ONLINE APPENDIX FOR SELECTIVE ATTENTION AND LEARNING

Joshua Schwartzstein
Dartmouth College

Appendix A: Further Definitions and Results

This appendix presents formal results that are referenced in the main text, as well as definitions that are used in the appendices. Proofs of all results are presented in Appendix C.

A.1. Supplementary Material to Section 2: Model

A.1.1. Prior. Here I give an alternative description of the agent's prior, which will be useful in presenting the proofs. The prior can compactly be expressed as

$$\mu(\theta) = \sum_{i \in \{X, \neg X\}} \sum_{j \in \{Z, \neg Z\}} \pi_{i,j} \mu^{i,j}(\theta).$$

Fix a model $M \in \mathscr{M}$ and define $c^M(x, \hat{z})$ as the set of covariates $(x', \hat{z}') \in \mathsf{X} \times \hat{\mathsf{Z}}$ such that, under that model, any y_t given covariates $(x_t, \hat{z}_t) = (x, \hat{z})$ is exchangeable with any $y_{t'}$ given covariates $(x_{t'}, \hat{z}_{t'}) = (x', \hat{z}')$; i.e., under model M, $\theta(x', \hat{z}') = \theta(x, \hat{z})$ with probability one if and only if $(x', \hat{z}') \in c^M(x, \hat{z})$. For example, under $M_{X,\neg Z}$ (where only x is important to predicting y), $c^{X,\neg Z}(x, \hat{z}) = \{(x', \hat{z}') \in \mathsf{X} \times \hat{\mathsf{Z}} : x' = x\}$ equals the set of covariates that agree on x. With a slight abuse of notation, label the common success probability across members of c^M under model M by $\theta(c^M)$. Intuitively, $c^M(x, \hat{z})$ equals the set of covariates that, under model M, can be lumped together with $(x, \hat{z})$ without affecting the accuracy of the agent's predictions.

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

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

E-mail: josh.schwartzstein@dartmouth.edu

A.1.2. Forecasts. This subsection describes the forecasts of an individual with selective attention in some detail (standard Bayesian forecasts are a special case) and presents some definitions which will be useful later.

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

$$\hat{E}[y|x,z,\hat{h}^t] = \sum_{M' \in \mathscr{M}} \hat{\pi}^t_{M'} E_\xi[\theta(c^{M'}(x,\hat{z}))|\hat{h}^t, M'], \tag{A.1}$$

where

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

$$\psi(\tilde{\theta}|\hat{h}^t, c^M) = \frac{\tilde{\theta}^{\kappa(c^M|\hat{h}^t)}(1-\tilde{\theta})^{N(c^M|\hat{h}^t)-\kappa(c^M|\hat{h}^t)}\psi(\tilde{\theta})}{\int \tau^{\kappa(c^M|\hat{h}^t)}(1-\tau)^{N(c^M|\hat{h}^t)-\kappa(c^M|\hat{h}^t)}\psi(\tau)d\tau}.$$

$N(c^M|\hat{h}^t)$ denotes the number of times the covariates have taken on some value $(x,\hat{z}) \in c^M$ along history $\hat{h}^t$ and $\kappa(c^M|\hat{h}^t)$ denotes the number of times that both the covariates have taken on such a value and $y=1$. I will sometimes abuse notation and write $N(x,\hat{z}|\hat{h}^t)$ and $\kappa(x,\hat{z}|\hat{h}^t)$ instead of $N(\{(x,\hat{z})\}|\hat{h}^t)$ and $\kappa(\{(x,\hat{z})\}|\hat{h}^t)$, respectively. Likewise, when convenient I will write $N(x|\hat{h}^t)$ instead of $N(\{(x',\hat{z}') : x' = x\}|\hat{h}^t)$, etc.

To illustrate, (A.1) takes a particularly simple form when $\psi(\theta) \sim \text{U}[0,1]$: $\hat{E}[y|x,z,\hat{h}^t]$ equals

$$\hat{\pi}^t_{X,Z}\frac{\kappa(x,\hat{z}|\hat{h}^t)+1}{N(x,\hat{z}|h^t)+2} + \hat{\pi}^t_{X,\neg Z}\frac{\kappa(x|\hat{h}^t)+1}{N(x|h^t)+2} + \hat{\pi}^t_{\neg X,Z}\frac{\kappa(\hat{z}|\hat{h}^t)+1}{N(\hat{z}|\hat{h}^t)+2} + \hat{\pi}^t_{\neg X,\neg Z}\frac{\bar{\kappa}(\hat{h}^t)+1}{t+1}, \tag{A.2}$$

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

For future reference,

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

where

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

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

$$\alpha_{i,j} = \frac{\pi_{i,j}}{\pi_{X,Z}} \tag{A.3}$$

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

A.2. Supplementary Material to Subsection 3.3: Systematically Biased Stereotypes and Beliefs

This section fleshes out the results from Subsection 3.3. First, selective attention may lead people to form persistent and systematically incorrect beliefs about what causes variation in the data. Second, such biased beliefs are quite robust in the sense that, even if somebody points out the possibility that some other variable is what is truly important, it will take the agent some time to evaluate this claim because he did not previously attend to the relationship between that variable and the outcome.

A.2.1. The Formation of Systematically Biased Beliefs. To formalize, it is useful to specialize to the case where x is a binary random variable and $\mathsf{X} = \{0,1\}$. Define

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

where

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

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

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

PROPOSITION A1. *Suppose $b_k \equiv b$, the agent settles on not encoding z, and $\mathsf{X} = \{0,1\}$. Then*

$$\hat{\mathscr{R}}_x(z') \equiv \hat{E}[y|x=1,z'] - \hat{E}[y|x=0,z'] = \mathscr{R}_x + \frac{\varphi}{\text{Var}(x)} \tag{A.4}$$

almost surely for all z'.

Proposition A1 says that when the agent settles on not encoding z, his limiting reaction to x conditional on $z = z'$, $\hat{\mathscr{R}}_x(z')$, differs from the standard Bayesian's, $\mathscr{R}_x(z')$, in two key ways corresponding to the two terms on the right hand side of (A.4). When $\varphi = 0$, the agent's limiting reaction reduces to the first term, $\mathscr{R}_x$: By persistently failing to encode z, the agent's limiting conditional reaction equals the standard Bayesian's limiting *average* reaction. Thinking of z as a situation, this is one of the distortions exploited in Mullainathan, Schwartzstein, and Shleifer (2008): By grouping distinct situations together in forming beliefs, an agent transfers the informational content of data across situations. For example, the agent may react to a piece of information which is uninformative in a given situation, z, because it is informative in another situation, z'.

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

It is easy to see from Proposition A1 that a selectively attentive agent who fails to learn to pay attention to z can either over- or underreact to x at z', depending on features of the joint distribution over (y,x,z). The simplest case to consider is the one where x is unimportant to predicting y and there can only be overreaction, the topic of the next corollary.

COROLLARY A1. *Suppose the conditions of Proposition A1 hold and, additionally, x is unimportant to predicting y. Then, so long as $\varphi \neq 0$,*

1. $|\hat{\mathscr{R}}_x(z')| = \frac{|\varphi|}{\text{Var}(x)} \neq 0$ *almost surely for all z': The agent overreacts to x and the extent of overreaction is increasing in $\frac{|\varphi|}{\text{Var}(x)}$.*
2. $\hat{\pi}_X^t \xrightarrow{a.s.} 1$*: The agent becomes certain that x is important to predicting y even though it is not.*

Corollary A1 considers the situation where x is completely unimportant to prediction and the selectively attentive agent settles on not encoding z. The first part says that, as a result of the possibility that the selectively attentive agent will

settle on not encoding z, he may come to overreact to x.[2] The degree to which the agent overreacts depends on the extent to which there is a tendency for z's that are associated with $x = 1$ to have relatively high (or low) corresponding success probabilities. Weakening this tendency will mitigate overreaction.

The second part of Corollary A1 says that, as a result of the possibility that the selectively attentive agent will settle on not encoding z, he may eventually become overconfident in the importance of x. This is true whenever z is associated with both x and y and the agent effectively suffers from omitted variable bias. Again, in this case, the agent mistakenly comes to view x as more than a proxy for selectively unattended to predictors.

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

A.2.2. The Robustness of Systematically Biased Beliefs. We have seen that selective attention allows for systematically biased beliefs to emerge and persist in the absence of communication across agents. What if an agent can credibly communicate the importance of a variable that the other has selectively failed to notice? A key feature of selective attention is that, 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.

