SELECTIVE ATTENTION AND LEARNING

Joshua Schwartzstein
Dartmouth College

Abstract
What do we notice and how does this affect what we learn and come to believe? I present a model of an agent who learns to make forecasts on the basis of readily available information, but is selective as to which information he attends to: he chooses whether to attend as a function of current beliefs about whether such information is predictive. If the agent does not attend to some piece of information, it cannot be recalled at a later date. He uses Bayes' rule to update his beliefs given attended-to information, but does not attempt to fill in missing information. The model demonstrates how selective attention may lead the agent to persistently fail to recognize important empirical regularities, make systematically biased forecasts, and hold incorrect beliefs about the statistical relationship between variables. In addition, it identifies factors that make such errors more likely or persistent. The model is applied to shed light on stereotyping and discrimination, persistent learning failures and disagreement, and the process of discovery. (JEL: C11, D01, D03, D83, D84)

1. Introduction

We learn to make forecasts through repeated observation. Consider an employer learning to predict worker productivity, a loan officer figuring out how to form expectations about trustworthiness and default, or a professor learning which teaching practices work best. Learning in this manner often relies on what we remember: characteristics of past workers, details of interactions with small business owners, teaching practices used in particular lectures. 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. Memory imperfections do not

The editor in charge of this paper was George-Marios Angeletos.

Acknowledgments: I am deeply grateful to Drew Fudenberg, Sendhil Mullainathan, and Andrei Shleifer for their generous guidance and encouragement throughout this project, and to Dan Benjamin, John Beshears, Pedro Bordalo, Ryan Bubb, Sylvain Chassang, Ian Dew-Becker, Ignacio Esponda, Nicola Gennaioli, Edward Glaeser, Lawrence Katz, Alex Kaufman, Scott Kominers, David Laibson, Ted O'Donoghue, Giacomo Ponzetto, Simone Schaner, Chris Snyder, Jeremy Stein, Rick Townsend, Timothy Vogelsang, Glen Weyl, Xiaoqi Zhu, three anonymous referees, an editor, and a co-editor for extremely helpful comments. This research was supported by the National Institute on Aging, Grant Number T32-AG000186 and by a National Science Foundation Graduate Research Fellowship.

E-mail: josh.schwartzstein@dartmouth.edu

Journal of the European Economic Association December 2014 12(6):1423–1452

DOI: 10.1111/jeea.12104

just stem from limited recall of information stored in memory; not all information will be attended to or encoded in the first place. It is hard or impossible to take note of all the characteristics of a worker, every detail of a face-to-face meeting, each aspect of how we teach. 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 of belief formation which highlights and draws out the implications of a key feature of what we notice in tasks of judgment and prediction: attention is *selective*, whereby we narrow our attention to variables currently believed to be informative relative to a prediction task (Kahneman 1973).[1] Rather than being endowed with "rational expectations" on what matters (e.g., Sims 2003, 2006; Gabaix 2013), an agent needs to *learn* which variables are worth attending to through experience. The model makes predictions about when the agent will in fact learn to attend to the right variables, when he will not, and how he may form systematically biased beliefs when he does not. The key insight is that inattention can compound itself: if the agent's prior does not indicate that he should attend to a variable, he may persistently fail to learn whether it is worth attending to. Such an agent may miss important empirical regularities and form incorrect beliefs about what causes variation in the data. Instead of necessarily learning to attend to important variables, the agent is biased towards coming to believe that what he attends to *is* important.

Section 2 sets up the model. An agent learns to predict binary outcome y given x and z, where x and z are finite random variables. Since the model involves a general forecasting task, it applies to a wide variety of situations: the agent could be an individual learning to predict others' behavior, an investor learning to predict whether an investment opportunity will be successful, a manager learning to predict output quality, and so on. The agent has a prior belief over whether x and/or z should be predictive of y. Additionally, conditional on being predictive, he has prior beliefs over *how* these variables predict y. A feature of the environment is that a standard Bayesian who attends to all details of events eventually learns which variables are predictive and makes asymptotically accurate forecasts, so any persistently biased forecasts stem from selective attention. To draw out the implications of such inattention in a simple manner, I consider what happens when the agent is selectively attentive to z: I assume the likelihood that the agent attends to z is increasing in the current probability he attaches to z being predictive of y, taking as given that the agent attends to y and x. In the baseline specification, the agent attends to z if and only if he places sufficient weight on it being predictive relative to some fixed cutoff, parameterizing the shadow

1. Schacter (2001) provides an overview of the evidence on memory limitations and, in particular, the second chapter explores research on the interface between attention and memory. See also DellaVigna (2009) for a recent survey of field evidence from economics on limited attention. I discuss the relationship between my model and the related economic literature in detail after presenting the model and results in full (Section 6), where such literature includes models of rational inattention (e.g., Sims 2003; Gabaix 2013; Woodford 2009), bandit problems and self-confirming equilibrium (e.g., Gittins 1979; Fudenberg and Levine 1993), coarse thinking (e.g., Jehiel 2005; Mullainathan, Schwartzstein, and Shleifer 2008), and confirmatory bias (e.g., Rabin and Schrag 1999).

cost of devoting attention.[2] The agent updates his beliefs using Bayes' rule, but, in the spirit of assumptions found in recent work modeling biases in information processing (e.g., Mullainathan 2002; Rabin and Schrag 1999), he is *naive* in the sense that he does not attempt to infer what z may have been. Instead, he uses an update rule which treats a missing value of z as a fixed but distinct nonmissing value. (Online Appendix B considers the alternative assumption that the agent is sophisticated.)

Section 3 draws out basic implications of the model. Due to selective attention, current beliefs affect which variables are attended to and, consequently, what is learned. Because of this dependence, the agent may persistently fail to pay attention to an important variable and, as a result, will not learn how it is related to the outcome of interest: under selective attention, an incorrect belief that z is not important is *self-confirming*. If we start off thinking it unlikely that food allergies are causing a headache, we are unlikely to track the relationship between what we eat and how we feel, and may fail to discover that such allergies are indeed the cause. This sheds light on evidence of people persistently failing to learn the importance of certain variables, including individuals neglecting the importance of the situation for determining every-day behavior (Ross 1977), small investors failing to appreciate the importance of analyst affiliation in interpreting investment recommendations (Malmendier and Shanthikumar 2007), or managers not recognizing how the cleanliness of the factory floor matters for productivity (Bloom et al. 2013). The model further predicts that such failures are more likely when the agent has less of an initial reason to suspect that z is predictive, matching evidence that people are less likely to notice relationships that prior theories do not deem plausible (Nisbett and Ross 1980).

Section 3 goes on to demonstrate that, when the agent settles on not attending to z, his limiting forecasts are as if he knows the true joint distribution over (y, x, z) but cannot observe z. As a result, a failure to learn to attend to a predictive variable feeds back to create a problem akin to omitted variable bias: by not learning to pay attention to a variable, an agent may persistently misreact to an associated variable, and naiveté implies that the agent can also misattribute cause to such a variable under the interpretation that an agent attributes cause whenever he maintains a belief that a variable is predictive, holding others fixed. However, such biased beliefs are *systematic*: because these beliefs must be consistent with the distribution over (y, x), whether or not the agent does misreact or misattribute cause, and the extent of his misreaction, depends completely on observable features of the environment. We may erroneously come to believe that a headache is caused by seasonal allergies rather than what we eat, but only if we tend to eat different foods across seasons. Moreover, such biased beliefs are *robust*: even if exogenous shocks lead the agent to begin attending to an important variable, he will have to track that variable for a long time to learn how it is related to the outcome of interest and whether it causes the variation previously

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

attributed to some other factor. If we go to a doctor to complain about the headache, we may not be able to answer whether it is particularly strong after eating certain foods, not having suspected a food allergy before. The model is illustrated with examples on the formation and persistence of systematically biased beliefs or stereotypes (Schaller and O'Brien 1992; Fiske and Taylor 2008).

To more formally study the robustness of incorrect beliefs stemming from selective attention, Section 4 extends the earlier analysis by assuming that there are random fluctuations in the shadow cost of devoting attention in a given period, where these fluctuations are such that the likelihood that the agent attends to z varies monotonically and continuously in the intensity of his belief in the importance of z. With the "continuous attention" assumptions, the agent will eventually learn to devote more and more attention to z, but this process may be very slow since the agent can only learn from information he attends to. The main result of this section concerns the speed of convergence, which increases in the degree to which the agent finds it difficult to explain what he observes without taking z into account. If knowledge of what we eat as well as the season does not add much explanatory power over a model that just includes the season, then it will take a particularly long time to learn to attend to what we eat. Since the degree of association between x and z both leads an agent to misreact to x when he fails to attend to z and take a long time to learn to attend to z, the same features that contribute to greater bias can make the bias more persistent.

The model is meant to apply to situations in which an agent needs to learn which variables are worth attending to in predicting some outcome of interest, and how those variables matter. Section 5 applies the model to analyze stereotyping and discrimination, the nature of learning failures and disagreement, as well as the process of discovery. Section 6 then goes on to discuss related literature and alternative approaches I could have taken. Section 7 concludes.

There are four online appendices: Online Appendix A contains further formal results that are referenced in the main text, Online Appendix B compares the naive and sophisticated versions of the model, Online Appendix C presents the proofs, and Online Appendix D presents further technical results which are useful for the proofs.

