ABSTRACT. What do we notice and how does this affect what we learn and come to believe? I present a model of an agent learning to forecast outcomes over time in which, because of limited cognitive resources, he engages in selective attention: he narrows his attention to event features currently believed to be informative relative to a prediction task. Because selective attention implies that certain event details may not be recorded in memory, the agent might never learn to pay attention to an important causal factor. As a result, he may persistently fail to recognize important empirical regularities, make biased forecasts, and hold incorrect beliefs about the causal relationship between variables. Importantly, the environment tightly restricts what forecasts and beliefs can persist. The model sheds light on the emergence and persistence of a set of systematic biases in inference, as well as on the formation and stability of erroneous group stereotypes.
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, a student learning what study techniques work best for her. Learning in this manner often relies on what we remember: characteristics of past workers, details of interactions with given small business owners, study techniques used on 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: the fact that not all information will be attended to or encoded in the first place is crucially important.\(^2\) It is hard or impossible to take note of all the characteristics of a given 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. 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.

I draw out the consequences of selective attention in a learning model. The first basic insight of the model is that when someone engages in selective attention, he may fail to learn to pay attention to an important causal factor and, as a result, will not learn how it is related to the output of interest.\(^3\)

---

\(^1\) 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 1995). 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. Beginning with Sims (2003), some economists have explored the consequences “rational inattention”, where agents optimize given a limited channel for absorbing information. I will return to those models later in this section.

\(^3\) 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 prior, while an agent who engages
The second 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, and the extent of misreaction, depends completely on observable features of the environment: these biases are *systematic*. Likewise, the environment tightly restricts his beliefs regarding the causal relationship between variables. Specifically, under the realistic assumption that people are not aware of how selective attention can result in biased inference, someone will attribute cause to a factor whenever he reacts to it. In other words, he will come to believe that a factor is causally related to an output even when it only proxies for selectively unattended to predictors.

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, over-confidence in the validity of salient cues, the overattribution of cause to salient event features, and the tendency to infer cause from correlation. By endogeneizing these and other biases as a consequence of selective attention, the model illuminates conditions under which we should expect them to emerge and persist. In addition, the model is applied to help understand the formation and stability of erroneous group stereotypes, and explores implications for discrimination.

To illustrate these ideas and the basic structure of the model, consider a simple example. 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, \( x \in \{\text{Not Prof, Prof}\} \), and whether the conversation will take place at a work or recreational setting, \( z \in \{\text{Work, Play}\} \). Independent of occupation, every individual is always friendly during recreation but never at work:

\[
E[y|\text{Prof, Play}] = E[y|\text{Not Prof, Play}] = 1 \\
E[y|\text{Prof, Work}] = E[y|\text{Not Prof, Work}] = 0.
\]

The likelihood of encountering different \((x, z)\), \(g(x, z)\), is given by Figure 1. The important feature of this joint distribution is that the student’s interactions with professors are relatively confined to work situations: \(g(\text{Work}|\text{Prof}) > g(\text{Work}|\text{Not Prof})\).

In selective attention uses prior knowledge to guide encoding. I discuss the relationship between confirmatory bias and selective attention in more detail in Section 3.

2
The student does not know $E[y|x,z]$, but learns to make predictions over time using events as mentally represented in memory. I assume that the student always notices and later recalls someone’s occupation and whether he acted friendly, but only notices and is later able to remember situational factors when he currently has some reason to believe that such factors influence friendliness: $x$ (occupation) is automatically encoded, or salient, so selective attention operates only on $z$ (situational factors).

Suppose the agent starts off not having reason to suspect that situational factors influence friendliness. As a result of selective attention, he then fails to learn that situational factors do in fact predict friendliness and continues ignoring them. Since the student persistently does not attend to or encode situational factors, assume that his forecast given $(Occupation, Situation)$ approaches the empirical frequency of friendliness given just $(Occupation)$. In the long run, the student will almost surely misforecast

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

$$
\hat{E}[y|Prof,\ Situation] = \frac{.1}{.1 + .25} = .29
$$

across situations by the strong law of large numbers.

This example has the following features in the long-run. First, the student mistakenly predicts that a professor is less likely to be friendly than a non-professor in a given situation because his interactions with professors are relatively confined to work situations: $\hat{E}[y|Prof,\ Situation] < \hat{E}[y|Not\ Prof,\ Situation]$. Second, he overreacts to the more inherently “salient” or automatically encoded feature (occupation) and underreacts to the less salient feature (situation). Third, while his limiting forecasts are biased conditional on observables, they are not arbitrary but are consistent with his effective observations: $\hat{E}[y|Occupation,\ Situation] = E[y|Occupation]$. Fourth, if the student does not recognize that selective attention results in a missing data problem that can result
in omitted variable bias, he may become more and more certain in a mistaken belief that whether or not an individual is a professor is a ceteris paribus (i.e., causal) predictor of friendliness.

Section 2 sets up the formal learning model, enabling me to generalize the above points and make them precise. As in the example, the 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., a person’s occupation) 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 can recall all details of events stored in memory eventually learns the true model and makes asymptotically accurate forecasts (Observation 1).

Section 3 introduces selective attention. The first basic assumption is that outcome variable \( y \) and input variable \( x \) are automatically attended to and encoded; selective attention operates only on \( z \). The interpretation is that some event features, like someone’s race, gender, or age, require less top-down attentional effort in order to attend to an encode, or are particularly salient (Fiske 1993).\(^4\) I model selective attention by assuming that the likelihood that the agent attends to and encodes \( z \) is increasing in the current probability he attaches to those mental models which specify \( z \) as being ceteris paribus predictive of \( y \). 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. 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 naïve 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 inferences. Instead, he uses an update rule which would be fully rational if his mentally represented history was complete.

I first establish some basic properties of the learning process. The agent eventually settles on how he mentally represents outcomes: from some period onward, 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 causally predictive, or cannot devote as much attention to learning to predict \( y \), e.g., he is under more time pressure or is otherwise more preoccupied (Proposition 2).\(^5\)

\(^4\)Alternatively, certain dimensions are chronically accessible (Fiske and Taylor 2008).
\(^5\)My paper is related to the literature on bandit problems (e.g., Gittens 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
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 easily and automatically encoded variables. For example, if the student settles on attending to the situation then he makes asymptotically accurate forecasts. If he does not, then his forecasts are biased given a situation but are consistent with the probability distribution over outcomes as averaged across work and recreation.\(^6\)

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 his posterior 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 \((y, x, \emptyset)\), then his limiting belief about whether \(z\) is causally predictive is unrestricted since he does not notice variation in \(z\), but, as a consequence of the naivete assumption, his limiting belief about whether \(x\) is causally predictive is restricted by whether \(x\) predicts \(y\) unconditional of \(z\): if it does, then the agent becomes convinced that \(x\) causally predicts \(y\). For example, if the student persistently attends to the situation, then he learns that situational factors but not occupation are causally predictive of friendliness. On the other hand, if he settles on failing to attend the situation, then the strength of his limiting belief about whether situational factors causally predict friendliness is influenced by his prior, but, over time, he will (wrongly) become more and more convinced that whether or not someone is a professor is a ceteris paribus (i.e., causal) predictor of friendliness since he does not appreciate that, having not attended to the situation, he cannot properly make such a causal inference.

---

\(^6\)When the agent settles on not encoding \(z\), say 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 taking coarse thinking or specific categorizations 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 what categorizations can persist and for when we should expect persistence.
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, I study how selective attention can result in misattribution of cause, as in the example where the student becomes convinced that whether someone is a professor influences whether he is friendly and does not just proxy for other important predictors (Proposition 6).

Through most of the paper, I make the stark but instructive assumption that the agent never attends to and encodes $z$ when he places little weight on mental models which specify $z$ as being important to prediction. More realistically, the agent will sometimes attend to $z$ in these circumstances, for example if he randomly has few other demands on his attention. In this case, even starting from a strong prior belief in mental models which exclude $z$ as a causal factor, the agent may receive enough disconfirming evidence that he will begin attending to $z$ with high probability within a reasonable time horizon: selective attention may slow down but not eliminate learning. To understand what features of the joint distribution over $(y, x, z)$ influence whether this will be the case, in Section 5 I study learning when there are random, momentary, 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 momentary fluctuations are such that the likelihood that the agent attends to $z$ varies monotonically and continuously in the likelihood he attaches to such processing being decision-relevant. Under this alternative specification, the agent will learn to devote more and more attention to $z$ and to properly incorporate information about $z$ in making predictions (Proposition 7), but he may fail to attend to $z$ and make biased forecasts with high probability for a long time. The main result of this section (Proposition 8) concerns the rate at which the agent begins attending to $z$: The speed of convergence increases in the degree to which the agent finds it difficult to explain what he observes without taking $z$ into account. Importantly, 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 $z$.

In Section 6, I apply the model to study out-group discrimination. Like in standard rational statistical discrimination models (Phelps 1972, Arrow 1973), discrimination is based on stereotypes built from experience, rather than tastes (Becker 1971). However, unlike in rational statistical discrimination models, people may persistently exaggerate mean differences across groups or even perceive mean differences when none exist because selective attention can result in a failure to take
certain information (e.g., situational factors) into account, resulting in coarse stereotypes. There is 
content to the model because observable features of the environment tightly restrict what misper-
ceptions 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. The implications of the model are consis-
tent with experimental evidence on stereotype formation and the model makes testable predictions 
regarding employer discrimination.

Section 7 considers some basic extensions of the model. 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 a long 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).\footnote{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.}

In another basic extension, 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.

I view my model as complementing recent papers which emphasize the importance of other 
cognitive limitations, such as imperfect recall or bounded memory (Mullainathan 2002, Gennaioli 
and Shleifer 2009), on economic decisions. The style of the paper is different from models of 
“rational inattention” (e.g., Sims 2003) and certain boundedly rational models (e.g., Rubinstein 
1998, Wilson 2003) in that it does not attempt to model optimal cognition, but specifies a tractable 
alternative guided by evidence from psychology.\footnote{One interpretation is that I view the allocation of cognitive resources as being governed by more automatic or heuristic than fully rational processes.} Perhaps a more substantive distinction is that I do 
not assume that the agent is endowed with perfect knowledge of what’s worth expending cognitive
resources on, but that he must learn this over time through repeated observation.\footnote{This paper can be viewed as working in the style of the subset of the literature on learning in games (e.g., Fudenberg and Kreps 1995) which exogenously specifies intuitive behavior rules of players rather than deriving those rules through utility maximization and allows players to start off favoring incorrect hypotheses regarding opponents’ play. It also shares similarities to recent models of costly information transmission (Gabaix et al. 2006, Gabaix and Laibson 2005), which recognize cognitive limitations but do not assume that agents optimize given those limitations.} Indeed, one basic insight of the analysis is that without expending the cognitive resources to attend to something, the agent will not learn that it’s worth attending to: limited cognitive resources can result in the agent never learning how to efficiently allocate those resources.

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 \( x \in X \) and \( z \in Z \) are both finite random variables, which 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., \( p_{\theta_0}(y = 1|x, z) = \theta_0(x, z) \)). Additionally, he knows the joint distribution \( g(x, z) \), which is positive for all \((x, z)\).\footnote{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 \( \theta_0 \).}

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, but not always, 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 above example where occupation is predictive of friendliness not controlling for situational factors. Formally, defining \( p_{\theta_0}(y = 1|x) \equiv \sum_{z'} \theta_0(x, z')g(z'|x) \), I make the following assumption.