Suppose $b_k \equiv b$, the agent starts off not encoding z ($\pi_Z < b$), and $\mathsf{X} = \{0,1\}$. At some large t, the agent begins attending to z because there is an unmodeled shock to his belief that z is important to predicting y (π_Z shifts to $\pi'_Z > \pi_Z$) or to the degree to which he is cognitively busy (b shifts to $b' < b$).[4] The main thing to note is that, even if this shock leads the agent to settle on encoding z and to make unbiased forecasts

2. Whenever $\mathsf{X} = \{0,1\}$ and x is unimportant to predicting y, Proposition A1 establishes that Assumption 2 holds if and only if $\varphi \neq 0$, so it is technically redundant to include this condition in the statement of Proposition A1; it is included for clarity.

3. Participants also retained more information about the conversationalist they attended to.

4. We can think of shocks to π_Z as resulting from a media report or something learned in a class and shocks to b as resulting from some (not modeled) reason why the agent would begin caring more about

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

$$\begin{aligned} E_{\xi}[\theta(1,z)|\hat{h}^t] - E_{\xi}[\theta(0,z)|\hat{h}^t] &\approx \pi'_Z[\tau-\tau] + (1-\pi'_Z)(E_{\theta_0}[y|x=1] - E_{\theta_0}[y|x=0]) \\ &= (1-\pi'_Z)(E_{\theta_0}[y|x=1] - E_{\theta_0}[y|x=0]), \end{aligned} \tag{A.5}$$

where $\tau = E_{\psi}[\theta]$ equals the prior success probability under density ψ. From (A.5), the agent's reaction to x in period t is approximately proportional to his reaction the period before when he did not attend to z. This is intuitive: By not having attended to z in the past he has not learned that z is important to predicting y or how z is important to predicting y. As a result, even when he attends to z, his forecast places substantial weight on the empirical frequency of $y=1$ given only (x).

A.3. Supplementary Material to Section 4: Continuous Attention

This section formalizes two observations from Section 4. First, when the continuous attention assumptions hold, the agent eventually learns to attend to z and to make accurate forecasts since he always attends to z with positive probability and almost surely encodes z an infinite number of times.

PROPOSITION A2. *Suppose the continuous attention assumptions hold. Then*

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

Second, even though the agent eventually learns to attend to z and to make accurate forecasts with arbitrarily large probability in the limit, he may continue not to attend to z and to make biased forecasts for a long time. In particular, a partial analog to Proposition 2 holds in this setting: the agent's ability to recognize empirical relationships within a reasonable time horizon depends on his prior.

PROPOSITION A3. *Suppose the continuous attention assumptions hold. Then, for all* $t \geq 2$*, the probability that the agent never encodes* z *before period* t *tends towards* 1 *as* $\pi_Z \to 0$.

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

Appendix B: Sophistication

This appendix describes how the predictions of the discrete attention model would change if the agent was assumed to be sophisticated rather than naive. Again, proofs of all formal results are presented in Appendix C.

There are two main differences if the agent is sophisticated, as will be more formally discussed below. First, a sophisticate may update his beliefs about the importance of variables he never attends to. For such an agent, a belief that z is unimportant to prediction is not necessarily self-confirming in the strong sense identified in Proposition 2. (Nevertheless, Proposition B2 below shows that a sophisticate with a "weak enough" prior belief in the importance of z will never encode z.) Second, the sophisticate entertains the possibility that unattended-to variables covary with those he attends to in a way that is responsible for the relationships he identifies. Such an agent cannot become wrongly certain of the importance of x: sophistication may alter long-run beliefs.[5]

B.1. When is a Belief that z is Unimportant to Prediction Self-Confirming for a Sophisticate?

DEFINITION B1. A belief that z is unimportant to prediction is *strongly self-confirming* if and only if the agent does not update his beliefs positively about the importance of z when he never encodes z; that is, if and only if $\hat{\pi}_Z^t \leq \pi_Z$ given every represented history of the form $(y_{t-1}, x_{t-1}, \varnothing, \ldots, y_1, x_1, \varnothing)$.

Lemma C3 shows that, when the agent is naive, a belief that z is unimportant to prediction is strongly self-confirming. This is not true when the agent is sophisticated.

PROPOSITION B1. *When the agent is sophisticated, there are primitives (i.e., values of $\psi(\cdot)$, π_X, π_Z, etc.) under which a belief that z is unimportant to prediction is not strongly self-confirming.*

The basic idea behind why the sophisticate may update his beliefs about the importance of never attended-to variables is that the variability of the outcome conditional on variables he attends to may provide a subjective signal indicating that unattended-to variables influence the success probability. Indeed, the proof of Proposition B1 provides an example. Nevertheless, even when the agent is sophisticated, a belief that z is unimportant to prediction is self-confirming in the weaker sense that the degree to which the agent can positively update his belief about the importance of z is bounded when he never encodes z. As a result, the sophisticate will never encode z when he starts off with a "weak enough" prior belief about its importance.

5. However, sophistication does not alter long-run forecasts in the sense that Proposition 3 continues to hold if the agent is sophisticated.

To formalize, we have the identity

$$\hat{\pi}_Z^t = \frac{\Lambda(\hat{h}^t)\pi_Z}{(\Lambda(\hat{h}^t)-1)\pi_Z+1},$$

where

$$\Lambda(\hat{h}^t) = \frac{1+\frac{1-\pi_X}{\pi_X}\mathscr{B}_{\neg X,Z}(\hat{h}^t)}{\mathscr{B}_{X,\neg Z}(\hat{h}^t)+\frac{1-\pi_X}{\pi_X}\mathscr{B}_{\neg X,\neg Z}(\hat{h}^t)} \tag{B.1}$$

can be thought of as a "likelihood ratio" comparing the likelihood of the mentally represented history given models where z is important to prediction as compared to the likelihood of that history given models where z is unimportant to prediction. (Unless otherwise noted, in this appendix all likelihoods and Bayes' factors are taken with respect to the sophisticated, rather than naive, agent's beliefs).

Given certain technical conditions, I will first establish an upper bound on Λ that holds uniformly across t and mentally represented histories where the agent has not encoded z through t. To elaborate on these technical conditions, define $\theta^X(x') = \sum_{z'} \theta(x',z')g(z'|x')$ as the success probability conditional on x' given θ and write $\theta^X = (\theta^X(x'))_{x'\in\mathsf{X}}$. Denote the c.d.f. over θ^X given model $M_{i,j}$ by $\Psi^X_{i,j}$ and note that both $\Psi^X_{X,Z}$ and $\Psi^X_{X,\neg Z}$ have associated densities $\psi^X_{X,Z}$ and $\psi^X_{X,\neg Z}$ which are bounded above. In the remainder of this appendix, make the further technical assumptions that $\Psi^X_{\neg X,Z}$ also has an associated density $\psi^X_{\neg X,Z}$ which is bounded above, positive at all values in the support of $\mu_{\neg X,Z}$, and continuous in a neighborhood of ${\theta_0}^X$ (the last two assumptions are not needed to prove Lemma B1, but will be convenient later).[6]

LEMMA B1. *There exists a constant $K<\infty$ such that, for the sophisticated agent,*

$$\Lambda(\hat{h}^t) \leq K$$

for all t and mentally represented histories of the form $\hat{h}^t = (y_{t-1},x_{t-1},\varnothing,\ldots,y_1,x_1,\varnothing)$.

It is almost immediate from Lemma B1 that a sophisticate with a weak enough prior belief about the importance of z will never encode z.

PROPOSITION B2. *Fix $b\in(0,1)$. There exists some $\bar{\pi}_Z \in (0,b)$ such that the sophisticated agent never encodes z if $\pi_Z < \bar{\pi}_Z$.*

6. When I say that θ^X is in the support of $\mu_{i,j}$ I mean that there exists a θ' in the support of $\mu_{i,j}$ such that $\theta'^X = \theta^X$. When there are only two values of x, two values of z, and $g(z'|x') \neq g(z'|x'')$ (as in the stereotyping example), we can use the change of variables formula to analytically derive that $\Psi^X_{\neg X,Z}$ has an associated density $\psi^X_{\neg X,Z}$ which is bounded above and positive at all values in the support of $\mu_{\neg X,Z}$.

B.2. What Does the Sophisticated Agent Come to Believe About the Importance of x?

Proposition 4 established that a naive agent who settles on not encoding z will almost surely become certain that x is important to predicting y so long as x is unconditionally (of z) predictive, where we mean the following when we say the agent becomes certain:

DEFINITION B2. The agent *becomes certain that x is important to predicting y* if and only if $\hat{\pi}_X^t \to 1$ as $t \to \infty$.

The conclusion of Proposition 4 does not hold when the agent is sophisticated: He will not erroneously become convinced in the importance of x if he can explain what he observes in a model where only z is important to prediction; that is, when omitted variable bias could fully explain the estimated relationship between y and x.

