Updating Beliefs when Evidence is Open to Interpretation: Implications for Bias and Polarization*

Roland G. Fryer, Jr., Philipp Harms, and Matthew O. Jackson†

This Version: October 2017
Forthcoming in Journal of the European Economic Association

Abstract

We introduce a model in which agents observe signals about the state of the world, some of which are open to interpretation. Our decision makers first interpret each signal and then form a posterior on the sequence of interpreted signals. This 'double updating' leads to confirmation bias and can lead agents who observe the same information to polarize. We explore the model's predictions in an on-line experiment in which individuals interpret research summaries about climate change and the death penalty. Consistent with the model, there is a significant relationship between an individual's prior and their interpretation of the summaries; and - even more striking - over half of the subjects exhibit polarizing behavior.

JEL classification numbers: D10, D80, J15, J71, I30
Keywords: beliefs, polarization, learning, updating, Bayesian updating, biases, discrimination, decision making

*We are grateful to Sandro Ambuehl, Ken Arrow, Tanaya Devi, Juan Dubra, Jon Eguia, Ben Golub, Richard Holden, Lawrence Katz, Peter Klibanoff, Scott Kominers, Franziska Michor, Giovanni Parmigiani, Matthew Rabin, Andrei Shleifer, and Joshua Schwartzein for helpful comments and suggestions. Adriano Fernandes provided exceptional research assistance.

†Fryer is at the Department of Economics, Harvard University, and the NBER, (rfryer@fas.harvard.edu); Jackson is at the Department of Economics, Stanford University, the Santa Fe Institute, and CIFAR (jacksonm@stanford.edu); and Harms is at the Department of Mathematical Stochastics, University of Freiburg, (philipp.harms@stochastik.uni-freiburg.de). We gratefully acknowledge financial support from the NSF grants SES-0961481 and SES-1155302, and ARO MURI Award No. W911NF-12-1-0509.

1 Introduction

Some argue that the world is becoming more polarized (Sunstein 2009, Brooks 2012). Consider, for instance, Americans' views on global warming. In a 2003 Gallup poll, 68 percent of self-identified Democrats believed that temperature changes over the past century could be attributed to human activity, relative to 52 percent of Republicans (Saad 2013). By 2013, these percentages had diverged to 78 percent and 39 percent. Between 2013 and 2014 – before and after the deaths of Michael Brown and Eric Garner at the hands of police officers – a Gallup poll found that the percentage of non-whites who believed that the honesty and ethical standards of police officers were "very high" or "high" fell from 45 percent to 23 percent while non-hispanic whites beliefs remained constant (Gallup 2014) [1].

We introduce a model that provides a simple foundation for why the above-described polarizations in beliefs should be observed from rational agents. In its simplest version, there are two possible states of nature A, B. An agent observes a series of signals a, b correlated with the state of nature. Some signals are ambiguous and come as ab.[2] The difference from standard Bayesian agents, is that we assume that an agent does not store a sequence such as $a, b, ab, ab, a, ab, b, b \ldots$ in memory, but *first* interprets the ab signals as they are seen according to some other rule, and then stores the interpretation in memory. Imagine the agent started by believing that A was more likely, then the sequence would be interpreted and stored in memory as $a, b, a, a, a, a, b, b \ldots$. By storing the full ambiguous sequence $a, b, ab, ab, a, ab, b, b \ldots$ the agent would see more evidence for b than a.

We consider three interpretation strategies: (i) a *maximum likelihood strategy* that matches the ambiguous signal to the likelier state based on the prior; (ii) a *randomized-interpretation strategy* that matches the ambiguous signal to the likelier state based on the prior with a fixed probability and randomly chooses with the complementary probability; and (iii) a *two-step strategy* that follows (ii) for a fixed number of initial periods, subsequently reverting to (i). The deviation of our model from full Bayesian updating is relatively minimal: under any strategy above agents interpret ambiguous signals before storing them and then base their posteriors on their interpretations rather than all of the original data.

We demonstrate that if a large enough share of experiences are open to interpretation

[1] The importance of polarizing beliefs in part derives from the conflict they induce. In influential papers, Esteban and Ray (1994) and Duclos, Esteban, and Ray (2004) derive measures of societal polarization and methods to estimate them. Relatedly, Jensen et al. (2012) derive measures of political polarization. Later work has linked increased polarization to conflict both theoretically (Esteban and Ray 2011) and empirically (Esteban, Mayoral, and Ray 2013). These papers motivate understanding the mechanisms that can lead opinions to diverge even in the face of common information, since such divergence can result in society-level disruptions, intolerance, and discrimination.

[2] The word 'ambiguous' is often used in the decision theory literature to refer to settings in which people do not have 'standard' expected utility functions with prior beliefs -either objective or subjective. Here, we use the term to refer to signals that are not definitive evidence in one direction or another.

and agents always use the maximum likelihood strategy, then two agents who have differing priors and who see *exactly* the same sequence of evidence can, with positive probability, end up polarized with one agent's posterior tending to place probability 1 on A and the other tending to place probability 1 on B. We also show that, under a randomized-interpretation strategy, for a large enough chance of following the prior, beliefs converge to the wrong state with positive probability. In stark contrast, under a two-step strategy, sufficiently patient agents approach the full information benchmark.

We extend the model to normally distributed signals and states. In that model, an agent's belief always converge to something that is biased towards the prior and early signals and different from the true state. As signals become more accurate compared to the prior, that bias decreases, and disappears in the limit of perfectly informative signals. With any fixed signal accuracy, all agents' limit beliefs can be strictly ordered based on their priors.

We explore the model's implications in an on-line experiment in which individuals read research summaries, and then interpret these and provide updated beliefs. This method of experimentation was pioneered by Lord, Ross, and Lepper (1979). Although similar in basic structure, our experiments do not preselect subjects based on extreme beliefs, and we test a fuller range of sorts of research summaries, including some that are more ambiguous. Perhaps most importantly, rather than having 48 undergraduate subjects, we have over 600 with much broader backgrounds. Our richer experimental design, allows us to examine many questions not studied before in this literature, and with significantly greater accuracy.[3]

Experiments were conducted through Amazon Mechanical Turk (MTurk). First, participants were asked four questions designed to check whether they were paying attention rather than blindly clicking through. Second, participants were presented with a question about their beliefs on the first topic (either climate change or the death penalty, the order of which was randomly assigned). Then, individuals read a series of summaries (redacted from abstracts and introductions of published academic papers, and edited to improve readability). After each summary, we asked participants whether they thought the summary provided evidence for or against the topic on a 16 point scale. Participants could not go back to previous screens, and the summaries and questions were on different screens. After all of the summaries were presented, we repeated the initial question on their beliefs about the topic. The participant then moved on to the next topic, which followed the same format. After reading and evaluating both sets of summaries, participants answered a number of questions on demographics. Payments for the survey ranged between \$0.40 and \$6.

The results of our experiment are largely in line with the predictions of the model. First, there is a significant correlation between an individual's prior belief and their interpretation

[3]Another important distinction is that we used real research articles rather than fictitious studies as the basis for our research summaries.

of evidence. Going from one extreme of the prior belief distribution to the other, implies a 0.8 standard deviation change in interpretation. Second, the bias in interpretation is stable in magnitude (varying insignificantly) when estimating the relationship between an individual's prior belief and their interpretation, conditioning on a variety of demographics, such as gender, education, income, and political affiliation.

Third, individuals react to the information given to them (even though it is, on average, neutral) and we find evidence that beliefs about climate change become significantly more extreme. A test of equality of distributions of subjects' prior and posterior beliefs has a p-value of 0.00 for climate change and 0.072 for death penalty. A variance ratio test has a p-value of 0.04 for climate change and 0.45 for the death penalty. Fourth, we also see evidence at the individual level of polarization: more than fifty percent of our sample (for at least one of the two topics) move their posterior belief further from the average belief after seeing the series of abstracts. This is consistent with our model, but would need to be embedded in a richer multi-dimensional space in which people's knowledge is sufficiently heterogenous and interacts with the current dimension (as in Benoit and Dubra 2014) to be rationalized by Bayesian updating[4].

Our experimental data are consistent with a rich literature in psychology which has long recognized humans' propensity to over-interpret evidence that reinforces their beliefs while disregarding contradictory information. A classic example is Darley and Gross (1983). A summary of many of these studies can be found in Appendix Table 1.

The paper proceeds as follows. Section 2 provides a brief literature review. Section 3 describes a framework for updating beliefs when evidence is unclear and uses this insight to understand potential mechanisms that drive polarization, Section 5 describes our experiments and Section 6 presents the results. The final section concludes. There are three appendices. Appendix A contains the proofs of all formal results as well as additional results. Appendix B describes the data collected and how we constructed the variables used in our analysis. Appendix C contains additional analysis of the data, the summaries used in the experiment, and the experimental instructions.