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

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

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):

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

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

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

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

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

so \( M_{\neg X,Z} \) is a \(|\hat{Z}|\) 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 \(|X| \times |\hat{Z}|\) parameter model; i.e., \( \mu^{X,Z}(\theta) \) places weight on

\(^{11}\)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.
<table>
<thead>
<tr>
<th>Models</th>
<th>Parameters</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_{\neg X, \neg Z}$</td>
<td>$\theta$</td>
<td>Neither $x$ nor $z$ predicts $y$</td>
</tr>
<tr>
<td>$M_{X, \neg Z}$</td>
<td>$(\theta(x'))_{x' \in X}$</td>
<td>Only $x$ predicts $y$</td>
</tr>
<tr>
<td>$M_{\neg X, Z}$</td>
<td>$(\theta(z'))_{z' \in \hat{Z}}$</td>
<td>Only $z$ predicts $y$</td>
</tr>
<tr>
<td>$M_{X, Z}$</td>
<td>$(\theta(x', z'))_{(x', z') \in X \times \hat{Z}}$</td>
<td>Both $x$ and $z$ predict $y$</td>
</tr>
</tbody>
</table>

**Figure 2:** Set of Mental Models

those vectors for which $\theta(x, z) \neq \theta(x, z')$ and $\theta(x', z'') \neq \theta(x'', z''')$ for some $x, x', x'', z, z', z''$. Figure 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, $\psi(\cdot)$.\footnote{I make a technical assumption on the density $\psi$ 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 $\psi$ 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

\[
\begin{align*}
\pi_{X,Z} &= \pi_X \pi_Z \\
\pi_{X,\neg Z} &= \pi_X (1 - \pi_Z) \\
\pi_{\neg X,Z} &= (1 - \pi_X) \pi_Z \\
\pi_{\neg X,\neg Z} &= (1 - \pi_X)(1 - \pi_Z)
\end{align*}
\]

for some $\pi_X, \pi_Z \in (0, 1]$. $\pi_X$ is interpreted as the subjective prior probability that $x$ is important to predicting $y$ (e.g., that occupation is important to predicting friendliness); $\pi_Z$ 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 up to period $t$ by

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

\footnote{I provide an alternative, more explicit, description of the agent’s prior in Appendix A.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 $\theta_0$ together with $g$. I denote this distribution by $P_{\theta_0}$.\(^{13}\)

The agent’s prior, together with $g$, generates a joint distribution over $\Theta, M,$ and $H$, where $\Theta$ is the set of all possible values of $\theta_0, M$ 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

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

\[ \overset{a.s.}{\to} \pi^t_{X,Z} \bar{y}_t(x, z) + \pi^t_{X,-Z} \bar{y}_t(x) + \pi^t_{-X,Z} \bar{y}_t(z) + \pi^t_{-X,-Z} \bar{y}_t \]

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 $z$.
- $\pi^t_{i,j} \equiv \Pr(M_{i,j}|h^t)$ 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),

$\overset{13}{P_{\theta_0}}$ is defined by setting

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

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

\[ \overset{14}{\Pr(h^t, \tilde{\Theta}, M) = \pi_M \int_{\Theta} \rho(h^t|\theta) \mu_M(d\theta)} \]

where

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

When $\psi(\theta) \sim U[0, 1]$, then, for any $t, h^t$, (2) is an accurate approximation of the agent’s period-$t$ forecast to order $1/N(x, z)$, where $N(x, z)$ equals the number of times $(x, z)$ has appeared along history $h^t$.\(^{15}\)
(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 \( y \), \( \pi_{tX,Z} \rightarrow 1 \)
2. Whenever \( x \) (unlike \( z \)) is unimportant to predicting \( y \), \( \pi_{t\neg X,Z} \rightarrow 1 \)

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

1. \( E[y|x,z,h^t] \rightarrow E_{\theta_0}[y|x,z] \) for all \((x,z)\), almost surely with respect to \( P_{\theta_0} \).
2. The agent learns the true model, almost surely with respect to \( P_{\theta_0} \).

**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.

### 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 research finding 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). Consequently, at a later date he 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 assume that both \( y \) and \( x \) are automatically attended to and encoded; only \( z \) is encoded with effort (i.e., selective attention operates only on \( z \)). To model selective attention, I assume that when the agent is cognitively busy, the likelihood that he attends to and encodes information along an effortful dimension \( z \) is increasing in the probability he attaches to such processing being decision-relevant. Formalizing these assumptions, the individual’s mental representation of the history is

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

where

\[
\hat{z}_k = \begin{cases} 
z_k & \text{if } e_k = 1 \text{ (the agent encodes } z_k) \\
\emptyset & \text{if } e_k = 0 \text{ (the agent does not encode } z_k) \end{cases}
\]

and

\[
e_k = \begin{cases} 
1 & \text{if } \hat{\pi}_k^k > b_k \\
0 & \text{if } \hat{\pi}_k^k \leq b_k 
\end{cases}
\]

\( e_k \in \{0, 1\} \) 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 \( \hat{\pi}_k^k \) 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 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

---

16To 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.
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 \( z \) more continuously varies 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 \( \xi : Z \times \hat{H} \rightarrow \Delta (Z \cup \{ \emptyset \}) \) 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 \( \hat{z}_k = \hat{z}' \in Z \cup \{ \emptyset \} \) given \( z \) and \( \hat{h}^k \). In other words, the encoding rule specifies how the agent encodes \( z \) given any history.

Assumption 4. The agent is naive in performing statistical inference: \( \hat{Z} = Z \cup \{ \emptyset \} \) and, for every \( \tilde{\Theta} \subset \Theta, M \in \mathcal{M}, \hat{h}^t \in \hat{H} \), he applies “likelihood function”

\[
Pr(\hat{h}^t | \tilde{\Theta}, M) \propto \frac{\int_{\Theta} \prod_{\tau=1}^{t-1} p_\theta(y_{\tau}|x_{\tau}, \hat{z}_\tau) \mu^M(d\theta)}{\int_{\Theta} \mu^M(d\theta)},
\]

where \( p_\theta(y = 1|x, \hat{z}) = \theta(x, \hat{z}) \) for all \((x, \hat{z}) \in X \times \hat{Z}\).

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

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

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

at each event \( E(\hat{h}^t) = \{ \hat{h}^{t\infty} : \hat{h}^{t\infty} = \hat{h}^t \} \). All remaining statements regarding almost sure convergence are with respect to this measure.
over $[0, 1]^{|X| \times |Z|}$ since he takes advantage of the structural relationship relating the success probability following missing versus non-missing values of $z$. For every $\tilde{\Theta} \subset \Theta, M \in \mathcal{M}$, and $\hat{h}^t \in \hat{H}$, such an agent applies “likelihood function”

\[
\Pr^S(\hat{h}^t | \tilde{\Theta}, M) \propto \frac{\int \prod_{\tau \in \mathcal{E}(t)} p_\theta(y_{\tau} | x_{\tau}, z_{\tau}) \prod_{\tau \notin \mathcal{E}(t)} p_\theta(y_{\tau} | x_{\tau}) \mu^M(d\theta)}{\int_{\Theta} \mu^M(d\theta)},
\]

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

Comparing the numerator in (6S) with the numerator in (6), we see that, 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.

I maintain the naivete assumption in what follows because it seems to be more realistic than the assumption that people are sophisticated.\(^{18}\) 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 $\emptyset$ as a non-missing value of $z$ in 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 $\Theta, M$, and $\hat{H}$ as generated by his prior together with $g$ and $\xi$.\(^{19}\) 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 there are many

\[^{18}\]See Mullainathan (2002) for evidence that people do not seem to correct for memory limitations when making inferences.

\[^{19}\]For any $\tilde{\Theta} \subset \Theta, M \in \mathcal{M}, \hat{h}^t \in \hat{H}$

\[
\Pr_\xi(\hat{h}^t, \tilde{\Theta}, M) = \pi_M \int_{\Theta} \rho_\xi(\hat{h}^t | \theta) \mu^M(d\theta)
\]

where

\[
\rho_\xi(\hat{h}^t | \theta) = \prod_{\tau=1}^{t-1} \theta(x_{\tau}, \hat{z}_\tau)^{y_{\tau}} (1 - \theta(x_{\tau}, \hat{z}_\tau))^{1-y_{\tau}} g_\xi(x_{\tau}, \hat{z}_\tau | \hat{h}^\tau)
\]

\[
g_\xi(x, \hat{z} | \hat{h}^t) = \sum_{z'} g(x, z') \xi(z', \hat{h}^t | \hat{z})
\]
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.\textsuperscript{20}

\textit{Discussion of assumptions.} It is worth discussing the assumptions underlying (3)-(5) in a bit more
detail. First, the assumption that information along certain stimulus dimensions require no effort to
attend to and encode is meant to capture 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).\textsuperscript{21} Additionally, the amount of effort
required to process and encode information along a stimulus dimension decreases with practice
(Bargh and Thein 1985). As a result, 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 and vice-
versa. For example, it may be useful to attend to an individual’s race during certain social, but not
economic, interactions (Fryer and Jackson 2008).\textsuperscript{22}

Second, note that, since $\emptyset \notin Z$, 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 situ-
utional 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).

\textsuperscript{20}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

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

for each $t$.

\textsuperscript{21}Researchers have identified “preconscious” 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).

\textsuperscript{22}An alternative interpretation for why $x$ is automatically encoded but $z$ is not is that, whereas $x$ is costless to observe,
there is a small cost associated with determining $z$. For example, it is easier to observe the price of a product than the
sales tax on that product (Chetty, Looney, and Kroft 2009).
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 ignores his assessment of by how much. I conjecture that my qualitative results for the discrete attention case would continue to hold if I relax this assumption. Intuitively, the only real change would be that the agent could not persistently encode \( z \) if it 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 the selectively attentive agent assigns to model \( M_{i,j} \) in period \( t \) is given by

\[
\hat{\pi}_{i,j}^t = \Pr_{\xi}(M_{i,j}|\hat{h}^t).
\]

As a result, the probability he assigns to \( z \) being important to predicting \( y \) in period \( t \) is

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

and the probability he assigns to \( x \) being important to predicting \( y \) in period \( t \) is

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

His period-\( t \) forecast of \( y \) given \( x \) and \( z \) is\(^{23}\)

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

which converges to

\[
\hat{\pi}_{X,Z}^t \bar{y}_t(x, \hat{z}) + \hat{\pi}_{X,-Z}^t \bar{y}_t(x) + \hat{\pi}_{-X,Z}^t \bar{y}_t(\hat{z}) + \hat{\pi}_{-X,-Z}^t \bar{y}_t
\]

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

\(^{23}\)I discuss the agent’s period-\( t \) forecast in greater detail in Appendix A.

\(^{24}\)When \( \psi(\theta) \sim U[0, 1] \), then, for all \( t, \hat{h}^t \), (8) is an accurate approximation of the agent’s period-\( t \) forecast to order \( 1 / N(x, \hat{z}) \), where \( N(x, \hat{z}) \) equals the number of times \((x, \hat{z})\) has appeared along history \( \hat{h}^t \).
frequency of \( y = 1 \) given \((\hat{z})\) (e.g., the empirical frequency of the individual being friendly only 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 will 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 that \( b_k \equiv b \) 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”

\[
\Lambda(h^t) = \frac{\Pr(\hat{h}^t|z \text{ important})}{\Pr(\hat{h}^t|z \text{ unimportant})} = \frac{\Pr(\hat{h}^t|M_{X,Z})\pi_X + \Pr(\hat{h}^t|M_{-X,Z})(1 - \pi_X)}{\Pr(\hat{h}^t|M_{X,-Z})\pi_X + \Pr(\hat{h}^t|M_{-X,-Z})(1 - \pi_X)}
\]

would have to get smaller and smaller to bring \( \hat{\pi}^t_Z \) below \( b \). But the probability that \( \Lambda(h^t) \) never drops below some cutoff \( \lambda \) tends towards one as \( \lambda \) 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 \( \emptyset \) as he would a specific 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:
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.

**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 \((x, z)\), \( \hat{E}[y|x, z, \tilde{h}^t] \) converges to \( E_{\theta_0}[y|x, z] \) almost surely.