2. Model

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 $x \in \mathrm{X}$ and $z \in \mathrm{Z}$ are finite random variables, which, unless otherwise noted, can each take on at least two values. For example, the agent could be an individual learning to predict a person's behavior (y) on the basis of information on their racial, gender, ethnic, occupational, or other group membership (x) as well as situational factors (z); or an investor learning to predict whether an investment will be successful (y) given an analyst's recommendation (x) and the

analyst's affiliation (z); or a manager learning to predict output quality (y) given how worker effort is monitored (x) and the tidiness of the work area (z), and so on.

Each period t, the agent: (i) observes some draw of (x, z), (x_t, z_t), from fixed distribution $g(x, z)$, (ii) gives his prediction of y, $\hat{y}_t$, to maximize $-(\hat{y}_t - y_t)^2$, and (iii) learns the true y_t. The agent knows that, given covariates (x, z), y is independently drawn from a Bernoulli distribution with fixed but *unknown* success probability $\theta_0(x, z)$ each period (i.e., $p_{\boldsymbol{\theta_0}}(y = 1|x, z) = \theta_0(x, z)$). Additionally, he knows the joint distribution $g(x, z)$, which is positive for all (x, z).[3]

I make an assumption on the (unknown) vector of success probabilities.

ASSUMPTION 1. z is important to predicting y: there exist x, z, z' such that $\theta_0(x, z) \neq \theta_0(x, z')$.

Later on, I sometimes consider the case where x is *unimportant* to predicting y, conditional on z, to highlight how selective attention to z can lead to biased beliefs, in particular an incorrect belief that x is important.[4] Either way, to limit the number of cases considered, I assume that the *unconditional* (of z) success probability depends on x, for example if whether a particular person has headaches is associated with the season, not controlling for what she eats. Formally, defining $p_{\boldsymbol{\theta_0}}(y = 1|x) \equiv \sum_{z'} \theta_0(x, z')g(z'|x)$, I make the following assumption.

ASSUMPTION 2. x is important to predicting y in the absence of conditioning on z: $p_{\boldsymbol{\theta_0}}(y = 1|x) \neq p_{\boldsymbol{\theta_0}}(y = 1|x')$ for some $x, x' \in \mathrm{X}$.

Prior. Since the agent does not know $\boldsymbol{\theta_0} = (\theta_0(x', z'))_{x' \in \mathrm{X}, z' \in \mathrm{Z}}$, he estimates it from the data using a hierarchical prior $\mu(\boldsymbol{\theta})$.[5] Specifically, he entertains and places positive probability on each of four different models of the world, $M \in \{M_{X,Z}, M_{\neg X,Z}, M_{X,\neg Z}, M_{\neg X,\neg Z}\} \equiv \mathcal{M}$. These models correspond to whether x and/or z are important to predicting y and each is associated with a prior distribution $\mu^{i,j}(\boldsymbol{\theta})$ ($i \in \{X, \neg X\}$, $j \in \{Z, \neg Z\}$) over vectors of success probabilities. The vector of success probabilities $\boldsymbol{\theta} = (\theta(x', z'))_{x' \in \mathrm{X}, z' \in \widehat{\mathrm{Z}}}$ has dimension $|\mathrm{X}| \times |\widehat{\mathrm{Z}}|$, where $\widehat{\mathrm{Z}} \supset \mathrm{Z}$. The importance of defining $\widehat{\mathrm{Z}}$ will be clear later on when describing selectively attentive forecasts, but, briefly, it will denote the set of ways in which a selectively attentive agent can encode or later recall z.

3. 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 θ_0.

4. Analogous to how we define the importance of z in Assumption 1, we say that x is important to predicting y if and only if there exist x, x', z such that $\theta_0(x, z) \neq \theta_0(x', z)$.

5. This 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. I provide an alternative, more explicit, description of the agent's prior in Online Appendix A.1.

TABLE 1. Set of models over which variables are predictive.

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

Under $M_{\neg X,\neg Z}$, the success probability $\theta(x,z)$ depends on neither x nor z:

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

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

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

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

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

so $M_{\neg X,Z}$ is a $|\widehat{\mathrm{Z}}|$-parameter model. Finally, under $M_{X,Z}$, $\theta(x,z)$ depends on both x and z so it is a $|\mathrm{X}| \times |\widehat{\mathrm{Z}}|$-parameter model. Table 1 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, $\psi(\cdot)$. I make a technical assumption on the density ψ which guarantees that a standard Bayesian will have correct beliefs in the limit (Diaconis and Freedman 1990; Fudenberg and Levine 2006), namely that the density ψ is *nondoctrinaire*: it is continuous and strictly positive.

Denote the prior probability placed on model $M_{i,j}$ by $\pi_{i,j}$ and assume the agent's prior subjective uncertainty over whether x is important is independent of that over whether z is important: suppose there exist $\pi_X, \pi_Z \in (0,1]$ such that $\pi_{X,Z} = \pi_X\pi_Z$, $\pi_{X,\neg Z} = \pi_X(1-\pi_Z)$, $\pi_{\neg X,Z} = (1-\pi_X)\pi_Z$, and $\pi_{\neg X,\neg Z} = (1-\pi_X)(1-\pi_Z)$, where π_X is interpreted as the subjective prior probability that x is important to predicting y and π_Z is interpreted as the subjective prior probability that z is important to predicting y.

2.2. Standard Bayesian

Denote the history through period t by

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

The probability of such a history is derived from the underlying probability distribution over infinite-horizon histories $h^\infty \in H^\infty$ as generated by θ_0 together with g, where this distribution is denoted by $P_{\boldsymbol{\theta_0}}$.

Since the agent does not know $\boldsymbol{\theta_0}$, he cannot update his beliefs using $P_{\boldsymbol{\theta_0}}$. Rather, the agent's prior, together with g, generates a joint distribution over $\Theta, \mathcal{M}$, and H, where Θ is the set of all possible values of $\boldsymbol{\theta_0}$, $\mathcal{M}$ is the set of possible models, and H is the set of all possible histories. Denote this distribution by $\Pr(\cdot)$, from which we derive the (standard) Bayesian's beliefs. His period-t forecast of y given x and z equals

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

where $\pi^t_{i,j} \equiv \Pr(M_{i,j}|h^t)$ equals the posterior probability placed on model $M_{i,j}$. It follows from well-known results (e.g., Diaconis and Freedman 1990) that, as a result of the nondoctrinaire assumption, the period-t likelihood the Bayesian attaches to $y = 1$ given x and z asymptotically approaches a weighted average of (i) the empirical frequency of $y = 1$ given (x, z), (ii) the empirical frequency of $y = 1$ given (x), the empirical frequency of $y = 1$ given (z), and the unconditional empirical frequency of $y = 1$.

The first observation characterizes further asymptotic properties of the standard Bayesian model, and makes use of the following definition.

DEFINITION 1. The agent *learns the true model* if

1. whenever x (in addition to z) is important to predicting y, $\pi^t_{X,Z} \to 1$,
2. whenever x (unlike z) is unimportant to predicting y, $\pi^t_{\neg X,Z} \to 1$.

OBSERVATION 1. Suppose the agent is a standard Bayesian. Then

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

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.[6] In this environment, any deviations from (asymptotically) perfect learning must stem from selective attention.

6. Interestingly, whenever x is unimportant to predicting y the Bayesian's posterior eventually places negligible weight on all models other than $\boldsymbol{M}_{\neg X,Z}$. 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).

2.3. 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, Pelham, and Krull 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), where encoding can roughly be thought of as storing into memory. Specifically, there is much experimental evidence that individuals narrow their attention to stimuli perceived to be important in performing a given task, and unattended-to stimuli are less likely to be remembered (e.g., Mack and Rock 1998; von Hippel et al. 1993). Consequently, at later date t, the agent may only have access to an incomplete mental representation of history h^t, denoted by $\hat{h}^t$.

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

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

where

$$\hat{z}_k = \begin{cases} z_k & \text{if } e_k = 1 \text{ (the agent encodes } z_k\text{)}, \\ \varnothing & \text{if } e_k = 0 \text{ (the agent does not encode } z_k\text{)} \end{cases} \tag{3}$$

and

$$e_k = \begin{cases} 1 & \text{if } \hat{\pi}_Z^k > b_k, \\ 0 & \text{if } \hat{\pi}_Z^k \leq b_k. \end{cases} \tag{4}$$

The $e_k \in \{0, 1\}$ 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 (it can also be thought of as capturing the shadow cost of devoting attention to z), and $\hat{\pi}_Z^k$ denotes the probability that the agent attaches to z being important to predicting y in period k. I assume that b_k is a random variable which is independent of (x_k, z_k) and independently drawn from a fixed and known distribution 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), (4) 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. To start, I assume that $b_k \equiv b$ for some $b \in (0, 1)$ so the agent is always somewhat busy, and, as a result, encodes z if and only if he believes sufficiently strongly that it aids in predicting y. In Section 4, I consider the case where there are random, momentary, fluctuations in the degree to which the agent is cognitively busy in a given period—that is, b_k is drawn according to a nondegenerate distribution. In this case, the likelihood

that the agent attends to z varies more continuously in the intensity of his belief that z is important to predicting y.