[4]More generally, no experiment can really reject Bayesian updating, as subjects could have priors in some much richer world-view that tell them if they see circumstances like those in the experiment, then it must have been generated by a posterior with the observed marginal distribution on the data. By embedding the experimental setting in a much richer world-view, Bayesian updating becomes non-falsifiable. Thus, we can reject that subjects had priors on just the one dimension that we asked opinions upon and model, but we cannot reject that they are updating in some richer way.

2 A Brief Review of the Literature

Rabin and Schrag (1999) provide a first decision-making model of confirmation bias.[5] In each period agents observe a noisy signal about the state of the world, at which point they update their beliefs. In their model, signals that are believed to be less likely are misinterpreted with an exogenous probability. Given the exogenous mistakes, agents can converge to incorrect beliefs if they misinterpret contradictory evidence sufficiently frequently. The model does not clarify the mechanism behind the misinterpretation.

Our model provides a foundation for the interpretation and storage of ambiguous information, a "why" behind long term bias, and how this can also lead to belief polarization.[6] We also explore the implications of our foundations in settings with continuous signal distributions, showing that bias *always* occurs, and depends on early signals (not just the prior).

Hellman and Cover (1970) provides insights into how restrictions on the updating process motivated by the psychology literature can yield biases in beliefs (see also Wilson 2014, Mullanaithan 2012, Gennaioli and Shleifer 2010, and Glaeser and Sunstein 2013)[7]. In their limited-memory model, agents observe a possibly infinite sequence of i.i.d. signals according to a time-invariant probability measure. Their goal is, given their sequence of signals, to identify one of two true states A, B. They model the decision problem of the agent as a $M-$state (possibly reducible) automaton, where M is finite. The automaton follows a Markov process driven by the underlying signal process.

The approximately optimal transition rule (i.e. minimizing the expected asymptotic proportion of errors) for the automaton first ranks memory states from 1 to M, with higher numbers indicating an increased likelihood that, without loss of generality, state A is true. Following a signal in favor of state $A(B)$, the machine shifts to the next higher (lower) state until an extreme states ($\{1, M\}$) is reached. If transition probabilities out of extreme states are close to zero relative to transition probabilities in interior states, near-optimality holds.

Wilson (2014) recasts Hellman and Cover (1970) under a more explicit decision-theoretic framework, allowing for the process to terminate probabilistically. She then derives fully constrained optimal rules (subject to the finiteness of the automaton): agents operate similarly to the Hellman and Cover (1970) automaton, but when optimizing facing a low probability of termination, will move only one memory state at a time (deterministically), will react only

[5] See also Dandekar, Goel and Lee (2013) for a network model of confirmatory bias.

[6] We focus on implications of such updating for biases in beliefs and polarization. Gilboa and Schmeidler (1993) and Siniscalchi (2011) provide more general foundations on belief updating with ambiguity.

[7] Related papers explicitly account for costly updating (e.g. Kominers, Mu, and Peysakhovich 2015), in which agents discard information whenever the cost of updating surpasses the perceived benefit; or for limited attention (e.g. Schwartzstein 2014). Such models can lead to persistent biases in updating, as not all information is incorporated. Our model is close to these models in terms of what drives biases in updating, although our results and the focus of our analysis are different.

to extreme signals, and will leave extreme states with very low probability. This can yield polarization under several scenarios. For instance, two people who receive the same set of signals but start at different memory states (optimal relative to their prior beliefs) can end up in different places for long periods of time, as they can get temporarily 'stuck' at one of the extreme states. More generally, as long as agents do not have identical priors and do not start in an extreme state, their beliefs can differ for long periods of time with positive probability, although beliefs will not polarize permanently.

Baliga, Hanany and Klibanoff (2013) provide a very different explanation for polarization based on ambiguity aversion. In their model, agents with different prior beliefs may update in different directions after observing some intermediate signals due to a "hedging effect" - agents wish to make predictions immune to uncertainty, and may be averse to different directions of ambiguity given their differing priors. Papers by Andreoni and Mylovanov (2012) and Benoit and Dubra (2014) are based on multidimensional uncertainty where observers may differ in terms of their knowledge about the different dimensions (which might be thought of as models of the phenomenon in question) leading them to update differently based on the same information. The major difference is that those models build from ambiguity or some other interacting dimensions of uncertainty that can cause agents to update in different directions even if they are Bayesian, whereas ours is based on a form of bounded rationality that can explain polarization entirely within a single ordered dimension of uncertainty.

Again, a critical distinction is that our model can result in permanent polarization, while models based on full rationality will converge to a common and accurate belief given rich enough observations. This distinction is also true when comparing our model to other boundedly rational models (e.g. Hellman and Cover 1970, Wilson 2014). In those models beliefs are ergodic: agents follow a similar irreducible and aperiodic Markov chain, simply starting in different states, and their limiting distributions would coincide, but their state at various times could diverge (infinitely often). Our analysis differs from those analyses as our agents become increasingly polarized and increasingly certain that they are each correct despite their disagreement, after seeing arbitrarily informative sequences.

Also, in the continuous version of our model, agents' beliefs always converge, and with probability 1, to a wrong posterior - something very different from the previous literature - so that two agents with different priors will always disagree in the limit. This also depends not just on the prior: early signals have a disproportionate sway in forming people's perceptions.

3 A Model of Updating Beliefs when Experiences are Open to Interpretation

There are two possible states of nature: $\omega \in \{A, B\}$. An agent observes a sequence of signals, s_t, one at each date $t \in \{1, 2, \ldots\}$. The signals lie in the set $\{a, b, ab, \emptyset\}$. A signal a is evidence that the state is A, a signal b is evidence that the state is B, a signal ab is open to interpretation, and the signals $\emptyset$ contain no information. Signals are conditionally (on the state) i.i.d.

With probability $1 - q$, there is no signal, denoted by $\emptyset$. With probability q, the signal realization starts out as either a or b, and matches the state with probability $p > 1/2$. However, prior to the agent's observation, there is an exogenous probability π that the signal realization gets distorted into an ambiguous ab. The symmetry between A and B both having the same fraction of a and b matching the state is simply for notational convenience, the assumption is not needed. Let $\lambda_0 \in (0, 1)$ be the agent's prior that $\omega = A$.

It follows directly from a standard law of large numbers argument (e.g. Doob's (1949) consistency theorem), that a Bayesian-updating agent who forms beliefs conditional upon the full sequence of signals has a posterior that converges to place probability 1 on the correct state, almost surely.

The signals ab are uninformative as they occur with a frequency $q\pi$ *regardless of the state*. Thus, a Bayesian updater who remembers all of the signals in their entirety ignores interactions that are open to interpretation.

3.1 Interpreting Ambiguous Signals

Although ambiguous signals are uninformative and should be ignored in the long run, there is a true signal underlying each one that can still be interpreted for the short run. An agent may be prone or required to make quick judgements immediately after perceiving ambiguous information in light of her current beliefs. We assume that she does not later go back and remember all the ambiguity that was present when updating her beliefs due to limited memory. The *interpretation* of ambiguous signals $s_t = ab$ as an a or b is based on the agent's experiences through time t.[8]

As we show below, this is approximately optimal if the agent is faced with making decisions in the short run that depend on how they interpret the situations they face. The bounded rationality relative to the longer run is that the agent updates based on the inter-

[8]This is a key departure from previous models. Agents in the Kominers, Mu, and Peysakhovich (2015) model may simply ignore the information if the cost of updating is larger than the perceived benefit. Similarly, individuals in the Schwartzstein (2014) framework may not notice ambiguous information due to selective attention.

preted signals, rather than simply using the interpretation of an ambiguous signal for the short-run interaction, but then not updating based on it.

3.2 An Example of Polarized Beliefs

Consider the following example. Suppose the true state is $\omega = A$, the probability that an unambiguous signal matches the state is $p = 2/3$, and the fraction of ambiguous signals is $\pi = 1/2$. The probability of observing a signal is $q = 4/5$. In such a case, the sequence of signals might look like: $a, ab, \emptyset, ab, b, a, ab, a, \emptyset, ab, ab, b, \ldots$ Consider the case such that the agent interprets $s_t = ab$ based on what is most likely under her current belief λ_{t-1}: the agent interprets $s_t = ab$ as a if $\lambda_{t-1} > 1/2$ and as b if $\lambda_{t-1} < 1/2$. The agent ignores non-signals.

Suppose that the agent's prior is $\lambda_0 = 3/4$. When faced with $s_1 = a$, the agent's posterior λ_1 becomes $6/7$. So, the agent updates beliefs according to Bayes' rule given the interpreted (remembered) signals.[9] Then when seeing $s_2 = ab$ the agent stores the signal as a, and ends up with a posterior of $\lambda_2 = 12/13$. The stored sequence is then $a, a, \emptyset, a, b, a, a, a, \emptyset, a, a, b, \ldots$, and the posterior at the end is very close to 1.

