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

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

March 2017

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 use Bayes' rule in an iterative way: first to interpret each signal and then to 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: the distance between their beliefs can grow after observing a common sequence of signals. Such updating is approximately optimal if agents must interpret ambiguous signals and sufficiently discount the future. If they are very patient but can only store interpretations of ambiguous signals, then a time-varying random interpretation rule (still double-updating) is approximately optimal. In a continuous (normally distributed) version of the model, we show that posterior beliefs never lose the influence of the prior and still always converge, but always converge to something that is influenced by the prior and early signals and so is wrong with probability one. Beliefs become arbitrarily accurate as the signal accuracy increases, but are always biased. We explore the model in an on-line experiment in which individuals interpret research summaries about climate change and the death penalty and report beliefs. Consistent with the model, not only is there a significant relationship between an individual's prior and their interpretation of the summaries; but more than half of the subjects exhibit polarizing behavior - shifting their beliefs further from the average belief after seeing the same summaries as all other subjects.

*We are grateful to Sandro Ambuehl, Ken Arrow, 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.

[†]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 Mathematics, ETH Zurich, (philipp.harms@math.ethz.ch). We gratefully acknowledge financial support from the NSF grants SES-0961481 and SES-1155302, and ARO MURI Award No. W911NF-12-1-0509.

JEL classification numbers: D10, D80, J15, J71, I30

Keywords: beliefs, polarization, learning, updating, Bayesian updating, biases, discrimination, decision making

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.[1] In a similar vein, there were large racial differences in reactions to the O.J. Simpson murder trial. Four days after the verdict was announced, 73 percent of whites but only 27 percent of blacks surveyed in a Gallup/CNN/USA Today poll believed that Simpson was guilty of murdering Nicole Brown Simpson and Ronald Goldman (Urschel 1995). Between 2013 and 2014 – before and after the deaths of Michael Brown and Eric Garner – 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).

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.[2] 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.

In this paper, we introduce a model that provides a simple foundation for why the above-described polarizations in beliefs should be observed from rational agents. In our model, agents interpret information according to Bayes' rule as it is received and store the interpreted signal in memory rather than storing the full information, and this simple modification of Bayesian updating can lead to increasing and extreme polarization. In particular, in our simplest version of the model there are two possible states of nature A, B and an agent observes a series of signals a, b that are correlated with the state of nature. Some of the

[1] Beliefs that the effects of global warming "have already begun to happen" and that "global warming will pose a serious threat to you or your way of life in your lifetime" show similar trajectories for both groups (Saad 2013).

[2] Relatedly, Jensen et al. (2012) derive measures of political polarization.

signals are ambiguous and come as ab and must be interpreted.[3] 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, interprets or *perceives* the ab signals as they are seen according to Bayes' rule (or some other rule), and then stores the interpretation in memory. So, if the agent started by believing that A was more likely, then the sequence would be interpreted/perceived 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 actually see more evidence for b than a, while interpreting signals according to Bayes' rule as they are received and then storing the interpretation leads to $a, b, a, a, a, a, b, b \ldots$ and more evidence for a than b.

This can lead people with slightly different priors to end up with increasingly different posteriors as they interpret the same string of ab's very differently, which then reinforces their differences in beliefs. As we show, if a large enough share of experiences are open to interpretation, 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 extend the model to admit continuous (normally distributed) signals and states. In that model an agent's belief always converge, but with probability 1 the beliefs 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 in the limit of perfectly informative signals the bias disappears, but with any fixed signal accuracy, all agents' limit beliefs can be strictly ordered based on their priors.

One view of our model is that it is explicit about an agent's *perception* of a situation. An agent perceives a given ambiguous signal based on their prior beliefs and then processes that perception when updating beliefs. Thus, it leads to a sort of double updating. The deviation of our model from full Bayesian updating is minimal: Bayes' rule is still used at all steps and agents are in a sense rational, but agents interpret ambiguous signals before storing them and then base their posteriors on their interpretations rather than all of the original data. Some of our results examine approximately optimal ways to interpret ambiguous signals. We show that the rule above, based on iterative use of Bayes' rule is approximately optimal if the agent is not too patient, and so cares enough about making correct decisions in the short run (or else begins with a strong prior). In contrast, if the agent is very patient, then the agent may wish to randomize interpretations and may do so in a manner that depends on the current belief. Thus, this sort of perception could have evolved for two reasons: first it saves on memory, but second it allows an agent to react quickly to a given situation by

[3] 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 - also fitting well with the word ambiguous.

making a quick judgement on a perceived situation. The cost of this quick interpretation is that the fuller information that was present is lost.

We explore the implications of the model by developing an on-line experiment in which individuals read research summaries (abstracted from published articles) and then interpret these and provide updated beliefs. This type of experiment was pioneered by Lord, Ross, and Lepper (1979), and 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. Additionally, rather than having 48 undergraduate subjects, we have over 600 (with much broader backgrounds); which, together with our richer experimental design, allows us to study many questions not studied before in this literature, and – importantly – with greater accuracy.[4]

Experiments were conducted through Amazon Mechanical Turk (MTurk) in which subjects were asked to read summaries of published research articles about climate change and the potential deterrent effects of the death penalty and then report their interpretations of the evidence and their updated beliefs. First, 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 were given a series of summaries to read, which were redacted from abstracts and introductions of published academic papers and edited to make more readable. After each summary, we asked the participant to decide whether he or she thought the summary provided evidence for or against the topic on a 16 point scale. Importantly, 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, depending on whether they completed the task (which typically took less than half an hour, so these were high wages by MTurk standards).

The results of our experiment are largely in line with the predictions of the model. We provide four sets of empirical observations.

First, there is a significant robust correlation between an individual's prior belief and their interpretation of evidence. The correlation between an individual's prior belief and their bias in interpretation of the summaries is typically around ten percent. For example, 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

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

estimating the relationship between an individual's prior belief and their interpretation conditioning on a variety of demographics: gender, education, income, and political affiliation.

