informative as observed variation in (Friendliness, Occupation|Work) or (Friendliness, Occupation|Play) in identifying a causal effect of whether an individual is a professor on friendliness.

5. Continuous attention

So far, I have made the stark but instructive assumption that the agent never attends to z when he places little weight on mental models which specify z as being important to prediction. It is perhaps more realistic to assume that the agent will attend to z with a probability that varies more continuously in the likelihood he attaches to such processing being decision-relevant (Kahneman 1973). I model this by assuming that there are random fluctuations in the degree to which the agent is cognitively busy in a given period.[33] Then, the likelihood that the agent attends to z will naturally vary in the intensity of his belief that z is important to prediction.

Formally, let $\eta(\hat{\pi}^k_Z) \equiv \text{Prob}[e_k = 1|\hat{\pi}^k_Z] = \text{Prob}[b_k < \hat{\pi}^k_Z]$ denote the likelihood that an agent pays attention to z in period k as a function of the probability he attaches in that period to z being important to predicting y. Before, I considered the case where $b_k \equiv b$ for some $b \in (0, 1)$. Now suppose that each b_k is independently drawn according to some fixed cumulative distribution

[33] One interpretation is that there are fluctuations in the "shadow cost" of devoting attention, where this cost may depend on the number and difficulty of other tasks faced by the agent, for example.

function F that makes $\eta(\cdot)$ continuously differentiable with $\eta'(\cdot) > 0$ and $\eta(1) = 1$. We say that *the continuous attention assumptions hold* whenever the b_k are drawn in this manner.

To take an example, the continuous attention assumptions hold if $b_k \overset{i.i.d}{\sim} \mathrm{U}[0,1]$. In this case, the likelihood that the agent attends to z as a function of $\hat{\pi}_Z^k$ is given by:

$$\eta(\hat{\pi}_Z^k) = \hat{\pi}_Z^k$$

for all $0 \leq \hat{\pi}_Z^k \leq 1$.

Proposition 6. *Suppose the continuous attention assumptions hold. Then*

(1) $\eta(\hat{\pi}_Z^t) \to 1$ *almost surely.*

(2) *For each* x, z*,* $\hat{E}[y|x,z,\hat{h}^t]$ *converges to* $E_{\boldsymbol{\theta_0}}[y|x,z]$ *in probability.*

The intuition for Proposition 6 is the following. Under the continuous attention assumptions, the agent always attends to z with positive probability and almost surely encodes z an infinite number of times. As a result, no matter his initial beliefs or the degree to which he initially attends to z, he will receive enough disconfirming evidence that he will learn that z is in fact important to predicting y, which will lead him to devote an arbitrarily large amount of attention to z and to make accurate forecasts with arbitrarily large probability in the limit.

Even though the agent eventually learns to attend to z and to make accurate forecasts with arbitarily large probability in the limit, he may continue not to attend to z and to make biased forecasts for a long time. In particular, note that, for large t, $\hat{E}[y|x,z,\hat{h}^t] \approx E_{\boldsymbol{\theta_0}}[y|x]$ in any period where the agent does not attend to z. To assess whether and when we should expect the agent to begin attending to z over some reasonable time horizon, I consider the rate at which the likelihood that he attends to z approaches 1. For the rest of this section, I assume that the agent eventually only considers the two models $M_{X,Z}$ and $M_{X,\neg Z}$, either because his prior places full weight on x being important to predicting y (i.e., $\pi_X = 1$) or because x is in fact important to predicting y. Making this assumption allows for the cleanest possible results. I get very similar but messier results for the general case.

Before going further, I should define what I mean by rate of convergence.

Definition 5. The asymptotic rate of convergence of a random variable $\mathcal{X}_t$ to $\mathcal{X}_0$ is $V(t)$ if there exists a strictly positive constant $C < \infty$ such that

$$\frac{|\mathcal{X}_t - \mathcal{X}_0|}{V(t)} \overset{a.s}{\to} C$$

Remark 1. If $\mathcal{X}_t$ converges to $\mathcal{X}_0$ with asymptotic rate $V(t)$ then $|\mathcal{X}_t - \mathcal{X}_0| = O(V(t))$ for large t almost surely. Also, $O(V(t))$ is the "best possible" (Ellison 1993) in the sense that there exist strictly positive constants c_1 and c_2 such that, almost surely, $c_1 V(t) \leq |\mathcal{X}_t - \mathcal{X}_0| \leq c_2 V(t)$ for large t.

It is reasonable to expect that the rate at which the agent learns to attend to z depends on the degree to which he has difficulty explaining observations without taking z into account. Put the other way around, the agent may continue not attending to z for a long time if he can accurately approximate the true distribution when he only takes x into account.

Formally, let $p_{\boldsymbol{\theta_0}}(y|x,z)$ denote the distribution of y conditional on both x and z given the true vector of success probabilities and $p_{\boldsymbol{\theta_0}}(y|x)$ denote the distribution of y conditional only on x given that vector. Define the *relative entropy distance*, d, between these two distributions as the average of the relative entropies between $p_{\boldsymbol{\theta_0}}(y|x',z')$ and $p_{\boldsymbol{\theta_0}}(y|x')$, where this average is taken over the probability mass function $g(x,z)$[34]:

$$d = \sum_{y,x,z} p_{\boldsymbol{\theta_0}}(y|x,z) g(x,z) \log \left(\frac{p_{\boldsymbol{\theta_0}}(y|x,z)}{p_{\boldsymbol{\theta_0}}(y|x)} \right). \tag{16}$$

d essentially measures the distance between $p_{\boldsymbol{\theta_0}}(y|x,z)$ and $p_{\boldsymbol{\theta_0}}(y|x)$, which can be thought of as a measure of how difficult it is for the agent to explain what he observes in the context of a model under which only x is important to prediction. d can also be thought of as a measure of the degree to which an agent, starting from a belief that z is unlikely to predict y, is "surprised" by what he observes when he encodes z.

Proposition 7. *Suppose the continuous attention assumptions hold and either (i)* $\pi_X = 1$ *or (ii)* x *is important to predicting* y*. Then* $\eta(\hat{\pi}_Z^t) \to 1$ *almost surely with an asymptotic rate of convergence* $e^{-d(t-1)}$.

[34] "Distance" d is called the conditional relative entropy in Cover and Thomas (2006).

For a brief sketch of the arguments involved in proving Proposition 7, the rate at which $\eta(\hat{\pi}_Z^k) \to 1$ is determined by the rate at which

$$\frac{\Pr_\xi(\hat{h}^t|M_{X,\neg Z})}{\Pr_\xi(\hat{h}^t|M_{X,Z})} \to 0. \tag{17}$$

Consider the simpler problem of determining the rate at which[35]

$$\frac{\Pr(h^t|\theta(x,z) = p_{\boldsymbol{\theta_0}}(y=1|x) \text{ for all } x,z)}{\Pr(h^t|\boldsymbol{\theta_0})} \to 0. \tag{18}$$

By the strong law of large numbers, $1/(t-1)$ times the log of (18) goes to $-d$. The proof applies similar logic to analyzing (17), which is more complicated because effective observations are not i.i.d. when the agent sometimes fails to encode z and because $\Pr_\xi(\hat{h}^t|M)$ integrates over parameters.

5.1. **Example continued.** Return to the earlier example and again suppose that an individual is always friendly during recreation but never at work ($p^H = 1$, $p^L = 0$). It is easy to calculate that, in this case,

$$\begin{aligned} d &= -\sum_x \sum_z g(x,z)\log(g(z|x)) \\ &= H(z|x), \end{aligned}$$

where $H(z|x)$ is the conditional entropy of z given x. It is well known that

$$H(z|x) = H(z) - I(z;x),$$

where

- $H(z) = -\sum_z g(z)\log(z)$ is the entropy of z = Situation, or a measure of the degree to which the student splits his time between work and recreation.
- $I(z;x) = \sum_{x,z} g(x,z)\log\frac{g(x,z)}{g(x)g(z)}$ is the mutual information between z = Situation and x = Occupation, which is a measure of the degree to which knowledge of whether an individual is a professor provides the agent with information regarding whether he is likely to encounter the individual during work or recreation; if occupation and situational factors

[35] Simpler problems along these lines have been studied by other economists in the past (e.g., Easley and O'Hara 1992).

are independent then $I(z;x) = 0$. Put differently, $I(z;x)$ is another measure of the degree of association between x and z.

Thus, fixing the degree to which the student splits his time between work and play (i.e., fixing $H(z)$), the rate at which the agent will learn to attend to situational factors is *decreasing* in the degree of association between occupation and situational factors (decreasing in $I(z;x)$). Combining this fact with the earlier analysis suggests that a student who has an even greater tendency to encounter professors more often during work than recreation (e.g., he is an undergraduate rather than graduate student) both has the potential to overreact to whether an individual is a professor to a greater extent and is less likely to begin attending to situational factors within a reasonable time horizon.

This example highlights what seems to be an important fact, namely that the extent to which the agent's reaction may be biased by failing to attend to z, which depends on the degree of "omitted variable bias", may be *negatively* related to the speed at which the agent learns to attend to z, which depends on the quality of feedback available to the agent when he encodes z. To see this simply, consider the limiting (albeit slightly unrealistic) case where the student encounters professors only at work and non-professors only during recreation. In this case, his reaction to whether an individual is a professor is maximally biased but his ability to learn that situational factors are important to predicting friendliness is minimized.

5.2. **Coarse stereotyping of out-group members.** Proposition 7 can be used to help understand why people often attend to less information in forming judgments concerning members of out-groups (i.e., groups to which they do not belong) and believe "they" are all alike (Fiske and Taylor 2008, page 261; Hilton and von Hippel 1996; Fryer and Jackson 2008).[36]

To illustrate, continue to consider the professor/friendliness example. In the context of this example, the interest is in characterizing conditions under which the student will be less likely to attend to situational factors when making predictions concerning professors. We must extend the

[36] Experimental participants can generate more subgroups when describing an in-group than an out-group (Park, Ryan, and Judd 1992), are more likely to generalize from the behavior of a specific group member to the group as a whole for out-groups (Quattrone and Jones 1980), and are less likely to recall individuating attributes (e.g., occupation) of an out-group member (Park and Rothbart 1982). On this last point, Park and Rothbart (1982) asked experimental participants to read a newspapertype story where the sex of the character was randomly assigned ("William Larsen, 27, risked his life to save a neighbor's child ..." versus "Barbara Martin, 27, risked her life ..."). Two days later participants were asked to recall the sex and occupation of the character. While there was no difference in recall of the sex of the in-group versus out-group protagonist, participants were more likely to recall the occupation of the in-group versus out-group protagonist.

example to allow for the possibility that the student attends to z with a probability that depends on whether he is forecasting the friendliness of a professor or non-professor. To do so in a simple manner, suppose the student separately learns to predict friendliness for non-professors (the "in-group") and professors (the "out-group"), y^{in} and y^{out}. Otherwise, the example is the same: Each period the student either encounters a non-professor, with probability $g(in)$, or a professor, with probability $1 - g(in)$. In a period where the student encounters an individual from group $j \in \{in, out\}$, he observes $z \in \mathsf{Z}$, drawn from $g(\cdot|j)$ (formally, X is a singleton), makes prediction $\hat{y}_t^j$, and finally learns the true value y_t^j.

For $j = in, out$, let

$$d(j) = \sum_{y,z} p_{\boldsymbol{\theta_0}}(y|j,z)g(z|j)\log\left(\frac{p_{\boldsymbol{\theta_0}}(y|j,z)}{\sum_{z'} p_{\boldsymbol{\theta_0}}(y|j,z')g(z'|j)}\right) \tag{19}$$

measure the difficulty the student has explaining observations without taking z into account for members of group $j \in \{in, out\}$ and let $\eta_k(j)$ denote the probability that the student attends to z in period k if, in that period, he encounters an individual of group membership j.

An obvious modification of Proposition 7 gives us that, under the continuous attention assumptions, the asymptotic rate of convergence of $\eta_t(j)$ to 1 is

$$e^{-d(j)g(j)(t-1)} \tag{20}$$

for $j = in, out$.

From (20), the speed with which the student learns to allocate attention to situational factors in making predictions concerning a member of group j is increasing in the frequency of interaction with members of group j, $g(j)$, and the degree to which it is difficult to explain observations without taking situational factors into account when interacting with members of group j, $d(j)$. The fact that this speed is increasing in $g(j)$ is intuitive: The speed with which an individual learns that a variable is important to prediction should be increasing in the frequency with which he obtains new observations. That it is increasing in $d(j)$ is also intuitive and follows from Proposition 7: It is not just the amount but the quality of contact which governs how quickly an agent will learn to attend to situational constraint information when predicting the behavior of members of group j.

To interpret this result further, consider the specific assumptions of the example: group membership does not predict y, z is binary, and $p_{\boldsymbol{\theta_0}}(y = 0|\text{Work}) = 1 = p_{\boldsymbol{\theta_0}}(y = 1|\text{Play})$. Given these assumptions, $d(j)$ equals

$$H_j(z) = -\sum_{z'} g(z'|j)\log(g(z'|j)),$$

or the entropy of z for group j. Thus, if situational factors do not vary much across encounters with members of group j (encounters are relatively homogeneous) then the agent will persistently ignore situational-constraint information even if such encounters are frequent.

A key implication of (20) is that the student will more quickly learn to attend to situational factors when making predictions regarding non-professors if and only if

$$g(in)d(in) > g(out)d(out). \tag{21}$$

Inequality (21) suggests that two factors are responsible for the tendency of individuals to attend to less information in assessing out-group members over some period of time: (i) interactions with members of an out-group tend to be less frequent and (ii) encounters with members of an out-group tend to be relatively homogeneous. The role of the first factor, relatively infrequent interactions, has been recognized by other economists as contributing to people holding relatively inaccurate beliefs about and/or persistently discriminating against members of an out-group (Fryer and Jackson 2008, Glaeser 2005).