In contrast, consider another agent whose prior is $\lambda_0 = 1/4$. When faced with $s_1 = a$, the agent's posterior λ_1 becomes $2/5$. Then when seeing $s_2 = ab$ the agent stores the signal as b, and ends up with a posterior $\lambda_2 = 1/4$. Now, the stored sequence is $a, b, \emptyset, b, b, a, b, a, \emptyset, b, b, b, \ldots$, and the posterior at the end is very close to 0.

Two agents observing exactly the same sequence with different (and non-extreme) priors polarize, that is, come to have increasingly different posteriors:[10]

$$\begin{array}{l} 3/4,\ 6/7,\ 12/13,\ 12/13,\ 24/25,\ 12/13,\ 24/25,\ 48/49,\ 96/97, \ldots \\ 1/4,\ 2/5,\ 1/4,\quad 1/4,\quad 1/7,\quad 1/13,\quad 1/7,\quad 1/13,\quad 1/7, \ldots \end{array}$$

3.3 Approximately Optimal Interpretations in the Face of Choosing an Action

In the above example, we considered situations where the agent interprets signals according to which state is more likely, which we referred to as the maximum likelihood rule. There is a tradeoff: under this rule, agents can react to the current situation, but then bias long-term learning. To formalize the tradeoff between immediate action and long-term learning under our limited-memory cognitive limitation, we extend the model to explicitly include

[9] In this case $\lambda_t = P(A|s_t = a, \lambda_{t-1}) = 2\lambda_{t-1}/(1 + \lambda_{t-1})$, and $\lambda_t = P(A|s_t = b, \lambda_{t-1}) = \lambda_{t-1}/(2 - \lambda_{t-1})$

[10] Proposition 3, below, shows the example occurs in a variety of settings, and not only do the agents become polarized, but they remain that way with positive probability, each converging to different beliefs.

the agent making choices over time, seeking to maximize the discounted sum of expected payoffs for correctly identifying the true state.

Imagine the agent has to choose either a or b at each date, including ones at which $s_t = ab$. She gets payoff of $u_t = 1$ if the current situation is correctly identified and $u_t = 0$ when a mistake is made.[11] When the signal is unambiguous, choice is easy. When the signal is ab, if the agent calls out a, then $u_t = 1$ with probability p if the state is A and probability $(1-p)$ if the state is B. Similarly, if the signal is ab and the agent calls out b, then $u_t = 1$ with probability $(1-p)$ if the state is A and probability p if the state is B.

The agent does not get immediate feedback on the payoffs. What is important, is that the agent must make a choice and remembers the choices made, but does not change beliefs in cases in which the decisions were incorrect. When there is no information - essentially no decision to be made, the agent needs not take any action.

To represent the full set of possible strategies that an agent might have for making interpretations (including strategies for agents with unbounded memories), we define histories. Let a history $h_t = (s_1, i_1; \ldots; s_t, i_t) \in \{\{\emptyset, a, b, ab\} \times \{\emptyset, a, b\}\}^t$ be a list of raw signals as well their interpretations through time t. Let $H_t = \{\{\emptyset, a, b, ab\} \times \{\emptyset, a, b\}\}^t$ be all the histories of length t and $H = \cup_t H_t$ be the set of all finite histories.

A *strategy* for the agent is a function σ that can depend on the history and the agent's beliefs and generates a probability that the current signal is interpreted as a: $\sigma : H \times [0,1] \to [0,1]$.[12] In particular, $\sigma(h_{t-1}, \lambda_0)$ is the probability that the agent interprets an ambiguous signal $s_t = ab$ as a conditional upon the history h_{t-1} and the initial prior belief λ_0. A *limited-memory strategy* is a strategy that depends only on interpreted and not on raw signals.[13]

An agent's expected payoffs can be written as: $U(\sigma, \delta, \lambda_0) = E\left(\sum_{t=1}^{\infty} \delta^t u_t(\sigma(h_{t-1}, \lambda_0))|\lambda_0\right)$.

The optimal strategy in the case of unconstrained memory is that of a fully rational Bayesian. It can be written solely as a function of the posterior belief, as the history that led to the posterior is irrelevant.

Our limited-memory strategies exhibit interesting history dependencies. A sequence can be reshuffled and will not affect Bayesian updating with unbounded memory. However, with bounded memory, the order of observation becomes important. Seeing a sequence where the a's all appear early tilts the prior towards the state A which then affects the interpretation of ab's towards a's. In contrast, reshuffling a sequence towards having the b's early has the opposite effect. For instance, in our example in Section 3.2, $a, ab, \emptyset, ab, b, a, ab, a, \emptyset, ab, ab, b$, was interpreted as $a, a, \emptyset, a, b, a, a, a, \emptyset, a, a, b, \ldots$ by anyone starting with a prior above $1/2$.

[11] Given the binary setting, the normalization is without loss of generality.

[12] Given that we allow the strategy to depend on the history, it is irrelevant whether we allow it to depend on the current posterior or original prior, as either can be deduced from the other given the history.

[13] A strategy is limited-memory if $\sigma(h_t) = \sigma(h'_t)$ whenever the even entries (the interpreted signals, i_t's) of h_t and h'_t coincide.

Suppose we reorder that original sequence to be $b, b, \emptyset, ab, ab, ab, ab, ab, \emptyset, a, a, a$. With the same prior in favor of A, but sufficiently close to 1/2, the interpretation would instead flip to be $b, b, \emptyset, b, b, b, b, b, \emptyset, a, a, a$, and end up pushing the beliefs towards B. So, the order in which a sequence of signals appear can now be consequential.

An optimal limited-memory strategy is one that adjusts the probability of interpreting a signal with the posterior, but appears difficult to derive in a closed form. Nonetheless, we can find strategies that approximate the optimal strategy when the agent is sufficiently patient or impatient. To this aim, we will compare several classes of strategies: (i) the approximately optimal strategy, (ii) ones that depend only on the time, and (iii) ones that involve randomizing in a simple manner based on the posterior.

In the latter class of strategies, the agent randomizes in interpreting ambiguous signals, but with a fixed probability that leans towards the posterior. Let $\gamma \in [0, 1]$ be such that the agent follows the posterior with probability γ and goes against the posterior with probability $1 - \gamma$. Under such a strategy, at time t with posterior λ_{t-1}, the agent interprets the unclear signal as a with probability $\gamma 1_{\lambda_{t-1} \geq .5} + (1-\gamma) 1_{\lambda_{t-1} < .5}$ and b with the remaining probability.[14] We denote this strategy by σ^γ.

The special case of $\gamma = 1$ corresponds to maximum likelihood interpretation, which we now show is an approximately optimal rule if the agent cares relatively more about the short run than the long run.

PROPOSITION **1** *If the agent's discount factor is small enough* or *the prior belief is close enough to either 0 or 1, then the maximum likelihood strategy is approximately optimal. That is, for any λ_0 and $\varepsilon > 0$, there exist $\overline{\delta}$ such that if $\delta < \overline{\delta}$, then $U(\sigma^1, \delta, \lambda_0) \geq U(\sigma, \delta, \lambda_0) - \varepsilon$ for all strategies σ. Moreover, the same statement holds for any δ for λ_0 that are close enough to 0 or 1.*

The intuition is straightforward. The underlying tension is between correctly calling the state in the short run, and interpreting signals over the long run. If the decision maker is sufficiently impatient then it is best to make correct decisions in the short run and not worry about long-run learning. The last part of the proposition shows that even if the agent is very patient, if the agent begins with a strong prior in one direction or the other (λ_0 near either 0 or 1), then maximum likelihood is again approximately optimal as the agent does not expect to learn much.

At the other extreme, by setting $\gamma = 1/2$, then the agent will learn the state with probability 1 in the long run. However, that is at the expense of making incorrect decisions at many dates. More generally, we can state the following proposition. Let us say that

[14] It is straightforward to make the updating function more continuous around .5, with no qualitative impact on the results.

there is *long-run learning* under the randomized-interpretation strategy σ^γ for some γ if the resulting beliefs λ_t converge to the true state almost surely.

PROPOSITION **2** *Consider a randomized-interpretation strategy σ^γ for some γ. If $\pi < \frac{p-1/2}{p}$, then ambiguous signals are infrequent enough so that there is long-run learning regardless of γ. If $\pi > \frac{p-1/2}{p}$:*

(a) then if $\gamma < \frac{p-1/2+\pi(1-p)}{\pi}$ there is long-run learning,

(b) but if $\gamma > \frac{p-1/2+\pi(1-p)}{\pi}$ then there is a positive probability that beliefs λ_t converge to the wrong state.