Third, 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 - providing significant evidence for a change in distribution for climate change and marginally significant evidence (depending on the significance level) for a change in distribution for the death penalty. Thus, they do react to the information given to them. Most importantly, we examine whether the distribution of beliefs becomes tighter or more extreme. A variance ratio test has a p-value of 0.04 for climate change and 0.45 for the death penalty - finding significant *aggregate* evidence for an increase in polarization of beliefs about climate change but not 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, that are viewed on average to be neutral. This is consistent with our simple model, but would need to be embedded in a richer multi-dimensional space in which people's knowledge is sufficiently heterogeneous and interacts with the current dimension (as in Benoit and Dubra 2014) to be rationalized by Bayesian updating.[5]

Our experimental data are also 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.[6] An example is Darley and Gross (1983). Subjects in their experiment are asked to rate a fictional student. Before rating the student, however, subjects view one of two videos of the student playing in her home. The first video shows the student in a poor, inner-city neighborhood, while the second depicts a middle-class suburban environment. Subjects shown the first video rate the student significantly lower. A subset of each group was also shown an identical video of the student in class, providing both correct and incorrect answers to a teacher's questions. Respondents who saw this second video diverged even further in their ratings. Subjects who viewed the low-quality (high-quality) neighborhood and the class video gave a lower (higher) rating than those who just saw the neighborhood video. The showing of the same video to the subjects caused them to further diverge in their grades on some of their opinions (on four out of eight dimensions).

[5] 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.

[6] A summary of many of these studies can be found in Appendix Table 1, and we review some of this evidence in more detail below.

With the lens of our model, the two neighborhood videos in Darley and Gross (1983) created different priors on the students' ability. This then affected the way that respondents interpreted the second video of students answering questions. A student with a lower prior on the talent of the student would put more emphasis on the incorrect answers in the second video, and a student with a higher prior would put more emphasis on the correct answers. This then leads to the observed polarization of some of the beliefs when rating the students. Bayes' rule with unbounded memory would lead to the opposite – seeing the same second video should, if anything, lead beliefs to get closer. With respect to the O. J. Simpson murder trial, people interpreted the same trial as either another example of an unfair legal system or another example of murder. A similar phenomenon may be responsible for diverging views on police ethical standards in America. The data are open to interpretation, and based on their differing past experiences and beliefs prior to the event people reached different conclusions.

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 4 presents a version of the model where the states and signals are normally distributed, Section 5 describes our experiments and Section 6 presents the data. 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. An online appendix contains additional analysis of the data, the summaries used in the experiment, and the experimental instructions.

2 A Brief Review of the Literature

Rabin and Schrag (1999) provide a first decision-making model of confirmation bias.[7] In each period agents observe a noisy signal about the state of the world, at which point they update their beliefs. Agents' perceptions are clouded by bias. In their model, signals that are believed to be less likely are misinterpreted with an exogenous probability. Given the exogenous mistakes introduced into the model, 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 that can be thought of as providing a "why" behind long term bias, and how this can also lead to belief polarization.[8] We also explore the implications of our foundations in

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

[8] Our focus is on the implications of such updating for biases in beliefs and polarization. For more general

settings with continuous signal distributions which provide some new results showing that bias *always* occurs (with probability one), and shows how it depends on early signals and not just the prior.

Hellman and Cover (1970) introduced a model of limited memory that 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)).[9] Agents observe a possibly infinite sequence of independent and identically distributed signals according to a time-invariant probability measure. Their goal is, given their sequence of signals, to identify which of two true states A and B they are in. They model the decision problem of the agent as a M-state (possibly reducible) automaton, where M is finite. The automaton follows a stationary updating procedure, transitioning from memory state to memory state based on the latest observation. Thus, the automaton follows a Markov process driven by the underlying signal process. They derive results about algorithms that approximately minimize the expected asymptotic proportion of errors for a given automaton.

The approximately optimal transition rule for the automaton is simple: first, rank 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 the one of the extreme states $\{1, M\}$ are reached. So long as the transition probabilities out of the extreme states are close to zero relative to the transition probabilities in the interior states, this near-optimality is ensured.

Wilson (2014) recasts the limited-memory approach of Hellman and Cover (1970) under a more explicit decision-theoretic framework, and allowing for the process to terminate probabilistically. She then derives fully constrained optimal rules (subject to the finiteness of the automaton). Wilson (2014) shows that agents operate similarly to the automaton described by Hellman and Cover (1970), but when optimizing facing a low probability of termination, agents will move only one memory state at a time (deterministically), will react only to extreme signals, and will leave the extreme states with very low probability. As Wilson (2014) points out, 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

foundations on the updating of beliefs based on ambiguous information see Gilboa and Schmeidler (1993) and Siniscalchi (2011).

[9] There are also related papers that explicitly account for costly updating (e.g., Kominers, Mu, and Peysakhovich (2015)), in which agents throw out information whenever the cost of updating is more than the perceived benefit, as well as models where agents have limited attention (e.g., Schwartzstein (2014)). Such models can lead to persistent biases in updating as not all information is incorporated. Our model is closer in spirit to these models in terms of what drives the bias in updating, although our results and the focus of our analysis are quite different.

can get temporarily ‘stuck’ at one of the extreme states. More generally, Wilson proves that as long as agents do not have identical priors and do not start at one of the extreme states, their beliefs can differ for long periods of time with positive probability, although the 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” - as agents wish to make predictions that are immune to uncertainty, and they 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 Bayesians, 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 such that agents maintain beliefs that are erroneous and become more convinced of those beliefs over time, 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, such as Hellman and Cover’s model and Wilson’s model. 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, decision makers always converge in beliefs, and with probability 1, converge 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, but also early signals have a disproportionate sway in forming people’s perceptions.

3 The Basic Model: Discrete States and Signals

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.