To the best of my knowledge, economists have ignored the second, the relative homogeneity of interactions, but the quality of interaction is an important determinant of the degree and persistence of intergroup bias (Allport 1954, Pettigrew 1998, Pettigrew and Tropp 2006). Proposition 7 as applied to this example suggests that individuals who encounter members of an out-group across more varied situations (e.g., both at work and in the neighborhood) are more likely to consider how members of the group act in particular situations, rather than on average, when forecasting behavior.

6. Illustrative examples

Economic models traditionally assume that people take all freely available information into account when forecasting the quality of an object to inform a decision. To take one example, a basic premise of existing rational statistical discrimination models is that employers optimally consider all easily observable information about a potential worker (from résumés, interviews, and recommendations) when making hiring decisions (e.g., Phelps 1972, Arrow 1973, Aigner and Cain

1977, Altonji and Pierret 2001). There are many real world examples that are hard to reconcile with this assumption, but are consistent with the selective attention model.

Consider the large mispricing of skills in the market for baseball players, popularly documented in Michael Lewis's *Moneyball* (Lewis 2003). Lewis (2003) argues that, historically, people who ran baseball teams had an incomplete picture of what makes a good baseball player, in particular a successful batter. Through following conventional advice, like "There is but one true criterion of skill at the bat, and that is the number of times bases are made on clean hits" (Lewis 2003, page 70), they ignored important components of batter skill, notably a batter's ability to get on base by receiving walks. Baseball statisticians like Bill James noted this deficiency (James 1982), but were largely dismissed by managers and others making hiring decisions. The managers appear to have been in error. Starting in the late 1990s, the Oakland Athletics began to focus on hiring players who excelled at getting on base.[37] Evidence suggests that, in doing so, they were able to build a very successful team at a relatively cheap price (Hakes and Sauer 2006). In 2002, the Athletics ranked twelfth in payroll in the American League (out of fourteen teams) but first in wins (Thaler and Sunstein 2004). However, this competitive advantage appears to have largely disappeared in recent years, in particular since the publication of *Moneyball*. Regression analysis indicates that, in 2004, wages reflected a player's ability to get on base (controlling for other factors), but did not before (Hakes and Sauer 2006).

Assuming Lewis (2003) is correct, selective attention provides an explanation for what happened. Though managers had access to freely available information proxying for the ability to take walks and get on base (e.g., on-base percentage, which takes walks into account), they did not carefully attend to such information.[38] In turn, they did not learn how important having disciplined batters could be to winning games. It was only later, after the importance of this skill was explicitly demonstrated by the Oakland Athletics and detailed in a popular book, that the market learned to pay attention and wages came to more closely reflect fundamental batter value.

Consider also the study by Malmendier and Shanthikumar (2007) on the tendency of small investors to take security analysts' stock recommendations literally. Affiliated analysts (i.e., those belonging to banks that have an underwriting relationship to firms they are reporting on) tend

[37] The Oakland Athletics topped the American League in walks in 1999 and 2001, ranked second or third in 2000, 2002 and 2004, and ranked fifth in 2003 (Hakes and Sauer 2006).

[38] On-base percentage is the fraction of plate appearances where the player reached base either through a walk or a hit. Walks do not figure into the classic batter statistic, a player's batting average.

to issue more favorable recommendations. For example, the modal recommendation is "buy" for affiliated analysts but "hold" for unaffiliated analysts. Malmendier and Shanthikumar (2007) find that large investors (e.g., pension funds) discount the recommendations of affiliated versus unaffiliated analysts. Small investors (e.g., individual investors), on the other hand, do not. This pattern of results is difficult to explain in a standard cost of information gathering framework, as small investors do not react differently to independent analysts' recommendations (i.e., those never involved in underwriting) even though members of this group often advertise their independence. It follows naturally, however, from the model of selective attention. By virtue of being relatively busy thinking about other things and having less precise knowledge about analysts' incentives, it is relatively unlikely that small investors will learn to attend to analyst affiliation or that affiliated analysts' recommendations should be relatively discounted.[39]

Also related are experimental findings on stereotype formation. Schaller and O'Brien (1992) asked experimental participants to judge the relative intelligence of two groups, A and B, on the basis of their performance on anagram tasks. Participants were presented with 50 observations, where each observation consisted of information on the group membership of a person (25 observations of each group), whether he solved or failed to solve the anagram, the actual anagram (some were five letters long and others were seven letters long), and the correct solution. The observations were constructed such that, conditioned on the length of the anagram, group A members solved more anagrams but, unconditionally, group B members solved more. After being presented with the observations, participants judged group B members to be more intelligent than group A members and predicted they would perform better if given the same anagram to solve in the future, presumably failing to take into account the correlation between group membership and anagram length. Consistent with this failure stemming from selective attention, manipulations designed to give participants more time to process each observation or to direct their attention towards anagram length through explicit instructions (prior to presenting the observations) facilitated more accurate judgments. In particular, when participants were both given more time as well as the instructions, they viewed A members as more intelligent and predicted that they would perform better if given the same anagram to solve.

[39] For related experimental evidence, see Cain, Loewenstein, and Moore (2005).

As a final example, Bertrand and Mullainathan (2004) find that, all else equal, résumé callbacks are more responsive to variables predictive of quality (e.g., years of experience, skills listed, existence of gaps in employment) for white sounding than for African-American sounding names. In a related pilot study, Bertrand et al. (2005) find that laboratory participants who report feeling more rushed are more likely to discriminate against African-American résumés. While there are several interpretations for these findings, one possibility is that screeners have coarser stereotypes for African Americans (Fiske and Taylor 2008, page 261; Hilton and von Hippel 1996; Fryer and Jackson 2008), which leads them to selectively attend to less information when screening their résumés.

7. Basic Extensions

In this Section, two basic extensions of the analysis are considered. In the first basic extension, I examine what happens if, after some amount of time, the agent begins attending to z because there is a shock to his belief that z is important to prediction. The main point is that, following such "debiasing", it will still take the agent time to learn to incorporate information about z in making predictions, since he did not notice z before. This is easiest to see in the context of the example of a doctor who brings up the possibility that food allergies could be causing an agent's headaches. Even if they are, the agent may need to keep a food diary for some time before learning which foods he should stay away from. This feature of the model helps clarify how its predictions will often differ from one in which an agent cannot attend to all available information when making a prediction, but can nonetheless recall such information if necessary later on (e.g., Hong, Stein and Yu 2007).[40] In the second, I show how selective attention can lead to asymptotic disagreement across agents who share a common prior and observe the same data when some agents can devote more attention than others to a prediction task.

7.1. **Debiasing.** Suppose $b_k \equiv b$, the agent starts off not encoding z ($\pi_Z < b$), and $\mathsf{X} = \{0, 1\}$. What happens if, at some large t, the agent begins attending to z because there is an unmodeled shock to his belief that z is important to predicting y (π_Z shifts to $\pi'_Z > \pi_Z$) or to the degree to which he is cognitively busy (b shifts to $b' < b$)?[41]

[40] Models like Hong, Stein and Yu's (2007) may be a better description of situations where past information (e.g., about firm earnings) is freely available in public records and tends to be revisited; mine may be a better description of situations where such information is not.

[41] Can think of shocks to π_Z as resulting from a media report or something learned in a class and shocks to b as resulting from some (not modeled) reason why the agent would begin caring more about predicting y (e.g., he begins

The main thing to note is that, even if this shock leads the agent to settle on encoding z and to make unbiased forecasts in the limit, he continues to make systematically biased forecasts for a long time. The reason is that it takes some time for the agent to learn how to use both x and z to make predictions since he has not attended to z in the past (there is "learning by encoding"). To see this, consider how the agent reacts to x given z at the time of the shock t. Since t is assumed to be large, $E_{\xi}[\theta(x,z)|M_{X,\neg Z},\hat{h}^t] \approx E_{\boldsymbol{\theta_0}}[y|x]$ and $\hat{\pi}^t_X \approx 1$ by the results of Section 3, so the agent's reaction to x given z in that period equals

$$E_{\xi}[\theta(1,z)|\hat{h}^t] - E_{\xi}[\theta(0,z)|\hat{h}^t] \approx \pi'_Z[\tau - \tau] + (1-\pi'_Z)(E_{\boldsymbol{\theta_0}}[y|x=1] - E_{\boldsymbol{\theta_0}}[y|x=0]) \tag{22}$$

$$= (1-\pi'_Z)(E_{\boldsymbol{\theta_0}}[y|x=1] - E_{\boldsymbol{\theta_0}}[y|x=0]), \tag{23}$$

where $\tau = E_{\psi}[\theta]$ equals the prior success probability under density ψ. From (23), the agent's reaction to x in period t is approximately proportional to his reaction the period before when he did not attend to z. This is intuitive: By not having attended to z in the past he has not learned that z is important to predicting y or how z is important to predicting y. As a result, even when he attends to z, his forecast places substantial weight on the empirical frequency of $y = 1$ given only (x).

7.2. **Disagreement.** Disagreement across agents arises naturally out of the model, even when agents share a common prior and observe the same information. Suppose that there are two agents, $i = 1, 2$, who can devote differing amounts of attention to the task of predicting y. Formally, let $b^i_k \equiv b^i$ denote the degree to which agent i is cognitively busy and suppose $b^1 \neq b^2$. Then, asymptotically, the two agents may react differently to pieces of information that are potentially informative about outcome variable y.

To see this clearly, suppose that the first agent can devote so much attention to the task at hand that she always encodes z ($b^1 = 0$), but the second is so consumed with other activities that she never encodes z ($b^2 = 1$). Further, suppose that $\mathsf{X} = \{0, 1\}$ and x is positively related to y conditional on each z': $\mathcal{R}_x(z') \geq 0$ for all z'. A straightforward application of Propositions 3 and 5 gives us that, starting from the same prior, the two agents may nevertheless asymptotically disagree about the sign of the relationship between x and y at all z' even after observing the same data; that is, we may have a situation where, in the limit, the first agent correctly reacts positively to

caring more about what leads to weight gain if he recently had a heart attack) or it becomes easier for the agent to attend to z (perhaps an attribute of a product is suddenly unshrouded).

$x = 1$ while the other incorrectly reacts negatively to $x = 1$. In particular, from Equation (10) it is easy to see that this will almost surely be the case whenever the omitted variable bias is sufficiently severe: $-\phi > \mathcal{R}_x \operatorname{Var}(x)$.

8. Conclusion

This paper has supplied a model of belief formation in which an agent is selective as to which information he attends. The central assumption of the model is that the likelihood that the agent encodes information along a dimension is increasing in the intensity of his belief that such information is predictive. I show that, as a consequence of selective attention, the agent may persistently fail to attend to an important predictor and hold incorrect beliefs about the statistical relationship between variables. In addition, I derive conditions under which such errors are more likely or persistent. Results match and shed light on several experimentally found biases in inference, including the difficulty people have in recognizing relationships that prior theories do not make plausible and the overattribution of cause to salient event features. Examples indicate that the model can be fruitfully applied to study a range of problems in stereotyping, persuasion, statistical discrimination, and other areas.

Appendix A. Prior, forecasts, and useful lemmas

A.1. **Prior.** I now give an alternative description of the agent's prior, which will be useful in presenting the proofs. The prior can compactly be expressed as $\mu(\boldsymbol{\theta}) = \sum_{i\in\{X,\neg X\}}\sum_{j\in\{Z,\neg Z\}} \pi_{i,j}\mu^{i,j}(\boldsymbol{\theta})$. Fix a model $M \in \mathcal{M}$ and define $c^M(x,\hat{z})$ as the set of covariates $(x',\hat{z}') \in \mathsf{X}\times\hat{\mathsf{Z}}$ such that, under that model, any y_t given covariates $(x_t,\hat{z}_t) = (x,\hat{z})$ is exchangeable with any $y_{t'}$ given covariates $(x_{t'},\hat{z}_{t'}) = (x',\hat{z}')$; i.e., under model M, $\theta(x',\hat{z}') = \theta(x,\hat{z})$ with probability one if and only if $(x',\hat{z}') \in c^M(x,\hat{z})$. For example, under $M_{X,\neg Z}$ (where only x is important to predicting y), $c^{X,\neg Z}(x,\hat{z}) = \{(x',\hat{z}') \in \mathsf{X}\times\hat{\mathsf{Z}} : x' = x\}$ equals the set of covariates that agree on x. With a slight abuse of notation, label the common success probability across members of c^M under model M by $\theta(c^M)$. Intuitively, $c^M(x,\hat{z})$ equals the set of covariates that, under model M, can be lumped together with $(x,\hat{z})$ without affecting the accuracy of the agent's predictions.

Let C^M denote the collection of c^M, so C^M is a partition of $\mathsf{X}\times\hat{\mathsf{Z}}$, and define $\boldsymbol{\Theta}(M) = [0,1]^{\#C^M}$ as the effective parameter space under model M with generic element $\boldsymbol{\theta}(M)$. μ^M is defined by the joint distribution it assigns to the $\#C^M$ parameters $\theta(c^M)$. These parameters are taken as independent with respect to μ^M and distributed according to density, $\psi(\cdot)$. To take an example, if $\psi(\theta) = \mathbf{1}_{\theta\in[0,1]}$, then $\theta(c^M) \sim \mathrm{U}[0,1]$ for each $M \in \mathcal{M}$, $c^M \in C^M$.

A.2. **Forecasts.** I will now describe the forecasts of an individual with selective attention in some detail (rational forecasts are a special case) and will present some definitions which will be useful later.