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, 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 described as possessing two “theories” useful in its ecology: (1) Distinctive gustatory cues, when followed by delayed gastric distress, should be considered suspect. (2) Distinctive spatial cues, when followed by immediate somatic pain, should be considered suspect. (Nisbett and Ross 1980, page 105)
If the agent settles on not encoding \( z \), then, for each \((x, z)\), \( \hat{E}[y|x, z, \hat{h}^t] \) converges to \( E_{\theta_0}[y|x] \) almost surely.

The intuition behind Proposition 3 is the following. If the agent settles on encoding \( z \) then, from some period onward, he finely represents each period’s outcome as \((y, x, z)\). On the other hand, if the agent settles on not encoding \( z \) then, from some period onward, he coarsely represents each period’s outcome as \((y, x, \emptyset)\) (this is coarser because \( \emptyset \) 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 easily and automatically encoded variables. 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 maintenance, 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 what 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 what 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.\(^{26}\)

\(^{26}\)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;
3.4. **Long-run beliefs given a mental representation.** In this subsection, I characterize the potential long-run beliefs of the selective attention agent.

**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 \( \hat{\pi}_X \overset{a.s.}{\to} 1 \) and, for large \( t \), \( \hat{\pi}_Z \leq b \).

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 he almost surely eventually places negligible weight on models where \( x \) is unimportant to predicting \( y \) because the unconditional success probability depends on \( x \) (recall Assumption 2), but the limiting behavior of \( \hat{\pi}_Z \) is largely unrestricted because, in the limit, he does not effectively 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 \( x \) predicts \( y \). This is because, by treating \( \emptyset \) 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 predicting \( y \). Put differently, the agent can identify \( \theta_0(x, \emptyset) - \theta_0(x', \emptyset) \) 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 \( z' \neq \emptyset \): The agent acts as if he believes that correlation implies causation.\(^{28}\)

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 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.

\(^{27}\)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 \( P_{\theta_0, \xi_0} \). 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.

\(^{28}\)The belief that correlation implies causation has been ranked as “probably among the two or three most serious and common errors of human reasoning” (Gould 1996, page 272).
4.1. Misreaction. In the long run, how will the selectively attentive agent misreact to $x$ when he fails to learn to pay attention to $z$? To study this question, it is useful to specialize to the case where $x$ is a binary random variable and $X = \{0, 1\}$. Define

$$R_x(z') = E_{\theta_0}[y|x = 1, z'] - E_{\theta_0}[y|x = 0, z']$$

$$R_x = E_z[R_x(z)|x = 1]$$

$$\phi = \text{Cov}_z(E_{\theta_0}[y|x = 0, z], g(x = 1|z)),$$

where

- $R_x(z')$ 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 $(x, z) = (0, z')$.

- $R_x$ 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 $x = 1$.\(^{29}\)

- $\phi$ 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 $X = \{0, 1\}$. Then

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

almost surely for all $z'$.

Proposition 5 says that when the agent settles on not encoding $z$, his limiting reaction to $x$ conditional on $z = z'$, $\hat{R}_x(z')$, differs from the standard Bayesian’s, $R_x(z')$, in two key ways

\(^{29}\) $R_x$ 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.
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, \( R_x \): By persistently failing to encode \( z \), the agent’s limiting 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 \( \phi \), which can be thought of as the magnitude and direction of omitted variable bias. A non-zero \( \phi \) 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 \( \hat{R}_x(z') \) and \( R_x(z') \) have the same sign. Then the agent overreacts to \( x \) at \( z' \) if \( |\hat{R}_x(z')| > |R_x(z')| \) and underreacts to \( x \) at \( z' \) if \( |\hat{R}_x(z')| < |R_x(z')| \). He overreacts to \( x \) if he overreacts to \( x \) at all \( z' \in Z \) and underreacts to \( x \) if he underreacts to \( x \) at all \( z' \in Z \).

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 \( R_x(z') = R_x \) for all \( z' \in Z \).\(^{30}\)

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

\[
R_x > R_x(z')
\]

\(^{30}\)Note that my definition of over- or underreaction only applies when \( \hat{R}_x(z') \) and \( R_x(z') \) 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

$$R_x < R_x(z').$$

(12)

To interpret conditions (11) and (12), suppose that $R_x(z) \geq 0$ for all $z \in Z$, so we can view $R_x(z)$ 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 by the literature on coarse thinking (e.g., Mullainathan 2002, 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:** $R_x(z') \equiv R_x$. Now consider the case where $R_x(z') = R_x$ for all $z' \in Z$. From Equation (10), the agent’s limiting reaction to $x$ in this case equals $\hat{R}_x(z') = R_x(z') + \phi/\text{Var}(x)$. To apply the definition of over- or underreaction, suppose that $\hat{R}_x(z')$ and $R_x$ 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 might 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. **Faulty belief in causation.** 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. Proposition 4 established that, in the limit, the agent will attribute cause to a factor whenever he reacts to it: he acts as if he believes correlation implies causation. Proposition 6 emphasizes a key 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.

**Proposition 6.** Suppose the conditions of Proposition 5 hold and, additionally, \( x \) is unimportant to predicting \( y \). Then, so long as \( \phi \neq 0 \),

1. \( |\hat{R}_x(z')| = \frac{|\phi|}{\text{Var}(x)} \neq 0 \) almost surely for all \( z' \): The agent overreacts to \( x \) and the extent of overreaction is increasing in \( \frac{|\phi|}{\text{Var}(x)} \).
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 \( y \), \( R_x(z') = 0 \) for all \( z' \) so \( R_x = 0 \). Then, by Proposition 5,

\[
\hat{R}_x(z') = E_{\theta_0}[y|x = 1] - E_{\theta_0}[y|x = 0] \\
= \frac{\phi}{\text{Var}(x)},
\]

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

Proposition 6 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 \( X = \{0, 1\} \) 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 6; it is included for clarity.
The second part of Proposition 6 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 \( g(Occupation, Situation) \) 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 \( (\mathcal{R}_{Occup}(Work) = \mathcal{R}_{Occup}(Play) = 0) \), and, by Proposition 4, will learn to place full weight on mental models which do not include whether an individual is a professor among causal factors influencing friendliness. On the other hand, if the agent settles on not attending to situational factors then Proposition 6 says that the student’s limiting reaction to whether the individual is a professor equals

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

\(^{32}\)Participants also retained more information about the conversationalist they attended to.
A simple calculation gives us that $\phi = (p^H - p^L)(g(\text{Prof})g(\text{Work}) - g(\text{Prof, Work}))$, so the agent’s limiting reaction is

$$R_{\text{Occup}}(\text{Situation}) = \frac{(p^H - p^L)(g(\text{Work}) - g(\text{Work|Prof}))}{(1 - g(\text{Prof}))}.$$

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 $g(\text{Work|Prof}) \neq g(\text{Work|Not Prof})$ and, in particular, will predict professors to be less friendly than others when $g(\text{Work|Prof}) > g(\text{Work|Not Prof})$.

In addition, whenever $g(\text{Work|Prof}) \neq g(\text{Work|Not Prof})$, Proposition 6 tells us that the agent 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 and encodes $z$ 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, momentary, 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 predicting $y$.

Formally, let $\eta(\hat{\pi}_z^k) \equiv \operatorname{Prob}[c_k = 1|\hat{\pi}_z^k] = \operatorname{Prob}[b_k < \hat{\pi}_z^k]$ 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 $b_k \overset{i.i.d.}{\sim} \mathcal{U}[0, 1]$. The likelihood that the agent attends to $z$ as a function of $\hat{\pi}_z^k$ is then

\(^{33}\)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.
given by:

\[ \eta(\hat{\pi}_Z^K) = \hat{\pi}_Z^K \]  

for all \( 0 \leq \hat{\pi}_Z^K \leq 1 \).

**Proposition 7.** Suppose that \( b_k \overset{i.i.d.}{\sim} U[0, 1] \). Then

1. \( \eta(\hat{\pi}_Z^K) \rightarrow 1 \) almost surely.
2. For each \( x, z \), \( \hat{E}[y|x, z, \hat{h}_t] \) converges to \( E_{\theta_0}[y|x, z] \) in probability.

The intuition for Proposition 7 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 the agent’s 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 arbitrarily 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 \( t \), \( \hat{E}[y|x, z, \hat{h}_t] \approx E_{\theta_0}[y|x] \) 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 any 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

\[
\frac{|\mathcal{X}_t - \mathcal{X}_0|}{V(t)} \xrightarrow{a.s.} C
\]

**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 \( c_1 V(t) \leq |X_t - X_0| \leq c_2 V(t) \) 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 might continue not attending to \( z \) for a long time if he can accurately approximate the true distribution when he only takes salient feature \( x \) into account.

Formally, let \( p_{\theta_0}(y|x, z) \) denote the conditional distribution of \( y \) given both \( x \) and \( z \) under the true model and \( p_{\theta_0}(y|x) \) denote the conditional distribution of \( y \) given only \( x \) under the true model. Define the relative entropy distance, \( d \), between these two distributions as the average of the relative entropies between \( p_{\theta_0}(y|x', z') \) and \( p_{\theta_0}(y|x') \), where this average is taken over the probability mass function \( g(x, z) \):

\[
d = \sum_{y,x,z} p_{\theta_0}(y|x, z) g(x, z) \log \left( \frac{p_{\theta_0}(y|x, z)}{p_{\theta_0}(y|x)} \right).
\]

(17)

\( d \) essentially measures the distance between \( p_{\theta_0}(y|x, z) \) and \( p_{\theta_0}(y|x) \), 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 8.** Suppose \( b_k \overset{i.i.d.}{\sim} U[0, 1] \) and either (i) \( \pi_X = 1 \) or (ii) \( x \) is important to predicting \( y \). Then \( \eta(\hat{\pi}_Z) \to 1 \) almost surely with an asymptotic rate of convergence \( e^{-d(t-1)} \).

For a brief sketch of the arguments involved in proving Proposition 8, note that the rate at which \( \eta(\hat{\pi}_Z) = \hat{\pi}_Z \to 1 \) is determined by the rate at which

\[
\frac{\Pr_{\xi}(\hat{h}^t|M_{X,-Z})}{\Pr_{\xi}(\hat{h}^t|M_{X,Z})} \to 0.
\]

(18)

Consider the simpler problem of determining the rate at which

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

(19)

\(^{34}\)“Distance” \( d \) is called the conditional relative entropy in Cover and Thomas (2006).

\(^{35}\)Simpler problems along these lines have been studied by other economists in the past (e.g., Easley and O’Hara 1992).
By the strong law of large numbers, $1/(t-1)$ times the log of (19) goes to $-d$. The proof applies similar logic to analyzing (18), 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 then easy to calculate that, in this case,

$$d = -\sum_x \sum_z g(x, z) \log(g(z|x))$$

$$= H(z|x),$$

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 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 agent 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 beliefs 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.

6. Discrimination

6.1. Overview. In this Section I apply the model to study out-group discrimination and relate the predictions of my model to those of existing models. Economic theories of discrimination typically fall into one of two major categories: taste-based and statistical discrimination models.\(^{36}\) Taste-based models (Becker 1971) emphasize preferences, and model discrimination as resulting from members of one group receiving disutility from interacting with members of another. Statistical models (Phelps 1972, Arrow 1973), on the other hand, emphasize uncertainty, and model discrimination as resulting from economic actors (typically employers) having imperfect information about the skills or behavior of others and optimally using all available information to make predictions.\(^{37}\)

In common with standard, “rational”, statistical discrimination models, discrimination in my model results from people having imperfect information about others. In contrast with such models, discrimination results (at least in part) from people failing to learn to optimally use all available information to make forecasts.\(^{38}\) The major distinguishing implications are that my model makes predictions about when and why people can persistently (i) misperceive mean differences across

\(^{36}\)Altonji and Blank (1999) provide a review.

\(^{37}\)One subset of such models (Phelps 1972, Arrow 1973) assume that group membership proxies for unobservable characteristics (e.g., skills) correlated with some outcome of interest (e.g., productivity), so people optimally use group membership information to make forecasts. Another subset of models (Aigner and Cain 1977, Lundberg and Startz 1983, Cornell and Welch 1996) assume that employers less precisely estimate the productivity of members of certain groups, perhaps because of cultural differences, which can lead to mean differences in earnings by group through a variety of channels (Altonji and Blank 1999).