For later reference, equations (3) and (4) (together with the agent's prior as well as an assumption about how b_k is distributed) implicitly define an *encoding rule* ξ: $\mathrm{Z} \times \hat{H} \to \Delta(\mathrm{Z} \cup \{\varnothing\})$ for the agent, where $\hat{H}$ denotes the set of all possible recalled histories and $\xi(z, \hat{h}^k)[\hat{z}']$ equals the probability (prior to b_k being drawn) that $\hat{z}_k = \hat{z}' \in \mathrm{Z} \cup \{\varnothing\}$ given z and $\hat{h}^k$. In other words, the encoding rule specifies how the agent encodes z given any history.[7]

How the Agent Treats Missing Information. To derive forecasts and beliefs given potentially incomplete history, $\hat{h}^t$, I need to specify how the agent treats missing values of z. I assume that he is naive and ignores any memory imperfections that result from selective attention when drawing inferences. I model this by assuming that the agent's prior treats missing and nonmissing information exactly the same: it treats $\varnothing$ as if it were a fixed but distinct nonmissing value.

ASSUMPTION 3. The agent is naive in performing statistical inference: his prior μ is over $[0, 1]^{|\mathrm{X}|\times|\widehat{\mathrm{Z}}|}$, where $\widehat{\mathrm{Z}} = \mathrm{Z} \cup \{\varnothing\}$ and the details of this prior are as specified in Section 2.1.

It is easiest to understand this assumption by comparing the naive agent with the more familiar sophisticated agent. In contrast to the naive agent, a sophisticated agent's prior only needs to be over $[0, 1]^{|\mathrm{X}|\times|\mathrm{Z}|}$ since he takes advantage of the structural relationship relating the success probability following missing and nonmissing values of z. Thus, whereas the naive agent treats missing and nonmissing values of z the exact same for purposes of inference, the sophisticated agent treats missing information differently than nonmissing information: He attempts to infer what missing data could have been when updating his beliefs.[8]

I maintain the naiveté assumption in the main text because it strikes me as more closely capturing the idea that the agent may be truly inattentive than the sophisticated

7. ξ, $\boldsymbol{\theta}_0$, and g generate a measure $P_{\boldsymbol{\theta}_0, \xi}$ over $\widehat{H}^\infty$, where $\hat{H}^\infty$ denotes the set of all infinite-horizon recalled histories. All remaining statements regarding almost sure convergence are with respect to this measure.

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

$$\Pr(\hat{h}^t | \tilde{\Theta}, \boldsymbol{M}) \propto \frac{\int_{\tilde{\Theta}} \prod_{\tau=1}^{t-1} p_{\boldsymbol{\theta}}(y_\tau | x_\tau, \hat{z}_\tau) \mu^M(d\boldsymbol{\theta})}{\int_{\tilde{\Theta}} \mu^M(d\boldsymbol{\theta})}, \tag{5}$$

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

$$\Pr{}^{\mathrm{S}}(\hat{h}^t | \tilde{\Theta}, \boldsymbol{M}) \propto \frac{\int \prod_{\tau \in \mathcal{E}(t)} p_{\boldsymbol{\theta}}(y_\tau | x_\tau, z_\tau) \prod_{\tau \notin \mathcal{E}(t)} p_{\boldsymbol{\theta}}(y_\tau | x_\tau) \mu^M(d\boldsymbol{\theta})}{\int_{\tilde{\Theta}} \mu^M(d\boldsymbol{\theta})}, \tag{5S}$$

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

alternative: when he does not attend to a variable, he both does not encode it and does not attempt to infer what it may have been. 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). Nevertheless, it is a useful exercise to draw out the implications of both models. I highlight which arguments and results rely on the naiveté assumption as they arise and, in Online Appendix B, I formally compare the models' implications.

While an individual treats $\varnothing$ as a fixed but distinct nonmissing value when drawing inferences, I assume that he is otherwise sophisticated in the sense that he "knows" the conditional likelihood of not encoding z given his encoding rule: his beliefs are derived from $\Pr_\xi(\cdot)$, which is the joint distribution over Θ, $\mathcal{M}$, and $\hat{H}$ as generated by his prior together with g and ξ. 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 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.

Beliefs and Forecasts. The probability that the selectively attentive agent assigns to model $M_{i,j}$ in period t is given by $\hat{\pi}^t_{i,j} = \Pr_\xi(M_{i,j}|\hat{h}^t)$, from which we derive the probability he assigns to z being important to predicting y, $\hat{\pi}^t_Z$, and the probability he assigns to x being important to predicting y, $\hat{\pi}^t_X$: $\hat{\pi}^t_Z = \Pr_\xi(M_{\neg X,Z}|\hat{h}^t) + \Pr_\xi(M_{X,Z}|\hat{h}^t)$ and $\hat{\pi}^t_X = \Pr_\xi(M_{X,\neg Z}|\hat{h}^t) + \Pr_\xi(M_{X,Z}|\hat{h}^t)$.

His period-t forecast of y given x and z is $\hat{E}[y|x,z,\hat{h}^t] = E_\xi[\theta(x,\hat{z})|\hat{h}^t]$, which almost surely approaches a weighted average of (i) the empirical frequency of $y = 1$ given $(x,\hat{z})$, (ii) the empirical frequency of $y = 1$ given (x), (iii) the empirical frequency of $y = 1$ given $(\hat{z})$, and (iv) the unconditional empirical frequency of $y = 1$. (See Online Appendix A.1 for more details on the agent's forecasts.) Importantly, in a given period t, the agent does not condition on z but on z as it is encoded in that period: encoding takes place before forecasting.

2.4. Discussion of Assumptions

It is worth discussing the assumptions underlying the model in a bit more detail. First, note the asymmetry between x and z: the agent is assumed to encode x regardless of his beliefs. There are several interpretations of this assumption. One is that x is some piece of "hard" information that is recorded and available to the agent whether or not he attends to it, whereas z is "soft" information that the agent needs to attend to in order to learn from. For example, x could include information on revenues in a company's earnings report, while z could include information on how management discusses this information in an earnings conference call with analysts. An alternative interpretation of this assumption is that it captures 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 (Bargh 1992). For example, there is much evidence that people instantly attend to and categorize others on the basis of age, gender, and race (Fiske 1993). While, under this interpretation, what makes some event features more automatically encoded than others lies outside the scope of the formal analysis, it is reasonable to expect that event features which are useful to making predictions and arriving at utility-maximizing decisions in many contexts are likely to attract attention, even when they may not be useful in the context under consideration. For example, gender may be salient in economic interactions because considering gender is useful in social interactions. Consistent with this idea of a spillover effect, the amount of effort required to process and encode information along a stimulus dimension decreases with practice (Bargh and Thein 1985).

Second, the formalization of selective attention (equation (4)) has the simplifying feature that whether the agent encodes z depends on his period-k belief about whether it is predictive but not his assessment of by how much. I conjecture that my qualitative results for the discrete attention case would continue to hold if I was to relax this assumption. Intuitively, the only real change would be that the agent could not persistently encode z if z is not *sufficiently* predictive, expanding the circumstances under which the agent's limiting forecasts and beliefs would be biased.

As a final note on the model's setup, it is clear that the model nests the standard Bayesian one as a special case: when the selectively attentive agent is never at all cognitively busy ($b_k \equiv 0$), then, each period, his forecasts and beliefs coincide with the standard Bayesian's.

3. Discrete Attention

To build intuition for the implications of the model, I first analyze the selective attention learning process for the discrete attention case—that is, where b_k is deterministic and the agent attends to z with probability 0 or 1 in a given period.

3.1. Long-Run Attention

The first result is that the agent eventually settles on how he mentally represents events, or, equivalently, on whether he does or does not encode z.

DEFINITION 2. The agent settles on encoding z if there exists some $\tilde{t}$ such that $e_k = 1$ for all $k \geq \tilde{t}$. The agent settles on not encoding z if there exists some $\tilde{t}$ such that $e_k = 0$ for all $k \geq \tilde{t}$.

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

To sketch the argument behind Proposition 1, suppose that, with positive probability, the agent does not settle on encoding or not encoding z, and condition on

this event. 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, and so on. It also 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 $\pi_Z \to 1$ (or $b \to 0$) the probability that the agent settles on encoding z tends towards* 1,
2. *when $\pi_Z < b$ the probability that the agent settles on not encoding z equals* 1.

The intuition behind Proposition 2 is the following. As $\pi_Z \to 1$ or $b \to 0$, the "likelihood ratio" relating the likelihood of the recalled history $\hat{h}^t$ under models where z is important to the likelihood of that history under models where z is unimportant would have to get smaller and smaller to bring $\hat{\pi}_Z^t$ below b. But the probability that this likelihood never drops below some cutoff λ tends towards one as λ approaches zero. In the other direction, when $\pi_Z < b$ the agent starts off not encoding z. In this case, he never updates his belief about whether z is important to predicting y and settles on not encoding z since, by treating $\varnothing$ as he would a distinct nonmissing value of z (the naiveté assumption), he forms beliefs as if there had 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, as fleshed out in Online Appendix B, this argument relies on the naiveté 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 unattended-to variable (z) that influences the success probability, though this mechanism is limited in the sense that even a sophisticate will never attend to z if his prior belief in the importance of z is *sufficiently* close to 0.[9]

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 predictive

9. The fact that even a sophisticate *may* not learn to attend to z from observing variation in y conditional on x is important to emphasize: while a sophisticate (unlike a naif) places some weight on this variation coming from variation in z, she also places some weight on it coming from natural variation resulting from the conditional (on x) success probability being bounded away from 0 or 1. As Online Appendix B makes clear, the latter possibility limits the degree to which she can update her beliefs about the importance of z when she never attends to it.

factor. This is consistent with evidence presented by Nisbett and Ross (1980, Chapter 5) who note that the likelihood that a relationship is detected is increasing in the extent to which prior "theories" put such a relationship on the radar screen.