PROPOSITION B3. *Suppose x is unimportant to predicting y (but is predictive unconditional of z) and the agent never encodes z because $\pi_Z < \bar{\pi}_Z$, where $\bar{\pi}_Z$ is defined as in Proposition B2. Then, while the naive agent would almost surely become certain that x is important to predicting y, the sophisticated agent almost surely would not.*

Appendix C: Proofs

See Appendix D for the development of results concerning the evolution of Bayes' factors in my model. I will frequently refer to results from that appendix in the proofs below.

C.1. Proofs of Results from Section 2: Model

Proof of Observation 1. To establish the first part of the observation, recall that the standard Bayesian's period t forecast satisfies

$$E[y|x,z,h^t] = \sum_{M' \in \mathscr{M}} \pi^t_{M'} E[\theta(c^{M'}(x,z))|h^t, M']. \tag{C.1}$$

Fix an $M \in \mathscr{M}$. Since the marginal prior density over $\theta(c^M(x,z))$ is non-doctrinaire under M, $E[\theta(c^M(x,z))|h^t, M] \overset{a.s.}{\to} E_{\theta_0}[y|c^M(x,z)]$ by Freedman (1963).

In addition, $E_{\theta_0}[y|c^M(x,z)] = E_{\theta_0}[y|x,z]$ for $M = M_{X,Z}$ as well as for $M_{\neg X,Z}$ when $M_{\neg X,Z}$ is the true model. Consequently, it is left to show that $\pi^t_{X,\neg Z}$ and $\pi^t_{\neg X,\neg Z}$ converge almost surely to zero and that $\pi^t_{\neg X,Z}$ converges almost surely to zero whenever $M_{\neg X,Z}$ is *not* the true model. But these statements follow immediately from Lemma D5.

The second part of the observation follows directly from Lemma D5. □

C.2. Proofs of Results from Section 3: Discrete Attention

Proof of Proposition 1. Suppose that, with positive probability under $P_{\theta_0,\xi}(\cdot)$, the agent does not settle on encoding or not encoding z (b must satisfy $0 < b < 1$). Label this event NS and condition on $\hat{h}^\infty \in NS$. Since the agent encodes z infinitely often conditional on NS, by Lemma D7 we must have $\hat{\pi}^t_Z \to 1$ with probability 1. As a result, with probability 1 there exists a $\tilde{t}$ such that $\hat{\pi}^t_Z > b$ for all $t \geq \tilde{t}$ so $e_t = 1$ for all $t \geq \tilde{t}$, a contradiction. □

A few Lemmas will be useful in establishing Proposition 2. First, define the likelihood ratio (or Bayes' factor) comparing the likelihood of a history under models where z is important to predicting y to the likelihood of that history under models where z is unimportant to predicting y, $\Lambda(h^t)$, as in Equation (B.1).

LEMMA C1. *$\pi_Z(h^t) > b$ if and only if*

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

Proof. Simple algebra. □

LEMMA C2. *For all $\varepsilon > 0$ there exists $\lambda > 0$ such that*

$$P_{\theta_0}\left(\min_{t' \geq 1} \Lambda(h^{t'}) > \lambda\right) \geq 1 - \varepsilon \tag{C.2}$$

Proof. Fix $\varepsilon > 0$. From Lemma D5, we know that $\mathscr{B}_{X,\neg Z}(h^t) \xrightarrow{a.s.} 0, \mathscr{B}_{\neg X,\neg Z}(h^t) \xrightarrow{a.s.} 0$. As a result,

$$\Lambda(h^t) = \frac{\left(1 + \frac{1+\pi_X}{\pi_X}\mathscr{B}_{\neg X,Z}(h^t)\right)}{\left(\mathscr{B}_{X,\neg Z}(h^t) + \frac{1-\pi_X}{\pi_X}\mathscr{B}_{\neg X,\neg Z}(h^t)\right)} \xrightarrow{a.s.} \infty.$$

Consequently, there exists a value $T \geq 1$ such that $P_{\theta_0}(\min_{t' \geq T} \Lambda(h^{t'}) \geq 1) > 1 - \varepsilon$ (see, for example, Lemma 7.2.10 in Grimmett and Stirzaker 2001).

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

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

we have

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

□

The next Lemma establishes some finite sample properties of Bayes' factors and beliefs when the agent never encodes z.

LEMMA C3. *If the (naive) agent has not encoded z up to period t, then*

$$\begin{aligned}
\mathscr{B}_{\neg X,Z}(\hat{h}^t) &= \mathscr{B}_{\neg X,\neg Z}(\hat{h}^t) \\
\mathscr{B}_{X,\neg Z}(\hat{h}^t) &= 1 \\
\pi_Z(\hat{h}^t) &= \pi_Z.
\end{aligned}$$

Proof. If the agent has not encoded z up to period t, then

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

Plugging these expressions into the definition of $\pi_Z(\hat{h}^t)$ yields

$$\begin{aligned}\pi_Z(\hat{h}^t) &= \frac{\pi_Z\left[1+\frac{1-\pi_X}{\pi_X}\mathscr{B}_{\neg X,Z}(\hat{h}^t)\right]}{1+\frac{1-\pi_X}{\pi_X}\mathscr{B}_{\neg X,Z}(\hat{h}^t)}\\ &= \pi_Z.\end{aligned}$$

□

Proof of Proposition 2.

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

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

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

By Lemma C1,

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

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

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

From Lemma C2 we know that there exists $\lambda(\varepsilon) > 0$ such that

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

so the result follows from setting π or b to satisfy

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

Part (2). It is left to show that the agent settles on not encoding z with probability 1 when $\pi_Z < b$. It is sufficient to show that, when $\pi_Z < b$, $\pi_Z(\hat{h}^t) = \pi_Z$ for all $t > 1$. But this follows from Lemma C3. □

Proof of Proposition 3. Part (1). Analagous to the proof of Observation 1.1 and hence omitted.
Part (2). If the agent settles on not encoding z then, by definition, there exists n such that $e_t = 0$ for all $t \geq n$. In any period $t \geq n$, the agent's expectation satisfies

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

Fix an $M \in \mathscr{M}$. Since the marginal prior density over $\theta(c^M(x,\varnothing))$ is non-doctrinaire under M, $E_\xi[\theta(c^M(x,\varnothing))|\hat{h}^t, M] \overset{a.s.}{\to} E_{\theta_0}[y|c^M(x,\varnothing)]$ by Freedman (1963), where $E_{\theta_0}[y|c^M(x,\varnothing)] = E_{\theta_0}[y|x]$ for $M \in \{M_{X,Z}, M_{X,\neg Z}\}$ and $E_{\theta_0}[y|c^M(x,\varnothing)] = E_{\theta_0}[y]$ for $M \in \{M_{\neg X,Z}, M_{\neg X,\neg Z}\}$.

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

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

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

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

$$\left(\frac{\int\prod_{k=n}^{t-1}\theta(x_k,\varnothing)^{y_k}(1-\theta(x_k,\varnothing))^{1-y_k}\mu^{\neg X,j}(d\theta|\hat{h}^n)}{\int\prod_{k=n}^{t-1}\theta(x_k,\varnothing)^{y_k}(1-\theta(x_k,\varnothing))^{1-y_k}\mu^{X,Z}(d\theta|\hat{h}^n)}\right) \tag{C.3}$$

converges to zero almost surely to establish such convergence of the Bayes' factor. But, noting that (i) (C.3) equals $\mathscr{B}^t_{\neg X,j}(p^0_0)$ for some uniformly non-doctrinaire $\mu^{\neg X,j}$, $\mu^{X,Z}$, and (ii)

$$(y_{t-1},x_{t-1},\hat{z}_{t-1},\ldots,y_n,x_n,\hat{z}_n)$$

is a random sample from p^0_0, the result follows from Lemma D6. □

Proof of Proposition 4. Part (1). Analagous to the proof of Observation 1.2 and hence omitted.
Part (2). The fact that $\hat{\pi}^t_X \overset{a.s.}{\to} 1$ when the agent settles on not encoding z follows immediately from Assumption 2 and Lemma D6. That $\hat{\pi}^t_Z \leq b$ for large t follows from the definition of settling on not encoding z and the encoding rule. □

C.3. Proofs of Results from Appendix A.2: Supplementary Results to Subsection 3.3

Proof of Proposition A1. This result follows from theorems in Samuels (1993), but, for completeness, I'll provide a proof.