\(^{38}\)An additional distinction, as emphasized in Proposition 6, is that the agent may become certain in a mistaken belief that race is causally related to outcomes of interest, while in a “rational” statistical discrimination world the agent should understand that race only proxies for important predictors.
groups, after controlling for freely available, but perhaps unattended to, information and (ii) attend to less information when making predictions about members of one group than when making predictions about members of another. Subsection 6.2 explores implication (i) and Subsection 6.3 explores implication (ii).

6.2. **Misperception of mean differences.** Let $x \in \{0, 1\}$ represent group identity (e.g., black/white, female/male), $z$ some potentially unattended to variable (e.g., situational factors, soft information), and $y$ some outcome of interest to be forecasted (e.g., friendliness, productivity), where $y = 1$ is interpreted as the more favorable outcome (e.g., more friendly, more productive). A key implication of earlier results is that limited and selective attention can lead to misperceptions regarding the degree to which group membership predicts $y$. To see this clearly, suppose $b_k \equiv b$, and the standard Bayesian’s limiting reaction to group membership is the same across $z'$ (generalizing the case where $x$ is unimportant to predicting $y$): $R_x(z') = R_x$ for all $z'$. Further, let $x = 1$ represent the in-group, and, for concreteness, suppose $R_x \geq 0$. Then, if the agent settles on not encoding $z$, his limiting reaction to group membership equals

$$R_x(z') = R_x(z') + \frac{\phi}{\text{Var}(x)}$$

by Proposition 5.

Equation (20) says that, when the selective attention agent settles on not encoding $z$, his limiting reaction to group membership, $\hat{R}_x(z')$, differs from that in a rational statistical discrimination world, $R_x(z')$, by the “omitted variable bias” term, $\frac{\phi}{\text{Var}(x)}$. When $\phi > 0$, such an agent misperceives the mean difference between groups as being larger than it is. Replacing “professors” with “out-group” and “non-professors” with “in-group” in the earlier illustrative example, the agent mistakenly perceives in-group members as being friendlier than out-group members as a result of persistently failing to attend to the situation. To take another example, there is evidence that the view that women are more communal and less assertive than men stems from the sex difference in the distribution of women and men into homemaker and employee roles (e.g., Eagly and Steffen 1984). To simplify the exposition, suppose that $\phi > 0$ for the rest of this Subsection.

Another implication of Equation (20) is that, unlike someone with invariant group preferences or a standard Bayesian, the selective attention agent will react less to group membership if $\phi$ is lower; i.e., if he is more likely to encounter members of an out-group in favorable contexts (e.g., in situations that encourage friendliness) or to encounter members of that group who score highly on
some observable but unattended to dimension (e.g., those from areas with good school systems). This point may provide a reason why celebrity role models can be beneficial (Fryer and Jackson 2008).

Finally, combined with Proposition 2, Equation (20) says that certain subtle factors or interventions that would have no effect in a “rational statistical discrimination” or “taste-based” world influence the degree of discrimination here. In particular, consistent with experimental evidence, when people have more time to process incoming information ($b$ is lower) or understand that it is useful to consider the power of the situation in predicting behavior and to categorize others on the basis of multiple dimensions ($\pi_Z$ is higher), it is more likely that the selective attention agent will learn to attend to $z$ (individuating characteristics, situational factors), reducing his limiting reaction to group membership by $\frac{\phi}{\text{var}(x)}$ (Schaller and O’Brien 1992, Bigler and Liben 1992).

6.3. **Coarse stereotyping of out-group members.** Discrimination may result not just from a perception of mean differences across groups, but from people more coarsely stereotyping members of an out-group and believing “they” are all alike (Fiske and Taylor 2008, page 261; Hilton and von Hippel 1996; Fryer and Jackson 2008).\(^{39}\) Indeed, many models of labor market discrimination are built off of related ideas that employers have more difficulty discriminating between members of an out-group as compared to members of an in-group (Aigner and Cain 1977, Cornell and Welch 1996), or more difficulty interpreting out-group members’ language or gestures (Lang 1986). The continuous attention version of my model provides a potential microfoundation for such assumptions, with implications for when they should be more or less relevant.

To allow for the possibility that the agent will more coarsely stereotype members of one group over some period of time, I assume that the agent separately learns to predict some outcome of interest for members of the in-group and out-group, $y^{in}$ and $y^{out}$. Otherwise, the model is the same as before: Each period the agent encounters either an in-group member, with probability $g(in)$, or an out-group member, with probability $1 - g(in)$. In a period where the individual encounters an individual from group $j \in \{in, out\}$, he observes $z \in Z$, drawn from $g(\cdot | j)$ (formally, $X$ is a singleton), makes prediction $\hat{y}^j_t$, and finally learns the true value $y^j_t$. As in the earlier examples, simplify by supposing that group membership is unimportant to predicting $y$. The process by

\(^{39}\)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).
which the agent separately learns to predict each $y^j$, $j = in, out$, is as specified in Section 5 on continuous attention.

For $j = in, out$, let

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

(21)

measure the difficulty the agent 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 agent attends to $z$ in period $k$ if, in that period, he encounters an individual of group membership $j$. An obvious modification of Proposition 8 gives the following result.

**Result 1.** Suppose the agent separately learns to make predictions for in-group and out-group members, that group membership is unimportant to predicting $y$, and $b_k \sim U[0, 1]$. Then, the asymptotic rate of convergence of $\eta_k(j)$ to 1 is $e^{-d(j)g(j)(t-1)}$ for $j = in, out$.

From Result 1, the speed with which the agent learns to allocate attention to $z$ 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 $z$ 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 8: It is not just the amount but the quality of contact which governs how quickly an agent will learn to attend to individuating information for members of group $j$. If $z$ does not vary much across encounters with members of group $j$ (encounters are relatively homogeneous) then the agent will persistently ignore individuating or situational-constraint information even if such encounters are frequent.\(^{40}\) To take an example, it

\[^{40}\text{A bit more formally, expand } d(j):\]

$$d(j) = H_j(z) - H_j(z|y)$$

(22)

where

- $H_j(z) = -\sum_{z'} g(z'|j) \log(g(z'|j))$ is the entropy of $z$ for group $j$, or a measure of the degree to which $z$ varies across encounters with members of group $j$.
- $H_j(z|y) = -\sum_{y'} \sum_{z'} p(y'|z')g(z'|j) \log \left( \frac{p(y'|z')g(z'|j)}{\sum_{z''} p(y'|z'')g(z''|j)} \right)$ is the conditional entropy of $z$ given outcome $y$ for group $j$. In the earlier example, where $z \in \{0, 1\}$ and $p_{\theta_0}(y = 0|z = 0) = 1 = p_{\theta_0}(y = 1|z = 1)$, we can ignore this term since, in that case, $H_j(z|y) = 0$.

From (22), we see that, fixing $H_j(z|y)$, the difficulty of explaining observations without taking $z$ into account for group $j$, $d(j)$, is increasing in the degree to which $z$ varies across encounters with members of group $j$, $H_j(z)$. 

---

35
seems unlikely that a white prison guard of a predominantly black prison will, on average, hold very nuanced views towards blacks.

A key implication of Result 1 is that the agent will more quickly learn to attend to $z$ for in-group than out-group members if and only if

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

Inequality (23) suggests that two factors are responsible for the tendency of individuals to more coarsely categorize members of an out-group 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). The model suggests that individuals who encounter members of an out-group across more varied situations (e.g., both at work and in the neighborhood) or who encounter more heterogeneous members of the out-group (e.g., across occupations or levels of social status) are more likely to consider the individuating characteristics of a given group member and consequently hold more accurate beliefs towards people from that group.

---

41 Consideration of the amount of inter-group interaction relates to experimental findings (reviewed in Fryer and Jackson 2008) that, while members of one ethnic group tend to find it more difficult to recognize faces of a different ethnic group than their own, this tendency is attenuated for individuals with more inter-group contact (e.g., Li, Dunning, and Malpass 1998).

42 Taking this factor into account may help explain why studies that only consider out-group size among potentially important environmental determinants of intergroup perceptions do not find a robust relationship between greater interracial contact between whites and blacks at the local level (e.g., town, neighborhood, metropolitan area) and the amount of negative stereotyping on the part of whites (Branton and Jones 2005, Oliver and Mendelberg 2000). It also may help explain why men tend to encode experiences involving women more coarsely than experiences involving men and vice-versa, even though any difference in the amount of within sex versus between sex contact should be negligible (Messick and Mackie 1989). To take an example, Park and Rothbart (1982) asked experimental participants to read a newspaper-type 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 is no difference in recall of the sex of the in-group versus out-group protagonist, participants were likely to recall the occupation of the in-group versus out-group protagonist.
members of that group (Cook 1985). This may provide an efficiency rationale for certain social policies which increase exposure to a diverse population of out-group members.\footnote{Many policies have been motivated or defended by the argument that increased intergroup contact reduces prejudice and racial tensions, including \textit{Brown v. Board of Education} (1954), the Fair Housing Act (1968), and workplace and university affirmative action programs (Dixon and Rosenbaum 2004).}

6.4. \textbf{Empirical predictions and possible test.} To review, the model can be applied to make several new predictions about the nature of discrimination. Subsection 6.2 emphasizes that, under conditions where people cannot devote much attention to learning to forecast some output, they may persistently fail to attend to predictive variables (e.g., situational factors) in which case they can exaggerate mean differences across groups. Such exaggeration will be lower when people tend to encounter out-group members in more favorable contexts or have more time to process information. There is experimental evidence consistent with this view of when and how erroneous stereotypes can be formed (e.g., Schaller and O’Brien 1992, Schaller et al. 1996, Eagly and Steffen 1984).\footnote{To take an example, 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. They 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. Control participants judged group \(B\) members to be more intelligent than group \(A\) members (presumably failing to take into account the correlation between group membership and anagram length). However, manipulations designed to give participants more time to process each observation or to direct their attention towards situational constraints through explicit instructions facilitated more accurate judgments.}

Subsection 6.3 emphasizes that people will more quickly learn to attend to individuating characteristics when making predictions about members of an in-group when encounters with in-group members are relatively frequent and heterogeneous. Bertrand and Mullainathan’s (2004) study helps illustrate a possible test of this result. They find that, all else equal, resume 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. Result 1 implies that this differential response should systematically vary across resume screeners. In particular, it predicts that this differential response should be attenuated for screeners who have encountered a more diverse population of blacks with respect to variables predictive of quality.

7. \textbf{Basic Extensions}

In this Section, I extend the analysis to study debiasing and disagreement.
7.1. **Debiasing and dilution.** To study debiasing, suppose the conditions of Proposition 6 hold, and, in addition, the agent starts off not encoding $z$ ($\pi_Z < b$). 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 $y$ ($\pi_Z$ shifts to $\pi_Z' > \pi_Z$) or to the degree to which he is cognitively busy ($b$ shifts to $b' < b$)?

The first 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, $E_{\xi}[^\theta(x, z) | M_{X,-Z}, \hat{h}'_t] \approx E_{\theta_0}[y|x]$ and $\hat{\pi}'_X \approx 1$ by Propositions 3 and 6, so the agent’s reaction to $x$ given $z$ in that period equals

\begin{equation}
|E_{\xi}[\theta(1, z)|\hat{h}'] - E_{\xi}[\theta(0, z)|\hat{h}']| \approx \pi'_Z[\tau - \tau] + (1 - \pi'_Z) \frac{|\phi|}{\text{Var}(x)}
\end{equation}

\begin{equation}
= (1 - \pi'_Z) \frac{|\phi|}{\text{Var}(x)}
\end{equation}

where $\tau = E_{\psi}[\theta]$ equals the prior success probability under density $\psi$. From (25), the agent overreacts to $x$ when he attends to $z$ in period $t$ if and only if he overreacts to $x$ when he does not attend to $z$ in period $t$. 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 $(x)$.