In particular, signals are independent over time, conditional upon the state. With a probability q a signal is observed and with a probability $1 - q$, independent of the state, the no signal is observed, denoted $\emptyset$. Thus, with probability q, the signal s_t starts out as either a or b. If the state is $\omega = A$ and the signal is not $\emptyset$, then the probability that the signal starts out as $s_t = a$ is $p > 1/2$. Likewise, if the state is $\omega = B$ and the signal is not $\emptyset$, then the probability that $s_t = b$ is $p > 1/2$. There is a probability π that a signal a or b comes together with the other signal and is ambiguous and appears as ab, while with probability $1 - \pi$ the signal stays as it was.

For example, A might be a world in which human activity is the cause of rising temperatures and B a world where human activity is not responsible. With probability q additional information is revealed, and with probability $1 - q$ there is no new information. In state A a fraction $p > 1/2$ of new information argues that human activity is the cause of rising temperatures, while in state B a fraction $p > 1/2$ of new information argues the opposite. The probability that the new information is ambiguous about human's effect on climate change is π so that it has aspects that could be interpreted as being "pro" human effect on climate change, but also have a "con" interpretation. With the remaining probability $1 - \pi$ the information is clear. Or, it might be that there are a sequence of studies regarding efficacy of the death penalty that come out, only q of which have any findings. Out of those, p provide clear findings and $1 - p$ of which are corrupt or flawed in state A and the reverse in state B.

The symmetry between A and B both having the same fraction of a and b matching the state is simply for notational convenience and making some of the calculations simpler. As is clear from the proof of the results in the appendix, the assumption is not needed.

$\lambda_0 \in (0, 1)$ is the agent's prior that $\omega = A$.

Throughout our analysis we assume that the prior, λ_0, is not 0 or 1, as otherwise learning is precluded. Similarly, we maintain the assumption that $\pi < 1$ as otherwise no information is ever revealed.

It follows directly from a standard law of large numbers argument (e.g., Doob's (1949) consistency theorem), that the agent's posterior beliefs converge to place weight one on the true state ω almost surely.

Observation **1** *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 about the state 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. For instance, one may want to behave differently depending on whether one believes that human activity is responsible for rising global temperatures, and so interpreting current information can be useful: the observer may want or need to react to current information.

Given that signals become ambiguous independently of the state, the conditional probability that the signal is a versus b is exactly the agent's belief entering the period.

A key element of our model is that at each date an agent perceives and interprets the signal for the day and stores it as only one bit of data - so remembers ambiguous signals in terms of what the agent believes/perceives to be the more likely signal to have generated the ab. That is, when the agent sees an ambiguous signal $s_t = ab$, the agent interprets and stores it as either a or b.[10]

This part of the model captures that an agent may be prone to making quick judgements - immediately perceiving ambiguous information in light of his or her current beliefs, or may have limited memory and cannot remember all the possible interpretations of each interaction and remembers his or her interpretation of each interaction rather than the full interaction. The *interpretation* of ambiguous signals $s_t = ab$ as an a or b is based on the agent's experiences through time t.

The alternative interpretation of the model is one in which agents must be quick to react to current situations, and so forming perceptions of current signals is valuable. The main constraint of the model is then that the agent does not later go back and remember all of the ambiguity that was present when updating his or her beliefs. So, the perception is what is then later processed when updating.

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

This not only seems both plausible and consistent with the evidence discussed in Section

[10] 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 Scwartzstein (2014) framework may not notice ambiguous information due to selective attention.

6.3, but also is consistent with our experiments discussed in Section 5.

3.1.1 A Parenting Example

Let us consider the following example to make our setting very concrete. A parent is faced with a sequence of things that they can either do or refrain from in raising their child. There are some situations in which they should definitely do something, a's say, and other situations that they should not do anything, b's say, and others that they cannot be sure what to do, ab's. The world is either A - that being a very active and engaged parent is better and so the majority of situations a parent should intervene, or B that being a more "free range" parent and, thus, not intervening too much is the better approach.

Imagine a parent starts with a prior over A and B and then starts getting a sequence of choices which require them to choose either to intervene or not. All parents make the "right" choices when seeing an a or b, but when faced with an ab they may make different choices based on their current beliefs. But they do make a choice to either act or not, and it is consequential, and there is a true state which would guide them if they knew it.

The key to the model is that they then remember over time the relative frequency with which they thought they should act versus not, but they don't recall which times they took action a because the signal was a versus a signal ab. More troubling, they eventually develop a long-term belief that is either that they should act in the majority of cases or that they should be laid back and act only in a minority of the cases. That belief becomes reinforced as they tilt one way or the other based on the strength of their initial beliefs and the string of a's, b's, or ab's. For instance, if their memory is that they chose to intervene in most cases, and so they then treat that as the sample based on which to draw inferences from their own past experiences, and they don't discriminate between the dates in which they had clear decisions and the dates upon which they chose to act simply based on their beliefs at the time. They might end up being correct or incorrect in their long term belief and in their choices of actions in the ambiguous states. Yet, all become convinced that they are the right kind of parent!

3.2 Polarized Beliefs

To illustrate this process and the polarization that it induces, let us pose another example and then show that this intuitive phenomenon can occur in many settings.

For the example, suppose that 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:

$$a, ab, \emptyset, ab, b, a, ab, a, \emptyset, ab, ab, b, \ldots$$

In particular, consider a 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.[11] Then when seeing $s_2 = ab$ the agent stores the signal as a, and ends up with a posterior of $\lambda_2 = 12/13$. In this case, the agent would store the sequence as

$$a, a, \emptyset, a, b, a, a, a, \emptyset, a, a, b, \ldots$$

and the posterior at the end of this sequence would already be 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 of $\lambda_2 = 1/4$. In this case, the agent would store the sequence as

$$a, b, \emptyset, b, b, a, b, a, \emptyset, b, b, b, \ldots$$

and the posterior at the end of this sequence would be very close to 0.