As just discussed, the case $\gamma = 1$ corresponds to a maximum-likelihood strategy, which is approximately optimal for a standard Bayesian. However, as exemplified in 3.2, our signal-interpreting agents may exhibit polarized beliefs after observing the same sequence of signals under a maximum likelihood strategy. This result generalizes as follows:

PROPOSITION **3** *Suppose that a nontrivial fraction of experiences are open to interpretation so that $\pi > \frac{p-1/2}{p}$. Consider two interpretative agents 1 and 2 who both use the maximum likelihood rule but have differing priors: agent 1's prior is that A is more likely (so 1 has a prior $\lambda_0 > 1/2$) and agent 2's prior is that B is more likely (so 2 has a prior $\lambda_0 < 1/2$). Let the two agents see exactly the same sequence of signals. With a positive probability that tends to 1 in π the two agents will end up polarized with 1's posterior tending to 1 and 2's posterior tending to 0. With a positive probability tending to 0 in π the two agents will end up with the same (possibly incorrect) posterior tending to either 0 or 1.*

The proof is based on the observation that when $\pi = 1$ and $\lambda_0 > 1/2$, then all signals are interpreted as a under the maximum likelihood storage rule. Moreover, the law of the belief process depends continuously on π.

Proposition 3 builds on Proposition 5(i) in Rabin and Schrag (1999). In their model, an agent ends up confirming the prior if there is a high enough fraction of times that the agents confirm their bias, and here it happens if enough signals are open to interpretation - as maximum likelihood and ambiguous signals (when the underlying signal was against the belief) are mathematically equivalent to reversing the belief in this two-state model. The above proposition then notes that this then implies that two different agents seeing the same sequence of signals can come to very different conclusions - so these sorts of models provide a foundation for polarization.

Since interpreting signals can lead to wrong long-term conclusions, maximum likelihood strategies in all time period would not be approximately optimal for patient agents. Appropriately randomized strategies allow for long-run learning and also avoid polarization. In

what follows, we consider a two-step rule that approximates a full information ideal benchmark. Our study of approximately optimal strategies further differentiates our approach from Rabin and Schrag (1999).[15]

3.4 Approximately Optimal Strategies: Two-Step Rules

Proposition 2 shows long-run learning under a randomized-interpretation strategy only occurs if randomization is sufficiently high. This can be very costly in the long run, as although the posterior converges, the agent is randomly interpreting ambiguous signals, even when she is almost certain of how they should be interpreted.

The optimal strategy should adjust the randomization with the belief: as the agent becomes increasingly sure of the true state, the agent should become increasingly confident in categorizing ambiguous signals, and use less randomization. The fully optimal strategy is difficult to characterize as it is the solution to an infinitely nested set of dynamic equations and we have not found a closed-form. Nonetheless, we can find a strategy that approximates the optimal strategy when the agent is sufficiently patient.

Consider a T-period *two-step* rule, defined as follows. For T periods the agent uses $\gamma = 1/2$, and after T periods the agent uses $\gamma = 1$. Let σ^T denote such a strategy. As a strong benchmark, let σ^{FI} denote the *full-information* strategy, where the agent actually knows whether the state is A or B - when seeing ab then always calls a if $\omega = A$ (or b if $\omega = B$). In this case, expected utility is independent of λ_0 and can be written as a function of the strategy and the discount factor alone: $U(\sigma^{FI}, \delta)$. This is a very stringent benchmark as it presumes information that the agent would never know even under the best circumstances. Even so, we can show that a simple two-step rule can approximate this benchmark.

PROPOSITION **4** *Sufficiently patient agents can get arbitrarily close to the full information payoff by using a two-step strategy: for any ϵ and λ_0, there exist T and $\overline{\delta}$ such that if $\delta > \overline{\delta}$, then $U(\sigma^T, \delta, \lambda_0) > U(\sigma^{FI}, \delta) - \epsilon$.*[16]

Putting these two results together, how a subjective degree of belief should change "rationally" to account for evidence which is open to interpretation depends, in important ways, on how patient a decision maker is. If they are sufficiently patient, a strategy that entails

[15] Our model and Rabin and Schrag (1999) (hereafter, RS) are exploring different foundations that allow for agents to maintain incorrect beliefs. RS presents a model of confirmatory bias, whereas our model studies a boundedly rational decision-theoretic foundation, following or not their belief process when forced to take actions at each point in time with a given probability. Our model proposes a cognitive framework of bounded rationality behind the q of RS, and also examines enrichments of that, as well as its very different implications in the continuum case.

[16] Proposition 4 also holds for a slightly different strategy: σ^x which is defined by setting $\gamma_t = 1/2$ until either $\lambda_t > 1 - x$ or $\lambda_t < x$, and then setting $\gamma_t = 1$ forever after. Instead of holding for a large enough T and δ, the proposition then holds for a small enough x and large enough δ.

randomization in the presence of unclear evidence for finite time and then following their maximum likelihood estimate thereafter approximates the full information outcome.[17]

In stark contrast, if agents sufficiently discount the future (or, equivalently, don't expect a large number of similar interactions in the future), they interpret ambiguous evidence as the state that has the highest maximum likelihood, given their prior belief. This leads to more informed (and fully optimal, though possibly mistaken) decisions in the short-run, but the potential of not learning over the longer-run.

An important remark is that we have specified two-step rules in terms of time. An alternative method is to do the following. If an agent's beliefs λ_t place at least weight ε on both states ($\lambda_t \in [\varepsilon, 1-\varepsilon]$), then randomize with equal probability on interpretations, and otherwise use the maximum likelihood method. This strategy does not require any attention to calendar time and also will approximate the full information strategy for small enough ε.

4 A Model with Normally Distributed Signals

To make our key insights transparent, the model above considered discrete signals {a,b,ab,$\emptyset$}. Complex information, such as the information contained in the research articles in the forthcoming experiment, is less of an "a" or "b" form, and is almost always open to interpretation. To illustrate how the updating that we have studied may work in such settings, we consider a model with normally distributed states and signals. This serves two purposes: first, a continuous model applies to many settings (e.g. our experiments below). Moreover, it should become clear that the insights obtained from this tractable version of the model will have analogs for often less tractable distributions.

The true state is denoted by $\mu \in \mathbb{R}$. An agent begins with a prior μ_0 which is the expectation of nature's mean based on a normal distribution over potential means with a variance σ_0^2. Signals are denoted s_t and are i.i.d. according to a normal distribution centered around the true mean μ and with variance σ_s^2: $N(\mu, \sigma_s^2)$. Let μ_t denote the posterior of a Bayesian updater after t signals, and σ_t^2 the associated variance.

An agent in our model first interprets the signal given his or her prior and then updates his or her beliefs. Everything else is done as in the Bayesian case. Let $\widehat{\mu}_t$ denote the posterior of an agent in our model, and $\widehat{s}_t$ the interpreted signal, where $\widehat{s}_t = \frac{\widehat{\mu}_{t-1}+x_t s_t}{1+x_t}$ and $\widehat{\mu}_t = \frac{\widehat{\mu}_{t-1}+x_t \widehat{s}_t}{1+x_t}$.

Thus, the agent first interprets the signal, moving it closer to the previous belief; then updates that belief, essentially weighting it twice. It is straightforward to demonstrate that an updater in our setting ends up overweighting the previous belief and underweighting the signal relative to a Bayesian updater.

[17] This result highlights a familiar tradeoff between exploitation (being correct in the short run) and experimentation (long-run learning) typical in the multi-armed bandit literature (Berry and Fristedt 1985).

PROPOSITION **5** *An agent's posterior $\widehat{\mu}_t$ is given by*

$$\widehat{\mu}_t = \left[\mu_0\left(\frac{\sigma_s^2}{\sigma_s^2+\sigma_0^2}\right) \; + \; \sum_{\tau=1}^{t} s_\tau \left(\frac{(\sigma_0^2\sigma_s^2)^2}{(\sigma_s^2+(\tau-1)\sigma_0^2+\sigma_0^2\sigma_s^2)^2}\right)\left(\frac{\sigma_s^2+\tau\sigma_0^2}{\sigma_s^2+(\tau+1)\sigma_0^2}\right)\right]\left[1+\frac{\sigma_0^2}{\sigma_s^2+t\sigma_0^2}\right].$$

Thus the agent places weight $\frac{\sigma_s^2}{\sigma_s^2+\sigma_0^2}\left[1+\frac{\sigma_0^2}{\sigma_s^2+t\sigma_0^2}\right]$ on the prior, which converges to $\frac{\sigma_s^2}{\sigma_s^2+\sigma_0^2}>0$ as t grows. The weight on any signal is decreasing in time, but does not vanish in the limit, converging to $\left(\frac{\sigma_0^2\sigma_s^2}{\sigma_s^2+(\tau-1)\sigma_0^2}\right)^2\left(\frac{\sigma_s^2+\tau\sigma_0^2}{\sigma_s^2+(\tau+1)\sigma_0^2}\right)$.

To highlight the contrast between the model's posterior and a standard Bayesian posterior[18], note that if $\sigma_0^2=\sigma_s^2$, then the former simplifies to

$$\widehat{\mu}_t = \left[\mu_0\frac{1}{2}+s_1\frac{1}{6}+s_2\frac{1}{12}+s_3\frac{1}{20}+\ldots+s_t\left(\frac{1}{(1+t)(2+t)}\right)\right]\left[1+\frac{1}{1+t}\right],$$

while the latter posterior simplifies to $\mu_t=\left[\mu_0+\sum_{\tau=1}^{t}s_\tau\right]\left[\frac{1}{t+1}\right]$.