The second thing to note is that, while the agent still overreacts to $x$ when he attends to $z$, we see from (25) that the magnitude of overreaction is immediately lower by a factor of $(1 - \pi'_Z)$: In the short-run, attending to $z$ will dilute the agent’s response to $x$.\(^{46}\) At first this may seem somewhat surprising since the weight attached to $z$ being important to predicting $y$ in period $t$ does

---

\(^{45}\)Can think of shocks to $\pi_Z$ 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 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).

\(^{46}\)This implication of the model is loosely related to experimental evidence that presenting an individual with many non-diagnostic pieces of information for prediction dilutes his response to diagnostic pieces of information (Nisbett and Ross 1980, pages 154 - 156). The model predicts a similar but distinct phenomenon: Unshrouding a previously unattended to but diagnostic piece of information will dilute an agent’s response to highly salient but non-diagnostic pieces of information.
not depend on whether or not the agent attends to \( z \) in that period. However, it follows from the fact that the agent’s best forecast of \( y \) given \((x_t, z_t) = (x', z')\) under the model in which both \( x \) and \( z \) are important to predicting \( y \) \((M_{X,Z})\) depends on whether or not he attends to \( z \) in period \( t \). When he does not, he considers what he has learned from past events in which the predictors where similarly mentally represented as \((x', \emptyset)\) and uses this information to make prediction \( E_\xi [\theta(x', \emptyset) | \hat{h}, M_{X,Z}] \approx \bar{y}(x', \emptyset | \hat{h}) \approx E_{\theta_0}[y|x'] \). When he does, he believes that he does not have access to any data relevant to predicting \( y \) if \( z \) is important to prediction since he never encoded \( z \) in the past (another consequence of the naivete assumption) and relies, instead, on his prior expectation, \( \tau \). Since this increased reliance on the prior is true across \( x \), the agent’s reaction to \( x \) is attenuated.

The point that, in the short run, the agent’s forecasts become more reliant on the prior (but not necessarily more precise) if he begins encoding \( z \) after having failed to encode \( z \) for some time is worth exploring in a bit more detail. When the agent begins attending to \( z \) he places probability \( \pi'_Z \) on the event that everything he has learned so far is unimportant to prediction since he never encoded \( z \) in the past (by the naivete assumption he does not understand that these observations would still be somewhat informative because of the structural relationship \( p_{\theta_0}(y = 1 | x, \emptyset) = \sum_{z' \in Z} \theta_0(x, z')g(z' | x) \)). To take an example, if the agent suddenly becomes convinced that situational factors are important to predicting friendliness (because he learned about the fundamental attribution error in Psychology 101 for example) then he believes that, with high probability, he can no longer accurately predict friendliness since he never previously attended to situational factors. In the short run, his forecasts will be closer to the prior: He acts as if he suffers from “information overload” (Eppler and Mengis 2004). However, if he continues encoding situational factors, then, in the long run, his forecasts will be unbiased by Proposition 3.

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 \( X = \{0, 1\} \) and \( x \) is positively related to \( y \) conditional on each \( z' \): \( R_x(z') \geq 0 \) 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 \( x = 1 \) 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: \( \phi < 0 \) and \( -\phi > R_x \text{Var}(x) \).

8. Conclusion

This paper has supplied a formal model of selective attention and learning which places structure on when and which details of experiences are likely to be attended to at the time of encoding, with implications for why and how judgments may end up being systematically biased. The central assumptions of the model are the following. First, certain event features, like whether or not someone is a professor, are more salient and automatically encoded than others, like situational factors. Second, mental models facilitate selective attention, whereby people are more likely to attend to and encode a feature if they place greater weight on mental models which identify that feature as being causally related or predictive of the output. Finally, people are naive and ignore memory imperfections that result from selective attention when drawing inferences.

The model sheds light on why and when people may persistently fail to attend to variables important to prediction, and consequently misreact and potentially mistakenly attribute cause to associated salient factors. It illuminates the conditions that make such a failure more or less persistent. And it is applied to help understand some important features of out-group discrimination, namely discrimination based on systematic misperceptions of mean differences across groups and coarse out-group stereotypes.
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 \( \mu(\theta) = \sum_{i \in \{X, \sim X\}} \sum_{j \in \{Z, \sim Z\}} \pi_{i,j} \mu^{i,j}(\theta) \). Fix a model \( M \in \mathcal{M} \) and define \( c^M(x, \hat{z}) \) as the set of covariates \( (x', \hat{z}') \in X \times \hat{Z} \) 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 \( M_{X, \sim Z} \) (where only \( x \) is important to predicting \( y \)), \( c^{X, \sim Z}(x, \hat{z}) = \{(x', \hat{z}') \in X \times \hat{Z} : x' = x \} \) 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 \( X \times \hat{Z} \), and define \( \Theta(M) = [0, 1]^{\#C^M} \) as the effective parameter space under model \( M \) with generic element \( \theta(M) \). \( \mu^M \) 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 \( \mu^M \) and distributed according to density, \( \psi(\cdot) \). To take an example, if \( \psi(\theta) = 1_{\theta \in [0, 1]} \), then \( \theta(c^M) \sim U[0, 1] \) for each \( M \in \mathcal{M} \), \( c^M \in C^M \).

A.2. Forecasts. I will now describe the forecasts of an individual with limited 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

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

where

\[
E_{\xi}[\theta(c^M)|\hat{h}^t, M] = \int \tilde{\theta} \psi(\tilde{\theta}|\hat{h}^t, c^M) d\tilde{\theta}
\]

\[
\psi(\tilde{\theta}|\hat{h}^t, c^M) = \frac{\tilde{\theta}^\kappa(c^M|\hat{h}^t) N(c^M|\hat{h}^t) - \kappa(c^M|\hat{h}^t) \psi(\tilde{\theta})}{\int \tau^\kappa(c^M|\hat{h}^t) N(c^M|\hat{h}^t) - \kappa(c^M|\hat{h}^t) \psi(\tau) d\tau}
\]

and \( N(c^M|\hat{h}^t) \) denotes the number of times the covariates have taken on some value \( (x, \hat{z}) \in c^M \) along history \( \hat{h}^t \); \( \kappa(c^M|\hat{h}^t) \) 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, (26) takes a particularly simple form when \(\psi(\theta) \sim U[0, 1]\):

\[
\hat{E}[y|x, z, \hat{h}^t] = \hat{\pi}_{t,x,z} \frac{\kappa(x, \hat{z}|\hat{h}^t) + 1}{N(x|\hat{h}^t) + 2} + \hat{\pi}_{t,x,-z} \frac{\kappa(\hat{z}|\hat{h}^t) + 1}{N(\hat{z}|\hat{h}^t) + 2} + \hat{\pi}_{t,-x,z} \frac{\kappa(x|\hat{h}^t) + 1}{N(x|\hat{h}^t) + 2} + \hat{\pi}_{t,-x,-z} \frac{\kappa(\hat{h}^t) + 1}{t + 1},
\]

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

For further reference,

\[
\hat{\pi}_{t,i,j} = \frac{\Pr{\xi}(M_{i,j}|\hat{h}^t)}{\sum_{i',j'} \Pr{\xi}(\hat{h}^t|M_{i',j'}) \pi_{i',j'}} = \frac{\alpha_{i,j} B_{i,j}^t}{\sum_{i',j'} \alpha_{i',j'} B_{i',j'}^t}
\]

where

\[
B_{i,j}^t = \frac{\Pr{\xi}(\hat{h}^t|M_{i,j})}{\Pr{\xi}(\hat{h}^t|M_{X,Z})} = \frac{\int \Pr{\xi}(\hat{h}^t|\theta) \mu^i_j(d\theta)}{\int \Pr{\xi}(\hat{h}^t|\theta) \mu^{X,Z}(d\theta)}
\]

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

\[
\alpha_{i,j} = \frac{\pi_{i,j}}{\pi_{X,Z}}
\]

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 \((y, x, \hat{z}) \in \{0, 1\} \times X \times \hat{Z}\). Define the Kullback Leibler distance between \(\hat{p}(y, x, \hat{z})\) and \(p_0(y, x, \hat{z})\)
as

\[(29) \quad d_K(\hat{p}, p_0) = \sum_{y, x, \hat{z}} p_0(y, x, \hat{z}) \log \left( \frac{p_0(y, x, \hat{z})}{\hat{p}(y, x, \hat{z})} \right) \]

with the convention that \(0 \log \left( \frac{0}{\hat{p}} \right) = 0\) for \(\hat{p} \geq 0\) and \(p_0 \log \left( \frac{p_0}{\hat{p}} \right) = \infty\) 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 \(\hat{p}(y, x, \hat{z}|\theta) = \theta(y, \hat{z})^y (1 - \theta(y, \hat{z}))^{1 - y} p_0(y, \hat{z})\) (sometimes abbreviated as \(\hat{p}_\theta(y, x, \hat{z})\)), where \(p_0(x, \hat{z}) = \sum_{y' \in \{0, 1\}} p_0(y', x, \hat{z})\). Define \(\hat{p}(y, x, \hat{z}|\theta(M)) = p_{\theta(M)}(y, x, \hat{z})\) in the obvious manner (\(\theta(M)\) is defined as in Subsection A.1) and let \(\theta(M) = \arg \min_{\theta(M) \in \Theta(M)} d_K(\hat{p}_\theta(M), p_0)\) denote a minimizer of the Kullback-Leibler distance between \(\hat{p}_\theta(M)(\cdot)\) and \(p_0(\cdot)\) among parameter values in the support of \(\mu^M(\cdot)\). Finally, define \(\delta_M = \delta_M(p_0) = d_K(\hat{p}_\theta(M), p_0)\).

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

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

\[(30) \quad -d_K(\hat{p}_\theta(M), p_0) = \sum_{y, x, \hat{z}} p_0(y|x, \hat{z}) p_0(x, \hat{z}) \log \left( \frac{\theta(c^M(x, \hat{z}))^y (1 - \theta(c^M(x, \hat{z})))^{1 - y}}{p_0(y|x, \hat{z})} \right) \]

\[(31) \quad = \sum_{c^M \in C^M} [p_0(y = 1|c^M)p_0(c^M) \log(\theta(c^M)) + p_0(y = 0|c^M)p_0(c^M) \log(1 - \theta(c^M))] - K \]

where \(K\) does not depend on \(\theta(M)\). It is routine to show that each term in the sum of (31) 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

\[(32) \quad \mathcal{I}_t(M) = \mathcal{I}(M|\hat{h}^t) = \int \prod_{k=1}^{t-1} \hat{p}(y_k, x_k, \hat{z}_k|\theta) \mu^M(d\theta) \]

as well as the predictive distribution

\[(33) \quad \hat{p}_t^M(y, x, \hat{z}) = \hat{p}_t^M(y, x, \hat{z}|\hat{h}^t) = \int \hat{p}(y, x, \hat{z}|\theta) \mu^M(d\theta|\hat{h}^t) \]

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

Note that it is possible for $\psi_{c^M}$ to vary with $c^M$ when $\mu^M$ is uniformly non-doctrinaire.

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

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

$p_0^\infty$ almost surely.