Let the true distribution over (y,x,z) be denoted by $p_0(\cdot)$ (the distribution generated by θ_0 and $g(\cdot)$) and let $\mathbb{E}$ denote the expectation operator under $p_0(\cdot)$. With this notation,

$$\begin{aligned}\mathscr{R}_x(z') &= \mathbb{E}[y|x=1,z'] - \mathbb{E}[y|x=0,z'] \\ \mathscr{R}_x &= \mathbb{E}[\mathscr{R}_x(z)|x=1] \\ \varphi &= \text{Cov}(\mathbb{E}[y|x=0,z], g(x=1|z)).\end{aligned}$$

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

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

□

Proof of Corollary A1. By the assumption that x is unimportant to predicting y, $\mathscr{R}_x(z') = 0$ for all z' so $\mathscr{R}_x = 0$. Then, by Proposition A1,

$$\begin{aligned}\hat{\mathscr{R}}_x(z') &= E_{\theta_0}[y|x=1] - E_{\theta_0}[y|x=0] \\ &= \frac{\varphi}{\text{Var}(x)},\end{aligned} \tag{C.4}$$

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

C.4. Proofs of Results from Sections 4 and A.3: Continuous Attention

LEMMA C4. *If the continuous attention assumptions hold then the agent almost surely encodes z an infinite number of times.*

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

$$1-\eta\left(\frac{b\pi_Z}{a(1-\pi_Z)+b\pi_Z}\right)<1, \tag{C.5}$$

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

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

and the result follows. □

Proof of Proposition A2.
Part (1): From Lemma C4 we know that the agent almost surely encodes z an infinite number of times. Combining this result with Lemma D7, we have that $\hat{\pi}_Z^t \to 1$ almost surely which implies that $\eta(\hat{\pi}_Z^t) \to 1$ almost surely.
Part (2): Fix (x,z) and $\varepsilon>0$. We want to show that

$$\lim_{t\to\infty} P_{\theta_0,\xi}(|\hat{E}[y|x,z,\hat{h}^t]-E_{\theta_0}[y|x,z]|>\varepsilon)=0. \tag{C.6}$$

Let us expand $P_{\theta_0,\xi}(|\hat{E}[y|x,z,\hat{h}^t]-E_{\theta_0}[y|x,z]|>\varepsilon)$, which equals

$$\begin{aligned} &P_{\theta_0,\xi}(|E_\xi[\theta(x,z)|\hat{h}^t]-E_{\theta_0}[y|x,z]|>\varepsilon)P_{\theta_0,\xi}(e_t=1)\\ +&P_{\theta_0,\xi}(|E_\xi[\theta(x,\varnothing)|\hat{h}^t]-E_{\theta_0}[y|x,z]|>\varepsilon)(1-P_{\theta_0,\xi}(e_t=1)). \end{aligned}$$

The expansion indicates that, to establish (C.6), it is sufficient to show two things:

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

A. follows from now familiar arguments applying the non-doctrinaire assumption, the strong law of large numbers, the consistency properties of Bayes' factors, and the fact that the agent encodes z an infinite number of times (Lemma C4). B. follows from the fact that $P_{\theta_0,\xi}(e_t=1)=E_{\theta_0,\xi}[\eta(\hat{\pi}_Z^t)]$ and $E_{\theta_0,\xi}[\eta(\hat{\pi}_Z^t)]\to 1$ because $\eta(\hat{\pi}_Z^t)$ is bounded and tends almost surely towards 1 by Proposition A2.1. □

Proof of Proposition A3. By Lemma C3, the probability that the agent never encodes z before period t is given by

$$(1-\eta(\pi_Z))^{t-1}, \tag{C.7}$$

which tends towards 1 as $\pi_Z \to 0$. □

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

LEMMA C5. *If the continuous attention assumptions hold then*

$$\frac{\#\mathscr{E}(t)}{t-1} \overset{a.s.}{\to} 1.$$

Proof. Define

$$\gamma^t = \frac{\#\mathscr{E}(t)}{t-1}.$$

I will apply a result from the theory of stochastic approximation to show that $\gamma^t \overset{a.s.}{\to} 1$ (Benaim 1999). We have

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

where

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

Note that

1. F is Lipschitz continuous and is defined on a compact set $[0,1]$
2. $E[\varepsilon_t|\hat{h}^t] = 0$ and $E(|\varepsilon_t|^2) < \infty$
3. $u_t \overset{a.s.}{\to} 0$

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

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

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

Note that the next few lemmas (Lemmas C6 - C8) rely heavily on results and notation from Appendix D.

LEMMA C6. $d = \delta_{X,\neg Z}(p_0^1)$

Proof. Apply Lemma D1 to get $\underline{\theta}(c^{X,\neg Z}(x,z)) = p_0^1(y|x) = p_{\theta_0}(y|x)$ for all x. The result then follows from the definition of $\delta_{X,\neg Z}(p_0^1)$. □

LEMMA C7. *If the continuous attention assumptions hold, then*

$$\frac{\mathscr{B}_{X,\neg Z}(\hat{h}^t)}{e^{-d(t-1)}} \xrightarrow{a.s.} K$$

for some K satisfying $0 < K < \infty$.

Proof. I will show that

$$\frac{1}{t-1}\log \mathscr{B}_{X,\neg Z}(\hat{h}^t) \to -d \tag{C.8}$$

almost surely. We can write

$$\mathscr{B}_{X,\neg Z}(\hat{h}^t) = \frac{\Pr_\xi(\hat{h}_1^t|M_{X,\neg Z})\Pr_\xi(\hat{h}_0^t|M_{X,\neg Z},\hat{h}_1^t)}{\Pr_\xi(\hat{h}_1^t|M_{X,Z})\Pr_\xi(\hat{h}_0^t|M_{X,Z})}, \tag{C.9}$$

where, recall,

$$\begin{aligned}\hat{h}_1^t &\equiv (y_\tau, x_\tau, \hat{z}_\tau)_{\tau<t:\hat{z}_\tau\neq\varnothing}\\ \hat{h}_0^t &\equiv (y_\tau, x_\tau, \hat{z}_\tau)_{\tau<t:\hat{z}_\tau=\varnothing}.\end{aligned}$$

From (C.9), we can write the left hand side of (C.8) as

$$\begin{aligned}&\frac{1}{t-1}\left[\log\left(\frac{\Pr_\xi(\hat{h}_1^t|M_{X,\neg Z})}{\prod_{k\in\mathscr{E}(t)}p_0^1(y_k,x_k,z_k)}\right)+\log\left(\frac{\Pr_\xi(\hat{h}_0^t|M_{X,\neg Z},\hat{h}_1^t)}{\prod_{k\notin\mathscr{E}(t)}p_0^0(y_k,x_k,\varnothing)}\right)\right]\\ &-\frac{1}{t-1}\left[\log\left(\frac{\Pr_\xi(\hat{h}_1^t|M_{X,Z})}{\prod_{k\in\mathscr{E}(t)}p_0^1(y_k,x_k,z_k)}\right)+\log\left(\frac{\Pr_\xi(\hat{h}_0^t|M_{X,Z})}{\prod_{k\notin\mathscr{E}(t)}p_0^0(y_k,x_k,\varnothing)}\right)\right]\end{aligned} \tag{C.10}$$

We know that the second term of (C.10) tends almost surely towards 0 as $t\to\infty$ by Lemma D2.

As a result, to establish (C.8) it remains to show that the first term tends almost surely towards $-d$. Rewrite this term as

$$\begin{aligned}&\frac{\#\mathscr{E}(t)}{t-1}\left[\frac{1}{\#\mathscr{E}(t)}\log\left(\frac{\Pr_\xi(\hat{h}_1^t|M_{X,\neg Z})}{\prod_{k\in\mathscr{E}(t)}p_0^1(y_k,x_k,z_k)}\right)\right]\\ &+\frac{t-1-\#\mathscr{E}(t)}{t-1}\left[\frac{1}{t-1-\#\mathscr{E}(t)}\log\left(\frac{\Pr_\xi(\hat{h}_0^t|M_{X,\neg Z},\hat{h}_1^t)}{\prod_{k\notin\mathscr{E}(t)}p_0^0(y_k,x_k,\varnothing)}\right)\right].\end{aligned} \tag{C.11}$$

By Lemmas D2, C5, and C6, (C.11) tends almost surely towards

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

as $t\to\infty$, which completes the proof. □

LEMMA C8. *If the continuous attention assumptions hold, then*

$$\frac{\mathscr{B}_{\neg X,\neg Z}(\hat{h}^t)}{e^{-d'(t-1)}} \xrightarrow{a.s.} K$$

for some $d' \geq d$ *and* K *satisfying* $0 < K < \infty$.