Thus, in this case, the Bayesian equally weights the prior and all subsequent signals, and none have any eventual weight. In contrast, in our model, the prior still holds weight and pulls signals towards it. Earlier signals are heavily weighted, as early beliefs are less entrenched. However, over time signals get interpreted more, thus matter less. [19]

5 An On-line Experiment

Amazon's Mechanical Turk (MTurk) provides access to a large and diverse pool of subjects that is increasingly being used for experiments.[20]

5.1 Survey Design and Format

We identified published research articles on two subjects: the death penalty and climate change. In particular, we identified articles that provided a variety of conclusions and would

[18] In this case, it is well-known that $\mu_t=\mu_0\frac{\sigma_s^2}{\sigma_s^2+t\sigma_0^2}+\sum_{\tau=1}^{t}s_\tau\frac{\sigma_0^2}{\sigma_s^2+t\sigma_0^2}$.

[19] The agent in this version of our model always has their beliefs converge, but with probability one the beliefs converge to something that is biased towards the prior and early signals. As signals become more accurate compared to the prior, that bias decreases, and in the limit the bias disappears, but with any fixed signal accuracy, the bias remains.

[20] See Horton, Rand, and Zeckhauser (2011) for some discussion of the advantages of using such online subject pools. A recent study compared MTurk samples with standard Internet samples and found similar gender and American vs. non-American distributions (Buhrmester, Kwang and Gosling 2011). However, a greater percentage of mTurk participants were non-White (36% vs 23%); the average MTurk participant was also older (32.8 vs 24.3 years) than the internet sample. Overall, MTurk participants were more diverse than standard internet samples and American college samples.

be easy to understand. After selecting articles, we redacted the abstracts and introductions into a short summary of each article - trying to keep each summary at more or less the same length and level of readability. We tested the summaries on Amazon Turk and had people rate whether they thought the summaries were 'pro' or 'con' or neutral/ambiguous on a 16 point scale.[21] For each topic, we chose two summaries that had been rated strongly pro, two summaries that had been rated strongly con, and two summaries that people rated as being in the middle.[22] We conducted all of the surveys via Qualtrics, a platform for designing and administering online surveys.

Our model of biased updating can be thought of in two parts, and we test each of these parts. In the first part, an agent sees ambiguous signals and interprets these signals according to the agent's prior beliefs. In the second part, the agent stores this signal and not the accompanying uncertainty and updates her beliefs based on the stored signal.

In particular, to do the testing, our 608 participants were presented with the six summaries mentioned above on each of the two issues. The participants were asked their beliefs regarding the issues before the summaries, then asked their interpretation of the summaries, and then again asked their beliefs after having read the summaries. The first part of the theory is examined by seeing how the prior beliefs influence the interpretation of the summaries, and the second part of the theory is examined by seeing how the posterior beliefs change in response to reading the summaries.

Each experiment began with four practice questions to get participants comfortable with the format of the questions. These questions also allowed us check if participants were reading the survey or just clicking through. Then participants were presented with a question about their beliefs on the topics (randomized in order across subjects):

- "Do you think the death penalty deters (stops) people from committing murder?"
- "Do you think human activity is the cause of rising temperatures?"

Respondents were asked to answer on a scale from -8 (I am certain that the death penalty does NOT deter people from committing murder/I am certain that human activity is NOT the cause of increasing temperatures) to 8 (I am certain that the death penalty DOES deter people from committing murder/I am certain that human activity IS the cause of increasing temperatures).

[21] Respondents were asked to answer the following question: "What kind of evidence does this source provide about whether the death penalty deters people from committing murders?" or "Which of the following best describes the summary?" Answer choices range from -8:"...the death penalty does NOT deter people from committing murder"/"This summary provides strong evidence that human activity is NOT the cause of increasing temperatures to +8:"...the death penalty DOES deter people from committing murder"/"This summary provides strong evidence that human activity IS the cause of increasing temperatures.

[22] See the Online Appendix for full surveys and exact text of the summaries.

Next, we presented the short summaries. After each summary, we asked the respondents to decide whether they thought the summary provided evidence for or against the topic. For example, the death penalty summaries were followed by the statement "This summary provides evidence that..." and had to select from a scale of -8 (The death penalty does NOT deter people from committing murder) to +8 (The death penalty DOES deter people from committing murder). Importantly, participants could not go back to previous screens and the summaries and questions were on different screens. This meant that participants were forced to answer based on their best recollection of the summary and could not go back to look for a definitive answer. After all the summaries were presented, we repeated the initial question on their beliefs about the topic. Then the participant moved on to the next topic, which followed the same format. After reading and evaluating both sets of summaries, participants answered a number of questions on demographics. Payments for the survey ranged between $0.40 and $6 for different variations.

5.2 More details on the design

A common concern with Amazon MTurk is that some respondents may not take a survey seriously and instead click through as quickly as possible to finish and earn money. We tried to limit this in two ways. First, we only accepted workers who had previous experience and had at least a 98 percent approval rating. This meant that they had familiarity with the system, and that almost every task that they had completed was considered by requestors to be satisfactory. Second, we used practice questions, which clearly had very simple right and wrong answers, to monitor whether or not people were paying attention. To address a concern that these questions might be too simple, we ran a set of treatments with a more selective screening mechanism: Before the survey began, these participants were presented with three short summaries on gluten sensitivity. After each summary, participants were asked two simple reading comprehension questions. Answering any of the six reading comprehension questions incorrectly resulted in a payment of $0.40 to $0.90 and ejection from the survey. If the participant answered all the questions correctly, they could advance to the actual survey. Participants received $6 if they passed the check questions and completed the entire survey. The majority of people who made it through the screening also completed the full survey. Since our payment was so high by Amazon MTurk standards, we were able to get a reasonable number of participants who attempted and completed the task.The final 127 subjects did not differ significantly from the 481 earlier ones (see Table 1 for details). Throughout the paper, we report results from both samples.

We also tried to prevent the same subjects from appearing more than once in the data. We clearly stated in our instructions that individuals who had already completed one of our surveys were not eligible to take others. We had a screening method that compared their

Amazon ID to all those that had been previously entered and only allowed them to continue if the ID had not been used. Despite this, some people entered incorrect IDs and were able to get around the screening method. If we were able to determine that an individual did this, we did not pay them, gave them a negative rating, and dropped them from our data set. Twenty-one individuals were dropped for (un)intentionally entering invalid IDs.

Table 1 presents our summary statistics. Our final dataset consists of 608 participants who entered valid Amazon MTurk IDs. Approximately 44 percent of participants were female, 80 percent were white, 38 percent were Christian, and 50 percent reported earning a college degrees. Ages in the sample range from 19 to 75, with an average age of 34 years. Average yearly income was approximately $30,000.

6 Experimental Results

6.1 Evidence on the Interpretation of Information

Recall that we labeled summaries as pro, con, or unclear based on our pilot results. In our experiment, of the six summaries that we used in each case, 1 and 2 were seen as pro (with a significantly positive average interpretation), 3 and 4 were seen as con (with a significantly negative average interpretation), and only one of 5 and 6 was seen as 'unclear' (having an average interpretation indistinguishable from 0). For Climate Change (Death Penalty), question 5 (6) was seen as con (significantly negative on average), while 6 (5) was indistinguishable from 0. We also report individual analyses for every question in Appendix Table 2, and the results are unaffected by this grouping.

Table 2 presents estimates of such a linear relationship between an individual's prior and their interpretation of the summaries. The estimating equations are of the form:

$$\text{Interpretation}_i = \alpha + X\beta + \gamma * \text{Prior Belief}_i + \epsilon_i,$$

where i indexes individuals, X are demographic controls. Odd numbered columns correspond to the full sample while even numbered columns correspond to the restricted sample (in which subjects had to answer several simple questions in order to be admitted). In terms of our model with normal signals, γ corresponds to $\frac{1}{1+x_t}$ and α corresponds to $\frac{x_t s}{1+x_t}$.

A starting point is to analyze the constants α, which provide information on how, on average, individuals interpreted the abstracts. For abstracts that were pro climate change, individuals rated them a 5.40 (out of 8). For abstracts that were evidence against climate change, the average rating was -3.52 (out of -8). The pro and con death penalty abstracts follow a similar pattern. The unclear abstract was insignificantly different from 0 in all cases.