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

1. $\mu^M(\{\theta: d_K(\hat{p}_\theta, p_0) < d\}) > 0$ only for, and for all, $d > \delta_M$.
2. $\lim_{t} \inf d_K(\hat{p}_M^t, p_0) \geq \delta_M$, $p_0^\infty$ almost surely
3. $\sup_t \text{Var}(\log(\mathcal{I}_{t+1}(M)/\mathcal{I}_t(M))) < \infty$

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

The third condition requires a bit more work to verify. Note that $\mathcal{I}_{t+1}(M) = \frac{\hat{p}_M^t(y_t, x_t, \hat{z}_t)}{p_0(y_t, x_t, \hat{z}_t)} \mathcal{I}_t(M)$ 

$\Rightarrow \log(\mathcal{I}_{t+1}(M)/\mathcal{I}_t(M)) = \log \left( \frac{\hat{p}_M^t(y_t, x_t, \hat{z}_t)}{p_0(y_t, x_t, \hat{z}_t)} \right)$, so condition (3) is equivalent to

$$\sup_t \text{Var} \left[ \log \left( \frac{\hat{p}_M^t(y_t, x_t, \hat{z}_t)}{p_0(y_t, x_t, \hat{z}_t)} \right) \right] < \infty$$

which can easily be shown to hold so long as

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

or

$$\sup_t E \left[ \log \left( \frac{\hat{p}_M^t(y|x, \hat{z})}{p_0(y|x, \hat{z})} \right)^2 \right] < \infty$$
for all \((y, x, \hat{z})\) which satisfy \(p_0(y, x, \hat{z}) > 0\).

To verify (37), 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

\[
q_t = \frac{\kappa_t + 1}{N_t + 2}
\]

(38)

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

\[
\hat{p}^M_t(y|x, \hat{z}) \rightarrow q_t^y(1 - q_t)^{1-y}
\]

(39)

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,

\[
| \log(\hat{p}^M_t(y|x, \hat{z}))^2 - \log(q_t^y(1 - q_t)^{1-y})^2 | < \epsilon
\]

for all \(t \geq n^{47}\) which implies that

\[
E[| \log(\hat{p}^M_t(y|x, \hat{z}))^2 - \log(q_t^y(1 - q_t)^{1-y})^2 |] < \epsilon
\]

(40)

for all \(t \geq n\). Since \(E[\log(\hat{p}^M_t(y|x, \hat{z}))^2] < \infty\) for all finite \(t\), to verify (37) it is sufficient to show that

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

(41)

by (40).

By symmetry, it is without loss of generality to verify (41) for the case where \(y = 1\). To this end,

\[
E[\log(q_t)^2] = E[E[\log(q_t)^2 | N_t]]
\]

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

\footnote{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
\[ \tilde{\theta} \equiv p_0(y = 1 | e^M(x, z)). \]

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 (41) and concludes the proof.

Define the Bayes’ factor conditional on \( p_0 \) as
\[ B_{i,j}(\hat{h}_t | p_0) = \frac{I_t(M_{i,j} | p_0)}{I_t(M_{X,Z} | p_0)} \]

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

**Lemma 3.** For all \( M_{i,j} \in \mathcal{M} \), \( p_0 \), and uniformly non-doctrinaire \( \mu_{i,j}, \mu^{X,Z} \),
\[ \frac{1}{t-1} \log B_{i,j}(p_0) \to \delta_{X,Z}(p_0) - \delta_{i,j}(p_0), \]
\( p_0^\infty \) almost surely.

**Proof.** Note that
\[ \frac{1}{t-1} \log B_{i,j}^t = \frac{1}{t-1} \log (I_t(M_{i,j})) - \frac{1}{t-1} \log (I_t(M_{X,Z})) \]
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 \( B_{i,j}(p_0) \to 0, p_0^\infty \) almost surely.

Remark 2 applies when \( \delta_{i,j}(p_0) > \delta_{X,Z}(p_0) \); what does the Bayes’ factor \( B_{i,j}(p_0) \) 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

$$
\phi(\kappa, N, \psi) = \int_0^1 \theta^\kappa (1 - \theta)^{N-\kappa} \psi(\theta) d\theta
$$

$$
\phi(\kappa, N) = \int_0^1 \theta^\kappa (1 - \theta)^{N-\kappa} d\theta = \text{Beta}(\kappa + 1, N - \kappa + 1)
$$

$$
\hat{q} = \frac{\kappa}{N}
$$

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

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

**Proof.** Note that for any non-doctrinaire $\psi$ 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 $\kappa$ 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.** $B_{X,-Z}(h^t) \to 0$ and $B_{-X,-Z}(h^t) \to 0$, $P_{\theta_0}$ almost surely. Additionally, if $x$ is important to predicting $y$, then $B_{-X,Z}(h^t) \to 0$, $P_{\theta_0}$ 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 $\delta_{X,-Z}(p_0^1) > \delta_{X,Z}(p_0^1), \delta_{-X,-Z}(p_0^1) > \delta_{X,Z}(p_0^1)$ and, whenever $x$ is important to predicting $y$, $\delta_{-X,Z}(p_0^1) > \delta_{X,Z}(p_0^1)$. 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

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

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

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

**Proof of Observation 1.2.** Lemma 5 shows that both \( B^t_{X', Z} \) and \( B^t_{X, Z} \) converge almost surely to zero and that \( B^t_{X, Z} \) 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 \( B^t_{X, Z} \) \( a.s. \to \infty \) whenever \( x \) is not important to predicting \( y \). First, write out the Bayes’ factor:

\[
B^t_{X, Z} = \frac{\Pr(h^t|M_{X, Z})}{\Pr(h^t|M_{X, Z})}
\]

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

From (47) it is sufficient to show that

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

for each \( z \in Z \).

Fix some \( z \). I will use Lemma 4 to estimate (48). Let \( \kappa_t = \kappa(z|h^t), N_t = N(z|h^t), \hat{q}_t = \frac{\kappa_t}{N_t}, \psi_t = \kappa_0(z', z|h^t), N_t' = N(z', z|h^t), \) and \( \hat{q}_t' = \frac{\kappa_t'}{N_t'} \).

Applying Lemma 4,

\[
\frac{\int_0^1 \theta^{\kappa_t}(1 - \theta)^{N_t - \kappa_t} \psi(\theta)d\theta}{\prod_{x'} \int_0^1 \theta^{\kappa_t'}(1 - \theta)^{N_t' - \kappa_t'} \psi(\theta)d\theta} \geq \frac{a\phi^*(\kappa_t, N_t)}{A^\#X \prod_{x'} \phi^*(\kappa_t', N_t')}
\]
for some constants $0 < a < A < \infty$. By the strong law of large numbers, the right hand side of (49) tends almost surely towards

$$C \sqrt{\prod_{t} N_t^{x_t}} \xrightarrow{a.s.} \infty$$

where $C$ is some positive constant independent of $t$. ■

B.2. Proof of Propositions from Section 3. I now present a series of Lemmas which will aid in the proof of propositions from Section 3. Define

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

(50)

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 establishes the almost sure limit of several Bayes’ factors when the agent never encodes $z$.

**Lemma 6.** Suppose $E_{\theta_0}[y|x] \neq E_{\theta_0}[y]$ for some $x \in X$. Then, for all uniformly non-doctrinaire $\mu^{\sim X, \sim Z}$, $\mu^{\sim X, Z}$, and $\mu^{X, Z}$, $B_{\sim X, \sim Z}(\hat{h}_t^t|p_0^0) \to 0$ and $B_{\sim X, Z}(\hat{h}_t^t|p_0^0) \to 0$, $(p_0^0) \to \infty$ almost surely.

**Proof.** Lemma 3 implies that it is sufficient to show that $\delta_{\sim X, \sim Z}(p_0^0) > \delta_{X, Z}(p_0^0)$ and $\delta_{X, Z}(p_0^0) > \delta_{X, Z}(p_0^0)$ whenever $E_{\theta_0}[y|x] \neq E_{\theta_0}[y]$ for some $x \in X$. 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

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

$$\pi_Z(\hat{h}_m^t|p_0^0) = \frac{1 + \frac{1-x}{x}B_{X, Z}(\hat{h}_m^t|p_0^0)}{1 + \frac{1-x}{x}B_{\sim X, Z}(\hat{h}_m^t|p_0^0) + \frac{1-x}{x}B_{X, Z}(\hat{h}_m^t|p_0^0) + \frac{(1-x)(1-x)}{x}B_{\sim X, \sim Z}(\hat{h}_m^t|p_0^0)}$$

(51)

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

(52)

$$B_{\sim X, Z}(\hat{h}_m^t|p_0^0) = 0$$

(53)

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

\[ B_{X,Z}(\hat{h}^t_m|p^0_0) = \frac{\int_0^1 \theta^c(m|\hat{h}^t_m)(1 - \theta)^N(m|\hat{h}^t_m) - \kappa(m|\hat{h}^t_m)\psi(\theta)d\theta}{\prod_x \int_0^1 \theta^c(x',m|\hat{h}^t_m)(1 - \theta)^N(x',m|\hat{h}^t_m) - \kappa(x',m|\hat{h}^t_m)\psi(\theta)d\theta} \]

\[ = \frac{\int_0^1 \theta^c(\hat{h}^t_m)(1 - \theta)^{t-1} - \kappa(\hat{h}^t_m)\psi(\theta)d\theta}{\prod_x \int_0^1 \theta^c(x'|\hat{h}^t_m)(1 - \theta)^N(x'|\hat{h}^t_m) - \kappa(x'|\hat{h}^t_m)\psi(\theta)d\theta} \]

\[ B_{X,\neg Z}(\hat{h}^t_m|p^0_0) = \frac{\prod_x \int_0^1 \theta^c(x'|\hat{h}^t_m)(1 - \theta)^N(x'|\hat{h}^t_m) - \kappa(x'|\hat{h}^t_m)\psi(\theta)d\theta}{\prod_x \int_0^1 \theta^c(x'|\hat{h}^t_m)(1 - \theta)^N(x'|\hat{h}^t_m) - \kappa(x'|\hat{h}^t_m)\psi(\theta)d\theta} \]

\[ = 1. \]

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

\[ \pi_Z(\hat{h}^t_m|p^0_0) = \frac{\pi_Z \left[ 1 + \frac{1 - \pi_X}{\pi_X} B_{X,Z}(\hat{h}^t_m|p^0_0) \right]}{1 + \frac{1 - \pi_X}{\pi_X} B_{X,\neg Z}(\hat{h}^t_m|p^0_0)} = \pi_Z. \]

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

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

(54) \[ \log(B_{i,X,Z}(\hat{h}^t)) \to -\infty \]

for each \( i \in \{X, \neg X\} \) with probability 1 conditional on the agent encoding \( z \) infinitely often.

Defining

\[ \hat{h}^t_1 \equiv (y_t, x_t, \hat{z}_t)_{t<\tau: \hat{z}_t \neq \emptyset} \]

\[ \hat{h}^t_0 \equiv (y_t, x_t, \hat{z}_t)_{t<\tau: \hat{z}_t = \emptyset} \]
we can write

\[ \mathcal{B}_{i,-z}^t = \frac{\Pr_X(\hat{h}_0|M_{i,-z}, \hat{h}_0)}{\Pr_X(\hat{h}_0|MX, Z)} \frac{\Pr_X(\hat{h}_1^t|M_{i,-z}, \hat{h}_1^t)}{\Pr_X(\hat{h}_1^t|M_{X,Z}, \hat{h}_0^t)} \]

so the LHS of (54) can be expressed as

\[
\log \left( \frac{\Pr_X(\hat{h}_0^t|M_{i,-z})}{\Pr_X(\hat{h}_0^t|MX, Z)} \right) + \log \left( \frac{\Pr_X(\hat{h}_1^t|M_{i,-z}, \hat{h}_0)}{\Pr_X(\hat{h}_1^t|M_{X,Z}, \hat{h}_0^t)} \right).
\]

When the agent fails to encode \( z \) only a finite number of times along a history, we can ignore the first term of (55) because it tends towards a finite value as \( t \to \infty \). Otherwise, Lemma 7 says that the first term of (55) 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 (55) tends towards \( -\infty \) with probability 1 in order to establish (54). This can be verified by showing that

\[
\limsup_t \frac{1}{\#\mathcal{E}(t)} \log \left( \frac{\Pr_X(\hat{h}_1^t|M_{i,-z}, \hat{h}_0)}{\prod_{\tau \in \mathcal{E}(t)} p_0^1(y_\tau, x_\tau, z_\tau)} \right) - \frac{1}{\#\mathcal{E}(t)} \log \left( \frac{\Pr_X(\hat{h}_1^t|M_{X,Z}, \hat{h}_0^t)}{\prod_{\tau \in \mathcal{E}(t)} p_0^1(y_\tau, x_\tau, z_\tau)} \right) < 0
\]

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

\[
p_0^1(y, x, z) = \theta_0(x, z)^y (1 - \theta_0(x, z))^{1-y} g(x, z)
\]

\( \mathcal{E}(t) = \{ \tau < t : \hat{z}_\tau \neq \emptyset \} \).