Proof. Let $d' = \delta_{\neg X,\neg Z}(p_0^1)$. Using analagous arguments to those in the proof of Lemma C7, we can show that

$$\frac{\mathscr{B}_{\neg X,\neg Z}(\hat{h}^t)}{e^{-\delta_{\neg X,\neg Z}(p_0^1)(t-1)}} \xrightarrow{a.s.} K$$

for some K satisfying $0 < K < \infty$. Since $\delta_{\neg X,\neg Z}(p_0^1) > \delta_{X,\neg Z}(p_0^1)$ (from the fact that adding more constraints weakly increases the minimized Kullback-Leibler divergence) and $\delta_{X,\neg Z}(p_0^1) = d$ (by Lemma C6), the result follows. □

LEMMA C9. *Suppose the continuous attention assumptions hold. If the asymptotic rate of convergence of* $\hat{\pi}_Z^t$ *to* 1 *is* $V(t)$ *then the asymptotic rate of convergence of* $\eta(\hat{\pi}_Z^t)$ *to* 1 *is* $V(t)$.

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

$$\frac{1-\hat{\pi}_Z^t}{V(t)} \xrightarrow{a.s.} C. \tag{C.12}$$

The goal is to show that (C.12) implies that there exists a strictly positive constant $C' < \infty$ such that

$$\frac{1-\eta(\hat{\pi}_Z^t)}{V(t)} \xrightarrow{a.s.} C'. \tag{C.13}$$

But (C.13) follows from (C.12) so long as there exists a strictly positive constant $K < \infty$ such that

$$\frac{1-\eta(\hat{\pi}_Z^t)}{1-\hat{\pi}_Z^t} \xrightarrow{a.s.} K, \tag{C.14}$$

which can easily be verified using l'Hopital's rule, together with the assumptions that $\eta' = f$ is continuous and strictly positive on $[0,1]$. □

Proof of Proposition 5. From Lemma C9, it is enough to show that

$$\frac{1-\hat{\pi}_Z^t}{e^{-d(t-1)}} \xrightarrow{a.s} C \tag{C.15}$$

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

Since

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

to demonstrate (C.15) it suffices to show that

$$\frac{e^{-d(t-1)}}{\frac{1-\pi_Z}{\pi_Z}\mathscr{B}_{X,\neg Z}(\hat{h}^t)+\frac{1-\pi_X}{\pi_X}\frac{1-\pi_Z}{\pi_Z}\mathscr{B}_{\neg X,\neg Z}(\hat{h}^t)} + \frac{\frac{1-\pi_X}{\pi_X}\mathscr{B}_{\neg X,Z}(\hat{h}^t)e^{-d(t-1)}}{\frac{1-\pi_Z}{\pi_Z}\mathscr{B}_{X,\neg Z}(\hat{h}^t)+\frac{1-\pi_X}{\pi_X}\frac{1-\pi_Z}{\pi_Z}\mathscr{B}_{\neg X,\neg Z}(\hat{h}^t)} \xrightarrow{a.s} c' \tag{C.16}$$

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

The first term on the left hand side of (C.16) converges almost surely to a positive finite constant by Lemmas C7 and C8. The second term on the left hand side of (C.16) converges almost surely to 0 whenever $\pi_X = 1$ (trivially) or x is important to predicting y (by Lemmas D3, C7, and C8). □

C.5. Proofs of Results from Appendix B: Sophistication

Proof of Proposition B1. I prove by example. We will consider the case where $\pi_X = 1$, so

$$\hat{\pi}_Z^t = \pi_Z \frac{\Pr(\hat{h}^t|M_{X,Z})}{\Pr(\hat{h}^t|M_{X,\neg Z})}.$$

It is enough to find values of primitives such that there exists some history of the form $\hat{h}^t = (y_{t-1}, x_{t-1}, \varnothing, \ldots, y_1, x_1, \varnothing)$, where, for the sophisticated agent,

$$\begin{aligned}
\rho &= \frac{\Pr(\hat{h}^t|M_{X,Z})}{\Pr(\hat{h}^t|M_{X,\neg Z})} \\
&= \frac{\int_\theta \prod_{\tau=1}^{t-1} p_\theta(y_\tau|x_\tau) d\mu^{X,Z}}{\int_\theta \prod_{\tau=1}^{t-1} p_\theta(y_\tau|x_\tau) d\mu^{X,\neg Z}} \\
&= \frac{\int_\theta \prod_{x'} (\sum_{z'} \theta(x',z') g(z'|x'))^{\kappa^t(x')} (1-\sum_{z'} \theta(x',z') g(z'|x'))^{N^t(x')-\kappa^t(x')} d\mu^{X,Z}}{\int_\theta \prod_{x'} \theta(x')^{\kappa^t(x')} (1-\theta(x'))^{N^t(x')-\kappa^t(x')} d\mu^{X,\neg Z}} \\
&> 1.
\end{aligned}$$

To this end, consider an example with $Z = \{1,2\}$, $g(1|x) = 1/2$ for all x, and $\psi(\theta) \sim \mathrm{U}[0,1]$. For any length-2 history of the form $(y_2, x, \varnothing, y_1, x, \varnothing)$, we have

$$\rho = \frac{\int (.5(\tau_1+\tau_2))^{y_1+y_2} (1-.5(\tau_1+\tau_2))^{2-y_1-y_2} d\tau_1 d\tau_2}{\int \tau^{y_1+y_2}(1-\tau)^{2-y_1-y_2} d\tau}.$$

Calculating ρ in this example for the case where $y_1 + y_2 = 1$ yields $\rho = 1.25$. □

Proof of Lemma B1. Note the inequality

$$\Lambda(\hat{h}^t) \leq \frac{1}{\mathscr{B}_{X,\neg Z}(\hat{h}^t)} + \frac{1-\pi_X}{\pi_X}\frac{\mathscr{B}_{\neg X,Z}(\hat{h}^t)}{\mathscr{B}_{X,\neg Z}(\hat{h}^t)}. \tag{C.17}$$

We will proceed by bounding the two terms on the right hand side of (C.17) above by a term that does not depend on $t, \hat{h}^t$.

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

$$\begin{aligned}\mathscr{B}_{X,\neg Z}(\hat{h}^t) &= \frac{\int \prod_{x'} \theta^X(x')^{\kappa^t(x')}(1-\theta^X(x'))^{N^t(x')-\kappa^t(x')}\psi^X_{X,\neg Z}(\theta^X)d\theta^X}{\int \prod_{x'} \theta^X(x')^{\kappa^t(x')}(1-\theta^X(x'))^{N^t(x')-\kappa^t(x')}\psi^X_{X,Z}(\theta^X)d\theta^X} \\ &\geq a_1,\end{aligned}$$

where $a_1 \in (0,\infty)$ is a constant that does not depend on $t, \hat{h}^t$. The inequality holds because $\psi^X_{X,\neg Z}(\theta^X) = \prod_{x'} \psi(\theta^X(x'))$ is bounded below by a positive constant and $\psi^X_{X,Z}(\theta^X)$ is bounded above by a positive constant. As a result, the first term on the right hand side of (C.17) is bounded above by a term that does not depend on $t, \hat{h}^t$.

Also, when $\hat{h}^t = (y_{t-1}, x_{t-1}, \varnothing, \ldots, y_1, x_1, \varnothing)$, we can write

$$\begin{aligned}\frac{\mathscr{B}_{\neg X,Z}(\hat{h}^t)}{\mathscr{B}_{X,\neg Z}(\hat{h}^t)} &= \frac{\int \prod_{x'} \theta^X(x')^{\kappa^t(x')}(1-\theta^X(x'))^{N^t(x')-\kappa^t(x')}\psi^X_{\neg X,Z}(\theta^X)d\theta^X}{\int \prod_{x'} \theta^X(x')^{\kappa^t(x')}(1-\theta^X(x'))^{N^t(x')-\kappa^t(x')}\psi^X_{X,\neg Z}(\theta^X)d\theta^X} \\ &\leq a_2,\end{aligned}$$

where $a_2 \in (0,\infty)$ is a constant that does not depend on $t, \hat{h}^t$. The inequality follows from the fact that $\psi^X_{X,\neg Z}$ is bounded below by a positive constant and $\psi^X_{\neg X,Z}$ is bounded above by a constant. We conclude that the second term on the right hand size of (C.17) is bounded above by a term that does not depend on $t, \hat{h}^t$, which completes the proof. □

Proof of Proposition B2. It is the case that

$$\hat{\pi}^t_Z \leq b \iff \Lambda(\hat{h}^t) \leq \frac{1-\pi_Z}{\pi_Z}\frac{b}{1-b},$$

so the agent will never encode z if

$$\Lambda(\hat{h}^t) \leq \frac{1-\pi_Z}{\pi_Z}\frac{b}{1-b} \tag{C.18}$$

for all $t, \hat{h}^t$.