Given the individual's prior, his period-t forecast given recalled history $\hat{h}^t$ is given by

$$\hat{E}[y|x,z,\hat{h}^t] = \sum_{M'\in\mathcal{M}} \hat{\pi}^t_{M'} E_\xi[\theta(c^{M'}(x,\hat{z}))|\hat{h}^t, M'], \tag{24}$$

where

$$\begin{aligned} E_\xi[\theta(c^M)|\hat{h}^t, M] &= \int \tilde{\theta}\psi(\tilde{\theta}|\hat{h}^t, c^M)d\tilde{\theta} \\ \psi(\tilde{\theta}|\hat{h}^t, c^M) &= \frac{\tilde{\theta}^{\kappa(c^M|\hat{h}^t)}(1-\tilde{\theta})^{N(c^M|\hat{h}^t)-\kappa(c^M|\hat{h}^t)}\psi(\tilde{\theta})}{\int \tau^{\kappa(c^M|\hat{h}^t)}(1-\tau)^{N(c^M|\hat{h}^t)-\kappa(c^M|\hat{h}^t)}\psi(\tau)d\tau}. \end{aligned}$$

$N(c^M|\hat{h}^t)$ denotes the number of times the covariates have taken on some value $(x,\hat{z}) \in c^M$ along history $\hat{h}^t$ and $\kappa(c^M|\hat{h}^t)$ denotes the number of times that both the covariates have taken on such a value and $y = 1$. I will sometimes abuse notation and write $N(x,\hat{z}|\hat{h}^t)$ and $\kappa(x,\hat{z}|\hat{h}^t)$ instead of

$N(\{(x,\hat{z})\}|\hat{h}^t)$ and $\kappa(\{(x,\hat{z})\}|\hat{h}^t)$, respectively. Likewise, when convenient I will write $N(x|\hat{h}^t)$ instead of $N(\{(x',\hat{z}'):x'=x\}|\hat{h}^t)$, etc.

To illustrate, (24) takes a particularly simple form when $\psi(\theta)\sim \mathrm{U}[0,1]$:

(25)

$$\hat{E}[y|x,z,\hat{h}^t]=\hat{\pi}^t_{X,Z}\frac{\kappa(x,\hat{z}|\hat{h}^t)+1}{N(x,\hat{z}|h^t)+2}+\hat{\pi}^t_{X,\neg Z}\frac{\kappa(x|\hat{h}^t)+1}{N(x|h^t)+2}+\hat{\pi}^t_{\neg X,Z}\frac{\kappa(\hat{z}|\hat{h}^t)+1}{N(\hat{z}|\hat{h}^t)+2}+\hat{\pi}^t_{\neg X,\neg Z}\frac{\bar{\kappa}(\hat{h}^t)+1}{t+1},$$

where $\bar{\kappa}(\hat{h}^t)=\sum_{x',\hat{z}'}\kappa(x',\hat{z}'|\hat{h}^t)$.

For future reference,

$$\begin{aligned}\hat{\pi}^t_{i,j}&=\Pr_\xi(M_{i,j}|\hat{h}^t)\\&=\frac{\Pr_\xi(\hat{h}^t|M_{i,j})\pi_{i,j}}{\sum_{i',j'}\Pr_\xi(\hat{h}^t|M_{i',j'})\pi_{i',j'}}\\&=\frac{\alpha_{i,j}\mathcal{B}^t_{i,j}}{\sum_{i',j'}\alpha_{i',j'}\mathcal{B}^t_{i',j'}}\end{aligned}$$

where

$$\begin{aligned}\mathcal{B}^t_{i,j}&=\frac{\Pr_\xi(\hat{h}^t|M_{i,j})}{\Pr_\xi(\hat{h}^t|M_{X,Z})}\\&=\frac{\int\Pr_\xi(\hat{h}^t|\boldsymbol{\theta})\mu^{i,j}(d\boldsymbol{\theta})}{\int\Pr_\xi(\hat{h}^t|\boldsymbol{\theta})\mu^{X,Z}(d\boldsymbol{\theta})}\end{aligned}$$

is the *Bayes factor* comparing model $M_{i,j}$ to model $M_{X,Z}$ (Kass and Raftery 1995 provide a review of Bayes factors) and

(26)
$$\alpha_{i,j}=\frac{\pi_{i,j}}{\pi_{X,Z}}$$

is the prior odds for $M_{i,j}$ against $M_{X,Z}$.

A.3. **Useful lemmas concerning the asymptotic properties of Bayes' factors.** Prior to presenting the remaining proofs, I establish several results which will be useful in establishing asymptotic properties of the Bayes' factors and will in turn aid in characterizing the agent's asymptotic forecasts and beliefs. Let $p_0(y,x,\hat{z})$ and $\hat{p}(y,x,\hat{z})$ denote probability mass functions over $(y,x,\hat{z})\in\{0,1\}\times\mathsf{X}\times\hat{\mathsf{Z}}$. Define the Kullback Leibler distance between $\hat{p}(y,x,\hat{z})$ and $p_0(y,x,\hat{z})$

as

$$d_K(\hat{p}, p_0) = \sum_{y,x,\hat{z}} p_0(y,x,\hat{z}) \log\left(\frac{p_0(y,x,\hat{z})}{\hat{p}(y,x,\hat{z})}\right) \tag{27}$$

with the convention that $0 \log\left(\frac{0}{\hat{p}}\right) = 0$ for $\hat{p} \geq 0$ and $p_0 \log\left(\frac{p_0}{0}\right) = \infty$ for $p_0 > 0$ (see, e.g., Cover and Thomas 2006).

For all $(y, x, \hat{z})$, assume that $\hat{p}(y, x, \hat{z})$ can be written as $\hat{p}(y, x, \hat{z}|\boldsymbol{\theta}) = \theta(x, \hat{z})^y(1-\theta(x, \hat{z}))^{1-y}p_0(x, \hat{z})$ (sometimes abbreviated as $\hat{p}_{\boldsymbol{\theta}}(y, x, \hat{z})$), where $p_0(x, \hat{z}) = \sum_{y' \in \{0,1\}} p_0(y', x, \hat{z})$. Define $\hat{p}(y, x, \hat{z}|\boldsymbol{\theta}(M)) = p_{\boldsymbol{\theta}(M)}(y, x, \hat{z})$ in the obvious manner ($\boldsymbol{\theta}(M)$ is defined as in Subsection A.1) and let $\underline{\boldsymbol{\theta}}(M) = \arg\min_{\boldsymbol{\theta}(M) \in \boldsymbol{\Theta}(M)} d_K(\hat{p}_{\boldsymbol{\theta}(M)}, p_0)$ denote a minimizer of the Kullback-Leibler distance between $\hat{p}_{\boldsymbol{\theta}(M)}(\cdot)$ and $p_0(\cdot)$ among parameter values in the support of $\mu^M(\cdot)$. Finally, define $\delta_M = \delta_M(p_0) = d_K(\hat{p}_{\underline{\boldsymbol{\theta}}(M)}, p_0)$.

Lemma 1. *For all $M \in \mathcal{M}$, p_0, and $c^M \in C^M$, $\underline{\theta}(c^M) = p_0(y = 1|c^M)$.*

Proof. Fix some $p_0(\cdot)$, M, and c^M.

$$-d_K(\hat{p}_{\boldsymbol{\theta}(M)}, p_0) = \sum_{y,x,\hat{z}} p_0(y|x,\hat{z})p_0(x,\hat{z}) \log\left(\frac{\theta(c^M(x,\hat{z}))^y(1-\theta(c^M(x,\hat{z})))^{1-y}}{p_0(y|x,\hat{z})}\right) \tag{28}$$

$$= \sum_{c^M \in C^M} \left[p_0(y=1|c^M)p_0(c^M)\log(\theta(c^M)) + p_0(y=0|c^M)p_0(c^M)\log(1-\theta(c^M))\right] - K \tag{29}$$

where K does not depend on $\boldsymbol{\theta}(M)$. It is routine to show that each term in the sum of (29) is maximized when $\theta(c^M) = p_0(y = 1|c^M)$, which concludes the proof. ■

Let $\hat{h}^t = (y_{t-1}, x_{t-1}, \hat{z}_{t-1}, \ldots, y_1, x_1, \hat{z}_1)$ be some random sample from $p_0(y, x, \hat{z})$. Define

$$\mathcal{I}_t(M) = \mathcal{I}(M|\hat{h}^t) = \int \frac{\prod_{k=1}^{t-1} \hat{p}(y_k, x_k, \hat{z}_k|\boldsymbol{\theta})}{\prod_{k=1}^{t-1} p_0(y_k, x_k, \hat{z}_k)} \mu^M(d\boldsymbol{\theta}) \tag{30}$$

as well as the predictive distribution

$$\hat{p}_t^M(y, x, \hat{z}) = \hat{p}^M(y, x, \hat{z}|\hat{h}^t) = \int \hat{p}(y, x, \hat{z}|\boldsymbol{\theta})\mu^M(d\boldsymbol{\theta}|\hat{h}^t). \tag{31}$$

Note that, while not explicit in the notation, both $\mathcal{I}_t(M)$ and $\hat{p}_t^M(\cdot)$ depend on p_0. To avoid confusion, I will sometimes make this dependence explicit by writing $\mathcal{I}_t(M|p_0)$ and $\hat{p}_t^M(\cdot|p_0)$.

It will be useful to establish some Lemmas with priors which are slightly more general than what has been assumed.

Definition 6. μ^M is *uniformly non-doctrinaire* if it makes each $\theta(c^M)$ independent with non-doctrinaire prior ψ_{c^M}.

Note that it is possible for ψ_{c^M} to vary with c^M when μ^M is uniformly non-doctrinaire.

Lemma 2. *For all $M \in \mathcal{M}, p_0$, and uniformly non-doctrinaire μ^M,*

$$\frac{1}{t-1} \log \mathcal{I}_t(M|p_0) \to -\delta_M(p_0), \tag{32}$$

p_0^∞ almost surely.

Proof. Fix some $M \in \mathcal{M}$, p_0, and uniformly non-doctrinaire μ^M. From Walker (2004, Theorem 2), it is sufficient to show that the following conditions hold:

(1) $\mu^M(\{\boldsymbol{\theta} : d_K(\hat{p}_{\boldsymbol{\theta}}, p_0) < d\}) > 0$ only for, and for all, $d > \delta_M$

(2) $\lim_t \inf d_K(\hat{p}_t^M, p_0) \geq \delta_M$, p_0^∞ almost surely

(3) $\sup_t \mathrm{Var}(\log(\mathcal{I}_{t+1}(M)/\mathcal{I}_t(M))) < \infty$

The "only for" part of the first condition holds trivially from the definition of δ_M and the "for all" part follows from the fact that $d_K(\hat{p}_{\boldsymbol{\theta}(M)}, p_0)$ is continuous in a neighborhood of $\underline{\boldsymbol{\theta}}(M)$ (since $\hat{p}_{\boldsymbol{\theta}(M)}(\cdot)$ is continuous in $\boldsymbol{\theta}(M)$) and $\mu^M(\cdot)$ places positive probability on all open neighborhoods in $\boldsymbol{\Theta}(M)$. The second condition also holds trivially since $d_K(\hat{p}_t^M, p_0) \geq \min_{\boldsymbol{\theta}(M) \in \boldsymbol{\Theta}(M)} d_K(\hat{p}_{\boldsymbol{\theta}(M)}, p_0) = \delta_M$ for all $t, \hat{h}^t$.

The third condition requires a bit more work to verify. Note that $\mathcal{I}_{t+1}(M) = \frac{\hat{p}_t^M(y_t, x_t, \hat{z}_t)}{p_0(y_t, x_t, \hat{z}_t)} \mathcal{I}_t(M)$ $\Rightarrow \log(\mathcal{I}_{t+1}(M)/\mathcal{I}_t(M)) = \log\left(\frac{\hat{p}_t^M(y_t, x_t, \hat{z}_t)}{p_0(y_t, x_t, \hat{z}_t)}\right)$, so condition (3) is equivalent to

$$\sup_t \mathrm{Var}\left[\log\left(\frac{\hat{p}_t^M(y_t, x_t, \hat{z}_t)}{p_0(y_t, x_t, \hat{z}_t)}\right)\right] < \infty \tag{33}$$

which can easily be shown to hold so long as

$$\sup_t E\left\{\sum_{y,x,\hat{z}} p_0(y, x, \hat{z}) \log\left(\frac{\hat{p}_t^M(y|x, \hat{z})}{p_0(y|x, \hat{z})}\right)^2\right\} < \infty \tag{34}$$

or

$$\sup_t E\left[\log\left(\hat{p}_t^M(y|x, \hat{z})\right)^2\right] < \infty \tag{35}$$

for all $(y, x, \hat{z})$ which satisfy $p_0(y, x, \hat{z}) > 0$.

To verify (35), fix some $(y, x, \hat{z})$ with $p_0(y, x, \hat{z}) > 0$ and let $N(c^M(x, \hat{z})|\hat{h}^t) = N_t$ denote the number of times the covariates have taken on some value $(x', \hat{z}') \in c^M(x, \hat{z})$ along history $\hat{h}^t$ and $\kappa(c^M(x, \hat{z})|\hat{h}^t) = \kappa_t$ denote the number of times both that the covariates have taken on such a value and $y = 1$. Then

$$q_t = \frac{\kappa_t + 1}{N_t + 2} \tag{36}$$

roughly equals the empirical frequency of $y = 1$ conditional on $(x', \hat{z}') \in c^M(x, \hat{z})$ up to period t.

An implication of the Theorem in Diaconis and Freedman (1990) is that

$$\hat{p}_t^M(y|x, \hat{z}) \rightarrow q_t^y (1 - q_t)^{1-y} \tag{37}$$

at a uniform rate across histories since the marginal prior density over $\theta(c^M(x, \hat{z}))$ is non-doctrinaire. Consequently, fixing an $\epsilon > 0$ there exists an $n > 0$ such that, independent of the history,

$$|\log(\hat{p}_t^M(y|x, \hat{z}))^2 - \log(q_t^y(1 - q_t)^{1-y})^2| < \epsilon$$

for all $t \geq n$[42] which implies that

$$E[|\log(\hat{p}_t^M(y|x, \hat{z}))^2 - \log(q_t^y(1 - q_t)^{1-y})^2|] < \epsilon \tag{38}$$

for all $t \geq n$. Since $E[\log(\hat{p}_t^M(y|x, \hat{z}))^2] < \infty$ for all finite t, to verify (35) it is sufficient to show that

$$\sup_t E[\log(q_t^y(1 - q_t)^{1-y})^2] < \infty \tag{39}$$

by (38).