The second term on the LHS of (56) tends towards 0 with probability 1 by Lemma 2.\(^{48}\) To complete the proof, it then remains to show that the first term on the LHS of (56) remains bounded away from 0 as \( t \to \infty \) for \( i \in \{X, \neg X\} \), or

\[
\limsup_t \frac{1}{\#\mathcal{E}(t)} \log \left( \frac{\Pr_X(\hat{h}_1^t|M_{i,-z}, \hat{h}_0)}{\prod_{\tau \in \mathcal{E}(t)} p_0^1(y_\tau, x_\tau, z_\tau)} \right) < 0.
\]

\(^{48}\)Note that

\[
\frac{1}{\#\mathcal{E}(t)} \log \left( \frac{\Pr_X(\hat{h}_1^t|M_{X,Z}, \hat{h}_0^t)}{\prod_{\tau \in \mathcal{E}(t)} p_0^1(y_\tau, x_\tau, z_\tau)} \right) = \frac{1}{\#\mathcal{E}(t)} \log \left( \frac{\Pr_X(\hat{h}_1^t|M_{X,Z})}{\prod_{\tau \in \mathcal{E}(t)} p_0^1(y_\tau, x_\tau, z_\tau)} \right)
\]

since, under \( \mu^{X,Z} \), subjective uncertainty regarding \( \theta(x, \emptyset) \) and \( \theta(x, z') \), \( z' \neq \emptyset \), is independent.
We can re-write the LHS of (57) as

\[ \frac{1}{\#\mathcal{E}(t)} \log \left( \prod_{x'} \prod_{z'} \int \frac{\theta(x', z')^{\kappa(x', z'|\hat{h}_t)} (1 - \theta(x', z'))^{N(x', z'|\hat{h}_t) - \kappa(x', z'|\hat{h}_t)} \mu^{\hat{h}_t} (d\theta | \hat{h}_0^t)}{\prod_{x'} \prod_{z'} \theta_0(x', z')^{\kappa(x', z'|\hat{h}_t)} (1 - \theta_0(x', z'))^{N(x', z'|\hat{h}_t) - \kappa(x', z'|\hat{h}_t)}} \right). \]

Since \( \mu^{\hat{h}_t} (\cdot | \hat{h}_0^t) \) places full support on vectors of success probabilities \( (\theta) \) with \( \theta(x, z) = \theta(x', z') \) for all \( x, z, z' \), we can bound (58) by noting that

\[
\prod_{x'} \prod_{z'} \int \frac{\theta(x', z')^{\kappa(x', z'|\hat{h}_t)} (1 - \theta(x', z'))^{N(x', z'|\hat{h}_t) - \kappa(x', z'|\hat{h}_t)} \mu^{\hat{h}_t} (d\theta | \hat{h}_0^t)}{\prod_{x'} \prod_{z'} \theta_0(x', z')^{\kappa(x', z'|\hat{h}_t)} (1 - \theta_0(x', z'))^{N(x', z'|\hat{h}_t) - \kappa(x', z'|\hat{h}_t)}} \leq \max_{\theta(0), \theta(1)} \left( \prod_{x'} \prod_{z'} \theta(x')^{\kappa(x', z'|\hat{h}_t)} (1 - \theta(x'))^{N(x', z'|\hat{h}_t) - \kappa(x', z'|\hat{h}_t)} \right)
\]

which implies that (58) is bounded above by

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

for all \( t, \hat{h}_t \). By the strong law of large numbers, expression (59) can be shown to tend towards

\[ -d_E (\hat{p}_g(M_{X, -Z}); P_0^1) < 0 \]

with probability 1 conditional on the agent encoding \( z \) infinitely often; this establishes (57) and completes the proof.

**Proof of Proposition 1.** Suppose that, with positive probability under \( P_{\theta_0, \xi} (\cdot) \), 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 \( \hat{\pi}_Z^t \rightarrow 1 \) with probability 1. As a result, with probability 1 there exists a \( \tilde{t} \) such that \( \hat{\pi}_Z^t > b \) for all \( t \geq \tilde{t} \). Since \( \tilde{t} \), \( e_t = 1 \) for all \( t \geq \tilde{t} \), a contradiction.

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

\[ \Lambda(h^t) \equiv \frac{1 + \frac{1 - \pi_X}{\pi_X} B_{X, -Z}(h^t)}{B_{X, -Z}(h^t) + \frac{1 - \pi_X}{\pi_X} B_{X, -Z}(h^t)} \]

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 \( \Lambda(h^t) > \frac{1 - \pi_Z}{\pi_Z} b \).

Proof.

\[
\pi_Z(h^t) > b \iff \frac{1 + \alpha_{-X,Z} B_{-X,Z}(h^t)}{1 + \alpha_{-X,Z} B_{-X,Z}(h^t) + \alpha_{-X,-Z} B_{-X,-Z}(h^t) + \alpha_{-X,-Z} B_{-X,-Z}(h^t)} > b \iff \\
\frac{\alpha_{X,-Z} B_{X,-Z}(h^t)}{\alpha_{X,-Z} B_{X,-Z}(h^t) + \alpha_{X,-Z} B_{X,-Z}(h^t) + \alpha_{X,-Z} B_{X,-Z}(h^t)} > \frac{b}{1 - b} \iff \\
\Lambda(h^t) = \frac{1 + \frac{1 - \pi_X}{\pi_X} B_{X,-Z}(h^t)}{B_{X,-Z}(h^t) + \frac{1 - \pi_X}{\pi_X} B_{X,-Z}(h^t)} > 1 - \frac{\pi_Z}{\pi_Z} \frac{b}{1 - b}.
\]

Lemma 10. For all \( \epsilon > 0 \) there exists \( \lambda > 0 \) such that

\[
(61) \quad P_{\theta_0} \left( \min_{t \geq 1} \Lambda(h^t) > \lambda \right) \geq 1 - \epsilon
\]

Proof. From Lemma 5, we know that \( B_{X,-Z}(h^t) \overset{a.s.}{\to} 0, B_{-X,-Z}(h^t) \overset{a.s.}{\to} 0 \). As a result, \( \Lambda(h^t) = \frac{1 + \frac{1 - \pi_X}{\pi_X} B_{X,-Z}(h^t)}{B_{X,-Z}(h^t) + \frac{1 - \pi_X}{\pi_X} B_{X,-Z}(h^t)} \overset{a.s.}{\to} \infty \). Consequently, for all \( \epsilon > 0 \) there exists a value \( T \geq 1 \) such that \( P_{\theta_0} \left( \min_{t \geq T} \Lambda(h^t) \geq 1 \right) > 1 - \epsilon \) (see, for example, Lemma 7.2.10 in Grimmett and Stirzaker 2001).

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

\[
\min_{h \in H} \min_{1 \leq k \leq T} \Lambda(h^k) > \lambda
\]

we have

\[
P_{\theta_0} \left( \min_{t \geq 1} \Lambda(h^t) > \lambda \right) \geq 1 - \epsilon.
\]

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 $\tilde{\pi}_Z^k$). Then, for all $k < t$, the marginal distribution over $(x_k, \hat{z}_k)$ is identically $p_0^k(x_k, \hat{z}_k)$ since $\xi(z, \hat{h}^k)[m] = 1$. 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.


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 $\epsilon$. 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 $P_{\theta_0,\xi}(\hat{h}^t) = P_{\theta_0}(h^t)$. As a result, it is sufficient to show that there exists $\pi_1 \in (0, 1)$ ($b_1 \in (0, 1)$) such that

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

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

By Lemma 9

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

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

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

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

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

so the result follows from setting $\pi$ or $b$ to satisfy

$$\frac{1 - \pi}{\pi} \frac{b}{1 - b} = \lambda(\epsilon) \Rightarrow \pi_1 = \frac{b}{b + \lambda(\epsilon)(1 - b)}$$

$$b_1 = \frac{\lambda(\epsilon) \pi}{\pi(\lambda(\epsilon) - 1) + 1}.$$

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. 

\[ \square \]

**Proof of Proposition 3. Part 1.** The proof that if the agent settles on encoding \( z \) then \( \hat{E}[y|x, z, \hat{h}^t] \) converges to \( E_{\theta_0}[y|x, z] \) almost surely is analogous to the proof of Observation 1.1 and is thus omitted.

**Part 2.** If the agent settles on not encoding \( z \) then, by definition, there exists \( n \) such that \( \epsilon_t = 0 \) for all \( t \geq n \). In any period \( t \geq n \), his expectation satisfies

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

Fix an \( M \in M \). Since the marginal prior density over \( \theta(c^{M}(x, \emptyset)) \) is non-doctrinaire under \( M \), \( E_\xi[\theta(c^{M}(x, \emptyset))|\hat{h}, M] \to \tilde{h}_{\theta}(c^{M}(x, \emptyset)) \) by Theorem 2.4 of Diaconis and Freedman (1990) and the strong law of large numbers, where \( E_{\theta_0}[y|c^{M}(x, \emptyset)] = E_{\theta_0}[y|x] \) for \( M \in \{M_{X,Z}, M_{X,-Z}\} \) and \( E_{\theta_0}[y|c^{M}(x, \emptyset)] = E_{\theta_0}[y] \) for \( M \in \{M_{-,X,Z}, M_{-,X,-Z}\} \).

If \( E_{\theta_0}[y|x] = E_{\theta_0}[y] \) (i.e., \( X \) is a singleton by Assumption 2), then we are done. Assume \( E_{\theta_0}[y|x] \neq E_{\theta_0}[y] \) for some \( x \). It is left to show that both \( \hat{\pi}_{-,X,Z}^{t} \) and \( \hat{\pi}_{-,X,-Z}^{t} \) converge almost surely to zero. Equivalently, it is left to show that both \( B_{-,X,Z}(\hat{h}^t) \) and \( B_{-,X,-Z}(\hat{h}^t) \) converge almost surely to zero.

For \( t \geq n \) and \( j \in \{-Z, Z\} \),

\[
B_{-,X,j}(\hat{h}^t) = \frac{Pr_\xi(\hat{h}^t|M_{-,X,j})}{Pr_\xi(\hat{h}^t|M_{X,Z})} = \frac{Pr_\xi(\hat{h}^t|M_{-,X,j}, \hat{h}^n) Pr_\xi(\hat{h}^n|M_{-,X,j})}{Pr_\xi(\hat{h}^t|M_{X,Z}, \hat{h}^n) Pr_\xi(\hat{h}^n|M_{X,Z})} = \left( \frac{\int \prod_{k=n}^{t-1} \theta(x_k, \emptyset) y_k (1 - \theta(x_k, \emptyset))^{1-y_k} \mu_{-,X,J}(d\theta|\hat{h}^n)}{\int \prod_{k=n}^{t-1} \theta(x_k, \emptyset) y_k (1 - \theta(x_k, \emptyset))^{1-y_k} \mu_{X,Z}(d\theta|\hat{h}^n)} \right) \frac{Pr_\xi(\hat{h}^n|M_{-,X,j})}{Pr_\xi(\hat{h}^n|M_{X,Z})}
\]
where \( \hat{h}_n^t = (y_{t-1}, x_{t-1}, m, \ldots, y_n, x_n, \emptyset) \). Since \( \frac{\Pr(h^n \mid M_{-X,j})}{\Pr(h^n \mid M_{X,Z})} \) is fixed for all \( t \geq n \), it is necessary and sufficient to show that

\[
\left( \frac{\int \prod_{k=n}^{t-1} \theta(x_k, \emptyset)^{y_k} (1 - \theta(x_k, \emptyset))^{1 - y_k} \mu^{-X,j}(d\theta | \hat{h}^n)}{\int \prod_{k=n}^{t-1} \theta(x_k, \emptyset)^{y_k} (1 - \theta(x_k, \emptyset))^{1 - y_k} \mu^{X,Z}(d\theta | \hat{h}^n)} \right)
\]

(62)

converges to zero almost surely to establish such convergence of the Bayes’ factor. But, noting that (62) equals \( B_{-X,j}^t(p_0^0) \) for some uniformly non-doctrinaire \( \mu^{-X,j}, \mu^{X,Z} \), and \((y_{t-1}, x_{t-1}, \hat{z}_{t-1}, \ldots, y_n, x_n, \hat{z}_n)\)
is a random sample from \( p_0^0 \), the result follows from Lemma 6.