From Lemma B1, we have that $\Lambda(\hat{h}^t) \leq K$ for some $K \in (0,\infty)$, where this bound holds for all $t, \hat{h}^t$. Thus, Condition (C.18) holds so long as

$$K \leq \frac{1-\pi_Z}{\pi_Z}\frac{b}{1-b} \iff \pi_Z \leq \frac{b/(1-b)}{K + b/(1-b)} \equiv \bar{\pi}_Z. \tag{C.19}$$

□

Proof of Proposition B3. The statement regarding the naive agent was established in Proposition 4, so we restrict attention to the sophisticate. We can write

$$\begin{aligned}\hat{\pi}_X^t &= \frac{1+\alpha_{X,\neg Z}\mathscr{B}_{X,\neg Z}(\hat{h}^t)}{1+\alpha_{X,\neg Z}\mathscr{B}_{X,\neg Z}(\hat{h}^t)+\alpha_{\neg X,Z}\mathscr{B}_{\neg X,Z}(\hat{h}^t)+\alpha_{\neg X,\neg Z}\mathscr{B}_{\neg X,\neg Z}(\hat{h}^t)} \\ &\leq \frac{1+\alpha_{X,\neg Z}\mathscr{B}_{X,\neg Z}(\hat{h}^t)}{1+\alpha_{X,\neg Z}\mathscr{B}_{X,\neg Z}(\hat{h}^t)+\alpha_{\neg X,Z}\mathscr{B}_{\neg X,Z}(\hat{h}^t)}.\end{aligned}$$

From the inequality, it suffices to show that, almost surely, $\mathscr{B}_{\neg X,Z}(\hat{h}^t)$ does not vanish as t gets large conditional on $\pi_Z < \bar{\pi}_Z$.[7]

When the agent never encodes z, we can write

$$\begin{aligned}\mathscr{B}_{\neg X,Z}(\hat{h}^t) &= \frac{\int \prod_{x'} \theta^X(x')^{\kappa^t(x')}(1-\theta^X(x'))^{N^t(x')-\kappa^t(x')}\psi^X_{\neg X,Z}(\theta^X)d\theta^X}{\int \prod_{x'} \theta^X(x')^{\kappa^t(x')}(1-\theta^X(x'))^{N^t(x')-\kappa^t(x')}\psi^X_{X,Z}(\theta^X)d\theta^X} \\ &= \frac{\left(\frac{\int \prod_{x'} \theta^X(x')^{\kappa^t(x')}(1-\theta^X(x'))^{N^t(x')-\kappa^t(x')}\psi^X_{\neg X,Z}(\theta^X)d\theta^X}{\int \prod_{x'} \theta^X(x')^{\kappa^t(x')}(1-\theta^X(x'))^{N^t(x')-\kappa^t(x')}d\theta^X}\right)}{\left(\frac{\int \prod_{x'} \theta^X(x')^{\kappa^t(x')}(1-\theta^X(x'))^{N^t(x')-\kappa^t(x')}\psi^X_{X,Z}(\theta^X)d\theta^X}{\int \prod_{x'} \theta^X(x')^{\kappa^t(x')}(1-\theta^X(x'))^{N^t(x')-\kappa^t(x')}d\theta^X}\right)} \\ &= \frac{E^\lambda[\psi^X_{\neg X,Z}(\theta^X)|\hat{h}^t]}{E^\lambda[\psi^X_{X,Z}(\theta^X)|\hat{h}^t]} \\ &\geq \frac{E^\lambda[\psi^X_{\neg X,Z}(\theta^X)|\hat{h}^t]}{\bar{g}}\end{aligned}$$

for some constant $\bar{g} \in (0,\infty)$, where E^λ is the expectation operator taken with respect to the Lebesgue prior over θ^X. The inequality follows from the fact that $\psi^X_{X,Z}$ is bounded above by some finite positive constant. From this inequality, it suffices to show that, almost surely, $E^\lambda[\psi^X_{\neg X,Z}(\theta^X)|\hat{h}^t]$ does not vanish.

From Freedman (1963), $\lambda_t \to \delta_{\theta_0{}^X}$, almost surely (weak-star), where λ_t denotes the posterior over θ^X given prior λ and history $\hat{h}^t$, and δ_{θ^X} denotes a point mass at θ^X. Since $\psi^X_{\neg X,Z}$ is continuous in a neighborhood of $\theta_0{}^X$ (by assumption), this implies that $E^\lambda[\psi^X_{\neg X,Z}(\theta^X)|\hat{h}^t] \to \psi^X_{\neg X,Z}(\theta_0{}^X)$ almost surely. Since $\theta_0{}^X$ is in the support of $\mu_{\neg X,Z}$ (because x is unimportant to predicting y), we have $\psi^X_{\neg X,Z}(\theta_0{}^X) > 0$ (by assumption). □

7. Recall that statements regarding almost sure convergence are with respect to the measure over infinite-horizon mentally represented histories as generated by θ_0, g, and the agent's encoding rule. When $\pi_Z < \bar{\pi}_Z$, the relevant measure is the distribution over $(y_\tau, x_\tau, \varnothing)_{\tau=1}^\infty$, as implied by θ_0, g.

Appendix D: Bayes' Factors

This appendix is broken up into four parts. First, I establish several results which are useful in establishing asymptotic properties of Bayes' factors generally, and in turn aid in characterizing the agent's asymptotic forecasts and beliefs. Next, I consider the evolution of Bayes' factors for three special cases: when the agent always encodes z, when the agent never encodes z, and when the agent encodes z infinitely often.

Let $p_0(y,x,\hat{z})$ and $\hat{p}(y,x,\hat{z})$ denote probability mass functions over $(y,x,\hat{z}) \in \{0,1\} \times \mathsf{X} \times \hat{\mathsf{Z}}$. Define the Kullback Leibler distance between $\hat{p}(y,x,\hat{z})$ and $p_0(y,x,\hat{z})$ as

$$d_K(\hat{p}, p_0) = \sum_{y,x,\hat{z}} p_0(y,x,\hat{z}) \log\left(\frac{p_0(y,x,\hat{z})}{\hat{p}(y,x,\hat{z})}\right) \tag{D.1}$$

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

For all $(y,x,\hat{z})$, assume that $\hat{p}(y,x,\hat{z})$ can be written as $\hat{p}(y,x,\hat{z}|\theta) = \theta(x,\hat{z})^y(1-\theta(x,\hat{z}))^{1-y}p_0(x,\hat{z})$ (sometimes abbreviated as $\hat{p}_\theta(y,x,\hat{z})$), where $p_0(x,\hat{z}) = \sum_{y' \in \{0,1\}} p_0(y',x,\hat{z})$. Define $\hat{p}(y,x,\hat{z}|\theta(M)) = p_{\theta(M)}(y,x,\hat{z})$ in the obvious manner ($\theta(M)$ is defined as in Subsection A.1.1) and let $\underline{\theta}(M) = \arg\min_{\theta(M)\in\Theta(M)} d_K(\hat{p}_{\theta(M)}, p_0)$ denote a minimizer of the Kullback-Leibler distance between $\hat{p}_{\theta(M)}(\cdot)$ and $p_0(\cdot)$ among parameter values in the support of $\mu^M(\cdot)$. Finally, define $\delta_M = \delta_M(p_0) = d_K(\hat{p}_{\underline{\theta}(M)}, p_0)$.

LEMMA D1. *For all $M \in \mathscr{M}$, p_0, and $c^M \in C^M$, $\underline{\theta}(c^M) = p_0(y=1|c^M)$.*

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

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

The right-hand-side of this expression can be re-written as

$$\sum_{c^M \in C^M} [p_0(y=1|c^M)p_0(c^M)\log(\theta(c^M)) + p_0(y=0|c^M)p_0(c^M)\log(1-\theta(c^M))] - K, \tag{D.2}$$

where K does not depend on $\theta(M)$. It is routine to show that each term in the sum of (D.2) is maximized when $\theta(c^M) = p_0(y=1|c^M)$, which concludes the proof. □

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

$$\mathscr{I}_t(M) = \mathscr{I}(M|\hat{h}^t) = \int \frac{\prod_{k=1}^{t-1} \hat{p}(y_k,x_k,\hat{z}_k|\theta)}{\prod_{k=1}^{t-1} p_0(y_k,x_k,\hat{z}_k)} \mu^M(d\theta) \tag{D.3}$$

as well as the predictive distribution

$$\hat{p}_t^M(y,x,\hat{z}) = \hat{p}^M(y,x,\hat{z}|\hat{h}^t) = \int \hat{p}(y,x,\hat{z}|\theta)\mu^M(d\theta|\hat{h}^t). \tag{D.4}$$

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

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

DEFINITION D1. μ^M is *uniformly non-doctrinaire* if it makes each $\theta(c^M)$ independent with non-doctrinaire prior ψ_{c^M}.

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

LEMMA D2. *For all $M \in \mathscr{M}$, p_0, and uniformly non-doctrinaire μ^M,*

$$\frac{1}{t-1}\log \mathscr{I}_t(M|p_0) \to -\delta_M(p_0), \tag{D.5}$$

p_0^∞ almost surely.