By symmetry, it is without loss of generality to verify (39) for the case where $y = 1$. To this end,

$$\begin{aligned} E[\log(q_t)^2] &= E[E[\log(q_t)^2|N_t]] \\ &= E\left[(1 + N_t)(1 - \tilde{\theta})^{N_t} \log\left(\frac{1}{2 + N_t}\right)^2\right] \end{aligned}$$

[42] Can show that this statement follows from Diaconis and Freedman's (1990) result using an argument similar to Fudenberg and Levine (1993, Proof of Lemma B.1).

where

$$\tilde{\theta} \equiv p_0(y = 1 | c^M(x, \hat{z})).$$

Now, since $\lim_{N\to\infty}(1 + N)(1 - \tilde{\theta})^N \log\left(\frac{1}{2+N}\right)^2 = 0$, there exists a constant $M < \infty$ such that

$$(1 + N)(1 - \tilde{\theta})^N \log\left(\frac{1}{2 + N}\right)^2 < M$$

for all N. As a result,

$$E\left[(1 + N_t)(1 - \tilde{\theta})^{N_t} \log\left(\frac{1}{2 + N_t}\right)^2\right] < M < \infty$$

for all t which verifies (39) and concludes the proof. ■

Define the Bayes' factor conditional on p_0 as

$$\mathcal{B}_{i,j}(\hat{h}^t | p_0) = \mathcal{B}^t_{i,j}(p_0) = \frac{\mathcal{I}_t(M_{i,j}|p_0)}{\mathcal{I}_t(M_{X,Z}|p_0)} \tag{40}$$

Note that $\mathcal{B}_{i,j}(\hat{h}^t) = \mathcal{B}_{i,j}(\hat{h}^t|p_0)$ for some p_0 whenever we can write $\Pr_\xi(\hat{h}^t|\boldsymbol{\theta}) = \prod_{k=1}^{t-1} \hat{p}(y_k, x_k, \hat{z}_k|\boldsymbol{\theta}) = \prod_{k=1}^{t-1} \theta(x_k, \hat{z}_k)^{y_k}(1 - \theta(x_k, \hat{z}_k))^{1-y_k} p_0(x_k, \hat{z}_k)$ for some p_0.

Lemma 3. *For all $M_{i,j} \in \mathcal{M}$, p_0, and uniformly non-doctrinaire $\mu^{i,j}, \mu^{X,Z}$,*

$$\frac{1}{t-1} \log \mathcal{B}^t_{i,j}(p_0) \to \delta_{X,Z}(p_0) - \delta_{i,j}(p_0), \tag{41}$$

p_0^∞ almost surely.

Proof. Note that

$$\frac{1}{t-1} \log \mathcal{B}^t_{i,j} = \frac{1}{t-1} \log\left(\mathcal{I}_t(M_{i,j})\right) - \frac{1}{t-1} \log\left(\mathcal{I}_t(M_{X,Z})\right) \tag{42}$$

so the result follows immediately from Lemma 2. ■

Remark 2. An immediate implication of Lemma 3 is that $\delta_{i,j}(p_0) > \delta_{X,Z}(p_0)$ implies $\mathcal{B}^t_{i,j}(p_0) \to 0$, p_0^∞ almost surely.

Remark 2 applies when $\delta_{i,j}(p_0) > \delta_{X,Z}(p_0)$; what does the Bayes' factor $\mathcal{B}^t_{i,j}(p_0)$ tend towards asymptotically when $\delta_{i,j}(p_0) = \delta_{X,Z}(p_0)$? I now present a Lemma (due to Diaconis and Freedman 1992) that will aid in estimating the Bayes' factor in this case and establishing asymptotic results.

First some definitions. Let $H(q)$ be the entropy function $q\log(q) + (1-q)\log(1-q)$ (set at 0 for $q = 0$ or 1) and define the following

$$\begin{aligned}
\phi(\kappa, N, \psi) &= \int_0^1 \theta^\kappa (1-\theta)^{N-\kappa} \psi(\theta) d\theta \\
\phi(\kappa, N) &= \int_0^1 \theta^\kappa (1-\theta)^{N-\kappa} d\theta = \text{Beta}(\kappa+1, N-\kappa+1) \\
\hat{q} &= \frac{\kappa}{N} \\
\phi^*(\kappa, N) &= \begin{cases} \frac{e^{NH(\hat{q})}}{\sqrt{N}} \sqrt{2\pi} \sqrt{\hat{q}(1-\hat{q})} & \text{for } 0 < \kappa < N \\ \frac{1}{N} & \text{for } \kappa = 0 \text{ or } N \end{cases}
\end{aligned}$$

Lemma 4. *For any non-doctrinaire $\psi(\cdot)$ there are $0 < a < A < \infty$ such that for all $N = 1, 2, \ldots$ and $\kappa = 0, 1, \ldots, N$, $a\phi^*(\kappa, N) < \phi(\kappa, N, \psi) < A\phi^*(\kappa, N)$.*

Proof. Note that for any non-doctrinaire ψ there exist constants b, B such that $0 < b \leq \psi(\theta) \leq B < \infty$ for all $\theta \in (0, 1)$. The result then follows from Lemma 3.3(a) in Diaconis and Freedman (1992). For a brief sketch, note that $b\phi(\kappa, N) \leq \phi(\kappa, N, \psi) \leq B\phi(\kappa, N)$. Now use Stirling's formula on $\phi(\kappa, N)$ for κ and $N - \kappa$ large. ■

Appendix B. Proofs

B.1. **Proofs of Observations.** In proving Observation 1 and some of the later propositions, I will make use of the following Lemma which establishes the almost sure limit of several Bayes' factors when the agent always encodes z (e.g., when he is a standard Bayesian).

Lemma 5. *$\mathcal{B}_{X,\neg Z}(h^t) \to 0$ and $\mathcal{B}_{\neg X,\neg Z}(h^t) \to 0$, $P_{\boldsymbol{\theta}_0}$ almost surely. Additionally, if x is important to predicting y, then $\mathcal{B}_{\neg X,Z}(h^t) \to 0$, $P_{\boldsymbol{\theta}_0}$ almost surely.*

Proof. When the agent always encodes z, each period's observation is independently drawn from $p_0^1(y, x, z) = \theta_0(x, z)^y (1 - \theta_0(x, z))^{1-y} g(x, z)$ for all (y, x, z). Then, Lemma 3 implies that it is sufficient to show that $\delta_{X,\neg Z}(p_0^1) > \delta_{X,Z}(p_0^1)$, $\delta_{\neg X,\neg Z}(p_0^1) > \delta_{X,Z}(p_0^1)$ and, whenever x is important to predicting y, $\delta_{\neg X,Z}(p_0^1) > \delta_{X,Z}(p_0^1)$. Can easily establish these inequalities by applying Lemma 1 for each $M \in \mathcal{M}$. ■

Proof of Observation 1.1. Recall that the standard Bayesian's period t forecast satisfies

$$E[y|x,z,h^t] = \sum_{M' \in \mathcal{M}} \pi^t_{M'} E[\theta(c^{M'}(x,z))|h^t, M'] \tag{43}$$

Fix an $M \in \mathcal{M}$. Since the marginal prior density over $\theta(c^M(x,z))$ is non-doctrinaire under M, $E[\theta(c^M(x,z))|h^t, M] \to \bar{y}_t(c^M(x,z)) \overset{a.s.}{\to} E_{\boldsymbol{\theta_0}}[y|c^M(x,z)]$ by Theorem 2.4 of Diaconis and Freedman (1990) and the strong law of large numbers ($\bar{y}_t(c^M)$ denotes the empirical frequency of $y = 1$ conditional on $(x,z) \in c^M$ up to period t).

In addition, $E_{\boldsymbol{\theta_0}}[y|c^M(x,z)] = E_{\boldsymbol{\theta_0}}[y|x,z]$ for $M = M_{X,Z}$ as well as for $M_{\neg X,Z}$ when $M_{\neg X,Z}$ is the true model. Consequently, it is left to show that $\pi^t_{X,\neg Z}$ and $\pi^t_{\neg X,\neg Z}$ converge almost surely to zero and that $\pi^t_{\neg X,Z}$ converges almost surely to zero whenever $M_{\neg X,Z}$ is *not* the true model. But these statements follow immediately from Lemma 5. ■

Proof of Observation 1.2. Lemma 5 shows that both $\mathcal{B}^t_{X,\neg Z}$ and $\mathcal{B}^t_{\neg X,\neg Z}$ converge almost surely to zero and that $\mathcal{B}^t_{\neg X,Z}$ converges almost surely to zero whenever x is important to predicting y. As a result, in order to prove that the standard Bayesian learns the true model almost surely it is left to show that $\mathcal{B}^t_{\neg X,Z} \overset{a.s.}{\to} \infty$ whenever x is not important to predicting y. First, write out the Bayes' factor:

$$\mathcal{B}^t_{\neg X,Z} = \frac{\Pr(h^t|M_{\neg X,Z})}{\Pr(h^t|M_{X,Z})} \tag{44}$$

$$= \prod_{z'} \frac{\int_0^1 \theta^{\kappa(z'|h^t)}(1-\theta)^{N(z'|h^t)-\kappa(z'|h^t)}\psi(\theta)d\theta}{\prod_{x'} \int_0^1 \theta^{\kappa(x',z'|h^t)}(1-\theta)^{N(x',z'|h^t)-\kappa(x',z'|h^t)}\psi(\theta)d\theta}. \tag{45}$$

From (45) it is sufficient to show that

$$\frac{\int_0^1 \theta^{\kappa(z|h^t)}(1-\theta)^{N(z|h^t)-\kappa(z|h^t)}\psi(\theta)d\theta}{\prod_{x'} \int_0^1 \theta^{\kappa(x',z|h^t)}(1-\theta)^{N(x',z|h^t)-\kappa(x',z|h^t)}\psi(\theta)d\theta} \overset{a.s.}{\to} \infty \tag{46}$$

for each $z \in \mathsf{Z}$.

Fix some z. I will use Lemma 4 to estimate (46). Let $\kappa_t = \kappa(z|h^t), N_t = N(z|h^t), \hat{q}_t = \frac{\kappa_t}{N_t}, \kappa_t^{x'} = \kappa(x',z|h^t), N_t^{x'} = N(x',z|h^t)$, and $\hat{q}_t^{x'} = \frac{\kappa_t^{x'}}{N_t^{x'}}$.

Applying Lemma 4,

$$\frac{\int_0^1 \theta^{\kappa_t}(1-\theta)^{N_t-\kappa_t}\psi(\theta)d\theta}{\prod_{x'} \int_0^1 \theta^{\kappa_t^{x'}}(1-\theta)^{N_t^{x'}-\kappa_t^{x'}}\psi(\theta)d\theta} \geq \frac{a\phi^*(\kappa_t, N_t)}{A^{\#\mathsf{X}} \prod_{x'} \phi^*(\kappa_t^{x'}, N_t^{x'})} \tag{47}$$

for some constants $0 < a < A < \infty$. By the strong law of large numbers, the right hand side of (47) tends almost surely towards

$$C\frac{\sqrt{\prod_{x'} N_t^{x'}}}{\sqrt{N_t}} \overset{a.s.}{\to} \infty$$

where C is some positive constant independent of t. ■

B.2. **Proofs of results from Section 3.** I now present a series of Lemmas which will aid in proving results from Section 3. Define

$$p_0^0(y, x, \hat{z}) = \begin{cases} \sum_{z' \in \mathsf{Z}} \theta_0(x, z')^y (1-\theta_0(x, z'))^{1-y} g(x, z') & \text{for each } y, x, \text{ and } \hat{z} = \varnothing \\ 0 & \text{for } \hat{z} \neq \varnothing \end{cases} \tag{48}$$

to equal the distribution over $(y, x, \hat{z})$ conditional on the agent not encoding z. Lemma 6 establishes the almost sure limit of several Bayes' factors when the agent never encodes z.

Lemma 6. *Suppose $E_{\boldsymbol{\theta}_0}[y|x] \neq E_{\boldsymbol{\theta}_0}[y]$ for some $x \in \mathsf{X}$. Then, for all uniformly non-doctrinaire $\mu^{\neg X, \neg Z}$, $\mu^{\neg X, Z}$, and $\mu^{X,Z}$, $\mathcal{B}_{\neg X, \neg Z}(\hat{h}^t|p_0^0) \to 0$ and $\mathcal{B}_{\neg X, Z}(\hat{h}^t|p_0^0) \to 0$, $(p_0^0)^\infty$ almost surely.*

Proof. Lemma 3 implies that it is sufficient to show that $\delta_{\neg X, \neg Z}(p_0^0) > \delta_{X,Z}(p_0^0)$ and $\delta_{\neg X, Z}(p_0^0) > \delta_{X,Z}(p_0^0)$ whenever $E_{\boldsymbol{\theta}_0}[y|x] \neq E_{\boldsymbol{\theta}_0}[y]$ for some $x \in \mathsf{X}$. Can easily verify these inequalities by applying Lemma 1 for each $M \in \mathcal{M}$. ■

The next Lemma establishes some finite sample properties of Bayes' factors when the agent never encodes z. First, define

$$\hat{h}_m^t = (y_{t-1}, x_{t-1}, \varnothing, \ldots, y_1, x_1, \varnothing)$$

$$\pi_Z(\hat{h}_m^t|p_0^0) = \frac{1 + \frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X, Z}(\hat{h}_m^t|p_0^0)}{1 + \frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X, Z}(\hat{h}_m^t|p_0^0) + \frac{1-\pi_Z}{\pi_Z}\mathcal{B}_{X, \neg Z}(\hat{h}_m^t|p_0^0) + \frac{(1-\pi_X)(1-\pi_Z)}{\pi_X \pi_Z}\mathcal{B}_{\neg X, \neg Z}(\hat{h}_m^t|p_0^0)}$$

Lemma 7. *For all t, $\hat{h}_m^t$, and ψ,*

$$\mathcal{B}_{\neg X, Z}(\hat{h}_m^t|p_0^0) = \mathcal{B}_{\neg X, \neg Z}(\hat{h}_m^t|p_0^0) \tag{49}$$

$$\mathcal{B}_{X, \neg Z}(\hat{h}_m^t|p_0^0) = 1 \tag{50}$$

$$\pi_Z(\hat{h}_m^t|p_0^0) = \pi_Z. \tag{51}$$

Proof.