Proof of Proposition 4. Part 1. Analagous to the proof of Observation 1.2 and hence omitted. Part 2. The fact that \( \hat{\pi}^t_X \xrightarrow{a.s.} 1 \) when the agent settles on not encoding \( z \) follows immediately from Assumption 2 and Lemma 6. That \( \hat{\pi}^t_Z \leq b \) for large \( t \) follows from the definition of settling on not encoding \( z \) and the encoding rule.

B.3. Proofs from Section 4.

Proof of Proposition 5. A version of this result appears in Samuels (1993), but, for completeness, I’ll provide a proof.\(^{49}\)

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

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

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

\[
\phi = \text{Cov}(\mathbb{E}[y \mid x = 0, z], g(x = 1 \mid z))
\]

\(^{49}\)My proof closely follows Samuels’s.
We can write \( \hat{E}[y|x = 1, z] - \hat{E}[y|x = 0, z] \) as

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

\[\blacksquare\]

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

**Lemma 12.** If \( b_k \overset{i.i.d.}{\sim} U[0, 1] \) 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

\[
\frac{a(1 - \pi z)}{a(1 - \pi z) + b\pi z} < 1,
\]

where \( a \) and \( b \) are positive constants (do not depend on \( k \)).\(^{50}\) As a result, the probability of never encoding \( z \) after \( t \) is less than the infinite product

\[
\left( \frac{a(1 - \pi z)}{a(1 - \pi z) + b\pi z} \right)^\infty = 0
\]

\(^{50}\)Straightforward computations establish that whenever the agent does not encode \( z \) between \( t \) and \( t + k \),

\[
1 - \hat{\pi}_{t+k} = \frac{1 - \pi_Z B_{X,z}(\hat{h}_1^{t+k}) + \frac{1 - \pi_Z}{\pi_Z} B_{X,z}(\hat{h}_0^{t+k}) B_{X,z}(\hat{h}_0^{t+k})}{1 + \frac{1 - \pi_Z}{\pi_Z} B_{X,z}(\hat{h}_1^{t+k}) + \frac{1 - \pi_Z}{\pi_Z} B_{X,z}(\hat{h}_0^{t+k}) + \frac{1 - \pi_Z}{\pi_Z} B_{X,z}(\hat{h}_0^{t+k}) B_{X,z}(\hat{h}_0^{t+k})},
\]

where

\[
B_{i,j}(\hat{h}_1^{t+k}) = \frac{\pi_x(h_1^{t+k})}{\pi_x(h_0^{t+k})} (\text{recall that } \hat{h}_1 = (y_\tau, x_\tau, \hat{z}_\tau)_{\tau < j: \hat{z}_\tau \neq \emptyset} \text{ and } \hat{h}_0 = (y_\tau, x_\tau, \hat{z}_\tau)_{\tau < j: \hat{z}_\tau = \emptyset})
\]

Upper bound (63) is derived by noting that, fixing \( t \), both \( B_{X,z}(\hat{h}_1^{t+k}) \) and \( B_{X,z}(\hat{h}_0^{t+k}) \) are bounded above by some finite positive constant \( a \) independent of \( k \) and history \( \hat{h}^{t+k} \). Likewise, fixing \( t \), \( B_{X,z}(\hat{h}_0^{t+k}) \) is bounded below by some finite positive constant \( b \) independent of \( k \) and history \( \hat{h}^{t+k} \). As a result, \( 1 - \hat{\pi}_{t+k} < \frac{a(1 - \pi_Z)}{a(1 - \pi_Z) + b\pi_Z} \).

Now take the supremum of this expression with respect to all possible values of \( B_{X,z}(\hat{h}_0^{t+k}) > 0 \) to get (63).

\[57\]
and the result follows.

**Proof of Proposition 7.**

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 $\hat{\pi}_Z^t \rightarrow 1$ almost surely which implies that $\eta(\hat{\pi}_Z^t) = \hat{\pi}_Z^t \rightarrow 1$ almost surely.

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

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

Expanding

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

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

to establish (64) it is sufficient to show that

- **A.** $E_{\xi}[\theta(x, z)|\hat{h}^t] \xrightarrow{a.s.} E_{\theta_0}[y|x, z]$  
- **B.** $P_{\theta_0, \xi}(e_t = 1) \rightarrow 1$

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 $P_{\theta_0, \xi}(e_t = 1) = E_{\theta_0, \xi}[\eta(\hat{\pi}_Z^t)]$ and $E_{\theta_0, \xi}[\eta(\hat{\pi}_Z^t)] \rightarrow 1$ because $\eta(\hat{\pi}_Z^t)$ is bounded and tends almost surely towards 1 by Proposition 7.1.

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

**Lemma 13.** If $b_k \sim U[0, 1]$ and $\hat{\pi}_Z^t \xrightarrow{a.s.} 1$ then $\frac{\#\mathcal{E}(t)}{t-1} \xrightarrow{a.s.} 1$.

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

$$
\gamma^t - \gamma^{t-1} = \frac{e_t - \gamma^{t-1}}{t-1}
$$

$$
= \frac{1}{t-1}(F(\gamma^{t-1}) + e_t + u_t),
$$

58
where

\[ F(\gamma^{t-1}) = 1 - \gamma^{t-1} \]

\[ \epsilon_t = \epsilon_t - \hat{\pi}_Z^t \]

\[ u_t = \hat{\pi}_Z^t - 1. \]

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. \( u_t \xrightarrow{a.s.} 0 \)

so Theorem A in Fudenberg and Takahashi (2007) tells us that, with probability 1, every \( \omega \)-limit of the process \( \{\gamma^t\} \) is connected and internally chain recurrent for \( \Phi \), where \( \Phi \) 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 \( \Phi \) is \( \{1\} \) by Liouville’s Theorem.

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

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

**Lemma 15.** \( \frac{B_{X,-Z}(\hat{h}_t)}{e^{-u_t(t-1)}} \xrightarrow{a.s.} K \) for some \( K \) satisfying \( 0 < K < \infty \).

**Proof.** I will show that

\[ \frac{1}{t-1} \log B_{X,-Z}(\hat{h}_t) \rightarrow -d \]

almost surely. We can write

\[ B_{X,-Z}(\hat{h}_t) = \frac{\Pr(\hat{h}_t^1|M_{X,-Z}) \Pr(\hat{h}_0^t|M_{X,-Z}, \hat{h}_1^t)}{\Pr(\hat{h}_1^t|M_{X,Z}) \Pr(\hat{h}_0^t|M_{X,Z})} \]

59
From (66), we can write the left hand side of (65) as

\[
\begin{align*}
\frac{1}{t-1} \left[ \log \left( \frac{\Pr(\hat{h}_t^t \mid M_{XZ})}{\prod_{k \in \mathcal{E}(t)} p_0^t(y_k, x_k, z_k)} \right) \right] + \log \left( \frac{\Pr(\hat{h}_t^t \mid M_{XZ}, \hat{h}_1^t)}{\prod_{k \notin \mathcal{E}(t)} p_0^t(y_k, x_k, \emptyset)} \right) \\
- \frac{1}{t-1} \left[ \log \left( \frac{\Pr(\hat{h}_t^t \mid M_{XZ})}{\prod_{k \in \mathcal{E}(t)} p_0^t(y_k, x_k, z_k)} \right) \right] + \log \left( \frac{\Pr(\hat{h}_t^t \mid M_{XZ})}{\prod_{k \notin \mathcal{E}(t)} p_0^t(y_k, x_k, \emptyset)} \right)
\end{align*}
\]  

(67)

We know that the second term of (67) tends almost surely towards 0 as \( t \to \infty \) by Lemma 2.

As a result, to establish (65) it remains to show that the first term tends almost surely towards \(-d\). Rewrite this term as

\[
\frac{\# \mathcal{E}(t)}{t-1} \left[ \frac{1}{\# \mathcal{E}(t)} \log \left( \frac{\Pr(\hat{h}_t^t \mid M_{XZ})}{\prod_{k \in \mathcal{E}(t)} p_0^t(y_k, x_k, z_k)} \right) \right] + \frac{t-1-\# \mathcal{E}(t)}{t-1} \left[ \frac{1}{t-1-\# \mathcal{E}(t)} \log \left( \frac{\Pr(\hat{h}_0^t \mid M_{XZ}, \hat{h}_1^t)}{\prod_{k \notin \mathcal{E}(t)} p_0^t(y_k, x_k, \emptyset)} \right) \right]
\]  

(68)

By Lemmas 2, 13, and 14, (68) tends almost surely towards 

\[ 1 \times -d + 0 \times 0 = -d \]

as \( t \to \infty \), which completes the proof.

\[ \blacksquare \]

**Lemma 16.** 
\( \frac{B_{-X,-Z}(\hat{h}_t^t)}{e^{-d(t-1)}} \xrightarrow{a.s.} K \) for some \( d' \geq d \) and \( K \) satisfying \( 0 < K < \infty \).

**Proof.** Let \( d' = \delta_{-X,-Z}(p_0^t) \). Using analogous arguments to those in the proof of Lemma 15, can show that 

\( \frac{B_{-X,-Z}(\hat{h}_t^t)}{e^{-\delta_{-X,-Z}(p_0^t)(t-1)}} \xrightarrow{a.s.} K \) for some \( K \) satisfying \( 0 < K < \infty \). Since \( \delta_{-X,-Z}(p_0^t) > \delta_{X,-Z}(p_0^t) \) (from the fact that adding more constraints weakly increases the minimized Kullback-Leibler divergence) and \( \delta_{X,-Z}(p_0^t) = d \) (by Lemma 14), the result follows.

\[ \blacksquare \]

**Proof of Proposition 8.** Recall that we want to show that

\[
\frac{1 - \hat{\pi}_Z^t}{e^{-d(t-1)}} \xrightarrow{a.s.} C
\]

(69)

for some positive constant \( C < \infty \).

Since

\[
1 - \hat{\pi}_Z^t = \frac{1}{1 + \frac{1 - \pi_X B_{-X,Z}(\hat{h}_t^t)}{\pi_Z B_{X,-Z}(\hat{h}_t^t) + \frac{1 - \pi_X B_{-X,Z}(\hat{h}_t^t)}{\pi_Z B_{X,-Z}(\hat{h}_t^t)}}}
\]

Since
to demonstrate (69) it suffices to show that

\[ e^{-d(t-1)} \frac{1 - \pi_Z}{\pi_Z} B_{X, -Z}(\hat{h}^t) + \frac{1 - \pi_X}{\pi_X} B_{\pi_{X, Z}}(\hat{h}^t) \]

\[ \xrightarrow{a.s.} c' \]

for some constant \( c' \) satisfying \( 0 < c' < \infty \).

The first term on the left hand side of (70) converges almost surely to a positive finite constant by Lemmas 15 and 16. The second term on the left hand side of (70) converges almost surely to 0 whenever \( \pi_X = 1 \) (trivially) or \( x \) is important to predicting \( y \) (by Lemmas 3, 15, and 16).