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

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

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

The third condition requires a bit more work to verify. Note that

$$\mathscr{I}_{t+1}(M) = \frac{\hat{p}_t^M(y_t, x_t, \hat{z}_t)}{p_0(y_t, x_t, \hat{z}_t)}\mathscr{I}_t(M) \Rightarrow \log(\mathscr{I}_{t+1}(M)/\mathscr{I}_t(M)) = \log\left(\frac{\hat{p}_t^M(y_t, x_t, \hat{z}_t)}{p_0(y_t, x_t, \hat{z}_t)}\right),$$

so condition (3) is equivalent to

$$\sup_t \mathrm{Var}\left[\log\left(\frac{\hat{p}_t^M(y_t, x_t, \hat{z}_t)}{p_0(y_t, x_t, \hat{z}_t)}\right)\right] < \infty \tag{D.6}$$

which can easily be shown to hold so long as

$$\sup_t E\left\{\sum_{y,x,\hat{z}} p_0(y, x, \hat{z}) \log\left(\frac{\hat{p}_t^M(y|x, \hat{z})}{p_0(y|x, \hat{z})}\right)^2\right\} < \infty \tag{D.7}$$

or

$$\sup_t E\left[\log\left(\hat{p}_t^M(y|x, \hat{z})\right)^2\right] < \infty \tag{D.8}$$

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

To verify (D.8), fix some $(y,x,\hat{z})$ with $p_0(y,x,\hat{z}) > 0$ and let $N(c^M(x,\hat{z})|\hat{h}^t) = N_t$ denote the number of times the covariates have taken on some value $(x',\hat{z}') \in c^M(x,\hat{z})$ along history $\hat{h}^t$ and $\kappa(c^M(x,\hat{z})|\hat{h}^t) = \kappa_t$ denote the number of times both that the covariates have taken on such a value and $y = 1$. Then

$$q_t = \frac{\kappa_t + 1}{N_t + 2} \tag{D.9}$$

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

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

$$|\hat{p}_t^M(y|x,\hat{z}) - q_t^y(1-q_t)^{1-y}| \to 0 \tag{D.10}$$

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

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

for all $t \geq n^8$ which implies that

$$E[|\log(\hat{p}_t^M(y|x,\hat{z}))^2 - \log(q_t^y(1-q_t)^{1-y})^2|] < \varepsilon \tag{D.11}$$

for all $t \geq n$. Since $E[\log(\hat{p}_t^M(y|x,\hat{z}))^2] < \infty$ for all finite t, to verify (D.8) it is sufficient to show that

$$\sup_t E[\log(q_t^y(1-q_t)^{1-y})^2] < \infty \tag{D.12}$$

by (D.11).

By symmetry, it is without loss of generality to verify (D.12) for the case where $y = 1$. To this end,

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

where

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

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

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

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

for all N. As a result,

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

for all t which verifies (D.12) and concludes the proof. □

Define the Bayes' factor conditional on p_0 as

$$\mathscr{B}_{i,j}(\hat{h}^t|p_0) = \mathscr{B}^t_{i,j}(p_0) = \frac{\mathscr{I}_t(M_{i,j}|p_0)}{\mathscr{I}_t(M_{X,Z}|p_0)} \tag{D.13}$$

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

LEMMA D3. *For all $M_{i,j} \in \mathscr{M}$, p_0, and uniformly non-doctrinaire $\mu^{i,j}, \mu^{X,Z}$,*

$$\frac{1}{t-1}\log\mathscr{B}^t_{i,j}(p_0) \to \delta_{X,Z}(p_0) - \delta_{i,j}(p_0), \tag{D.14}$$

p_0^∞ almost surely.

Proof. Note that

$$\frac{1}{t-1}\log\mathscr{B}^t_{i,j} = \frac{1}{t-1}\log\left(\mathscr{I}_t(M_{i,j})\right) - \frac{1}{t-1}\log\left(\mathscr{I}_t(M_{X,Z})\right) \tag{D.15}$$

so the result follows immediately from Lemma D2. □

REMARK D1. An immediate implication of Lemma D3 is that $\delta_{i,j}(p_0) > \delta_{X,Z}(p_0)$ implies $\mathscr{B}^t_{i,j}(p_0) \to 0$, p_0^∞ almost surely.

Remark D1 applies when $\delta_{i,j}(p_0) > \delta_{X,Z}(p_0)$; what does the Bayes' factor $\mathscr{B}^t_{i,j}(p_0)$ tend towards asymptotically when $\delta_{i,j}(p_0) = \delta_{X,Z}(p_0)$? I now present a Lemma (due to Diaconis and Freedman 1992) that will aid in estimating the Bayes' factor in this case and establishing asymptotic results. First some definitions. Let $H(q)$ be the entropy function $q\log(q) + (1-q)\log(1-q)$ (set at 0 for $q = 0$ or 1) and define

the following

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

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

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

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

LEMMA D4. *For any non-doctrinaire $\psi(\cdot)$ there are $0<a<A<\infty$ such that for all $N=1,2,\ldots$ and $\kappa=0,1,\ldots,N$, $a\varphi^*(\kappa,N) < \varphi(\kappa,N,\psi) < A\varphi^*(\kappa,N)$.*

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

D.1. Bayes' Factors When the Agent Always Encodes z

LEMMA D5. *$\mathscr{B}_{X,\neg Z}(h^t) \to 0$ and $\mathscr{B}_{\neg X,\neg Z}(h^t) \to 0$, P_{θ_0} almost surely. Additionally, if x is important to predicting y, then $\mathscr{B}_{\neg X,Z}(h^t) \to 0$, P_{θ_0} almost surely; otherwise, $\mathscr{B}_{\neg X,Z}(h^t) \to \infty$, P_{θ_0} almost surely.*

Proof. First, I will establish that $\mathscr{B}^t_{X,\neg Z}$ and $\mathscr{B}^t_{\neg X,\neg Z}$ converge almost surely to zero, and that $\mathscr{B}^t_{\neg X,Z}$ converges almost surely to zero whenever x is important to predicting y. When the agent always encodes z, each period's observation is independently drawn from $p_0^1(y,x,z) = \theta_0(x,z)^y(1-\theta_0(x,z))^{1-y}g(x,z)$ for all (y,x,z). Then, Lemma D3 implies that it is sufficient to show that $\delta_{X,\neg Z}(p_0^1) > \delta_{X,Z}(p_0^1)$, $\delta_{\neg X,\neg Z}(p_0^1) > \delta_{X,Z}(p_0^1)$ and, whenever x is important to predicting y, $\delta_{\neg X,Z}(p_0^1) > \delta_{X,Z}(p_0^1)$. We can easily establish these inequalities by applying Lemma D1 for each $M \in \mathscr{M}$.

It is left to show that $\mathscr{B}^t_{\neg X,Z} \xrightarrow{a.s.} \infty$ whenever x is not important to predicting y. First, write out the Bayes' factor:

$$\mathscr{B}^t_{\neg X,Z} = \frac{\Pr(h^t|M_{\neg X,Z})}{\Pr(h^t|M_{X,Z})} \tag{D.16}$$

$$= \prod_{z'} \frac{\int_0^1 \theta^{\kappa(z'|h^t)}(1-\theta)^{N(z'|h^t)-\kappa(z'|h^t)}\psi(\theta)d\theta}{\prod_{x'}\int_0^1 \theta^{\kappa(x',z'|h^t)}(1-\theta)^{N(x',z'|h^t)-\kappa(x',z'|h^t)}\psi(\theta)d\theta}. \tag{D.17}$$

From (D.17) it is sufficient to show that

$$\frac{\int_0^1 \theta^{\kappa(z|h^t)}(1-\theta)^{N(z|h^t)-\kappa(z|h^t)}\psi(\theta)d\theta}{\prod_{x'}\int_0^1 \theta^{\kappa(x',z|h^t)}(1-\theta)^{N(x',z|h^t)-\kappa(x',z|h^t)}\psi(\theta)d\theta} \xrightarrow{a.s.} \infty \tag{D.18}$$

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

Fix some z. I will use Lemma D4 to estimate (D.18). Let $\kappa_t = \kappa(z|h^t), N_t = N(z|h^t), \hat{q}_t = \kappa_t/N_t, \kappa_t^{x'} = \kappa(x',z|h^t), N_t^{x'} = N(x',z|h^t)$, and $\hat{q}_t^{x'} = \kappa_t^{x'}/N_t^{x'}$.