Table 2 also reports the influence of an individual's prior belief on the interpretation of

the summaries. If γ is significantly positive, this is evidence of updating of the summaries in the direction of the prior. The influence of the prior belief on interpretations of summaries concerning climate change is 0.13 (0.03) for 'pro' abstracts, 0.07 (0.02) for 'con' abstracts and 0.08 (0.03) for unclear abstracts (standard errors in parentheses). Similar coefficients for summaries concerning the deterrent effects of death penalty are 0.08 (0.02) for 'pro' abstracts, 0.09 (0.02) for 'con' abstracts, and 0.05 (0.02) for ambiguous summaries. These were all highly significant in the full sample (mostly at the 99 percent level, and in one case at the 95 percent level) and also of similar magnitude and significance in the restricted sample, excluding the pro death penalty case - for such models, the coefficients of prior beliefs on interpretations were insignificant for the restricted sample.

The differences in interpretations indicate that even 'pro' and 'con' articles are open to some differences in interpretation. In terms of the scaling, it is hard to know how to interpret the relative movements from a 5 to a 7 compared to the movement from a 0 to a 1, and so it is difficult to interpret the differences in coefficients by whether the article is ambiguous or pro/con.

Table 3 estimates similar regressions but includes interactions terms: male/female, republican/democrat/independent, and attained/did not attain a college degree. The coefficient between prior belief and interpretation is stable across demographic characteristics.

6.2 Evidence on Belief Updating

The above results show that subjects interpret the summaries with a significant bias (relative to the average interpretation) in the direction of their priors. This is consistent with our model, which is built on people interpreting any evidence that is open to interpretation based on their current beliefs. This is also consistent with Bayes' rule.

The second part of our theory is then that people use the interpreted signals, rather than the raw (ambiguous) signals, when they update to reach their posterior beliefs. This provides a sort of "double updating" in the direction of their prior, which is what leads to a possible polarization of beliefs when faced with similar evidence.

To investigate this we examine how the posterior beliefs differ from the prior beliefs. First, we note that the distribution of posterior beliefs is statistically different from the distribution of prior beliefs. Using a Kolmogorov-Smirnov test, a non-parametric test of equality of distributions, we estimate a p-value of 0.000 for climate change and 0.072 for the death penalty, implying that there is a statistically significant change in distribution for climate change, but only marginally so for death penalty. This difference is not conclusive evidence in favor of the model, since after seeing information people would update under both the Bayesian and our model and under many other models, and so it is not unexpected to see a difference in distributions.

Now let us examine predictions of our model that differ from many other models (e.g., fully rational Bayesian updating or some equal averaging of signals). It is possible to have polarization under our model: i.e., to see an increase in the variance in the posterior beliefs relative to the prior beliefs even after all individuals process the same information, even if the average signal lies between the priors. In particular, our model could have the posterior move away from the average signal. This is not possible if subjects are fully rational Bayesian updaters *and* view the world within the one dimension of our model and the experiment. However, if one enriches the dimensionality of the space and adds heterogeneity in past experience on those dimensions and correlations across dimensions (in the way explained by Benoit and Dubra 2014) then one could rationalize the data via Bayesian updating.[23]

To examine this, we first conduct a standard variance ratio test (comparing the ratio of the variance of the posterior to the prior distribution). We find a p-value of 0.03 for climate change and 0.45 for death penalty, suggesting an increase in variance for the posterior beliefs with regards to climate change but not for posterior beliefs about the death penalty.

A limitation of a test that just looks at overall variance is that individuals could be heterogeneous. For instance, it is still possible the overall variance does not increase because some people are standard Bayesians, or simple belief averagers, who move closer to the mean, while others exhibit polarizing behaviors more consistent with our model. Digging deeper, we investigate what fraction of the subjects are polarizing.

Let $Prior_i$ and $Post_i$ denote i's prior and posterior beliefs, respectively (i.e., their answers to the belief questions before and after reading the summaries). Given that the average interpretation of all summaries was close to 0 (-0.138 for climate change and -0.261 for death penalty on the 16 point scale), if we take this to be a not-too-biased estimate of the actual signals, then any subject who behaved in a Bayesian manner (or who weighted summaries equally and mixed them with the prior) would have a posterior weakly closer to 0 than the prior. The fraction for whom $|Post_i| > |Prior_i|$ is 22.9 percent for climate change, 31.3 percent for death penalty, and 45.6 percent for at least one topic.[24]

[23]Bayesian updating is not falsifiable if one allows for richer priors over unobserved state spaces. Effectively, any data in any experiment can be rationalized via a prior that has the particular experiences of the subject in the experiment as being associated with states in a way that leads to that subject's posterior. Benoit and Dubra's construction is a clever example of how this can work fairly naturally in a simple context, but the point holds very generally - and one can never rule out Bayesian updating. Here we find the current model to provide a more direct explanation for the data, but that is subjective.

[24] The same holds if we do this relative to the average posterior belief $\overline{Post} = \frac{\sum_i Post_i}{608}$. We can examine how many subjects have $|Post_i - \overline{Post}| > |Prior_i - \overline{Post}|$. The fraction of subjects for who this holds is 32.9 percent for climate change, 32.9 percent for death penalty, and 55.1 percent for at least one topic.

6.3 Further Evidence Consistent with the Model

Appendix Table 1 summarizes a variety of previous studies in which identical information given to subjects in experimental settings resulted in increased polarization of beliefs – individuals expressing more confidence in their initial beliefs. The seminal paper in this literature is Lord, Ross, and Lepper (1979), who provided experimental subjects with two articles on the deterrent effects of capital punishment. The first article argued that capital punishment has a deterrent effect on crime, while the second argued there was no relationship. The authors observe both biased assimilation (subjects rate the article reinforcing their viewpoint as more convincing) and polarization (subjects express greater confidence in their original beliefs). There are follow-up studies with a similar design of presenting two essays with different viewpoints that used larger and more balanced subject pools, since Lord, Ross, and Lepper (1979) had a small sample and had selected subjects with strong prior opinions. Miller et al. (1993) found that the extent of polarization depended on whether it was self-reported or viewed via the subject's own subsequent writings. Kuhn and Lao (1996) found heterogeneity in the reactions depending on the subjects initial opinions, and found that some subjects polarized while others were more engaged in interpreting the writings.

Other studies found evidence of polarization consistent with our theory: using opinions of nuclear power (Plous 1991), homosexual stereotypes (Munro and Ditto 1997), perceptions of fictional brands (Russo, Meloy, and Medvec 1998), theories of the assassination of John F. Kennedy (McHoskey 2002), the perceived safety of nanotechnology (Kahan et al 2007), and the accuracy of statements made by contemporary politicians (Nyahn and Reifler 2010).

Nyhan, Reifler, and Ubel (2013) provide a recent example relating to political beliefs regarding health care reform. They conduct an experiment to determine if more aggressive media fact-checking can correct the (false) belief that the Affordable Care Act would create "death panels." Participants from an opt-in Internet panel were randomly assigned to either a control group in which they read an article on Sarah Palin's claims about "death panels" or a treatment group in which the article also contained corrective information refuting Palin.

Consistent with the maximum likelihood storage rule, Nyhan, Reifler, and Ubel (2013) find that the treatment reduced belief in death panels and strong opposition to the Affordable Care Act among those who viewed Palin unfavorably and those who view her favorably but have low political knowledge. However, identical information served to *strengthen* beliefs in death panels among politically knowledgeable Palin supporters.

Our comparatively large and heterogeneous subject pool, and mixture of pro, con, and ambiguous articles, provide us with necessary acuity in measuring the reactions of subjects to the different types of articles as a function of the subjects' characteristics and initial opinions. We also pay special attention to the interpretation of ambiguous evidence by prior belief and how this correlates with changes in beliefs. This allows us to test our theory and

distinguish it from Bayesian updating within the one dimensional setting. Thus, although the design of our experiments is similar to predecessors, our analysis is not.

6.4 Further Biases

There is one additional issue gleaned from our experiments that may be of interest for further research. We note the following anomaly in individual updating behavior.

Table 4 estimates the coefficient that relates prior beliefs to the interpretation of ambiguous evidence for different portions of the prior belief distribution. In particular, we report coefficients for individuals with positive priors (between 0 and +8) and negative priors (between -8 and 0), separately.

Table 4 presents some challenges for any existing model of interpretation of information. When viewing 'pro' summaries, individuals who reported 'pro' priors have a large and significant coefficient in terms of how they bias the interpretation – two to three times the full sample coefficient and statistically significant – but the coefficients for individuals who reported being 'con' is insignificant. Similarly, for the 'con' summaries, we see a stronger bias for people with con priors than for those with pro priors. If this pattern holds up to further investigation (it may be underpowered), it is inconsistent with the Bayesian foundation that our model presumes (and is also inconsistent with random updating as in Rabin and Schrag (1999)). This suggests that people may take evidence that is consistent with their priors and bias it towards their priors (in a way consistent with Bayesian updating), but *not* bias information that is contradictory to their priors or at least process it quite differently.