$$\begin{aligned}
\mathcal{B}_{\neg X,Z}(\hat{h}^t_m|p^0_0) &= \frac{\int_0^1 \theta^{\kappa(m|\hat{h}^t_m)}(1-\theta)^{N(m|\hat{h}^t_m)-\kappa(m|\hat{h}^t_m)}\psi(\theta)d\theta}{\prod_{x'}\int_0^1 \theta^{\kappa(x',m|\hat{h}^t_m)}(1-\theta)^{N(x',m|\hat{h}^t_m)-\kappa(x',m|\hat{h}^t_m)}\psi(\theta)d\theta} \\
&= \frac{\int_0^1 \theta^{\bar{\kappa}(\hat{h}^t_m)}(1-\theta)^{t-1-\bar{\kappa}(\hat{h}^t_m)}\psi(\theta)d\theta}{\prod_{x'}\int_0^1 \theta^{\kappa(x'|\hat{h}^t_m)}(1-\theta)^{N(x'|\hat{h}^t_m)-\kappa(x'|\hat{h}^t_m)}\psi(\theta)d\theta} \\
\mathcal{B}_{\neg X,\neg Z}(\hat{h}^t_m|p^0_0) &= \frac{\int_0^1 \theta^{\bar{\kappa}(\hat{h}^t_m)}(1-\theta)^{t-1-\bar{\kappa}(\hat{h}^t_m)}\psi(\theta)d\theta}{\prod_{x'}\int_0^1 \theta^{\kappa(x'|\hat{h}^t_m)}(1-\theta)^{N(x'|\hat{h}^t_m)-\kappa(x'|\hat{h}^t_m)}\psi(\theta)d\theta} \\
&= \mathcal{B}_{\neg X,Z}(\hat{h}^t_m|p^0_0) \\
\mathcal{B}_{X,\neg Z}(\hat{h}^t_m|p^0_0) &= \frac{\prod_{x'}\int_0^1 \theta^{\kappa(x'|\hat{h}^t_m)}(1-\theta)^{N(x'|\hat{h}^t_m)-\kappa(x'|\hat{h}^t_m)}\psi(\theta)d\theta}{\prod_{x'}\int_0^1 \theta^{\kappa(x'|\hat{h}^t_m)}(1-\theta)^{N(x'|\hat{h}^t_m)-\kappa(x'|\hat{h}^t_m)}\psi(\theta)d\theta} \\
&= 1.
\end{aligned}$$

Plugging these expressions into the definition of $\pi_Z(\hat{h}^t_m|p^0_0)$ yields

$$\begin{aligned}
\pi_Z(\hat{h}^t_m|p^0_0) &= \frac{\pi_Z\left[1+\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,Z}(\hat{h}^t_m|p^0_0)\right]}{1+\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,Z}(\hat{h}^t_m|p^0_0)} \\
&= \pi_Z.
\end{aligned}$$

■

Lemma 8. *Suppose that, with positive probability under $P_{\boldsymbol{\theta_0},\xi}(\cdot)$, the agent encodes z infinitely often. Conditional on the agent encoding z infinitely often, $\hat{\pi}^t_Z \to 1$ almost surely.*

Proof. I want to show that, conditional on the agent encoding z infinitely often, $\mathcal{B}_{X,\neg Z}(\hat{h}^t) \to 0$ and $\mathcal{B}_{\neg X,\neg Z}(\hat{h}^t) \to 0$ with probability 1. Equivalently, I will establish that

$$\log(\mathcal{B}_{i,\neg Z}(\hat{h}^t)) \to -\infty \tag{52}$$

for each $i \in \{X, \neg X\}$.

Defining

$$\begin{aligned}
\hat{h}^t_1 &\equiv (y_\tau, x_\tau, \hat{z}_\tau)_{\tau<t:\hat{z}_\tau\neq\varnothing} \\
\hat{h}^t_0 &\equiv (y_\tau, x_\tau, \hat{z}_\tau)_{\tau<t:\hat{z}_\tau=\varnothing}
\end{aligned}$$

we can write

$$\mathcal{B}^t_{i,\neg Z} = \frac{\Pr_\xi(\hat{h}^t_0|M_{i,\neg Z})\Pr_\xi(\hat{h}^t_1|M_{i,\neg Z},\hat{h}^t_0)}{\Pr_\xi(\hat{h}^t_0|M_{X,Z})\Pr_\xi(\hat{h}^t_1|M_{X,Z},\hat{h}^t_0)}$$

so the LHS of (52) can be expressed as

$$\log\left(\frac{\Pr_\xi(\hat{h}^t_0|M_{i,\neg Z})}{\Pr_\xi(\hat{h}^t_0|M_{X,Z})}\right) + \log\left(\frac{\Pr_\xi(\hat{h}^t_1|M_{i,\neg Z},\hat{h}^t_0)}{\Pr_\xi(\hat{h}^t_1|M_{X,Z},\hat{h}^t_0)}\right). \tag{53}$$

When the agent fails to encode z only a finite number of times along a history, we can ignore the first term of (53) because it tends towards a finite value as $t \to \infty$. Otherwise, Lemma 7 says that the first term of (53) is identically 0 for $i = X$, as well as for $i = \neg X$ when X is a singleton; Lemma 6 (together with Assumption 2) says that the first term tends towards $-\infty$ with probability 1 for $i = \neg X$ when X contains at least two elements. As a result, no matter which case we are in it is sufficient to show that the second term of (53) tends towards $-\infty$ with probability 1 in order to establish (52). This can be verified by showing that

$$\limsup_t \frac{1}{\#\mathcal{E}(t)}\log\left(\frac{\Pr_\xi(\hat{h}^t_1|M_{i,\neg Z},\hat{h}^t_0)}{\prod_{\tau\in\mathcal{E}(t)} p^1_0(y_\tau,x_\tau,z_\tau)}\right) - \frac{1}{\#\mathcal{E}(t)}\log\left(\frac{\Pr_\xi(\hat{h}^t_1|M_{X,Z},\hat{h}^t_0)}{\prod_{\tau\in\mathcal{E}(t)} p^1_0(y_\tau,x_\tau,z_\tau)}\right) < 0 \tag{54}$$

with probability 1 for $i \in \{X, \neg X\}$, where

$$p^1_0(y,x,z) = \theta_0(x,z)^y(1-\theta_0(x,z))^{1-y}g(x,z)$$
$$\mathcal{E}(t) = \{\tau < t : \hat{z}_\tau \neq \varnothing\}.$$

The second term on the LHS of (54) tends towards 0 with probability 1 by Lemma 2.[43] To complete the proof, it then remains to show that the first term on the LHS of (54) remains bounded away from 0 as $t \to \infty$ for $i \in \{X, \neg X\}$, or

$$\limsup_t \frac{1}{\#\mathcal{E}(t)}\log\left(\frac{\Pr_\xi(\hat{h}^t_1|M_{i,\neg Z},\hat{h}^t_0)}{\prod_{\tau\in\mathcal{E}(t)} p^1_0(y_\tau,x_\tau,z_\tau)}\right) < 0. \tag{55}$$

[43] Note that

$$\frac{1}{\#\mathcal{E}(t)}\log\left(\frac{\Pr_\xi(\hat{h}^t_1|M_{X,Z},\hat{h}^t_0)}{\prod_{\tau\in\mathcal{E}(t)} p^1_0(y_\tau,x_\tau,z_\tau)}\right) = \frac{1}{\#\mathcal{E}(t)}\log\left(\frac{\Pr_\xi(\hat{h}^t_1|M_{X,Z})}{\prod_{\tau\in\mathcal{E}(t)} p^1_0(y_\tau,x_\tau,z_\tau)}\right)$$

since, under $\mu^{X,Z}$, subjective uncertainty regarding $\theta(x,\varnothing)$ and $\theta(x,z'), z' \neq \varnothing$, is independent.

We can re-write the LHS of (55) as

$$(56)\qquad \frac{1}{\#\mathcal{E}(t)}\log\left(\frac{\prod_{x'}\prod_{z'}\int\theta(x',z')^{\kappa(x',z'|\hat{h}_1^t)}(1-\theta(x',z'))^{N(x',z'|\hat{h}_1^t)-\kappa(x',z'|\hat{h}_1^t)}\mu^{i,\neg Z}(d\boldsymbol{\theta}|\hat{h}_0^t)}{\prod_{x'}\prod_{z'}\theta_0(x',z')^{\kappa(x',z'|\hat{h}_1^t)}(1-\theta_0(x',z'))^{N(x',z'|\hat{h}_1^t)-\kappa(x',z'|\hat{h}_1^t)}}\right).$$

Since $\mu^{i,\neg Z}(\cdot|\hat{h}_0^t)$ places full support on vectors of success probabilities ($\boldsymbol{\theta}$) with $\theta(x,z)=\theta(x,z')$ for all x,z,z', we can bound (56) by noting that

$$\begin{aligned}\prod_{x'}\prod_{z'}\int\theta(x',z')^{\kappa(x',z'|\hat{h}_1^t)}(1-\theta(x',z'))^{N(x',z'|\hat{h}_1^t)-\kappa(x',z'|\hat{h}_1^t)}\mu^{i,\neg Z}(d\boldsymbol{\theta}|\hat{h}_0^t)&\\ \le\max_{\theta(0),\theta(1)}\prod_{x'}\prod_{z'}\theta(x')^{\kappa(x',z'|\hat{h}_1^t)}(1-\theta(x'))^{N(x',z'|\hat{h}_1^t)-\kappa(x',z'|\hat{h}_1^t)}&\\ =\prod_{x'}\prod_{z'}{\frac{\kappa(x'|\hat{h}_1^t)}{N(x'|\hat{h}_1^t)}}^{\kappa(x',z'|\hat{h}_1^t)}\left(1-\frac{\kappa(x'|\hat{h}_1^t)}{N(x'|\hat{h}_1^t)}\right)^{N(x',z'|\hat{h}_1^t)-\kappa(x',z'|\hat{h}_1^t)}&\end{aligned}$$

which implies that (56) is bounded above by

$$(57)\qquad \frac{1}{\#\mathcal{E}(t)}\log\left(\frac{\prod_{x'}\prod_{z'}{\frac{\kappa(x'|\hat{h}_1^t)}{N(x'|\hat{h}_1^t)}}^{\kappa(x',z'|\hat{h}_1^t)}\left(1-\frac{\kappa(x'|\hat{h}_1^t)}{N(x'|\hat{h}_1^t)}\right)^{N(x',z'|\hat{h}_1^t)-\kappa(x',z'|\hat{h}_1^t)}}{\prod_{x'}\prod_{z'}\theta_0(x',z')^{\kappa(x',z'|\hat{h}_1^t)}(1-\theta_0(x',z'))^{N(x',z'|\hat{h}_1^t)-\kappa(x',z'|\hat{h}_1^t)}}\right).$$

for all $t,\hat{h}^t$. By the strong law of large numbers, expression (57) can be shown to tend towards $-d_K(\hat{p}_{\underline{\boldsymbol{\theta}}(M_{X,\neg Z})},p_0^1)<0$ with probability 1 conditional on the agent encoding z infinitely often; this establishes (55) and completes the proof. ■

Proof of Proposition 1. Suppose that, with positive probability under $P_{\boldsymbol{\theta_0},\xi}(\cdot)$, the agent does not settle on encoding or not encoding z (b must satisfy $0<b<1$). Label this event NS and condition on $\hat{h}^\infty\in NS$. Since the agent encodes z infinitely often conditional on NS, by Lemma 8 we must have $\hat{\pi}_Z^t\to 1$ with probability 1. As a result, with probability 1 there exists a $\tilde{t}$ such that $\hat{\pi}_Z^t>b$ for all $t\ge\tilde{t}$ so $e_t=1$ for all $t\ge\tilde{t}$, a contradiction. ■

A few Lemmas will be useful to establish Proposition 2. First, define

$$(58)\qquad \Lambda(h^t)\equiv\frac{1+\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,Z}(h^t)}{\mathcal{B}_{X,\neg Z}(h^t)+\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,\neg Z}(h^t)},$$

which can be thought of as a likelihood ratio (or Bayes' factor) comparing the likelihood of a history under models where z is important to predicting y versus the likelihood of that history under models where z is unimportant to predicting y.

Lemma 9. *$\pi_Z(h^t) > b$ if and only if $\Lambda(h^t) > \frac{1-\pi_Z}{\pi_Z}\frac{b}{1-b}$.*

Proof.

$$\pi_Z(h^t) > b \iff$$

$$\frac{1+\alpha_{\neg X,Z}\mathcal{B}_{\neg X,Z}(h^t)}{1+\alpha_{\neg X,Z}\mathcal{B}_{\neg X,Z}(h^t)+\alpha_{X,\neg Z}\mathcal{B}_{X,\neg Z}(h^t)+\alpha_{\neg X,\neg Z}\mathcal{B}_{\neg X,\neg Z}(h^t)} > b \iff$$

$$\frac{1+\alpha_{\neg X,Z}\mathcal{B}_{\neg X,Z}(h^t)}{\alpha_{X,\neg Z}\mathcal{B}_{X,\neg Z}(h^t)+\alpha_{\neg X,\neg Z}\mathcal{B}_{\neg X,\neg Z}(h^t)} > \frac{b}{1-b} \iff$$

$$\frac{1+\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,Z}(h^t)}{\frac{1-\pi_Z}{\pi_Z}\mathcal{B}_{X,\neg Z}(h^t)+\frac{(1-\pi_X)(1-\pi_Z)}{\pi_X\pi_Z}\mathcal{B}_{\neg X,\neg Z}(h^t)} > \frac{b}{1-b} \text{ (recall } \alpha_{i,j} = \frac{\pi_{i,j}}{\pi_{X,Z}}\text{)} \iff$$

$$\Lambda(h^t) = \frac{1+\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,Z}(h^t)}{\mathcal{B}_{X,\neg Z}(h^t)+\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,\neg Z}(h^t)} > \frac{1-\pi_Z}{\pi_Z}\frac{b}{1-b}$$

■

Lemma 10. *For all $\epsilon > 0$ there exists $\lambda > 0$ such that*

$$P_{\boldsymbol{\theta_0}}\left(\min_{t'\geq 1}\Lambda(h^{t'}) > \lambda\right) \geq 1-\epsilon \tag{59}$$

Proof. Fix $\epsilon > 0$. From Lemma 5, we know that $\mathcal{B}_{X,\neg Z}(h^t) \overset{a.s.}{\to} 0, \mathcal{B}_{\neg X,\neg Z}(h^t) \overset{a.s.}{\to} 0$. As a result, $\Lambda(h^t) = \frac{\left(1+\frac{1+\pi_X}{\pi_X}\mathcal{B}_{\neg X,Z}(h^t)\right)}{\left(\mathcal{B}_{X,\neg Z}(h^t)+\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,\neg Z}(h^t)\right)} \overset{a.s.}{\to} \infty$. Consequently, there exists a value $T \geq 1$ such that $P_{\boldsymbol{\theta_0}}(\min_{t'\geq T}\Lambda(h^{t'}) \geq 1) > 1-\epsilon$ (see, for example, Lemma 7.2.10 in Grimmett and Stirzaker 2001).