Applying Lemma D4,

$$\frac{\int_0^1 \theta^{\kappa_t}(1-\theta)^{N_t-\kappa_t}\psi(\theta)d\theta}{\prod_{x'}\int_0^1 \theta^{\kappa_t^{x'}}(1-\theta)^{N_t^{x'}-\kappa_t^{x'}}\psi(\theta)d\theta} \geq \frac{a\varphi^*(\kappa_t,N_t)}{A^{\#\mathsf{X}}\prod_{x'}\varphi^*(\kappa_t^{x'},N_t^{x'})} \tag{D.19}$$

for some constants $0 < a < A < \infty$. By the strong law of large numbers, the right hand side of (D.19) tends almost surely towards

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

where C is some positive constant independent of t. □

D.2. Bayes' Factors When the Agent Never Encodes z

Define

$$p_0^0(y,x,\hat{z}) = \begin{cases} \sum_{z' \in \mathsf{Z}} \theta_0(x,z')^y(1-\theta_0(x,z'))^{1-y}g(x,z') & \text{for each } y,x, \text{ and } \hat{z} = \varnothing \\ 0 & \text{for } \hat{z} \neq \varnothing \end{cases} \tag{D.20}$$

to equal the distribution over $(y,x,\hat{z})$ conditional on the agent not encoding z. Lemma D6 establishes the almost sure limit of several Bayes' factors when the agent never encodes z.

LEMMA D6. *Suppose $E_{\theta_0}[y|x] \neq E_{\theta_0}[y]$ for some $x \in \mathsf{X}$. Then, for all uniformly non-doctrinaire $\mu^{\neg X,\neg Z}, \mu^{\neg X,Z}$, and $\mu^{X,Z}$, $\mathscr{B}_{\neg X,\neg Z}(\hat{h}^t|p_0^0) \to 0$ and $\mathscr{B}_{\neg X,Z}(\hat{h}^t|p_0^0) \to 0$, $(p_0^0)^\infty$ almost surely.*

Proof. Lemma D3 implies that it is sufficient to show that $\delta_{\neg X,\neg Z}(p_0^0) > \delta_{X,Z}(p_0^0)$ and $\delta_{\neg X,Z}(p_0^0) > \delta_{X,Z}(p_0^0)$ whenever $E_{\theta_0}[y|x] \neq E_{\theta_0}[y]$ for some $x \in \mathsf{X}$. We can easily verify these inequalities by applying Lemma D1 for each $M \in \mathscr{M}$. □

D.3. Bayes' Factors When the Agent Encodes z Infinitely Often

LEMMA D7. *Suppose that, with positive probability under $P_{\theta_0,\xi}(\cdot)$, the agent encodes z infinitely often. Conditional on the agent encoding z infinitely often, $\mathscr{B}_{X,\neg Z}(\hat{h}^t) \to 0$ and $\mathscr{B}_{\neg X,\neg Z}(\hat{h}^t) \to 0$ almost surely.*

Proof. I will establish that, almost surely,

$$\log(\mathscr{B}_{i,\neg Z}(\hat{h}^t)) \to -\infty \tag{D.21}$$

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

Defining

$$\hat{h}_1^t \equiv (y_\tau, x_\tau, \hat{z}_\tau)_{\tau<t:\hat{z}_\tau \neq \varnothing}$$
$$\hat{h}_0^t \equiv (y_\tau, x_\tau, \hat{z}_\tau)_{\tau<t:\hat{z}_\tau = \varnothing}$$

we can write

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

so the LHS of (D.21) can be expressed as

$$\log\left(\frac{\Pr_\xi(\hat{h}_0^t | M_{i,\neg Z})}{\Pr_\xi(\hat{h}_0^t | M_{X,Z})}\right) + \log\left(\frac{\Pr_\xi(\hat{h}_1^t | M_{i,\neg Z}, \hat{h}_0^t)}{\Pr_\xi(\hat{h}_1^t | M_{X,Z}, \hat{h}_0^t)}\right). \tag{D.22}$$

When the agent fails to encode z only a finite number of times along a history, we can ignore the first term of (D.22) because it tends towards a finite value as $t \to \infty$. Otherwise, Lemma C3 says that the first term of (D.22) is identically 0 for $i = X$, as well as for $i = \neg X$ when X is a singleton; Lemma D6 (together with Assumption 2) says that the first term tends towards $-\infty$ with probability 1 for $i = \neg X$ when X contains at least two elements. As a result, no matter which case we are in it is sufficient to show that the second term of (D.22) tends towards $-\infty$ with probability 1 in order to establish (D.21). This can be verified by showing that

$$\limsup_t \frac{1}{\#\mathscr{E}(t)} \log\left(\frac{\Pr_\xi(\hat{h}_1^t | M_{i,\neg Z}, \hat{h}_0^t)}{\prod_{\tau \in \mathscr{E}(t)} p_0^1(y_\tau, x_\tau, z_\tau)}\right) - \frac{1}{\#\mathscr{E}(t)} \log\left(\frac{\Pr_\xi(\hat{h}_1^t | M_{X,Z}, \hat{h}_0^t)}{\prod_{\tau \in \mathscr{E}(t)} p_0^1(y_\tau, x_\tau, z_\tau)}\right) < 0 \tag{D.23}$$

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

$$p_0^1(y, x, z) = \theta_0(x, z)^y (1 - \theta_0(x, z))^{1-y} g(x, z)$$
$$\mathscr{E}(t) = \{\tau < t : \hat{z}_\tau \neq \varnothing\}.$$

The second term on the LHS of (D.23) tends towards 0 with probability 1 by Lemma D2.[9] To complete the proof, it then remains to show that the first term on the LHS of (D.23) remains bounded away from 0 as $t \to \infty$ for $i \in \{X, \neg X\}$, or

$$\limsup_t \frac{1}{\#\mathscr{E}(t)} \log\left(\frac{\Pr_\xi(\hat{h}_1^t | M_{i,\neg Z}, \hat{h}_0^t)}{\prod_{\tau \in \mathscr{E}(t)} p_0^1(y_\tau, x_\tau, z_\tau)}\right) < 0. \tag{D.24}$$

9. Note that

$$\frac{1}{\#\mathscr{E}(t)} \log\left(\frac{\Pr_\xi(\hat{h}_1^t | M_{X,Z}, \hat{h}_0^t)}{\prod_{\tau \in \mathscr{E}(t)} p_0^1(y_\tau, x_\tau, z_\tau)}\right) = \frac{1}{\#\mathscr{E}(t)} \log\left(\frac{\Pr_\xi(\hat{h}_1^t | M_{X,Z})}{\prod_{\tau \in \mathscr{E}(t)} p_0^1(y_\tau, x_\tau, z_\tau)}\right)$$

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

We can re-write the LHS of (D.24) as

$$\frac{1}{\#\mathscr{E}(t)}\log\left(\frac{\prod_{x'}\prod_{z'}\int\theta(x',z')^{\kappa(x',z'|\hat{h}_1^t)}(1-\theta(x',z'))^{N(x',z'|\hat{h}_1^t)-\kappa(x',z'|\hat{h}_1^t)}\mu^{i,\neg Z}(d\theta|\hat{h}_0^t)}{\prod_{x'}\prod_{z'}\theta_0(x',z')^{\kappa(x',z'|\hat{h}_1^t)}(1-\theta_0(x',z'))^{N(x',z'|\hat{h}_1^t)-\kappa(x',z'|\hat{h}_1^t)}}\right). \tag{D.25}$$

Since $\mu^{i,\neg Z}(\cdot|\hat{h}_0^t)$ places full support on vectors of success probabilities (θ) with $\theta(x,z)=\theta(x,z')$ for all x,z,z', we can bound (D.25) by noting that

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

which implies that (D.25) is bounded above by

$$\frac{1}{\#\mathscr{E}(t)}\log\left(\frac{\prod_{x'}\prod_{z'}\frac{\kappa(x'|\hat{h}_1^t)}{N(x'|\hat{h}_1^t)}^{\kappa(x',z'|\hat{h}_1^t)}\left(1-\frac{\kappa(x'|\hat{h}_1^t)}{N(x'|\hat{h}_1^t)}\right)^{N(x',z'|\hat{h}_1^t)-\kappa(x',z'|\hat{h}_1^t)}}{\prod_{x'}\prod_{z'}\theta_0(x',z')^{\kappa(x',z'|\hat{h}_1^t)}(1-\theta_0(x',z'))^{N(x',z'|\hat{h}_1^t)-\kappa(x',z'|\hat{h}_1^t)}}\right). \tag{D.26}$$

for all $t,\hat{h}^t$. By the strong law of large numbers, expression (D.26) can be shown to tend towards $-d_K(\hat{p}_{\underline{\theta}(M_{X,\neg Z})},p_0^1)<0$ with probability 1 conditional on the agent encoding z infinitely often; this establishes (D.24) and completes the proof. □