Proposition 2 also illustrates how the degree to which an agent is cognitively busy (the level of b) when learning to predict an outcome 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; Gilbert, Pelham, and Krull 1988).

Finally, since the comparative statics in Proposition 2 hold uniformly in $\boldsymbol{\theta_0}$, we see that a prior belief that z is unlikely to be important to prediction can be *self-confirming,* even when z is *very* predictive. While the continuous attention version of the model will slightly qualify this conclusion, the model highlights how selective attention can lead the agent to persistently miss big empirical regularities.

3.2. Long-Run Forecasts and Beliefs

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 section, I characterize the potential long-run forecasts, and then go on to characterize the potential long-run beliefs over models of which variables are important.

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

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

For the intuition behind Proposition 3, if the agent settles on encoding z, then, from some period on, he finely represents each period's outcome as (y, x, z). On the other hand, if he settles on not encoding z, then, from some period on, he incompletely represents each period's outcome as $(y, x, \varnothing)$. Either way, his asymptotic forecasts will be consistent with the true probability distribution over outcomes as he represents them (his effective observations).

Together with Proposition 1, Proposition 3 implies that forecasts converge and there is structure to any limiting biased forecasts: such forecasts can persist only if they are consistent with the true probability distribution over (y, x). This observation will be important in considering how the model can help explain *systematically* biased beliefs (as well as stereotypes).

Next, consider the agent's long-run beliefs over models of which variables are important.

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 he does not learn the true model: specifically, $\hat{\pi}^t_X \overset{a.s.}{\to} 1$ and, for large t, $\hat{\pi}^t_Z \leq b$.*

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. The second part says that when the agent settles on not encoding z, then, almost surely, he eventually places negligible weight on models where x is unimportant to predicting y because the *unconditional* success probability depends on x (recall Assumption 2). On the other hand, the limiting behavior of $\hat{\pi}^t_Z$ is largely unrestricted because he effectively does not observe any variation in z. Although the agent "knows" that he sometimes cannot recall z and does not have access to all data, he still becomes convinced that x is important to predicting y. This is because, by *treating* $\varnothing$ as a nonmissing value of z (the naiveté assumption), he believes he has access to all *relevant* data necessary to determine whether x is important to prediction. Put differently, the agent can identify $E_{\boldsymbol{\theta}_0}[y|x] - E_{\boldsymbol{\theta}_0}[y|x']$ for all x, x', which he considers the same as being able to identify $E_{\boldsymbol{\theta}_0}[y|x, z'] - E_{\boldsymbol{\theta}_0}[y|x', z']$ for all x, x' and any $z' \neq \varnothing$. This result, which relies on the naiveté assumption (see Online Appendix B), can be interpreted as saying that the agent sometimes acts as if he believes that correlation implies cause. However, since the agent only converges on this belief when he settles on not encoding z, the model makes predictions about when the agent will in fact make such an error: when he persistently fails to attend to a causal factor.

3.3. Systematically Biased Stereotypes and Beliefs

An application of the results so far is that selective attention may lead people to form persistent and systematically incorrect beliefs about what causes variation in the data. To see this, consider the following stylized example on stereotyping. (For related experimental evidence, see Schaller and O'Brien 1992.) Suppose an agent repeatedly faces the task of predicting whether individuals will act friendly in conversation, $y \in \{0, 1\}$, conditional on information about a given person's group membership, $x \in \{A, B\}$, and whether the conversation will take place at a work or recreational situation, $z \in \{\text{Work}, \text{Play}\}$. (If it helps, group membership can be thought of as male/female, in-group/out-group, black/white, student/professor, etc.) The agent's encounters with group B members are relatively confined to work situations: $g(\text{Work}|B) > g(\text{Work}|A)$. Independent of group membership, every individual is always friendly during recreation but never at work (the situation completely determines behavior):

$$E_{\boldsymbol{\theta}_0}[y|A, \text{Play}] = E_{\boldsymbol{\theta}_0}[y|B, \text{Play}] = 1,$$
$$E_{\boldsymbol{\theta}_0}[y|A, \text{Work}] = E_{\boldsymbol{\theta}_0}[y|B, \text{Work}] = 0.$$

TABLE 2. Likelihood that agent interacts with member of group $x \in \{A, B\}$ in situation $z \in$ {Work, Play}.

	A	B
Work	0.25	0.25
Play	0.4	0.1

Suppose, however, that the selectively attentive agent (incorrectly) starts off believing that situational factors are unlikely to matter ($\pi_Z < b$), so Proposition 2 implies that he settles on not attending to such factors. The agent will consequently mistakenly come to believe that group-B members are less friendly than group-A members because he tends to encounter them in situations that discourage friendliness. To illustrate, Proposition 3 implies that when the likelihood of encountering different (x, z), $g(x, z)$, is given as in Table 2, the agent comes to misforecast

$$\lim_{t\to\infty} \hat{E}[y|A, \text{Situation}, \hat{h}^t] = \tfrac{0.4}{0.4+0.25} = 0.62,$$

$$\lim_{t\to\infty} \hat{E}[y|B, \text{Situation}, \hat{h}^t] = \tfrac{0.1}{0.1+0.25} = 0.29$$

across situations. He will thus overreact to group membership.[10] Moreover, Proposition 4 tells us that the agent will become overconfident in having identified a relationship between group membership and friendliness even though he "knows" that he sometimes does not attend to situational factors. Again, the reason is that, by the naiveté assumption, he treats the mentally represented history as if it were complete. In particular, he mistakenly treats observed variation in (Friendliness, Group|Real-World Interaction) as being equally informative as observed variation in (Friendliness, Group|Work) or (Friendliness, Group|Play) in identifying a causal effect of group membership on friendliness: he comes to believe that group identity truly matters—that is, that it is more than a proxy for selectively unattended-to predictors.

To take another example, consider the case of childbed (or puerperal) fever. In the mid-19th century, this was the leading cause of maternal deaths in hospitals (Gawande 2004). The cause: a failure of doctors to wash their hands before coming in contact with the mothers (Nuland 2004; Gawande 2004). Rather than focusing on this explanation, there were other hypotheses for what caused mothers to become ill that stemmed from the popular "miasmic" or "bad air" theory of disease, namely that childbed fever was caused by mothers inhaling foul air identified by bad smells (Halliday 2001; Nuland 2004). The model helps understand this example. Doctors did not have a compelling theory for why handwashing would matter (the germ theory of disease had not been discovered), but did have a theory that bad smells could matter (the miasmic theory of disease was popular). As a result, they did not attend much to cleanliness when

10. In this example, selective attention results in a persistent bias related to the Fundamental Attribution Error (Ross 1977; Gilbert and Malone 1995).

attempting to uncover what caused mothers to get sick but did attend to the presence of bad smells, which facilitated the persistence of incorrect theories that foul air was to blame in the case of childbed fever. Indeed, these incorrect theories appear to have had explanatory power when factors like handwashing were not controlled for: women delivered by midwives at home were sixty times less likely to die of childbed fever than were women delivered by male doctors at the (poorly smelling) hospital (Levitt and Dubner 2009).

The intuition behind such examples is that a failure to learn to pay attention to a variable endogenously 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.[11] Online Appendix A.2 presents formal results along these lines and, in particular, Proposition A.1 adapts results from the statistics literature (Samuels 1993), to develop a formula that relates the magnitude of the resulting bias to features of the joint distribution over (y, x, z). Such a formula highlights that when false beliefs stem from selective attention, they are *systematically* biased: the model both makes predictions about false beliefs that can persist, as well as false beliefs that cannot. In the stereotyping example, a false belief that cannot persist is that group-B members are almost never friendly, since such a forecast is inconsistent with any coarse representation of outcomes. On the other hand, a false belief that group-B members are friendly only around 30% of the time during recreation *can* persist because such a prediction is consistent with actual outcomes as averaged across work and recreation. Similarly, the idea that foul air was to blame in causing childbed fever may have persisted over other (wrong) theories because of its apparent explanatory power when factors like handwashing were not attended to.

A further insight is that false beliefs resulting from selective attention are robust in that even if an agent can credibly communicate the importance of a variable that the other has selectively failed to notice, then, following such "debiasing", it will still take the agent time to learn to incorporate information about that variable in making predictions, since he did not keep track of it before, and to mitigate his misreaction to associated variables. To illustrate, even if someone later brings up the potential importance of the situation in driving friendliness, the agent will not immediately be able to recognize that it is what truly drives behavior because he did not previously encode the relationship between friendliness and the situation. (This idea is fleshed out more formally in Online Appendix A.2.) Returning to the example of childbed fever, the belief that bad smells were to blame persisted long after Ignac Semmelweis uncovered the relationship between hand-washing and maternal deaths. Rather than leading doctors to have an "a-ha" moment (i.e., "it's true that mothers have been less likely to get sick when I had clean hands prior to coming in contact with them"), they were largely dismissive for many years (Nuland 2004). While presumably doctors were exposed to a reasonable amount of variation in the cleanliness of their hands or

11. These results relate to experimental findings, as described in Online Appendix A.2, 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).

others' hands prior to treatment, absent the germ theory of disease, selective attention may have led them to filter out the relevant data when forming beliefs.

To better understand the robustness of incorrect beliefs resulting from selective attention, I turn to analyzing a more "continuous" notion of attention.