Since, in addition, there exists λ $(0 < \lambda < 1)$ such that

$$\min_h \min_{1\leq k\leq T}\Lambda(h^k) > \lambda$$

we have

$$P_{\boldsymbol{\theta_0}}(\min_{t'\geq 1}\Lambda(h^{t'}) > \lambda) \geq 1-\epsilon.$$

■

Lemma 11. *If $\hat{\pi}_Z^k < b$ for all $k < t$ $(t > 1)$ then $\hat{\pi}_Z^t = \pi_Z$.*

Proof. Suppose that $\pi_{Z,\xi}(\hat{h}^k) < b$ for all $k < t$ ($\pi_{Z,\xi}(\hat{h}^k)$ is long-hand for $\hat{\pi}_Z^k$). Then, for all $k < t$, the marginal distribution over $(x_k, \hat{z}_k)$ is identically $p_0^0(x_k, \hat{z}_k)$ since $\xi(z, \hat{h}^k)[\varnothing] = 1$. As a result, $\pi_{Z,\xi}(\hat{h}^t) = \pi_Z(\hat{h}_m^t | p_0^0) = \pi_Z$, where the last equality follows from Lemma 7. ■

Proof of Proposition 2. **Part 1**.

First I show that, for all $\epsilon > 0$, there exists $\pi_1 \in (0, 1)$ (or $b_1 \in (0, 1)$) such that the agent settles on encoding z with probability at least $1 - \epsilon$ for all $\pi_Z \geq \pi_1$ ($b \leq b_1$). Fix ϵ. Note that, whenever $\hat{\pi}_Z^k > b$ for all $k < t$, $\hat{h}^t = h^t$, $\hat{\pi}_Z^t = \pi_Z(h^t)$, and $P_{\boldsymbol{\theta_0},\xi}(\hat{h}^t) = P_{\boldsymbol{\theta_0}}(h^t)$. As a result, it is sufficient to show that there exists $\pi_1 \in (0, 1)$ ($b_1 \in (0, 1)$) such that

$$P_{\boldsymbol{\theta_0}}(\min_{t' \geq 1} \pi_Z(h^{t'}) > b) \geq 1 - \epsilon$$

whenever $\pi_Z \geq \pi_1$ ($b \leq b_1$).

By Lemma 9,

$$\begin{aligned} \pi_Z(h^t) > b &\iff \\ \Lambda(h^t) &> \frac{1 - \pi_Z}{\pi_Z} \frac{b}{1 - b}. \end{aligned}$$

Consequently, $P_{\boldsymbol{\theta_0}}(\min_{t' \geq 1} \pi_Z(h^{t'}) > b) \geq 1 - \epsilon$ if and only if

$$P_{\boldsymbol{\theta_0}}\left(\min_{t' \geq 1} \Lambda(h^{t'}) > \frac{1 - \pi_Z}{\pi_Z} \frac{b}{1 - b}\right) \geq 1 - \epsilon.$$

From Lemma 10 we know that there exists $\lambda(\epsilon) > 0$ such that

$$P_{\boldsymbol{\theta_0}}\left(\min_{t' \geq 1} \Lambda(h^{t'}) > \lambda(\epsilon)\right) \geq 1 - \epsilon,$$

so the result follows from setting π or b to satisfy

$$\begin{aligned} \frac{1 - \pi}{\pi} \frac{b}{1 - b} &= \lambda(\epsilon) \Rightarrow \\ \pi_1 &= \frac{b}{b + \lambda(\epsilon)(1 - b)} \\ b_1 &= \frac{\lambda(\epsilon)\pi}{\pi(\lambda(\epsilon) - 1) + 1}. \end{aligned}$$

Part 2.

It is left to show that, for all $\epsilon > 0$, there exists $\pi_2 \in (0,1)$ such that the agent settles on not encoding z with probability at least $1-\epsilon$ for all $\pi_Z \leq \pi_2$ ($b \geq b_2$). It is sufficient to show that, when $\pi_Z < b$, $\pi_{Z,\xi}(\hat{h}^t) = \pi_Z$ for all $t > 1$. But this follows from Lemma 11. ■

Proof of Proposition 3. **Part 1**. Analagous to the proof of Observation 1.1 and hence omitted.

Part 2. If the agent settles on not encoding z then, by definition, there exists n such that $e_t = 0$ for all $t \geq n$. In any period $t \geq n$, the agent's expectation satisfies

$$\begin{aligned}\hat{E}[y|x,z,\hat{h}^t] &= E_\xi[\theta(x,\varnothing)|\hat{h}^t] \\ &= \sum_{M'\in\mathcal{M}} \hat{\pi}^t_{M'} E_\xi[\theta(c^{M'}(x,\varnothing))|\hat{h}^t, M'].\end{aligned}$$

Fix an $M \in \mathcal{M}$. Since the marginal prior density over $\theta(c^M(x,\varnothing))$ is non-doctrinaire under M, $E_\xi[\theta(c^M(x,\varnothing))|\hat{h}^t, M] \to \bar{y}_t(c^M(x,\varnothing)) \overset{a.s.}{\to} E_{\boldsymbol{\theta_0}}[y|c^M(x,\varnothing)]$ by Theorem 2.4 of Diaconis and Freedman (1990) and the strong law of large numbers, where $E_{\boldsymbol{\theta_0}}[y|c^M(x,\varnothing)] = E_{\boldsymbol{\theta_0}}[y|x]$ for $M \in \{M_{X,Z}, M_{X,\neg Z}\}$ and $E_{\boldsymbol{\theta_0}}[y|c^M(x,\varnothing)] = E_{\boldsymbol{\theta_0}}[y]$ for $M \in \{M_{\neg X,Z}, M_{\neg X,\neg Z}\}$.

If $E_{\boldsymbol{\theta_0}}[y|x] = E_{\boldsymbol{\theta_0}}[y]$, then we are done. Assume $E_{\boldsymbol{\theta_0}}[y|x] \neq E_{\boldsymbol{\theta_0}}[y]$ for some x. It is left to show that both $\hat{\pi}^t_{\neg X,Z}$ and $\hat{\pi}^t_{\neg X,\neg Z}$ converge almost surely to zero. Equivalently, it is left to show that both $\mathcal{B}_{\neg X,Z}(\hat{h}^t)$ and $\mathcal{B}_{\neg X,\neg Z}(\hat{h}^t)$ converge almost surely to zero.

For $t \geq n$ and $j \in \{\neg Z, Z\}$,

$$\begin{aligned}\mathcal{B}_{\neg X,j}(\hat{h}^t) &= \frac{\Pr_\xi(\hat{h}^t|M_{\neg X,j})}{\Pr_\xi(\hat{h}^t|M_{X,Z})} \\ &= \frac{\Pr_\xi(\hat{h}^t_n|M_{\neg X,j},\hat{h}^n)\Pr_\xi(\hat{h}^n|M_{\neg X,j})}{\Pr_\xi(\hat{h}^t_n|M_{X,Z},\hat{h}^n)\Pr_\xi(\hat{h}^n|M_{X,Z})} \\ &= \left(\frac{\int\prod_{k=n}^{t-1}\theta(x_k,\varnothing)^{y_k}(1-\theta(x_k,\varnothing))^{1-y_k}\mu^{\neg X,j}(d\boldsymbol{\theta}|\hat{h}^n)}{\int\prod_{k=n}^{t-1}\theta(x_k,\varnothing)^{y_k}(1-\theta(x_k,\varnothing))^{1-y_k}\mu^{X,Z}(d\boldsymbol{\theta}|\hat{h}^n)}\right)\frac{\Pr_\xi(\hat{h}^n|M_{\neg X,j})}{\Pr_\xi(\hat{h}^n|M_{X,Z})},\end{aligned}$$

where $\hat{h}^t_n = (y_{t-1}, x_{t-1}, \varnothing, \ldots, y_n, x_n, \varnothing)$. Since $\frac{\Pr_\xi(\hat{h}^n|M_{\neg X,j})}{\Pr_\xi(\hat{h}^n|M_{X,Z})}$ is fixed for all $t \geq n$, it is necessary and sufficient to show that

$$\left(\frac{\int\prod_{k=n}^{t-1}\theta(x_k,\varnothing)^{y_k}(1-\theta(x_k,\varnothing))^{1-y_k}\mu^{\neg X,j}(d\boldsymbol{\theta}|\hat{h}^n)}{\int\prod_{k=n}^{t-1}\theta(x_k,\varnothing)^{y_k}(1-\theta(x_k,\varnothing))^{1-y_k}\mu^{X,Z}(d\boldsymbol{\theta}|\hat{h}^n)}\right) \tag{60}$$

converges to zero almost surely to establish such convergence of the Bayes' factor. But, noting that (i) (60) equals $\mathcal{B}^t_{\neg X,j}(p_0^0)$ for some uniformly non-doctrinaire $\mu^{\neg X,j}$, $\mu^{X,Z}$, and (ii) $(y_{t-1}, x_{t-1}, \hat{z}_{t-1}, \ldots, y_n, x_n, \hat{z}_n)$ is a random sample from p_0^0, the result follows from Lemma 6. ■

Proof of Proposition 4. **<u>Part 1</u>**. Analagous to the proof of Observation 1.2 and hence omitted.
<u>Part 2</u>. The fact that $\hat{\pi}^t_{\mathsf{X}} \overset{a.s.}{\to} 1$ when the agent settles on not encoding z follows immediately from Assumption 2 and Lemma 6. That $\hat{\pi}^t_{\mathsf{Z}} \leq b$ for large t follows from the definition of settling on not encoding z and the encoding rule. ■

B.3. **Proofs of results from Section 4.**

Proof of Proposition 5. A version of this result appears in Samuels (1993), but, for completeness, I'll provide a proof.[44]

Let the true distribution over (y, x, z) be denoted by $p_0(\cdot)$ (the distribution generated by $\boldsymbol{\theta_0}$ and $g(\cdot)$) and let $\mathbb{E}$ denote the expectation operator under $p_0(\cdot)$. With this notation,

$$\mathcal{R}_x(z') = \mathbb{E}[y|x=1, z'] - \mathbb{E}[y|x=0, z']$$

$$\mathcal{R}_x = \mathbb{E}[\mathcal{R}_x(z)|x=1]$$

$$\phi = \mathrm{Cov}(\mathbb{E}[y|x=0, z], g(x=1|z))$$

From Proposition 3, $\hat{E}[y|x=1, z] - \hat{E}[y|x=0, z]$ almost surely equals

$$\begin{aligned}
&\mathbb{E}[y|x=1] - \mathbb{E}[y|x=0] \\
&= \mathbb{E}[\mathbb{E}[y|x,z]|x=1] - \mathbb{E}[\mathbb{E}[y|x,z]|x=0] \\
&= \frac{\mathbb{E}[\mathbb{E}[y|x=1,z]g(x=1|z)](1-g(x=1)) - \mathbb{E}[\mathbb{E}[y|x=0,z]g(x=0|z)]g(x=1)}{g(x=1)(1-g(x=1))} \\
&= \frac{\mathbb{E}[\mathcal{R}_x(z)g(x=1|z)](1-g(x=1))}{g(x=1)(1-g(x=1))} + \frac{\mathbb{E}[\mathbb{E}[y|x=0,z](g(x=1|z)-g(x=1))]}{g(x=1)(1-g(x=1))} \\
&= \mathcal{R}_x + \frac{\phi}{g(x=1)(1-g(x=1))} \\
&= \mathcal{R}_x + \frac{\phi}{\mathrm{Var}(x)}
\end{aligned}$$

■

[44] My proof closely follows Samuels's.

B.4. **Proofs of results from Section 5.**

Lemma 12. *If the continuous attention assumptions hold then the agent almost surely encodes z an infinite number of times.*

Proof. Fix some t. I will show that the probability of never encoding z after t is bounded above by 0. Independent of the history before t, the probability of not encoding z at $t+k$ ($k>0$) given not having encoded z between t and $t+k$ is strictly less than

$$1-\eta\left(\frac{b\pi_Z}{a(1-\pi_Z)+b\pi_Z}\right)<1, \tag{61}$$

where a and b are positive constants (do not depend on k).[45] As a result, the probability of never encoding z after t is less than the infinite product

$$\left(1-\eta\left(\frac{b\pi_Z}{a(1-\pi_Z)+b\pi_Z}\right)\right)^{\infty}=0$$

and the result follows. ■

Proof of Proposition 6.

Part (1): From Lemma 12 we know that the agent almost surely encodes z an infinite number of times. Combining this result with Lemma 8, we have that $\hat{\pi}_Z^t \to 1$ almost surely which implies that $\eta(\hat{\pi}_Z^t)\to 1$ almost surely.