Two agents observing exactly the same sequence with different (and non-extreme) priors, come to have increasingly different posteriors. Their posteriors are

$$\begin{array}{llllllllllll}
3/4, & 6/7, & 6/7, & 12/13, & 24/25, & 12/13, & 24/25, & 48/49, & 48/49, & 96/97, & 192/193, & 384/385, \ldots \\
1/4, & 2/5, & 2/5, & 1/4, & 1/7, & 1/13, & 1/7, & 1/13, & 1/13, & 1/7, & 1/13, & 1/25, \ldots
\end{array}$$

Thus, we see a clear polarization from two agents observing exactly the same sequence of information. Proposition 3, below, shows that 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.

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

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

is a tradeoff: if an agent stores a signal in the way that is viewed as most likely given the current beliefs, it can help in reacting to the current situation, but then the agent biases long-term learning.

To explore this tradeoff, consider a situation in which the agent must take an action based on the current interaction. For instance, in the context of our parenting example, at each date the agent must react appropriately to being faced with a chance to take some action (for instance, setting a rule or prohibiting the child from taking some action, etc.).

Note that it is the particular situation/signal that must be appropriately interpreted, not the state. The probability of the state is useful in making a decision, but it is the particular encounters that the agent must make decisions about. In the case of an ambiguous encounter, the agent must take an action. The key assumption is that if the agent decides that it is best to respond to $s_t = ab$ at time t in a manner consistent with a signal of a, then the agent must store the signal as an a. The agent cannot treat the signal in one manner for actions and then remember it differently. This is in line with our limited memory assumption.[12]

To formalize the tradeoff between immediate action and long-term learning under our limited-memory cognitive limitation, we extend the model to explicitly include the agent making explicit choices over time, seeking to maximize the discounted sum of expected payoffs for correctly identifying the true state.

In particular, the agent has choose either a or b at each date, including ones at which $s_t = ab$. At dates in which the signal is unambiguous, that choice is easy, but not when the signal is ambiguous. She gets payoff of $u_t = 1$ if the current situation is correctly identified and $u_t = 0$ when a mistake is made.[13] In particular, 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 ambiguous 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.

As in the parenting example, the agent does not get immediate feedback on the payoffs - which are not learned until years later. 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. So, this may cover parenting, global warming, deterrence effects of crime sentencing, and so forth.

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

[12] It is also important that the agent either not observe the payoffs or not update based on them. This might be due to not getting feedback - as in taking actions that might raise or lower global warming but do not provide instant information - or just remembering beliefs and not the sequence of payoffs associated with the actions.

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

interpretations (including strategies for agents with unbounded memories), we first 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: $\sigma : H \times [0, 1] \to [0, 1]$.[14] 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.[15]

An agent's expected payoffs can be written as:

$$U(\sigma, \delta, \lambda_0) = E\left(\sum_{t=1}^{\infty} \delta^t u_t(\sigma(h_{t-1}, \lambda_0)) \middle| \lambda_0\right).$$

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. For instance, 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 $a, ab, \emptyset, ab, b, a, ab, a, \emptyset, ab, ab, b$, 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. 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 (enormously) 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

[14] 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.

[15] 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.

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 a parameter 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.[16] We denote this strategy by σ^γ.

The special case of $\gamma = 1$ corresponds to maximum likelihood interpretation. As we show now, maximum likelihood is an approximately optimal method of interpretation 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 it is clear that 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 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,

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

(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, that is, the agent always follow their updated priors. For a standard Bayesian agent, maximum-likelihood would be an optimal strategy in this setting. However, as exemplified in 3.2, our signal-interpreting agents may exhibit polarized beliefs after observing the same sequence of signals when adopting 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.

Once an agent is constrained to interpret signals, there is then a tradeoff. Interpreting signals is good in the short run, but can lead to wrong long-term conclusions. hence maximum-likelihood strategy in all time periods would not be approximately optimal for patient agents. Appropriately randomized strategies allow for long-run learning and also avoid polarization. This is what we explore next.

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).[17]

[17] Our model and Rabin and Schrag (1999) (hereafter, RS) are exploring different foundations that allow

3.4 Approximately Optimal Strategies: Two-Step Rules

Proposition 2 shows that long run learning under a randomized-interpretation strategy only occurs if the randomization is sufficiently high (γ is sufficiently low). This can be very costly in the long run, as although the posterior converges, it happens because the agent is randomly interpreting ambiguous signals, even when the agent 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 easily 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. We will show that with sufficient patience, such a strategy is approximately optimal.

As a strong benchmark, let σ^{FI} denote the *full-information* strategy, where the agent actually knows whether the state is A or B, and so when seeing ab then always calls a if $\omega = A$ (or b if $\omega = B$). In this case, the 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 variation on the two-step rule. That is, for any ϵ and λ_0, there exist T and $\overline{\delta}$ such that if $\delta > \overline{\delta}$, then $U(\sigma^T, \delta, \lambda_0) > U(\sigma^{FI}, \delta) - \epsilon$.*

Thus, with sufficient patience a very simple strategy approaches the full information benchmark.[18]

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

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 as a potential underlying mechanism behind the q of RS, and also examines enrichments of that, as well as its very different implications in the continuum case.

[18] 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.

ways, on how patient a decision maker is. If they are sufficiently patient, a strategy that entails randomization in the presence of unclear evidence for finite time and then following their maximum likelihood estimate thereafter approximates the full information Bayesian outcome.[19]

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 The Extended Model: Continuous States and Signals

To make the ideas and key insights as transparent as possible, the model we discussed above considered discrete signals of the form $\{a, b, ab, \emptyset\}$. Yet, complex information, such as the information contained in a research article or a news broadcast, is less clearly of an "a" or "b" form, and is almost always open to interpretation.