4. Continuous Attention

So far, I have made the assumption that the agent never attends to z when he places little weight on models which specify z as being important to prediction. Perhaps, instead, the agent attends to z with a probability that varies more continuously in the likelihood he attaches to such processing being decision relevant (Kahneman 1973). I model this by assuming that there are random fluctuations in the degree to which the agent is cognitively busy in a given period.[12] Then, the likelihood that the agent attends to z will naturally vary in the intensity of his belief that z is important to prediction.

Formally, let $\eta(\hat{\pi}_Z^k) \equiv \text{Prob}[e_k = 1|\hat{\pi}_Z^k] = \text{Prob}[b_k < \hat{\pi}_Z^k]$ 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 each b_k is independently drawn according to some fixed cumulative distribution function F with support on $[0, 1]$. F is assumed to have an associated density function f that is continuous and strictly positive on $[0, 1]$. We say that *the continuous attention assumptions hold* whenever the b_k are drawn in this manner. To take an example, the continuous attention assumptions hold if $b_k \overset{i.i.d.}{\sim} \text{U}[0, 1]$. In this case, the likelihood that the agent attends to z as a function of $\hat{\pi}_Z^k$ is given by $\eta(\hat{\pi}_Z^k) = \hat{\pi}_Z^k$ for all $0 \leq \hat{\pi}_Z^k \leq 1$.

Under the continuous attention assumptions, the agent always attends to z with positive probability and almost surely encodes z an infinite number of times. In Online Appendix A.3, I show how this implies that, no matter the agent's initial beliefs or the degree to which he initially attends to z, he will eventually 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. Nevertheless, he may continue not to attend to z and to make biased forecasts for a long time. Online Appendix A.3 also establishes that a partial analog to Proposition 2 is true when the continuous attention assumptions hold: for all $t \geq 2$, the probability that the agent never encodes z before period t tends towards 1 as $\pi_Z \to 0$. In other words, the agent's ability to recognize empirical relationships within a reasonable time horizon still depends on his prior.

The main benefit of considering the continuous attention assumptions is that it allows us to consider which features of the joint distribution over observables influence

12. 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.

whether we should expect the agent to begin attending to a predictor with high probability within a reasonable time horizon. To this end, I consider the rate at which the likelihood that the agent 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 similar but messier results for the general case.

Intuitively, we say that some random variable $\mathcal{X}_t$ converges to random variable $\mathcal{X}_0$ at some rate $V(t)$, where $V(t)$ tends towards 0 as $t \to \infty$, if $|\mathcal{X}_t - \mathcal{X}_0|$ approaches zero like $V(t)$ does. To develop analytic results, we focus on the asymptotic rate of convergence.

DEFINITION 3. The asymptotic rate of convergence of a random variable $\mathcal{X}_t$ to $\mathcal{X}_0$ is $V(t)$ if there exists a strictly positive constant $C < \infty$ such that

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

The rate at which the agent learns to attend to z depends on the degree to which he has difficulty explaining observations without taking z into account. Put the other way around, the agent may continue not attending to z for a long time if he can accurately approximate the true distribution when he only takes x into account. Formally, define the *relative entropy distance*, d, between $p_{\boldsymbol{\theta}_0}(y|x,z)$ and $p_{\boldsymbol{\theta}_0}(y|x)$ as the average of the relative entropies between these distributions, where this average is taken over the probability mass function $g(x,z)$:

$$d = \sum_{y,x,z} p_{\boldsymbol{\theta}_0}(y|x,z) g(x,z) \log \left(\frac{p_{\boldsymbol{\theta}_0}(y|x,z)}{p_{\boldsymbol{\theta}_0}(y|x)} \right). \tag{6}$$

PROPOSITION 5. *Suppose the continuous attention assumptions hold and either (i) $\pi_X = 1$ or (ii) x is important to predicting y. Then $\eta(\hat{\pi}_Z^t) \to 1$ almost surely with an asymptotic rate of convergence $e^{-d(t-1)}$, where d is the relative entropy distance between $p_{\boldsymbol{\theta}_0}(y|x,z)$ and $p_{\boldsymbol{\theta}_0}(y|x)$, defined as in condition* (6).

To briefly sketch the arguments involved in proving Proposition 5, an initial observation is that the rate at which $\eta(\hat{\pi}_Z^t) \to 1$ is determined by the rate at which $\rho(\hat{h}^t) \equiv \Pr(\hat{h}^t|M_{X,\neg Z}) / \Pr(\hat{h}^t|M_{X,Z}) \to 0$. The problem is then to show that

$$\frac{1}{t-1} \log(\rho(\hat{h}^t)) \overset{a.s.}{\to} -d, \tag{7}$$

which is demonstrated via two intermediate results. First, the rate of convergence of $\rho(\hat{h}^t) \to 0$ is the same as that of $\rho(h^t) \to 0$, which is largely established using stochastic approximation techniques. Second, the rate of convergence of $\rho(h^t) \to 0$ is the same as that of

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

which follows from a recent result of Walker (2004). (Lemma D2 in Online Appendix D shows how this result applies to the current setting.) From these two facts, (7) reduces to showing that

$$\frac{1}{t-1}\log\left(\frac{\Pr(h^t|\theta(x,z) = p_{\boldsymbol{\theta_0}}(y=1|x) \text{ for all } x,z)}{\Pr(h^t|\boldsymbol{\theta_0})}\right) \overset{a.s.}{\to} -d,$$

which is a consequence of the strong law of large numbers.

One implication of Proposition 5 is that the same features that contribute to greater bias can make the bias more persistent. Return to the stylized stereotyping example and continue to suppose that an individual is always friendly during recreation but never at work. It is 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 $H(z) = -\sum_z g(z)\log(g(z))$ is the entropy of $z = \textit{Situation}$, or a measure of the degree to which the agent splits his time between work and recreation, and $I(z;x) = \sum_{x,z} g(x,z)\log\left(g(x,z)/(g(x)g(z))\right)$ is the mutual information between $z = \textit{Situation}$ and $x = \textit{Group}$, which is a measure of the degree to which knowledge of group membership provides the agent with information regarding whether he is likely to encounter the individual during work or recreation.

Thus, in this example, fixing the degree to which the agent splits his time between work and recreation (i.e., fixing $H(z)$), the rate at which he will learn to attend to situational factors is *decreasing* in the degree of association between group and situational factors (decreasing in $I(z;x)$). Combining this fact with the earlier analysis suggests that an agent who has an even greater tendency to encounter group B members more often during work than recreation both has the potential to overreact to group membership to a greater extent and is less likely to begin attending to situational factors within a reasonable time horizon. This example highlights that the extent to which the agent's reaction may be biased by failing to attend to z, which depends on the degree of "omitted variable bias", may be *negatively* related to the speed at which the agent learns to attend to z, which depends on the quality of feedback available to the agent when he encodes z.

Proposition 5 also helps understand why certain incorrect beliefs are *so* persistent: they have significant explanatory power (formally, low d). While the presence of bad smells did not cause childbed fever, bad smells were likely highly associated with hospital delivery by doctors with unclean hands rather than home delivery by midwives

with relatively clean hands. In such an example, even if someone occasionally attends to the true causal factor, it takes many observations to recognize that a theory involving that factor does better at explaining the data than the prevailing belief.

5. Other Applications

This section discusses some applications of the analysis, where I begin by fleshing out the stereotyping example used to illustrate the model. I then go on to discuss a more elaborate set of applications.

Stereotyping and Discrimination. One observation is that selective attention can lead to discriminatory behavior which by some measures *appears* to reflect prejudice or a taste for interacting with members of certain groups (Becker 1971), as opposed to statistical discrimination as typically conceptualized (e.g., Phelps 1972; Arrow 1973).[13] Returning to the stylized stereotyping example in the discrete attention case, suppose that after forming beliefs about the relationship between friendliness, group membership, and the situation, an employer chooses who to hire for a job where friendliness matters—for example, the position of a sales clerk—and for simplicity assume this is all that matters. Given the previous assumptions, even when the situation completely determines behavior, the selectively attentive employer who persistently fails to attend to the situation forms beliefs that group-A members are friendlier in every situation. As a result, he will strictly prefer to hire members of group A, and the employer will discriminate against members of group B. Moreover, an analyst who observes this discrimination and can observe how friendly members of each group would be on the job will conclude that it cannot reflect rational statistical discrimination since members of both groups would perform equally well. Rather, she may infer that the employer acts as if he receives particular disutility from hiring group-B members.

Despite this apparent similarity, selectively attentive discrimination has distinct implications. First, because the environment shapes expectations, the quality of intergroup contact matters. An employer who has less of a tendency to encounter group-B members in situations that discourage friendliness relative to members of group A will display less discrimination, both because he will react to group membership to a lesser extent (Proposition 3 and Online Appendix Proposition A.1) and, under the continuous attention assumptions, because he will more quickly learn to attend to situational factors in predicting friendliness (Proposition 5).[14] Second, since the way the agent encodes information as he forms beliefs is important, the model predicts that

13. Recall that taste-based models of discrimination (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.

14. Some economic models emphasize how the *frequency* of interactions with group members can influence the extent of discrimination (Fryer and Jackson 2008; Glaeser 2005), but to the best of my

interventions that can guide attention in this stage through influencing the prior may attenuate stereotyping and discrimination later on. Such interventions need not try to de-emphasize the importance of group identity, but rather to highlight the importance of associated causal variables. In the example, an intervention that gets eventual employers to attend to the power of the situation (i.e., by increasing π_Z) can counteract stereotyping and discrimination down the road.