Although this has some resemblance to motivated beliefs (e.g. Eil and Rao 2011, Möbius et al. 2014) there are some potentially intriguing differences. One is that the information is not about the subject and may not have any impact on their self-image. The other is that here subjects did not ignore contrary information, but in fact that they treated it more accurately, while they shifted information that is in agreement with their beliefs.[25] Thus, there could be quite subtle or complex mechanisms at work in terms of how people process information that contradicts their beliefs compared to that which confirms their beliefs. This is an interesting subject for further research.

7 Concluding Remarks

Polarization of beliefs has been documented by both economists and psychologists. To date, however, there has been little understanding of the underlying mechanisms that lead

[25]So, this is a different effect, for instance, from Eil and Rao's finding that subjects tend to process agreeable information like Bayesians, but dismiss contradictory information.

to such polarization, and especially why increasing polarization occurs even though agents are faced with access to the same information. To fill this void, we illuminate a simple idea: when evidence is open to interpretation, then a straightforward – and constrained optimal – rule can lead individuals to polarize when their information sets are identical. We also test the mechanism of our model in an on-line experiment and find results that are largely in line with our model's predictions.

In addition, it follows directly from our model that agents who are forced to crystalize their beliefs in the face of signals that are open to interpretation will polarize faster (and in more situations) than those who only infrequently have to react to signals that are open to interpretation. This provides a testable empirical prediction for future experiments.

Again, our experiments cannot reject models such as those of Baliga, Hanany and Klibanoff (2013), Andreoni and Mylovanov (2012), and Benoit and Dubra (2014), since those models involve aspects of the subjects that are not observable. This presents an interesting challenge for future research, and suggests further experiments in which subjects may be tested on other dimensions of their thinking which might impact how they update their beliefs on some subject.

Beyond polarization, our adaptation of Bayes' rule has potentially important implications for other information-based models such as discrimination. For instance, using our updating rule, statistical discrimination can persist in a Coute and Loury (1993) model with infinite signals. Moreover, our model implies that policies designed to counteract discrimination will have to account for the possibility that employers may act on how they perceive signals about applicants when hiring, rather than fully accounting for all the ambiguity in those signals when forming beliefs.

References

[1] Andreoni, James and Tymofiy Mylovanov (2012) "Diverging Opinions," *American Economic Journal: Microeconomics*, 4(1): 209-232.

[2] Baliga, Sandeep, Eran Hanany, and Peter Klibanoff "Polarization and Ambiguity," forthcoming: *American Economic Review*.

[3] Benoît, Jean-Pierre and Juan Dubra (2014) "A Theory of Rational Attitude Polarization," London Business School.

[4] Berry, Donald A. and Bert Fristedt. 1985. *Bandit Problems: Sequential Allocation of Experiments.* Monographs on Statistics and Applied Probability. London; New York. Chapman/Hall.

[5] Brooks, David. 2012, June 1. "The Segmentation Century." *The New York Times*, p. A27. Retrieved from `http://www.nytimes.com/2012/06/01/opinion/brooks-the-segmentation-century.html?_r=0`

[6] Buhrmester, Michael, Kwang, Tracy and Samuel D. Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?' emphPerspectives on Psychological Science, 6(1): 3-5.

[7] Coate, Steven, and Glenn Loury. 1993. "Will Affirmative Action Policies Eliminate Negative Stereotypes." *American Economic Review*, 83(5): 1220-40

[8] Dandekar, Pranav, Ashish Goel, and David T. Lee. (2013) "Biased assimilation, homophily, and the dynamics of polarization," *Proceedings of the National Academy of Sciences*, www.pnas.org/cgi/doi/10.1073/pnas.1217220110

[9] Darley, John M. and Paget H. Gross. 1983. "A Hypothesis-Confirming Bias in Labeling Effects." *Journal of Personality and Social Psychology*, 44(1): 20-33.

[10] Doob, J.L. "Application of the theory of martingales." 1949. In *Le Calcul des Probabilités et ses Applications, Colloques Internationaux du Centre National de la Recherche Scientifique*, 13: 2327.

[11] Duclos, Jean-Yves, Joan Esteban, and Debraj Ray. 2004. "Polarization: Concepts, Measurement, Estimation," *Econometrica* 72(6): 1737-1772.

[12] Eil, David and Justin Rao (2011) "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics* 3(2): 114 - 138.

[13] Esteban, Joan, Laura Mayoral, and Debraj Ray. 2013. "Ethnicity and Conflict: An Empirical Study." *American Economic Review*, forthcoming.

[14] Esteban, Joan, and Debraj Ray. 1994. "On the Measurement of Polarization." *Econometrica*, 62(4): 819-851.

[15] Esteban, Joan, and Debraj Ray. 2011. "Linking Conflict to Inequality and Polarization." *American Economic Review*, 101(4): 1345-74.

[16] Gennaioli, Nicola and Andrei Shleifer. 2010. "What Comes to Mind," *The Quarterly Journal of Economics*, 125 (4): 1399-1433.

[17] Gilboa, Itzhak and David Schmeidler. 1993. "Updating Ambiguous Beliefs," *The Journal of Economic Theory.* 59 (1): 33-49.

[18] Glaeser, Edward, and Cass Sunstein. "Why Does Balanced News Produce Unbalanced Views?" NBER Working Paper No. 18975.

[19] Hellman, Martin A. and Thomas M. Cover. 1970. "Learning with Finite Memory," *The Annals of Mathematical Statistics*, Vol. 41, No. 3 (June), pp. 765 - 782.

[20] Horton, John J., David G. Rand, and Richard J. Zeckhauser (2011) "The online laboratory: Conducting experiments in a real labor market," *Experimental Economics* 14 (3): 399-425.

[21] Jensen, Jacob, Ethan Kaplan, Suresh Naidu, and Laurence Wilse-Samson. 2012. "Political Polarization and the Dynamics of Political Language: Evidence from 130 years of Partisan Speech." *Brookings Papers on Economic Activity*, Fall.

[22] Lord, Charles, Lee Ross and Mark Lepper. 1979. "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence." *Journal of Personality and Social Psychology*, 37(11): 2098-2109.

[23] Kahan, Dan M., Paul Slovic, Donald Braman, John Gastil, and Geoffrey L. Cohen. 2007. "Affect , Values, and Nanotechnology Risk Perceptions: An Experimental Investigation." GWU Legal Studies Research Paper No. 261; Yale Law School, Public Law Working Paper No. 155; GWU Law School Public Law Research Paper No. 261; 2nd Annual Conference on Empirical Legal Studies Paper. Available at SSRN: `http://ssrn.com/abstract=968652`

[24] Kominers, Scott, Xiaosheng Mu, and Alexander Peysakhovich. 2015. "Paying (for) Attention: The Impact of Information Processing Costs on Bayesian Inference." mimeo.

[25] Kuhn, Deanna, and Joseph Lao. 1996. "Effects of Evidence on Attitudes: Is Polarization the Norm?" *Psychological Science* Vol. 7, No. 2, pp. 115-120

[26] McHoskey, John W. 2002. "Case Closed? On the John F. Kennedy Assassination: Biased Assimilation of Evidence and Attitude Polarization." *Basic and Applied Social Psychology*, 17(3): 395-409.

[27] Miller, Arthur G. , John W. McHoskey, Cynthia M. Bane, and Timothy G. Dowd. 1993. "The Attitude Polarization Phenomenon: Role of Response Measure, Attitude Extremity, and Behavioral Consequences of Reported Attitude Change," *Journal of Personality and Social Psychology*, Vol. 64, No. 4, 561-574.

[28] Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat, (2014) "Managing Self-Confidence," mimeo.

[29] Mullainathan, Sendhil. 2002. "A Memory-Based Model of Bounded Rationality." *Quarterly Journal of Economics*, 117(3): 735-774.

[30] Munro, Geoffrey D. and Peter H. Ditto. 1997. "Biased Assimilation, Attitude Polarization, and Affect in Reactions to Stereotype-Relevant Scientific Information." *Personality and Social Psychology Bulletin*, 23(6): 636-653.

[31] Nyhan, Brendan and Jason Reifler. 2010. "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior*, 32:303-330.

[32] Nyhan, Brendan, Jason Reifler, and Peter Ubel. "The Hazards of Correcting Myths about Health Care Reform." *Medical Care* 51(2): 127-132.

[33] Plous, Scott. 1991. "Biases in the Assimilation of Technological Breakdowns: Do Accidents Make Us Safer?" *Journal of Applied Social Psychology*, 21(13): 1058-1082.

[34] Rabin, Matthew and Joel L. Shrag. 1999. "First Impressions Matter: A Model of Confirmatory Bias," *The Quarterly Journal of Economics* 114(1): 37-82.

[35] Russo, J. Edward, Margaret G. Meloy, and Victoria Husted Medvec. 1998. "Predecisional Distortion of Product Information." *Journal of Marketing Research*, 35(4): 438-452.

[36] Saad, Lydia. 2013, April 9. "Republican Skepticism Toward Global Warming Eases." Gallup Politics, `http://www.gallup.com/poll/161714/republican-skepticism-global-warming-eases.aspx`. Retrieved April 27, 2013.