To illustrate how the updating that we have studied works in settings with many possible values we consider a model with normally distributed states and signals. This applies to many settings, and approximates many others (including the setting of the experiments below). It should also become clear that the insights obtained from this tractable version of the model would have analogs for other, 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 independently and identically distributed 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

[19] 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).

variance. A Bayesian updater would have posterior

$$\mu_t = \frac{\mu_{t-1} + x_t s_t}{1 + x_t}, \tag{1}$$

where

$$x_t = \frac{\sigma_{t-1}^2}{\sigma_s^2} \qquad \text{and} \qquad \sigma_{t-1}^2 = \frac{\sigma_0^2 \sigma_s^2}{\sigma_s^2 + (t-1)\sigma_0^2}.$$

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. Then,

$$\widehat{s}_t = \frac{\widehat{\mu}_{t-1} + x_t s_t}{1 + x_t},$$

and then

$$\widehat{\mu}_t = \frac{\widehat{\mu}_{t-1} + x_t \widehat{s}_t}{1 + x_t}.$$

Thus, the agent first interprets the signal, which moves it closer to the previous belief, and then updates that belief, essentially weighting the previous belief twice. This updating rule can be rewritten in terms of the actual signal rather than the interpreted signal as:

$$\widehat{\mu}_t = \frac{\widehat{\mu}_{t-1}[(1+x_t)^2 - x_t^2] + x_t^2 s_t}{(1+x_t)^2}. \tag{2}$$

This shows that an updater in our setting ends up overweighting the previous belief and underweighting the signal relative to a Bayesian updater.

This is a tractable setting and we can solve for the posterior at any point in time as a function of the original prior and sequence of signals as shown in the following proposition.

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

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

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. It also has weight on any given signal that is decreasing in time, but does not vanish in the limit (converging to $\left(\frac{\sigma_0^2}{\sigma_s^2+\tau\sigma_0^2}\right)^2 \left(\frac{\sigma_s^2+\tau\sigma_0^2}{\sigma_s^2+(\tau+1)\sigma_0^2}\right)$).

It is useful to contrast the posterior under the model with a Bayesian agent's posterior, which is

$$\mu_t = \mu_0 \frac{\sigma_s^2}{\sigma_s^2 + t\sigma_0^2} + \sum_{\tau=1}^{t} s_\tau \frac{\sigma_0^2}{\sigma_s^2 + t\sigma_0^2}.$$

To get a better feeling for the contrast between the model's posterior and a Bayesian posterior, note that if $\sigma_0^2 = \sigma_s^2$ then the posterior simplifies to be

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

while the Bayesian 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 would equally weight the prior and all subsequent signals and none would have any eventual weight. In contrast, in our model the prior still holds weight and pulls signals towards it. Earlier signals end up with more weight, as early beliefs are less entrenched. However, over time signals get interpreted more and thus end up mattering less.

Notice: 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.

5 An On-line Experiment

In this section, we describe the on-line experiment designed to test the mechanisms of our model.

5.1 Amazon Mechanical Turk

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

[20] See Horton, Rand, and Zeckhauser (2011) for some discussion of the advantages of using such online subject pools.

5.2 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 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.

In particular, each experiment began with four practice questions to get participants comfortable with the format of the questions. These questions also allowed us to see if participants were reading the survey or just clicking through. See the Online Appendix for the practice questions. Then participants were presented with a question about their beliefs on the topics (which were 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

[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.

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).

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.

5.3 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 as much as possible in two main 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, a popular and controversial topic. 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. Fortunately, data from this selective sample mirrored that from the larger more general survey (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 either intentionally or unintentionally entering invalid IDs.

5.4 The Subjects

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). In particular, for Climate Change, question 5 was seen as con (significantly negative on average), while 6 was indistinguishable from 0. For the Death Penalty questions, question 6 was seen as con, while 5 was indistinguishable from 0. Thus, for Climate Change we group 1,2 as pro, 3,4,5 as con, and 6 as unclear; while for the Death Penalty we group 1,2 as pro, 3,4,6 as con, and 5 as unclear. We also report individual analyses for every question in the Appendix (see 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

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

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) for comparison.

In terms of our model with normal signals, γ corresponds to $\frac{1}{1+x_t}$ from and α corresponds to $\frac{x_t s}{1+x_t}$.

A starting point for understanding the results from the experiments is to analyze the constants, α, from the regressions. This provides valuable information on how, on average, individuals interpreted the abstracts. For the abstracts that were pro climate change, individuals rated them a 5.40 (out of 8) on the pro scale. For the 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, via the γ parameters. If the parameter 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 (.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 (.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, exccept for the pro death penalty case for which 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 for various characteristics: for instance, male/female, republican/democrat/independent, and attained a college degree/did not attain a college degree.

The coefficient between prior belief and interpretation is stable across demographic characteristics.

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 on-subsamples. 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.

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 distribution 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

[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.

penalty.

A limitation of a test that just looks at overall variance is that individuals could be heterogeneous and so, for instance, it is still possible that many individuals are polarizing but that 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]

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 subject pools and used a more balanced subject pool, since Lord, Ross, and Lepper (1979) had a small sample and had selected subjects with strong prior opinions. One of those follow ups, 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. Another follow-up study, 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.

Several other studies have found evidence of polarization that is consistent with our

[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.

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

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.

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, while there appears to be a difference in the treatment of information based on whether it is in agreement or contradicts subjects' beliefs, 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 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 – iterative application of Bayes' 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 (through deliberate action or social interactions) 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 inter-

[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.

esting 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, it is straightforward to show that statistical discrimination can persist in a world 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] Fryer, Roland, and Matthew O. Jackson. 2008. "A Categorical Model of Cognition and Biased Decision Making," *The B.E. Journal of Theoretical Economics*, Volume 8, Issue 1, Article 6.

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

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

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

[20] 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.

[21] Hoeffding, Wassily. 1963. "Probability Inequalities for Sums of Bounded Random Variables." *Journal of the American Statistical Association*, 58(301): 13-30.

[22] 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.

[23] Jacod, Jean, and Albert Shiryaev. 2003. *Limit Theorems for Stochastic Processes.* Berlin: Springer-Verlag.

[24] 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.

[25] 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.

[26] 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`

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

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

[29] 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.

[30] 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.

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

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

[33] 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.

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

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

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

[37] Pew Research Center, August 2014, “Stark Racial Divisions in Reactions to Ferguson Police Shooting”.

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

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

[40] 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.