Understanding Learning Failures. The model also sheds light on why even very experienced agents may be far from the productivity frontier. To take a simple example, Bloom et al. (2013) find that many firms fail to adopt the recommended practice of maintaining the shop floor clear of waste and obstacles. This is puzzling from the perspective of standard learning models in which the key input is readily available data (e.g., Besley and Case 1993, 1994; Jovanovic and Nyarko 1996; Foster and Rosenzweig 1995), since there presumably exists a lot of natural variation in the cleanliness of the factory floor that firms could have used to estimate the relationship between cleanliness and productivity. It follows naturally, however, from the model of selective attention, under the assumption that many managers would not have had a strong prior reason to believe in the importance of this relationship, and consequently would not have attended to cleanliness in forming beliefs about the determinants of output quality.[15]

This example indicates how selectively attentive agents may end up taking suboptimal actions relative to if they attended to the right variables when forming beliefs (or had rational expectations). More substantively, by endogenizing this process, the model allows for comparative statics on what predicts or influences such failures. First, as we saw earlier, prior beliefs or good theories on which variables are important matter, as is the degree of cognitive busyness or the (shadow) cost of devoting attention (Proposition 2). Second, because the model emphasizes the interplay between attention and memory—$\hat{z} = \varnothing$ when the agent does not attend—it is possible to measure which variables the agent attends to through measuring his recall, and consequently to predict the relationships he will take into account when making decisions. For example, the model predicts that a manager who can more accurately recall precisely when the factory floor has been clean in the past, or, more specifically, how being clean has historically covaried with output quality, is more likely to be optimizing along this dimension. Third, an analyst may be able to detect learning failures just by examining

knowledge economic models have largely neglected how the *quality* of that interaction can matter, though this is emphasized in some of the psychology literature (Allport 1954; Pettigrew 1998; Pettigrew and Tropp 2006).

15. To see this, let y denote the quality of the output ($y = 1$ represents high quality) and $z \in \{\text{clean, not clean}\}$, where a manager has many observations of (y, z) prior to making some decision that influences the likelihood that the floor will be clean, and $g(z') > 0$ for $z' = \text{clean, not clean}$. (For simplicity, we are ignoring x.) Propositions 2 and 3 imply that even if the data reveal that it is worthwhile to keep the factory floor clean, since $E_{\theta_0}[y|\text{clean}] > E_{\theta_0}[y|\text{not clean}]$, the manager may fail to recognize this fact when his prior puts sufficiently small weight on the cleanliness of the factory floor being important to the quality of the firm's output, and consequently could make a suboptimal decision on how much effort to devote to keeping the floor clean.

data the agent has available to him, as well as to influence the agent's behavior by highlighting unattended-to relationships in those data. Returning once again to the management example, an outside observer, for example a consultant, may be able to observe the factory operations and recognize—using *the same* data that the manager has available to him—that the manager is leaving money on the table by not devoting effort to keeping the factory floor clean. Further, by presenting the manager with summary information about the relationship between the firm's output quality and the cleanliness of the factory floor, the observer may be able to influence the manager's beliefs and decisions.

Building from the theoretical framework in this paper, Hanna, Mullainathan, and Schwartzstein (2014) explore the latter two predictions in greater detail in a model of technology adoption and use, and find empirical support in the context of a field experiment with seaweed farmers that tests these and other predictions.[16]

Disagreement. Suppose there are two individuals $i = 1, 2$ who separately form beliefs about the relationship between y, x, and z. They have access to the same data when forming beliefs, but $i = 1$ begins with a stronger belief that z is important to predicting y than $i = 2$: $\pi_Z^1 > \pi_Z^2$, where π_Z^i denotes individual i's prior belief that z is important. Supposing the individuals are selectively attentive and the process by which they form beliefs is as described by the discrete attention version of the model, then Proposition 2 suggests that $i = 1$ is more likely to learn that z is important than $i = 2$, while Proposition 3 implies that their limiting reactions to new information can differ. (We would reach a similar conclusion by supposing that $i = 1$ is less cognitively busy than $i = 2$.) In this manner, the model may help understand why people may persistently react differently to the same information, in contrast to more standard models which famously predict that agreement should be the norm (e.g., Savage 1954; Blackwell and Dubins 1962), at least given sufficiently rich data to learn from.[17]

To take an example, consider Malmendier and Shanthikumar's (2007) finding that small investors take security analysts' stock recommendations more literally than large investors. Affiliated analysts—that is, those belonging to banks that have an

16. Briefly summarizing the experimental findings presented in Hanna et al. (2014), the survey data indicate that the seaweed farmers do not attend to pod size, a particular input dimension, as a vast majority of farmers did not know what size they use and would not hazard a guess about the optimal size. Further, at baseline, a given farmer used a wide variety of sizes across pods at a given point of time, suggesting they had the data to estimate the relationship between size and yields (at least within the support of those already used). Consistent with the prediction that agents are less likely to optimize dimensions they reveal they do not pay attention to, experimental trials suggest that farmers are particularly far from optimizing pod size, even within the support of sizes they use at baseline. Consistent with the prediction that highlighting unattended-to relationships in the data can induce learning, presenting farmers with information about the relative performance of different sizes on their plots impacted their behavior.

17. Recent models by Andreoni and Mylovanov (2012) and Acemoglu, Chernozhukov, and Yildiz (2009) show how heterogeneous prior beliefs or private information can lead to persistent disagreement when, even after infinite observations, public information is insufficiently rich to identify the true underlying state of nature. In contrast, my model shows how disagreement can persist even after people observe rich enough data to identify this state.

underwriting relationship to firms they are reporting on—tend to issue more favorable recommendations than unaffiliated analysts. Large investors (e.g., pension funds) relatively discount the recommendations of affiliated analysts; small investors (e.g., individual investors), on the other hand, do not. This pattern of results is difficult to solely explain in a standard cost of information gathering framework, as small investors do not react more to independent analysts' recommendations—that is, those never involved in underwriting—even though members of this group often advertise their independence. However, the model of selective attention provides a natural possible explanation.[18] By virtue of being relatively busy thinking about other things and having less precise knowledge about analysts' incentives, it is relatively unlikely that small investors—who correspond to $i = 2$ in the previous notation—will learn to attend to analyst affiliation or that affiliated analysts' recommendations should be relatively discounted. Instead, in the limit, Proposition 3 implies that such investors may respond the same way to affiliated and unaffiliated analysts' recommendations, while large investors will react more to the recommendations of unaffiliated analysts.

Extensions of the broader analysis could consider the degree to which disagreement persists when individuals can communicate in some manner, for example if they can observe each other's forecasts. While such an extension lies outside the scope of the formal analysis, one observation is that, even in this situation, disagreement is likely to persist for some time for reasons similar to those laid out in the discussion of robustness in Section 3.3: even if observing someone else's forecasts leads an agent to realize that he is not attending to an important variable, he will then need to start attending to the relationship between the variable and an outcome of interest for some time before he learns how they are in fact related (perhaps with the help of observing the other agent's forecasts over time). This process may take even longer when there is more than one z variable, as in this case the agent does not automatically know which variable to attend to if he learns that he previously failed to attend to some important variables.

The Process of Discovery. Since the model endogenizes which variables agents attend to—rather than take this as a given—it may shed light on the process of discovery. In particular, the model points to the importance of improved theories of which variables matter: some shock that gets people to start attending to an important variable (e.g., a new theory or study that influences π_Z) can initiate a period where people start gathering more and more information about how that variable relates to an outcome of interest and to progressively more accurate forecasts and efficient decisions. The earlier discussion on how the germ theory of disease was important for getting doctors to recognize the importance of practices like washing their hands (even though they

18. In the notation of the model, we could think of $y \in \{0, 1\}$ as the quality of an investment ($y = 1$ is better than $y = 0$), $x \in \{\text{Buy, Hold, Sell}\}$ as an analyst's recommendation about the investment, and $z \in \{\text{Affiliated, Unaffiliated}\}$ as the analyst's affiliation, where, for example, the true relationship between these variables could have the property that $E_{\theta_0}[y|x, \text{ Affiliated}]$ is independent of the recommendation, x, while $E_{\theta_0}[y|\text{Buy, Unaffiliated}] > E_{\theta_0}[y|\text{Hold, Unaffiliated}] > E_{\theta_0}[y|\text{Sell, Unaffiliated}]$.

were long aware of the *possibility* that practices like hand washing could matter) is one such example.

Likewise, the model suggests a role for studies or reports that document previously unattended-to relationships. In the context of cross-sectional asset pricing, Nagel (2012) argues that the model of selective attention may help explain both the abnormal return predictability associated with certain predictors, since investors may persistently fail to attend to such predictors, and evidence that this predictability can go down after it is publicized in academic studies (e.g., McClean and Pontiff 2012). In effect, studies may help investors fill in some of the gaps in $\hat{h}^t$ (or allow them to better estimate the importance of some z variables, despite such gaps).

In addition to making predictions about how improved theories of which variables matter lead to new discoveries, the continuous attention formulation of the model makes predictions on when we should expect theories involving more variables to gain prominence: when they add significant explanatory power given the underlying environment. One potentially testable implication is that changes in the environment—shifts in $\boldsymbol{\theta_0}$ or $g(x, z)$—can predictably alter the speed with which agents learn that it is worthwhile to attend to z, through influencing the relative entropy distance between $p_{\boldsymbol{\theta_0}}(y|x, z)$ and $p_{\boldsymbol{\theta_0}}(y|x)$, and thereby the degree to which agents have difficulty explaining what they observe without taking z into account.