Part (2): Fix (x,z) and $\epsilon>0$. Want to show that

$$\lim_{t\to\infty} P_{\boldsymbol{\theta_0},\xi}(|\hat{E}[y|x,z,\hat{h}^t]-E_{\boldsymbol{\theta_0}}[y|x,z]|>\epsilon)=0 \tag{62}$$

[45] Straightforward computations establish that whenever the agent does not encode z between t and $t+k$,

$$\hat{\pi}_Z^{t+k}=\frac{1+\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,Z}(\hat{h}_1^t|\hat{h}_0^{t+k})\mathcal{B}_{\neg X,Z}(\hat{h}_0^{t+k})}{1+\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,Z}(\hat{h}_1^t|\hat{h}_0^{t+k})\mathcal{B}_{\neg X,Z}(\hat{h}_0^{t+k})+\frac{1-\pi_Z}{\pi_Z}\mathcal{B}_{X,\neg Z}(\hat{h}_1^t|\hat{h}_0^{t+k})+\frac{1-\pi_X}{\pi_X}\frac{1-\pi_Z}{\pi_Z}\mathcal{B}_{\neg X,\neg Z}(\hat{h}_1^t|\hat{h}_0^{t+k})\mathcal{B}_{\neg X,Z}(\hat{h}_0^{t+k})},$$

where $\mathcal{B}_{i,j}(\hat{h}_1^t|\hat{h}_0^{t+k})\equiv\frac{\Pr_\xi(\hat{h}_1^t|M_{i,j},\hat{h}_0^{t+k})}{\Pr_\xi(\hat{h}_1^t|M_{X,Z},\hat{h}_0^{t+k})}$ (recall that $\hat{h}_1^j=(y_\tau,x_\tau,\hat{z}_\tau)_{\tau<j:\hat{z}_\tau\neq\varnothing}$ and $\hat{h}_0^j=(y_\tau,x_\tau,\hat{z}_\tau)_{\tau<j:\hat{z}_\tau=\varnothing}$). Upper bound (61) is derived by noting that, fixing t, both $\mathcal{B}_{X,\neg Z}(\hat{h}_1^t|\hat{h}_0^{t+k})$ and $\mathcal{B}_{\neg X,\neg Z}(\hat{h}_1^t|\hat{h}_0^{t+k})$ are bounded above by some finite positive constant a independent of k and history $\hat{h}^{t+k}$. Likewise, fixing t, $\mathcal{B}_{\neg X,Z}(\hat{h}_1^t|\hat{h}_0^{t+k})$ is bounded below by some finite positive constant $\tilde{b}$ independent of k and history $\hat{h}^{t+k}$. As a result,

$$\hat{\pi}_Z^{t+k}>\frac{1+\frac{1-\pi_X}{\pi_X}\tilde{b}\mathcal{B}_{\neg X,Z}(\hat{h}_0^{t+k})}{1+\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,Z}(\hat{h}_0^{t+k})\left[\tilde{b}+\frac{1-\pi_Z}{\pi_Z}a\right]+\frac{1-\pi_Z}{\pi_Z}a}.$$

Now take the infimum of the right hand side of this expression with respect to all possible values of $\mathcal{B}_{\neg X,Z}(\hat{h}_0^{t+k})$ to get (61).

Expanding

$$P_{\boldsymbol{\theta_0},\xi}(|\hat{E}[y|x,z,\hat{h}^t] - E_{\boldsymbol{\theta_0}}[y|x,z]| > \epsilon) = P_{\boldsymbol{\theta_0},\xi}(|E_\xi[\theta(x,z)|\hat{h}^t] - E_{\boldsymbol{\theta_0}}[y|x,z]| > \epsilon)P_{\boldsymbol{\theta_0},\xi}(e_t = 1)$$
$$+ P_{\boldsymbol{\theta_0},\xi}(|E_\xi[\theta(x,\varnothing)|\hat{h}^t] - E_{\boldsymbol{\theta_0}}[y|x,z]| > \epsilon)(1 - P_{\boldsymbol{\theta_0},\xi}(e_t = 1)),$$

to establish (62) it is sufficient to show that

A. $E_\xi[\theta(x,z)|\hat{h}^t] \overset{a.s.}{\to} E_{\boldsymbol{\theta_0}}[y|x,z]$

B. $P_{\boldsymbol{\theta_0},\xi}(e_t = 1) \to 1$

A. follows from now familiar arguments applying the non-doctrinaire assumption, the strong law of large numbers, the consistency properties of Bayes' factors, and the fact that the agent encodes z an infinite number of times (Lemma 12). B. follows from the fact that $P_{\boldsymbol{\theta_0},\xi}(e_t = 1) = E_{\boldsymbol{\theta_0},\xi}[\eta(\hat{\pi}_Z^t)]$ and $E_{\boldsymbol{\theta_0},\xi}[\eta(\hat{\pi}_Z^t)] \to 1$ because $\eta(\hat{\pi}_Z^t)$ is bounded and tends almost surely towards 1 by Proposition 6.1. ■

The next Lemma demonstrates that the fraction of time that the agent spends encoding z tends towards 1 assuming continuous attention. Recall that $\mathcal{E}(t) = \{\tau < t : \hat{z}_\tau \neq \varnothing\}$ denotes the number of times the agent encodes z prior to period t.

Lemma 13. *If the continuous attention assumptions hold then* $\frac{\#\mathcal{E}(t)}{t-1} \overset{a.s.}{\to} 1$.

Proof. Define $\gamma^t = \frac{\#\mathcal{E}(t)}{t-1}$. I will apply a result from the theory of stochastic approximation to show that $\gamma^t \overset{a.s.}{\to} 1$ (Benaim 1999). We have

$$\begin{aligned}\gamma^t - \gamma^{t-1} &= \frac{e_t - \gamma^{t-1}}{t-1} \\ &= \frac{1}{t-1}(F(\gamma^{t-1}) + \epsilon_t + u_t),\end{aligned}$$

where

$$\begin{aligned}F(\gamma^{t-1}) &= 1 - \gamma^{t-1} \\ \epsilon_t &= e_t - \eta(\hat{\pi}_Z^t) \\ u_t &= \eta(\hat{\pi}_Z^t) - 1.\end{aligned}$$

Note that

(1) F is Lipschitz continuous and is defined on a compact set $[0, 1]$

(2) $E[\epsilon_t|\hat{h}^t] = 0$ and $E(|\epsilon_t|^2) < \infty$

(3) $u_t \overset{a.s.}{\to} 0$

so Theorem A in Fudenberg and Takahashi (2008) tells us that, with probability 1, every ω-limit of the process $\{\gamma^t\}$ is connected and internally chain recurrent for Φ, where Φ is the continuous time semi-flow induced by

$$\dot{\gamma}(t) = F(\gamma(t)).$$

Since $F'(\gamma) = -1 < 0$ and the unique steady state of the continuous time process is $\gamma = 1$, the only connected and internally chain recurrent set for Φ is $\{1\}$ by Liouville's Theorem. ■

Lemma 14. $d = \delta_{X,\neg Z}(p_0^1)$

Proof. Apply Lemma 1 to get $\underline{\theta}(c^{X,\neg Z}(x,z)) = p_0^1(y|x) = p_{\boldsymbol{\theta_0}}(y|x)$ for all x. The result then follows from the definition of $\delta_{X,\neg Z}(p_0^1)$. ■

Lemma 15. *If the continuous attention assumptions hold, then* $\frac{\mathcal{B}_{X,\neg Z}(\hat{h}^t)}{e^{-d(t-1)}} \overset{a.s.}{\to} K$ *for some* K *satisfying* $0 < K < \infty$.

Proof. I will show that

$$\frac{1}{t-1}\log \mathcal{B}_{X,\neg Z}(\hat{h}^t) \to -d \tag{63}$$

almost surely. We can write

$$\mathcal{B}_{X,\neg Z}(\hat{h}^t) = \frac{\Pr_\xi(\hat{h}_1^t|M_{X,\neg Z})\Pr_\xi(\hat{h}_0^t|M_{X,\neg Z},\hat{h}_1^t)}{\Pr_\xi(\hat{h}_1^t|M_{X,Z})\Pr_\xi(\hat{h}_0^t|M_{X,Z})} \tag{64}$$

From (64), we can write the left hand side of (63) as

$$\begin{aligned} &\frac{1}{t-1}\left[\log\left(\frac{\Pr_\xi(\hat{h}_1^t|M_{X,\neg Z})}{\prod_{k\in\mathcal{E}(t)} p_0^1(y_k,x_k,z_k)}\right) + \log\left(\frac{\Pr_\xi(\hat{h}_0^t|M_{X,\neg Z},\hat{h}_1^t)}{\prod_{k\notin\mathcal{E}(t)} p_0^0(y_k,x_k,\varnothing)}\right)\right] \\ &-\frac{1}{t-1}\left[\log\left(\frac{\Pr_\xi(\hat{h}_1^t|M_{X,Z})}{\prod_{k\in\mathcal{E}(t)} p_0^1(y_k,x_k,z_k)}\right) + \log\left(\frac{\Pr_\xi(\hat{h}_0^t|M_{X,Z})}{\prod_{k\notin\mathcal{E}(t)} p_0^0(y_k,x_k,\varnothing)}\right)\right] \end{aligned} \tag{65}$$

We know that the second term of (65) tends almost surely towards 0 as $t \to \infty$ by Lemma 2.

As a result, to establish (63) it remains to show that the first term tends almost surely towards $-d$. Rewrite this term as

(66)
$$\frac{\#\mathcal{E}(t)}{t-1}\left[\frac{1}{\#\mathcal{E}(t)}\log\left(\frac{\Pr_\xi(\hat{h}_1^t|M_{X,\neg Z})}{\prod_{k\in\mathcal{E}(t)} p_0^1(y_k,x_k,z_k)}\right)\right]+\frac{t-1-\#\mathcal{E}(t)}{t-1}\left[\frac{1}{t-1-\#\mathcal{E}(t)}\log\left(\frac{\Pr_\xi(\hat{h}_0^t|M_{X,\neg Z},\hat{h}_1^t)}{\prod_{k\notin\mathcal{E}(t)} p_0^0(y_k,x_k,\varnothing)}\right)\right]$$

By Lemmas 2, 13, and 14, (66) tends almost surely towards

$$1\times -d+0\times 0=-d$$

as $t\rightarrow\infty$, which completes the proof. ■

Lemma 16. *If the continuous attention assumptions hold, then* $\frac{\mathcal{B}_{\neg X,\neg Z}(\hat{h}^t)}{e^{-d'(t-1)}}\overset{a.s.}{\rightarrow} K$ *for some* $d'\geq d$ *and* K *satisfying* $0<K<\infty$.

Proof. Let $d'=\delta_{\neg X,\neg Z}(p_0^1)$. Using analagous arguments to those in the proof of Lemma 15, can show that $\frac{\mathcal{B}_{\neg X,\neg Z}(\hat{h}^t)}{e^{-\delta_{\neg X,\neg Z}(p_0^1)(t-1)}}\overset{a.s.}{\rightarrow} K$ for some K satisfying $0<K<\infty$. Since $\delta_{\neg X,\neg Z}(p_0^1)>\delta_{X,\neg Z}(p_0^1)$ (from the fact that adding more constraints weakly increases the minimized Kullback-Leibler divergence) and $\delta_{X,\neg Z}(p_0^1)=d$ (by Lemma 14), the result follows. ■

Lemma 17. *Suppose the continuous attention assumptions hold. If the asymptotic rate of convergence of* $\hat{\pi}_Z^t$ *to* 1 *is* $V(t)$ *then the asymptotic rate of convergence of* $\eta(\hat{\pi}_Z^t)$ *to* 1 *is* $V(t)$.

Proof. Suppose the asymptotic rate of convergence of $\hat{\pi}_Z^t$ to 1 is $V(t)$. Then, by definition, there must exist a strictly positive constant $C<\infty$ such that

$$\frac{1-\hat{\pi}_Z^t}{V(t)}\overset{a.s.}{\rightarrow} C. \tag{67}$$

The goal is to show that (67) implies that there exists a strictly positive constant $C'<\infty$ such that

$$\frac{1-\eta(\hat{\pi}_Z^t)}{V(t)}\overset{a.s.}{\rightarrow} C'. \tag{68}$$

But (68) follows from (67) so long as there exists a strictly positive constant $K<\infty$ such that

$$\frac{1-\eta(\hat{\pi}_Z^t)}{1-\hat{\pi}_Z^t}\overset{a.s.}{\rightarrow} K, \tag{69}$$

which can easily be verified using the continuous differentiability of $\eta(\cdot)$ and l'Hopital's rule. ■

Proof of Proposition 7. From Lemma 17, it is enough to show that

$$\frac{1-\hat{\pi}_Z^t}{e^{-d(t-1)}} \overset{a.s}{\to} C \tag{70}$$

for some strictly positive constant $C < \infty$.

Since

$$1-\hat{\pi}_Z^t = \frac{1}{1+\frac{1+\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,Z}(\hat{h}^t)}{\frac{1-\pi_Z}{\pi_Z}\mathcal{B}_{X,\neg Z}(\hat{h}^t)+\frac{1-\pi_X}{\pi_X}\frac{1-\pi_Z}{\pi_Z}\mathcal{B}_{\neg X,\neg Z}(\hat{h}^t)}},$$

to demonstrate (70) it suffices to show that

$$\frac{e^{-d(t-1)}}{\frac{1-\pi_Z}{\pi_Z}\mathcal{B}_{X,\neg Z}(\hat{h}^t)+\frac{1-\pi_X}{\pi_X}\frac{1-\pi_Z}{\pi_Z}\mathcal{B}_{\neg X,\neg Z}(\hat{h}^t)} + \frac{\frac{1-\pi_X}{\pi_X}\mathcal{B}_{\neg X,Z}(\hat{h}^t)e^{-d(t-1)}}{\frac{1-\pi_Z}{\pi_Z}\mathcal{B}_{X,\neg Z}(\hat{h}^t)+\frac{1-\pi_X}{\pi_X}\frac{1-\pi_Z}{\pi_Z}\mathcal{B}_{\neg X,\neg Z}(\hat{h}^t)} \overset{a.s}{\to} c' \tag{71}$$

for some constant c' satisfying $0 < c' < \infty$.

The first term on the left hand side of (71) converges almost surely to a positive finite constant by Lemmas 15 and 16. The second term on the left hand side of (71) converges almost surely to 0 whenever $\pi_X = 1$ (trivially) or x is important to predicting y (by Lemmas 3, 15, and 16). ■

References

Aigner, D.J. and G.G. Cain, “Statistical Theories of Discrimination in Labor Markets,” *Industrial and Labor Relations Review*, 1977, pp. 175–187.