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

[42] Siniscalchi, Marciano. 2011. “Dynamic Choice Under Ambiguity.” *Theoretical Economics* 6, 379-421.

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

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

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

8 Appendix A: Proofs of Propositions

For all of the results below, we consider the case in which $q = 1$ as the other case is exactly the same simply ignoring dates in which no signal is observed, as the agent takes no action and does not update on those dates, and there are infinitely many dates in which signals are observed, almost surely.

Proof of Proposition 1.

Let us first show for any λ_0 and $\varepsilon > 0$, there exist $\overline{\delta}$ such that if $\delta \leq \overline{\delta}$, then $U(\sigma^1, \delta, \lambda_0) \geq U(\sigma, \delta, \lambda_0) - \varepsilon$ for all strategies σ. Recall that

$$U(\sigma, \delta, \lambda_0) = E\left(\sum_{t=1}^{\infty} \delta^t u_t(\sigma(h_{t-1}, \lambda_0)) \middle| \lambda_0\right).$$

Write

$$U(\sigma, \delta, \lambda_0) = E\left(u_1(\sigma(\emptyset, \lambda_0))|\lambda_0\right) + E\left(\sum_{t=2}^{\infty} \delta^t u_t(\sigma(h_{t-1}, \lambda_0)) \middle| \lambda_0\right).$$

The basic idea is that as $\delta \to 0$, the future does not matter and the decision maker only needs to maximize the current period's payoff which amounts to choosing the most likely interpretation. Note that $E\left(\sum_{t=2}^{\infty} \delta^t u_t(\sigma(h_{t-1}, \lambda_0))|\lambda_0\right)$ lies in the interval $[-\frac{\delta}{1-\delta}, \frac{\delta}{1-\delta}]$ lies within $[-\varepsilon, \varepsilon]$ if $\delta \leq \overline{\delta} = \frac{\varepsilon}{1+\varepsilon}$. Thus, if $\delta \leq \overline{\delta}$, then

$$U(\sigma^1, \delta, \lambda_0) - U(\sigma, \delta, \lambda_0) \geq E\left(u_1(\sigma^1(\emptyset, \lambda_0))\middle|\lambda_0\right) - E\left(u_1(\sigma(\emptyset, \lambda_0))|\lambda_0\right) - \varepsilon. \tag{3}$$

Next, note that

$$E\left(u_1(\sigma(\emptyset, \lambda_0))|\lambda_0\right) = E[p \Pr[i_1 = \omega] + (1-p) \Pr[i_1 \neq \omega]|\lambda_0].$$

Since $p > 1/2$, the maximizing solution is to set i_1 to match the most likely state ω given λ_0, and so σ^1 is optimal for the first period optimization. This implies that

$$E\left(u_1(\sigma^1(\emptyset, \lambda_0))\middle|\lambda_0\right) \geq E\left(u_1(\sigma(\emptyset, \lambda_0))|\lambda_0\right).$$

Thus, from (5) it follows that if $\delta \leq \overline{\delta}$, than

$$U(\sigma^1, \delta, \lambda_0) \geq U(\sigma, \delta, \lambda_0) - \varepsilon. \tag{4}$$

Next, let us now show that it is possible to choose $\overline{\delta}$ such that it approaches 1 as λ_0 approaches 0 or 1. For any δ, there exists $T(\delta)$ such that the expected sum of discounted utilities past time $T(\delta)$ amounts to less than $\varepsilon/2$ and so the utility is captured in the first

$T(\delta)$ periods:[26]

$$U(\sigma, \delta, \lambda_0) \geq E\left(\sum_{t=1}^{T(\delta)} \delta^t u_t(\sigma(h_{t-1}, \lambda_0)) \middle| \lambda_0 \right) - \varepsilon/2.$$

Next, note that the expression

$$E\left(\sum_{t=1}^{T(\delta)} \delta^t u_t(\sigma^1(h_{t-1}, \lambda_0)) \right)$$

is continuous in λ_0 including the extreme points of $\lambda_0 \in \{0, 1\}$ for any given δ. Note also that σ^1 is the optimal strategy if $\lambda_0 = 1$, since then the expected payoff in any given period (independent of the history) is simply the probability that the interpretation is A times p plus the probability that the interpretation is the interpretation is B times $1 - p$. This is maximized by setting the interpretation to A. Similarly if If $\lambda_0 = 0$ the optimal strategy is to interpret things as B in any given period. Thus, maximum likelihood storage rule, σ^1, is optimal for $\lambda_0 \in \{0, 1\}$. Given the continuity, it is within $\varepsilon/2$ of being optimal for any λ_0 close enough to 1 or 0. So, for any δ we can find λ_0 close enough to 1 or 0 for which

$$U(\sigma^1, \delta, \lambda_0) \geq U(\sigma, \delta, \lambda_0) - \varepsilon. \tag{5}$$

which completes the proof. ■

Proof of Proposition 2. We first state an auxiliary result, from Hoeffding (1963), that is useful in proving Proposition 2.

Lemma 1 **(Hoeffding's inequality)** *If $X_1, \ldots, X_t$ are independent and $a_i \leq X_i \leq b_i$ for $i = 1, 2, \ldots, t$, then for $\delta > 0$,*

$$P\left(\sum_{i=1}^{t} \big(X_i - E(X_i)\big) \geq t\epsilon \right) \leq e^{-2t^2\epsilon^2 / \sum_{i=1}^{t}(b_i - a_i)}.$$

Let $n(\lambda)$ be the number of b interpreted signals minus the number of a interpreted signals needed to reach the frontier where $\lambda_t = 1/2$ starting from $\lambda_0 = \lambda$, i.e.,

$$n(\lambda) = \left\lfloor \frac{\log\left(\frac{\lambda}{1-\lambda}\right)}{\log\left(\frac{p}{1-p}\right)} \right\rfloor.$$

The $\lfloor \cdot \rfloor$ reflects starting from a prior below 1/2, and otherwise it would be rounded up.

[26] An upper bound is to set $\frac{\delta^T}{1-\delta} = \varepsilon/2$.