6. Relationship to Existing Literature and Alternative Approaches

There is a large and growing literature that aims to draw out the implications of inattention in economic settings. Many of the models in the literature start from the premise that people are "rationally inattentive" and optimally allocate their limited attention given the underlying stochastic environment (Sims 2003, 2006; Peng and Xiong 2006; Mackowiak and Wiederholt 2009; Gabaix 2013; Woodford 2009, 2012a, 2012b). In particular, these models have been closed by assuming that people have rational expectations about what information is worth attending to. My model emphasizes that assumptions of rational expectations are particularly hard to justify when agents are inattentive: inattention itself limits agents' ability to learn what *is* worth attending to, and people may persistently fail to attend to very important variables. My model can be viewed as complementing existing models of rational inattention, as it addresses questions that those models (by nature) have been silent on, such as the process by which people can learn to attend to the right variables, and how their forecasts and beliefs may be persistently biased when they do not.[19]

19. Less closely related are recent models in which agents initially attend to all features of the environment, but then, ex post, disproportionately focus on certain salient features when making decisions (Bordalo, Gennaioli, and Shleifer 2012; Bordalo, Gennaioli, and Shleifer 2013; Koszegi and Szeidl 2013), as well as economic models of "motivated learning" (e.g., Compte and Postlewaite 2004; Gottlieb 2010; Karlsson, Loewenstein, and Seppi 2009), which assume preferences over beliefs.

The logic of the model is similar to that highlighted in the literature on bandit problems (e.g., Gittins 1979) and self-confirming equilibrium (e.g., Fudenberg and Levine 1993), which show that it is possible for individuals to maintain incorrect beliefs about the payoff consequences of actions that have rarely been tried and for these beliefs, in turn, to support suboptimal actions. A key distinction is that beliefs are consistent with available data in such models—the constraint is data collection—while beliefs are only consistent with data as encoded in mine—the constraint is data processing. As a result, my model helps understand why agents can have incorrect beliefs that persist in the face of contradictory data.

In this manner, my model is also related to Rabin and Schrag's (1999) model of confirmatory bias, or the tendency of individuals to misinterpret new information as supporting previously held hypotheses, 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. (See also Wilson 2003.) The predictions of my model are sharper since the logic of confirmatory bias does not by itself pin down which incorrect beliefs we can expect to persist. For example, if an individual begins with a belief that members of some group 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 group members) and become more and more convinced in this incorrect hypothesis. However, under the model of selective attention, such an incorrect belief cannot persist because evidence is filtered at the level of models of *which* factors influence an outcome and not at the level of hypotheses about *how* those factors influence an outcome. 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. This feature of the model drives why incorrect stereotypes and beliefs resulting from selective attention are *systematic*, yielding predictions on how they should vary as a function of the joint distribution over observables (Proposition 3 and Online Appendix Proposition A.1). For example, as discussed previously, the model predicts that a person will react less to group membership in a fixed situation when forecasting friendliness (and discriminate less) if she has less of a relative tendency to encounter members of certain groups in situations that discourage friendliness.

This last feature also connects the model to the literature on coarse thinking. Because the selectively attentive agent's limiting forecasts are consistent with outcomes as she represents them, her limiting forecasts when she settles on not encoding z are mathematically equivalent to those of a coarse thinker who groups all values of z—for example, situational factors—together into the same category (or "analogy class") and applies the same model of inference across members of that category (Mullainathan 2000; Jehiel 2005; Mullainathan, Schwartzstein and Shleifer 2008). Rather than take coarse thinking as given as in much of the previous literature (e.g., Eyster and Rabin 2005; Ettinger and Jehiel 2010; Fryer and Jackson 2008; Esponda 2008, Esponda and Pouzo 2012), I endogenize it as a potential limiting outcome (or approximate outcome

over a reasonable time horizon) of a learning process given cognitive frictions.[20] Under this interpretation, my model then implies further comparative statics predictions about when coarse categorizations emerge and persist, for example when the agent's prior says that an important variable is unlikely to be predictive and when, on average, the agent can reasonably approximate what he observes without taking that variable into account.[21]

Similarly, the selectively attentive agent's limiting forecasts also coincide with those of an agent who may not be able attend to all available information when making a prediction, but can nevertheless recall such information if necessary later on (e.g., Hong, Stein and Yu's 2007 model of paradigm shifts). A key difference is that, unlike models such as Hong, Stein and Yu's, which may be a better description of some situations where past information (e.g., about firm earnings) is freely available in public records and tends to be revisited, a selectively attentive agent will not have a "eureka" moment if someone convincingly brings up the potential importance of an unattended-to variable; rather, following the discussion of robustness in Section 3.3 he will subsequently need to learn how to interpret data along previously unattended-to dimensions. My model predicts that new (accepted) theories of which variables matter can *gradually* lead to the recognition of relationships that in principle could have been discovered using previously available data.[22]

As a final note, I would like to compare the model with closely related approaches I could have taken. An alternative formalization of how the agent represents information when he does not attend combined with naiveté would hold that when the agent does not attend to z, then he fills z in randomly—for example, $\hat{z}$ is drawn from a uniform distribution over Z—but he updates as if he has perfect recall. That is, rather than incompletely representing his experiences, the agent could fill in missing details and

20. Some of these papers complement mine by highlighting specific z that people may not take into account (Eyster and Rabin 2005; Esponda 2008; Esponda and Pouzo 2012). For example, Esponda (2008) draws out the implications of people failing to take selection into account when forming beliefs about the best action to take in certain adverse selection settings, whereby they are assumed to persistently neglect the correlation between actions of other players (e.g., the asking price) and payoff-relevant uncertainty (e.g., the value of an object).

21. Models of coarse thinking are often explicitly or implicitly justified as limit points of incomplete learning processes (e.g., Jehiel 2005, p. 88). My model sheds light on the validity of this justification. Other approaches to endogenize coarse thinking include Al-Najjar and Pai (2013), who envision such thinking as arising from an attempt to avoid overfitting data. In contrast to Al-Najjar and Pai, my model allows for coarse thinking to persist even with unlimited data. I do not view my model as providing a complete "theoretical foundation" for coarse thinking: learning-based approaches, including my model, have difficulty capturing certain compelling examples which reflect the use of categories or associations that, on some level, people *know* are not appropriate for a task at hand. For example, people appear to be overly sensitive to the month in predicting the temperature, whereby they overpredict the difference in temperature between February 24 and March 4 and underpredict that between March 4 and March 14 (Krueger and Clement 1994). I suspect this does not stem from a mistaken belief that, fixing the month, the day is not very important in predicting the temperature, but from a more basic cognitive operation involving the use of categories.

22. This feature of the model also distinguishes it from Aragones et al. (2005) which posits that people may not learn empirical relationships because discovering regularities in existing knowledge is computationally complex.

remember distorted versions. Such a model should also capture the intuition that an agent is less likely to learn the importance of variables that he does not attend to, since, when $\hat{z}$ is drawn at random from Z (and independently of y, x), then there will not be a systematic relationship between $\hat{z}$ and y. However, the predictions of such a model would in some ways be more extreme: the agent would become more and more confident in the belief that z is unimportant the longer he does not attend to z, while my model predicts that the agent's beliefs in the importance of z can be flat when he does not attend to z. Another formalization would hold that when the agent starts off with a sufficiently strong belief that z is not important, then, in addition to not attending to z, he subsequently updates his beliefs as if the z dimension *does not even exist*. The most straightforward implementations of this assumption would have difficulty allowing an agent who starts off not attending to z to ever learn about its importance, as he can under the continuous attention assumptions. Modeling the agent as entertaining the possibility that z matters, even if he does not attend, more readily accommodates such possibilities.[23]

7. Conclusion

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

References

Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz (2009). "Fragility of Asymptotic Agreement under Bayesian Learning." Working paper, MIT.

Al-Najjar, Nabil and Mallesh Pai (2013). "Coarse Decision Making and Overfitting." Working paper, University of Pennsylvania.

Allport, Gordon W. (1954). *The Nature of Prejudice*. Addison-Wesley.

23. A more subtle variant of the model would hold that the agent is naive in treating missing information in periods that he does not attend to z, but is sophisticated in periods that he does attend. Since I mostly focus on asymptotic results, this variant should not change the qualitative predictions of the model and, in particular, I conjecture that all of the formal propositions would continue to hold under this alternative formulation.