Allport, G.W., *The Nature of Prejudice*, Reading, MA: Addison-Wesley, 1954.

Altonji, J.G. and R.M. Blank, “Race and Gender in the Labor Market,” *Handbooks In Economics*, 1999, *5* (3 Part C), 3143–3260.

Aragones, E., I. Gilboa, A. Postlewaite, and D. Schmeidler, “Fact-Free Learning,” *American Economic Review*, 2005, *95* (5), 1355–1368.

Arrow, K., “The Theory of Discrimination,” *Discrimination in Labor Markets*, 1973, pp. 3–33.

Barberis, N., A. Shleifer, and R. Vishny, “A Model of Investor Sentiment,” *Journal of Financial Economics*, 1998, *49* (3), 307–343.

Bargh, J.A., “The Ecology of Automaticity: Toward Establishing the Conditions Needed to Produce Automatic Processing Effects,” *The American Journal of Psychology*, 1992, pp. 181–199.

______ and R.D. Thein, “Individual Construct Accessibility, Person Memory, and the Recall-Judgment Link: The Case of Information Overload,” *Journal of Personality and Social Psychology*, 1985, *49* (1), 129–1.

Becker, G.S., *The Economics of Discrimination*, University of Chicago Press, 1971.

Benaim, M., “Dynamics of Stochastic Approximation Algorithms,” *Seminaire de probabilites XXXIII. Berlin: Springer. Lect. Notes Math*, 1999, *1709*, 1–68.

Bertrand, M. and S. Mullainathan, “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment On Labor Market Discrimination,” *American Economic Review*, 2004, pp. 991–1013.

______, D. Chugh, and S. Mullainathan, “Implicit Discrimination,” *American Economic Review*, 2005, *95* (2), 94–98.

Bigler, R.S. and L.S. Liben, “Cognitive Mechanisms in Children’s Gender Stereotyping: Theoretical and Educational Implications of a Cognitive-Based Intervention,” *Child Development*, 1992, pp. 1351–1363.

Branton, R.P. and B.S. Jones, “Reexamining Racial Attitudes: The Conditional Relationship Between Diversity and Socioeconomic Environment,” *American Journal of Political Science*, 2005, pp. 359–372.

Cain, D.M., G. Loewenstein, and D.A. Moore, “The Dirt On Coming Clean: Perverse Effects of Disclosing Conflicts of Interest,” *The Journal of Legal Studies*, 2005, *34* (1), 1–25.

Camerer, C.F. and G. Loewenstein, “Behavioral Economics: Past, Present, Future,” 2003, *Advances in Behavioral Economics.*

Chetty, R., A. Looney, and K. Kroft, “Salience and Taxation: Theory and Evidence,” *American Economic Review*, 2009, *99* (4), 1145–1177.

Cook, S.W., “Experimenting On Social Issues: The Case of School Desegregation,” *American Psychologist*, 1985, *40* (4), 452–460.

Cornell, B. and I. Welch, “Culture, Information, and Screening Discrimination,” *Journal of Political Economy*, 1996, pp. 542–571.

Cover, T.M. and J.A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 2006.

DellaVigna, S., "Psychology and Economics: Evidence From the Field," *NBER Working Paper*, 2007.

Diaconis, P. and D. Freedman, "On the Uniform Consistency of Bayes Estimates for Multinomial Probabilities," *The Annals of Statistics*, 1990, pp. 1317–1327.

______ **and DA Freedman**, "Nonparametric Binary Regression: A Bayesian Approach," *The Annals of Statistics*, 1993, pp. 2108–2137.

Dixon, J.C. and M.S. Rosenbaum, "Nice to Know You? Testing Contact, Cultural, and Group Threat Theories of Anti-Black and Anti-Hispanic Stereotypes," *Social Science Quarterly*, 2004, *85* (2), 257–280.

Dow, J., "Search Decisions with Limited Memory," *The Review of Economic Studies*, 1991, *58* (1), 1–14.

Eagly, A.H. and V.J. Steffen, "Gender Stereotypes Stem From the Distribution of Women and Men Into Social Roles.," *Journal of Personality and Social Psychology*, 1984, *46* (4), 735–754.

Easley, D. and M. O'Hara, "Time and the Process of Security Price Adjustment," *Journal of Finance*, 1992, pp. 577–605.

Ellison, G., "Learning, Local Interaction, and Coordination," *Econometrica*, 1993, pp. 1047–1071.

Eppler, M.J. and J. Mengis, "The Concept of Information Overload: A Review of Literature From Organization Science, Accounting, Marketing, MIS, and Related Disciplines," *The Information Society*, 2004, *20* (5), 325–344.

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

Eyster, E. and M. Rabin, "Cursed Equilibrium," *Econometrica*, 2005, pp. 1623–1672.

Falkinger, J., "Attention Economies," *Journal of Economic Theory*, 2007, *133* (1), 266–294.

Fiedler, K., "Beware of Samples! A Cognitive-Ecological Sampling Approach to Judgment Biases," *Psychological Review*, 2000, *107* (4), 659–676.

Fiske, S.T., "Social Cognition and Social Perception," *Annual Review of Psychology*, 1993, *44* (1), 155–194.

______ **and S.E. Taylor**, *Social Cognition: From Brains to Culture*, McGraw-Hill Higher Education, 2008.

______**, D.A. Kenny, and S.E. Taylor**, "Structural Models for the Mediation of Salience Effects On Attribution," *Journal of Experimental Social Psychology*, 1982, *18* (2), 105–127.

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

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

______ **and** ______, "Superstition and Rational Learning," *American Economic Review*, 2006, pp. 630–651.

______ **and D.M. Kreps**, "Learning in Extensive-Form Games I. Self-Confirming Equilibria," *Games and Economic Behavior*, 1995, *8* (1), 20–55.

______ **and S. Takahashi**, "Heterogeneous Beliefs and Local Information in Stochastic Fictitious Play," *Games and Economic Behavior*, 2008.

Gabaix, X. and D. Laibson, "Bounded Rationality and Directed Cognition," *Mimeo*, 2005.

———, ———, G. Moloche, and S. Weinberg, “Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model,” *American Economic Review*, 2006, pp. 1043–1068.

Gennaioli, N. and A. Shleifer, “What Comes to Mind,” *mimeo*, 2009.

Gilbert, D.T., B.W. Pelham, and D.S. Krull, “On Cognitive Busyness: When Person Perceivers Meet Persons Perceived,” *Journal of Personality and Social Psychology*, 1988, *54* (5), 733–740.

Gilboa, I. and D. Schmeidler, “Case-Based Decision Theory,” *Quarterly Journal of Economics*, 1995, pp. 605–639.

Gilovich, T., R. Vallone, and A. Tversky, “The Hot Hand in Basketball: On the Misperception of Random Sequences,” *Cognitive Psychology*, 1985, *17* (3), 295–314.

Gittins, JC, “Bandit Processes and Dynamic Allocation Indices,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 1979, pp. 148–177.

Glaeser, E.L., “The Political Economy of Hatred,” *Quarterly Journal of Economics*, 2005, *120* (1), 45–86.

Gould, S.J., *The Mismeasure of Man*, Norton New York, 1996.

Grimmett, G. and D. Stirzaker, *Probability and Random Processes*, Oxford University Press, USA, 2001.

Hakes, J.K. and R.D. Sauer, “An Economic Evaluation of the Moneyball Hypothesis,” *The Journal of Economic Perspectives*, 2006, *20* (3), 173–186.

Hilton, J.L. and W. Von Hippel, “Stereotypes,” *Annual Review of Psychology*, 1996, *47* (1), 237–271.

Hong, H., J.C. Stein, and J. Yu, “Simple Forecasts and Paradigm Shifts,” *The Journal of Finance*, 2007, *62* (3), 1207–1242.

James, B., *The Bill James Baseball Abstract 1982*, Ballentine, New York, 1982.

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

Kahneman, D., *Attention and Effort*, Prentice-Hall Englewood Cliffs, NJ, 1973.

Kass, R.E. and A.E. Raftery, “Bayes Factors.,” *Journal of the American Statistical Association*, 1995, *90* (430).

Lang, K., “A Language Theory of Discrimination,” *Quarterly Journal of Economics*, 1986, pp. 363–382.

Lewis, M., *Moneyball: The Art of Winning an Unfair Game*, WW Norton & Company, 2003.

Li, J., C. Dunning, and R. Malpass, “Cross-Racial Identification Among European-Americans: Basketball Fandom and the Contact Hypothesis. Working Paper,” 1998.

Lundberg, S.J. and R. Startz, “Private Discrimination and Social Intervention in Competitive Labor Market,” *American Economic Review*, 1983, pp. 340–347.

Mack, A. and I. Rock, “Inattentional Blindness: Perception Without Attention,” *Visual Attention*, 1998, *8*, 55–76.

Mackowiak, B. and M. Wiederholt, “Optimal Sticky Prices Under Rational Inattention,” *American Economic Review*, 2009, *99* (3), 769–803.

Malmendier, U. and D. Shanthikumar, “Are Small Investors Naive About Incentives?,” *Journal of Financial Economics*, 2007, *85* (2), 457–489.

Mankiw, N.G. and R. Reis, “Sticky Information Versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve,” *Quarterly Journal of Economics*, 2002, *117* (4), 1295–1328.

Messick, DM and DM Mackie, “Intergroup Relations,” *Annual Review of Psychology*, 1989, *40* (1), 45–81.

Miller, G.A., "The Magical Number Seven, Plus or Minus Two: Some Limits On Our Capacity for Information Processing," *Psychological Review*, 1956, *63* (2), 81–97.

Mullainathan, S., "Thinking Through Categories," *mimeo, MIT*, 2000.

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

———, J. Schwartzstein, and A. Shleifer, "Coarse Thinking and Persuasion," *Quarterly Journal of Economics*, 2008, *123* (2), 577–619.

Nisbett, R.E. and L. Ross, *Human Inference: Strategies and Shortcomings of Social Judgment*, Prentice Hall, 1980.

Oliver, J.E. and T. Mendelberg, "Reconsidering the Environmental Determinants of White Racial Attitudes," *American Journal of Political Science*, 2000, pp. 574–589.

Park, B. and M. Rothbart, "Perception of Out-Group Homogeneity and Levels of Social Categorization: Memory for the Subordinate Attributes of in-Group and Out-Group Members," *Journal of Personality and Social Psychology*, 1982, *42* (6), 1051–1068.

———, CS Ryan, and CM Judd, "Role of Meaningful Subgroups in Explaining Differences in Perceived Variability for in-Groups and Out-Groups," *Journal of Personality and Social Psychology*, 1992, *63* (4), 553–567.

Peng, L. and W. Xiong, "Limited Attention and Asset Prices," *Journal of Financial Economics*, 2006, *80* (3), 563–602.

Peski, M., "Prior Symmetry, Categorization and Similarity-Based Reasoning," *mimeo*, 2007.

Pettigrew, T.F., "Intergroup Contact Theory," *Annual Review of Psychology*, 1998, *49* (1), 65–85.

——— and L.R. Tropp, "A Meta-Analytic Test of Intergroup Contact Theory," *Journal of Personality and Social Psychology*, 2006, *90* (5), 751.

Phelps, E.S., "The Statistical Theory of Racism and Sexism," *American Economic Review*, 1972, pp. 659–661.

Piccione, M. and A. Rubinstein, "On the Interpretation of Decision Problems with Imperfect Recall," *Games and Economic Behavior*, 1997, *20* (1), 3–24.

Quattrone, G.A. and E.E. Jones, "The Perception of Variability Within in-Groups and Out-Groups: Implications for the Law of Small Numbers," *Journal of Personality and Social Psychology*, 1980, *38* (1), 141–52.

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

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

Reis, R., "Inattentive Producers," *Review of Economic Studies*, 2006, *73* (3), 793–821.

Rubinstein, A., *Modeling Bounded Rationality*, MIT Press, 1998.

Samuels, M.L., "Simpson's Paradox and Related Phenomena," *Journal of the American Statistical Association*, 1993, pp. 81–88.

Schacter, D.L., *The Seven Sins of Memory: How the Mind Forgets and Remembers*, Houghton Mifflin Company, 2001.

Schaller, M. and M. O'Brien, "'Intuitive Analysis of Covariance' and Group Stereotype Formation," *Personality and Social Psychology Bulletin*, 1992, *18* (6), 776.

———, C.H. Asp, M.C. Roseil, and S.J. Heim, "Training in Statistical Reasoning Inhibits the Formation of Erroneous Group Stereotypes," *Personality and Social Psychology Bulletin*, 1996, *22* (8), 829.

Shapiro, J., "A 'Memory-Jamming' Theory of Advertising," *University of Chicago mimeograph*, 2006.

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

———, "Rational Inattention: Beyond the Linear-Quadratic Case," *American Economic Review*, 2006, pp. 158–163.

Taylor, S.E. and S.T. Fiske, "Point of View and Perceptions of Causality," *Journal of Personality and Social Psychology*, 1975, *32* (3), 439–445.

Thaler, R.H. and C.R. Sunstein, "Market Efficiency and Rationality: The Peculiar Case of Baseball," *Michigan Law Review*, 2004, *102* (6), 1390–1403.

von Hippel, W., J. Jonides, J.L. Hilton, and S. Narayan, "Inhibitory Effect of Schematic Processing On Perceptual Encoding," *Journal of Personality and Social Psychology*, 1993, *64*, 921–921.

Walker, S.G., "Modern Bayesian Asymptotics," *Statistical Science*, 2004, pp. 111–117.

Wardrop, R.L., "Simpson's Paradox and the Hot Hand in Basketball," *American Statistician*, 1995, pp. 24–28.

Weber, R., C. Camerer, Y. Rottenstreich, and M. Knez, "The Illusion of Leadership: Misattribution of Cause in Coordination Games," *Organization Science*, 2001, pp. 582–598.

Wilson, A., "Bounded Memory and Biases in Information Processing," *NAJ Economics*, 2002, *5* (3).