The process $n_t = n(\lambda_t)$ is a random walk in the integers such that n_t is increased by 1 every time there is an interpreted a signal and decreased by 1 every time there is an interpreted b signal. The conditional laws given the states A and B are denoted by P_A and P_B, respectively, and E_A and E_B are the corresponding expectations.

(a) First, note that if $(1-\pi)p + \pi(1-\gamma) > 1/2$ (which is rewritten as $\gamma < \frac{1/2-(1-p)(1-\pi)}{\pi}$), then even if all of the unclear signals are incorrectly interpreted, the majority of signals will still match the true state. Therefore, if the true state is A, then the increments $\Delta n_t = n_{t+1} - n_t$ are positive in expectation, i.e., $E_A(\Delta n_t) > 0$. Moreover, they have bounded first and second moments. It follows from the strong law of large numbers that $(n_t - E_A(n_t))/t$ converges to zero P_A-a.s., which implies that $n_t \to \infty$ P_A-a.s. and $\lambda_t \to 1$ P_A-a.s. The P_B-a.s. convergence of λ_t to zero is proven in a similar same way.

Note that this is automatically satisfied if $\pi < (p-1/2)/p$ for any γ, which establishes the first sentence of the proposition.

(b) Now suppose that $(1-\pi)p + \pi(1-\gamma) < 1/2$ and assume that the true state is B. We claim that P_B assigns positive probability to the event $\lambda_t \to 1$, which coincides with the event $n_t \to \infty$. First, we note that n_t reaches any preset level with positive probability if t is large enough. Therefore, it is sufficient to prove the proposition for n_0 large. Whenever n_t is positive, it is more likely to increase than to decrease, i.e., $P_B(\Delta n_t = 1) \equiv z > 1/2$. As long as this is the case, $E_B(\Delta n_t) = 2z - 1 > 0$ and Hoeffding's inequality states that for any $\epsilon > 0$,

$$P_B\big(n_t - n_0 \leq (2z - 1 - \epsilon)t\big) \leq e^{-t\epsilon^2/2}.$$

Setting $\epsilon = (2z-1)/2$ leads to the bound

$$P_B\big(n_t \leq (z - /2)t\big) \leq P_B\big(n_t - n_0 \leq (z - 1/2)t\big) \leq e^{-t(z-1/2)^2/2}.$$

When n_0 is large, n_t cannot immediately fall below $(z-1/2)t$. More specifically, this is impossible for $t \leq \lfloor n_0/(z+1/2) \rfloor$. It follows that

$$P_B\big(\forall t : n_t > (z - 1/2)t\big) = 1 - P_B\big(\exists t : n_t \leq (z - 1/2)t\big) \geq 1 - \sum_{t > \lfloor n_0/(z+1/2) \rfloor} e^{-t(z-1/2)^2/2}.$$

The last expression is positive if n_0 is large enough. This proves that

$$P_B\big(\lim_{t\to\infty} \lambda_t = 1\big) = P_B\big(\lim_{t\to\infty} n_t = \infty\big) > 0.$$

It can be shown in a similar way that P_A assigns positive probability to the event $\lambda_t \to 0$.

■

Proof of Proposition 4. In the limit, the belief process places a.s. weight 1 on the true state of Nature when the $\gamma = 1/2$ rule is used, as shown in proposition 2 (and could also be deduced from Levy's 0-1 law and Martingale convergence of beliefs). Therefore, the belief process remains eventually on the correct side of the frontier. Formally, under the $\gamma = 1/2$ rule, the random time

$$S = \inf\{t \geq 0 : \forall s \geq t : 1_{\lambda_t > 1/2} = \omega\}$$

is P_ω-a.s. finite for any $\omega \in \{A, B\}$. Therefore, for any T,

$$U(\sigma^T, \delta, \lambda) \geq E\left(1_{S \leq T} \sum_{t=T}^{\infty} \delta^t u_t(\sigma^T(h_{t-1}, \lambda_{t-1}))\right) = E\left(1_{S \leq T} \sum_{t=T}^{\infty} \delta^t u_t(\sigma^{FI}(\omega))\right)$$
$$> E\left(\sum_{t=T}^{\infty} \delta^t u_t(\sigma^{FI}(\omega))\right) - \epsilon/2 > E\left(\sum_{t=0}^{\infty} \delta^t u_t(\sigma^{FI}(\omega))\right) - \epsilon = U(\sigma^{FI}, \delta) - \epsilon.$$

In the above equation, the first relation holds because some non-negative terms are dropped, the second relation holds because the agent calls out the right state past time S, the third relation holds for large enough T because $P(S \leq T) \to 1$ as $T \to \infty$, and the fourth relation holds for δ close enough to 1 because then the first T stages do not matter relative to the rest. ■

Proof of Proposition 5: Solving (2) iteratively, it follows that

$$\widehat{\mu}_t = \mu_0 \left(\times_{\tau=1}^{t} W_\tau\right) + \sum_{\tau=1}^{t} s_\tau \frac{x_\tau^2}{(1 + x_\tau)^2} \left(\times_{\tau'=\tau+1}^{t} W_{\tau'}\right) \tag{6}$$

where

$$W_\tau = \left[\frac{(1 + x_\tau)^2 - x_\tau^2}{(1 + x_\tau)^2}\right].$$

Note that since $x_t = \frac{\sigma_{t-1}^2}{\sigma_s^2}$ and $\sigma_{t-1}^2 = \frac{\sigma_0^2 \sigma_s^2}{\sigma_s^2 + (t-1)\sigma_0^2}$, it follows that $x_t = \frac{\sigma_0^2}{\sigma_s^2 + (t-1)\sigma_0^2}$, and so

$$W_\tau = \frac{(\sigma_s^2 + \tau\sigma_0^2)^2 - (\sigma_0^2)^2}{(\sigma_s^2 + \tau\sigma_0^2)^2} = \frac{(\sigma_s^2 + (\tau - 1)\sigma_0^2)(\sigma_s^2 + (\tau + 1)\sigma_0^2)}{(\sigma_s^2 + \tau\sigma_0^2)^2}.$$