[37] Schwartzstein, Joshua. 2014. "Selective Attention and Learning." *Journal of the European Economic Association* 12(6): 1423-1452.

[38] Siniscalchi, Marciano. 2011. "Dynamic Choice Under Ambiguity." *Theoretical Economics* 6, 379-421.

[39] Sunstein, Cass. 2009. *Going to Extremes: How Like Minds Unite and Divide*. Oxford University Press.

[40] Urschel, Joe. 1995, October 9. "Poll: A Nation More Divided." *USA Today*, p. 5A. Retrieved from LexisNexis Academic database, April 22, 2013.

[41] Wilson, Andrea. 2014. "Bounded Memory and Biases in Information Processing." *Econometrica*, 82(6): 2257-2294.

Table 1: Summary Statistics

	Non-Restricted Sample	Restricted Sample	Difference p-val	Full Sample
	(1)	(2)	(3)	(4)
Female	0.450	0.418	0.032 0.532	0.443
Age	34.3	34.6	0.032 0.798	34.4
White	0.807	0.756	0.051 0.207	0.796
Non-Christian	0.618	0.630	-0.012 0.813	0.621
College Grad	0.489	0.551	-0.062 0.210	0.502
Employed	0.632	0.638	-0.006 0.905	0.633
Yearly Income	\$29,567	\$29,500	\$67 0.978	\$29,553
Hourly Wage	\$14.76	\$12.84	\$1.92 0.137	\$14.35
Contin. USA	0.946	0.835	0.111***	0.923
Urban	0.258	0.299	0.000 -0.041	0.266
Suburban	0.555	0.528	0.349 0.027	0.549
Rural	0.185	0.173	0.580 0.012	0.183
Democrat	0.405	0.465	0.760 -0.060	0.418
Republican	0.150	0.197	0.230 -0.047	0.160
Independent	0.424	0.323	0.197 0.101**	0.403
Observations	481	127	0.039 608	608

Notes: This table presents summary statistics for the participants in the final sample. Column (1) contains those participants who took the survey and were not screened with reading comprehension questions. Column (2) contains those participants who were presented with surveys that tested reading comprehension and successfully answered all test questions. Column (3) reports the difference between the estimates in columns (1) and (2) and the p-values from a test of equal means. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 2: Regressions of Interpretations on Priors

	Climate Change		Death Penalty	
	(1)	(2)	(3)	(4)
Pro Summaries				
Prior Belief	0.132***	0.125*	0.077***	0.028
	(0.030)	(0.076)	(0.021)	(0.048)
Constant	5.398***	5.385***	4.552***	4.522***
	(0.163)	(0.383)	(0.097)	(0.204)
Observations	1,216	254	1,216	254
Con Summaries				
Prior Belief	0.072***	0.104***	0.088***	0.096*
	(0.018)	(0.036)	(0.021)	(0.052)
Constant	-3.941***	-3.937***	-3.519***	-3.259***
	(0.090)	(0.177)	(0.099)	(0.260)
Observations	1,824	381	1,824	381
Unclear Summaries				
Prior Belief	0.079***	0.217***	0.051**	0.061
	(0.029)	(0.067)	(0.020)	(0.049)
Constant	0.200	0.036	-0.118	0.053
	(0.141)	(0.334)	(0.090)	(0.228)
Observations	608	127	608	127

Notes: This table presents estimates of the influence of prior beliefs on interpretation of the summaries by category of summary. Columns (1) and (3) contain those participants who took the survey and were not screened with reading comprehension questions. Columns (2) and (4) contain those participants who were presented with surveys that tested reading comprehension and successfully answered all test questions. Pro summaries include summaries 1 and 2 for both climate change and death penalty. Con summaries include summaries 3,4,5 for climate change and summaries 3,4 and 6 for death penalty. Unclear summaries include summary 6 for climate change and summary 5 for death penalty. See Appendix Table 2 for individual summary results. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively. All standard errors are clustered at the individual level.

Table 3: Regressions with Interaction Terms Based on Demographics

	Gender			*Education*			*Polit. Affil.*			
	Male	Female	p-val	College	No College	p-val	Democ	Repub	Ind.	p-val
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Climate Change										
Pro Summaries	0.135***	0.110***		0.073**	0.217***		0.194***	0.162*	0.061*	
	(0.046)	(0.038)	0.664	(0.031)	(0.055)	0.024	(0.052)	(0.084)	(0.037)	0.097
	656	522		606	610		508	194	490	
Con Summaries	0.094***	0.071***		0.050**	0.088***		0.080**	0.116**	0.065**	
	(0.024)	(0.024)	0.496	(0.022)	(0.028)	0.295	(0.037)	(0.050)	(0.027)	0.655
	984	783		909	915		762	291	735	
Unclear Summaries	0.124***	0.014		0.051**	0.113***		0.095*	0.173*	0.024	
	(0.042)	(0.035)	0.044	(0.038)	(0.044)	0.287	(0.049)	(0.089)	(0.035)	0.202
	328	261		303	305		254	97	245	
Panel B: Death Penalty										
Pro Summaries	0.081***	0.062*		0.060**	0.098***		0.032*	0.104**	0.099***	
	(0.030)	(0.032)	0.650	(0.029)	(0.031)	0.362	(0.032)	(0.049)	(0.037)	0.297
	656	522		606	610		508	194	490	
Con Summaries	0.103***	0.055		0.099***	0.075**		0.070**	0.111*	0.084**	
	(0.027)	(0.034)	0.268	(0.028)	(0.032)	0.561	(0.030)	(0.059)	(0.033)	0.814
	984	783		909	915		762	291	735	
Unclear Summaries	0.048*	0.036		0.075***	0.026**		0.047**	0.089*	0.025	
	(0.028)	(0.029)	0.752	(0.026)	(0.031)	0.223	(0.033)	(0.053)	(0.028)	0.559
	328	261		303	305		254	97	245	

Notes: This table presents estimates of the effects of prior beliefs on interpretation and divides the data based onsubsamples. Column (1) includes males in our sample, column (2) contains females, column (4) contains college graduates, column (5) contains those who did not attain college degrees and columns (7), (8) and (9) include dempocrats, republicans and independents, respectively. Columns (3), (6) and (10) report p-values resulting from a test of equal coefficients between the gender, educational attainment, and political afilliation subgroups, respectively. Pro summaries include summaries 1 and 2 for both climate change and death penalty. Con summaries include summaries 3,4,and 5 for climate change and summaries 3,4 and 6 for death penalty. Unclear summaries include summary 6 for climate change and summary 5 for death penalty. See Appendix Table 1 for individual summary results. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively. All standard errors are clustered at the individual level.

Table 4: Regressions Broken Down by Whether the Prior Was Pro or Con

	Climate Change			*Death Penalty*		
	All	Pro Prior	Con Prior	All	Pro Prior	Con Prior
	(1)	(2)	(3)	(4)	(5)	(6)
Pro Summaries						
Coeff on Prior	0.132***	0.281***	-0.157	0.077***	0.233***	-0.087
	(0.030)	(0.036)	(0.151)	(0.021)	(0.059)	(0.074)
Constant	5.398***	4.639***	4.653***	4.552***	4.090***	3.679***
	(0.163)	(0.203)	(0.668)	(0.097)	(0.228)	(0.391)
N	1,216	1,082	134	1,216	532	684
Con Summaries						
Coeff on Prior	0.072***	0.038	0.061	0.088***	0.089	0.192***
	(0.018)	(0.033)	(0.073)	(0.021)	(0.075)	(0.055)
Constant	-3.941***	-3.764***	-4.177***	-3.519***	-3.579***	-2.917***
	(0.090)	(0.166)	(0.384)	(0.099)	(0.237)	(0.304)
N	1,824	1,623	201	1,824	798	1,026
Unclear Summaries						
Coeff on Prior	0.079***	0.110***	0.039	0.051**	0.204***	0.062
	(0.029)	(0.039)	(0.126)	(0.020)	(0.072)	(0.042)
Constant	0.200	0.041	0.153	-0.118	-0.668***	0.023
	(0.141)	(0.191)	(0.542)	(0.090)	(0.196)	(0.218)
N	608	541	67	608	266	342

Notes: This table presents estimates of the effects of prior beliefs on interpretation and divides the data based on prior belief. Column (1) is for the full sample. Columns (2) and (5) contain those individuals who reported having a belief greater than or equal to 0 on a scale of -8 to 8 for the respective topic. Columns (3) and (6) contain those individuals who reported having a belief less than 0 on a scale of -8 to 8 for the respective topic. Pro summaries include summaries 1 and 2 for both climate change and death penalty. Con summaries include summaries 3,4,and 5 for climate change and summaries 3,4 and 6 for death penalty. Unclear summaries include summary 6 for climate change and summary 5 for death penalty. See Appendix Table 2 for individual summary results. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively. All standard errors are clustered at the individual level.