Andreoni, James and Tymofiy Mylovanov (2012). "Diverging Opinions." *American Economic Journal: Microeconomics*, 4, 209–232.
Aragones, Enriqueta, Itzhak Gilboa, Andrew Postlewaite, and David Schmeidler (2005). "Fact-Free Learning." *American Economic Review*, 95(5), 1355–1368.
Arrow, Kenneth (1973). "The Theory of Discrimination." In *Discrimination in Labor Markets*, edited by O. Ashenfelter and A. Rees. Princeton University Press, pp. 3–33.
Bargh, John A. (1992). "The Ecology of Automaticity: Toward Establishing the Conditions Needed to Produce Automatic Processing Effects." *American Journal of Psychology*, 181–199.
Bargh, John A. and Roman D. Thein (1985). "Individual Construct Accessibility, Person Memory, and the Recall–Judgment Link: The Case of Information Overload." *Journal of Personality and Social Psychology*, 49, 1129–1146.
Becker, Gary S. (1971). *The Economics of Discrimination*. University of Chicago Press.
Besley, Timothy and Anne Case (1993). "Modeling Technology Adoption in Developing Countries." *American Economic Review*, 83(2), 396–402.
Besley, Timothy and Anne Case (1994). "Diffusion as a Learning Process: Evidence from HYV Cotton." Working paper, Princeton University.
Blackwell, David and Lester Dubins (1962). "Merging of Opinions with Increasing Information." *Annals of Mathematical Statistics*, 33, 882–886.
Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts (2013). "Does Management Matter? Evidence from India." *Quarterly Journal of Economics*, 128, 1–51.
Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2012). "Salience Theory of Choice Under Risk." *Quarterly Journal of Economics*, 127, 1243-1285.
Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2013). "Salience and Consumer Choice." *Journal of Political Economy*, 121, 803–843.
Compte, Olivier and Andrew Postlewaite (2004). "Confidence-Enhanced Performance." *American Economic Review*, 94(5), 1536–1557.
Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory*. Wiley-Interscience.
DellaVigna, Stefano (2009). "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature*, 47, 315–372.
Diaconis, Persi and David Freedman (1990). "On the Uniform Consistency of Bayes Estimates for Multinomial Probabilities." *Annals of Statistics*, 18, 1317–1327.
Esponda, Ignacio (2008). "Behavioral Equilibrium in Economies with Adverse Selection." *American Economic Review*, 98(4), 1269–1291.
Esponda, Ignacio and Demian Pouzo (2012). "Learning Foundation for Equilibrium in Voting Environments with Private Information." Working paper, NYU.
Ettinger, David and Philippe Jehiel (2010). "A Theory of Deception." *American Economic Journal: Microeconomics*, 2, 1–20.
Eyster, Erik and Matthew Rabin (2005). "Cursed Equilibrium." *Econometrica*, 73, 1623–1672.
Fiske, Susan T. (1993). "Social Cognition and Social Perception." *Annual Review of Psychology*, 44, 155–194.
Fiske, Susan T. and Shelley E. Taylor (2008). *Social Cognition: From Brains to Culture*. McGraw-Hill Higher Education.
Foster, Andrew D. and Mark R. Rosenzweig (1995). "Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture." *Journal of Political Economy*, 103, 1176–1209.
Fryer, Roland and Matthew Jackson (2008). "A Categorical Model of Cognition and Biased Decision-Making." *B. E. Journal of Theoretical Economics*, 8, 1–44.
Fudenberg, Drew and David K. Levine (1993). "Self-Confirming Equilibrium." *Econometrica*, 61, 523–545.
Fudenberg, Drew and David K. Levine (2006). "Superstition and Rational Learning." *American Economic Review*, 96(3) 630–651.
Gabaix, Xavier (2013). "A Sparsity-Based Model of Bounded Rationality." Working paper, NYU.
Gabaix, Xavier and David Laibson (2005). "Bounded Rationality and Directed Cognition." Working paper, Harvard University.

Gabaix, Xavier, David Laibson, Guillermo Moloche, and Stephen Weinberg (2006). "Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model." *American Economic Review*, 96(4), 1043–1068.
Gawande, Atul (2004). "On Washing Hands." *New England Journal of Medicine*, 350, 1283–1286.
Gilbert, Daniel T. and Patrick S. Malone (1995). "The Correspondence Bias." *Psychological Bulletin*, 117, 21–38.
Gilbert, Daniel T., Brett W. Pelham, and Douglas S. Krull (1988). "On Cognitive Busyness: When Person Perceivers Meet Persons Perceived." *Journal of Personality and Social Psychology*, 54, 733–740.
Gittins, John C. (1979). "Bandit Processes and Dynamic Allocation Indices." *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 148–177.
Glaeser, Edward L. (2005). "The Political Economy of Hatred." *Quarterly Journal of Economics*, 120, 45–86.
Gottlieb, Daniel (2010). "Will You Never Learn? Self Deception and Biases in Information Processing." Working paper, University of Pennsylvania.
Halliday, Stephen (2001). "Death and Miasma in Victorian London: An Obstinate Belief." *British Medical Journal*, 323(7327), 1469–1471.
Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein (2014). "Learning Through Noticing: Theory and Experimental Evidence in Farming." *Quarterly Journal of Economics*, forthcoming (doi:10.1093/qje/qju015).
Hong, Harrison, Jeremy C. Stein, and Jialin Yu (2007). "Simple Forecasts and Paradigm Shifts." *Journal of Finance*, 62, 1207–1242.
Jehiel, Philippe (2005). "Analogy-Based Expectation Equilibrium." *Journal of Economic Theory*, 123, 81–104.
Jovanovic, Boyan and Yaw Nyarko (1996). "Learning by Doing and the Choice of Technology." *Econometrica*, 64, 1299–1310.
Kahneman, Daniel (1973). *Attention and Effort*. Prentice-Hall.
Karlsson, Niklas, George Loewenstein, and Duane Seppi (2009). "The Ostrich Effect: Selective Attention to Information." *Journal of Risk and Uncertainty*, 38, 95–115.
Kass, Robert E. and Adrian E. Raftery (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90, 773–795.
Koszegi, Botond and Adam Szeidl (2013). "A Model of Focusing in Economic Choice." *Quarterly Journal of Economics*, 128, 53–104.
Krueger, Joachim and Russell W. Clement (1994). "Memory-based Judgments About Multiple Categories: A Revision and Extension of Tajfel's Accentuation Theory." *Journal of Personality and Social Psychology*, 67, 35–47.
Levitt, Steven D. and Stephen J. Dubner (2009). *SuperFreakonomics*. William Morrow.
Mack, Arien and Irvin Rock (1998). "Inattentional Blindness: Perception Without Attention." *Visual Attention*, 8, 55–76.
Mackowiak, Bartosz and Mirko Wiederholt (2009). "Optimal Sticky Prices Under Rational Inattention." *American Economic Review*, 99(3), 769–803.
Malmendier, Ulrike and Devin Shanthikumar (2007). "Are Small Investors Naive About Incentives?"*Journal of Financial Economics*, 85, 457–489.
McLean, R. David and Jeffrey Pontiff (2012). "Does Academic Research Destroy Stock Return Predictability?" Working paper, Boston College.
Mullainathan, Sendhil (2000). "Thinking Through Categories." Working paper, MIT.
Mullainathan, Sendhil (2002). "A Memory-Based Model of Bounded Rationality." *Quarterly Journal of Economics*, 117, 735–774.
Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer (2008). "Coarse Thinking and Persuasion." *Quarterly Journal of Economics*, 123, 577–619.
Nagel, Stefan (2012). "Empirical Cross-sectional Asset Pricing." NBER Working Paper No. 18554.
Nisbett, Richard E. and Lee Ross (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall.

Nuland, Sherwin B. (2004). *The Doctors' Plague: Germs, Childbed Fever, and the Strange Story of Ignac Semmelweis*. W. W. Norton.
Peng, Lin and Wei Xiong (2006). "Limited Attention and Asset Prices." *Journal of Financial Economics*, 80, 563–602.
Pettigrew, Thomas F. (1998). "Intergroup Contact Theory." *Annual Review of Psychology*, 49, 65–85.
Pettigrew, Thomas F. and Linda R. Tropp (2006). "A Meta-Analytic Test of Intergroup Contact Theory." *Journal of Personality and Social Psychology*, 90, 751.
Phelps, Edmund S. (1972). "The Statistical Theory of Racism and Sexism." *American Economic Review*, 62(4), 659–661.
Rabin, Matthew and Joel L. Schrag (1999). "First Impressions Matter: A Model of Confirmatory Bias." *Quarterly Journal of Economics*, 114, 37–82.
Ross, Lee (1977). "The Intuitive Psychologist and his Shortcomings: Distortions in the Attribution Process." *Advances in Experimental Social Psychology*, 10, 173–220.
Samuels, Myra L. (1993). "Simpson's Paradox and Related Phenomena." *Journal of the American Statistical Association*, 88, 81–88.
Savage, Leonard (1954). *The Foundations of Statistics*. Dover.
Schacter, Daniel L. (2001). *The Seven Sins of Memory: How the Mind Forgets and Remembers*. Houghton Mifflin.
Schaller, Mark and Meredith O'Brien (1992). "'Intuitive Analysis of Covariance' and Group Stereotype Formation." *Personality and Social Psychology Bulletin*, 18, 776.
Sims, Christopher A. (2003). "Implications of Rational Inattention." *Journal of Monetary Economics*, 50, 665–690.
Sims, Christopher A. (2006). "Rational Inattention: Beyond the Linear-Quadratic Case." *American Economic Review*, 96(2), 158–163.
von Hippel, William, John Jonides, James L. Hilton, and Sowmya Narayan (1993). "Inhibitory Effect of Schematic Processing On Perceptual Encoding." *Journal of Personality and Social Psychology*, 64, 921–921.
Walker, Stephen G. (2004). "Modern Bayesian Asymptotics." *Statistical Science*, 19, 111–117.
Wilson, Andrea (2003). "Bounded Memory and Biases in Information Processing." Working paper, Princeton University.
Woodford, Michael (2009). "Information-Constrained State-Dependent Pricing." *Journal of Monetary Economics*, 56, S100–S124.
Woodford, Michael (2012a). "Inattentive Valuation and Reference-dependent Choice." Working paper, Columbia University.
Woodford, Michael (2012b). "Prospect Theory as Efficient Perceptual Distortion." *American Economic Review*, 102(3), 41–46.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the pulisher's website:

Online Appendix A: Further Definitions and Results
Online Appendix B: Sophistication
Online Appendix C: Proofs
Online Appendix D: Bayes Factors