Thus,

$$\left(\times_{\tau=1}^{t} W_\tau\right) = \frac{(\sigma_s^2 + 0\sigma_0^2)(\sigma_s^2 + (t + 1)\sigma_0^2)}{(\sigma_s^2 + 1\sigma_0^2)(\sigma_s^2 + t\sigma_0^2)}$$

or

$$\left(\times_{\tau=1}^{t} W_\tau\right) = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_0^2} \left[1 + \frac{\sigma_0^2}{\sigma_s^2 + t\sigma_0^2}\right].$$

Similarly,

$$\left(\times_{\tau'=\tau+1}^{t} W_\tau\right) = \frac{\sigma_s^2 + \tau\sigma_0^2}{\sigma_s^2 + (\tau+1)\sigma_0^2}\left[1 + \frac{\sigma_0^2}{\sigma_s^2 + t\sigma_0^2}\right].$$

Substituting for these expressions and x_τ into (6), we obtain:

$$\widehat{\mu}_t = \mu_0\left(\frac{\sigma_s^2}{\sigma_s^2+\sigma_0^2}\right)\left[1+\frac{\sigma_0^2}{\sigma_s^2+t\sigma_0^2}\right]+\sum_{\tau=1}^{t} s_\tau\left(\frac{\sigma_0^2}{\sigma_s^2+\tau\sigma_0^2}\right)^2\left(\frac{\sigma_s^2+\tau\sigma_0^2}{\sigma_s^2+(\tau+1)\sigma_0^2}\right)\left[1+\frac{\sigma_0^2}{\sigma_s^2+t\sigma_0^2}\right]. \tag{7}$$

(7) gives the expression claimed in the proposition. The expression for the Bayesian updater is standard:

$$\mu_t = \mu_0\frac{\sigma_s^2}{\sigma_s^2+t\sigma_0^2} + \sum_{\tau=1}^{t} s_\tau\frac{\sigma_0^2}{\sigma_s^2+t\sigma_0^2},$$

and completes the proof. ▮

Proof of Proposition 3. The argument in the proof of proposition 2b shows that when $\pi < 1$, then the values 0 and 1 occur with positive probability as the limit of the belief process λ_t as $t \to \infty$. It remains to show that $\lambda_t \to 1$ with probability tending to one as $\pi \to 1$ if $\lambda_0 > 1/2$. So let us assume $\lambda_0 > 1/2$. Obviously, in the extreme case that $\pi = 1$, all signals are interpreted as a and therefore, the probability that $\lambda_t \to 1$ equals 1. This probability depends continuously on the parameter π by Lemma 2. ■

Lemma **2** *Under any strategy, the belief process Λ is continuous in the total variation norm with respect to the parameters p and π.*

Proof. It is equivalent to show the proposition for the point process n_t defined in the proof of Proposition 2 instead of the process λ_t. Let $p^k \to p$, $\pi^k \to \pi$, and let P^k and P be the corresponding laws of the process n_t. Furthermore, let $\mathcal{F}_t$ be the sigma algebra generated by $n_0, \dots, n_t$ and P_t^k the restriction of P^k to $\mathcal{F}_t$. It follows directly from the Chapman-Kolmogorov equations or from Jacod and Shiryaev (2003, corollary V.4.39a) applied to the point process $(n_t - n_0 + t)/2$ that the total variation of the signed measure $P_t^k - P_t$ tends to zero, i.e.,

$$\|P_t^k - P_t\| = \sup\{|P_t^k(\phi) - P_t(\phi)| : \phi\ \mathcal{F}_t\text{-measurable function on } \Omega \text{ with } |\phi| \le 1\} \to 0.$$

The restriction that ϕ is $\mathcal{F}_t$-measurable can be removed by an approximation argument: for any $\mathcal{F}_t$-measurable function ϕ_t, one has

$$\begin{aligned}|P^k(\phi) - P(\phi)| &\le |P^k(\phi) - P^k(\phi_t)| + |P^k(\phi_t) - P(\phi_t)| + |P(\phi_t) - P(\phi)| \\ &\le |P^k(\phi) - P^k(\phi_t)| + \|P^k - P\| + |P(\phi_t) - P(\phi)|.\end{aligned} \tag{8}$$

Setting k large enough, $\|P_t^k - P_t\|$ can be made smaller than $\epsilon/3$. Then ϕ_t can be set equal to the $\mathcal{F}_t$-conditional expectation of ϕ under the measure $(P^k + P)/2$. It follows that $\phi_t \to \phi$ a.s. under P^k and P. By the dominated convergence theorem, the first and third term in the right-hand side of equation (8) are smaller than $\epsilon/3$ when t is large enough. It follows that (8) is arbitrarily small for large enough values of k. Thus P^k converges to P in the total variation norm. ■

9 Appendix B: Data Appendix

Interpretation of Summaries
These variables are on a scale of -8 to 8.

- -8 implies "I am certain that the death penalty does NOT deter people from committing murder"
- 0 implies "I am not certain whether the death penalty deters people from committing murder"
- 8 implies "I am certain thatt he death penalty DOES deter people from committing murder."

We used the raw numbers that participants entered. Climate change questions were worded identically.

Prior Beliefs
These variables are on a scale of -8 to 8. To the scale is identical to that described above. We used the raw numbers that participants entered.

Gender
This question was a free response. We coded the variable as a 1 if participants entered "F", "Female", "Woman" etc. We coded the variable as 0 if the participant entered "Man", "M", "Male" etc. We coded missing gender as 1 if participants left the answer blank, or entered something that was not decipherable, and 0 otherwise.

Additional Demographic Indicators: For each of the following variable categories, participants were required to select an answer to continue. Each of the variables corresponds to a single answer choice within the category that was offered to participants. We coded these variables as 0 if the participant did not select the corresponding box, and 1 otherwise.

Race/Ethnicity
Answer choices include: Black, Chinese, Indian, Other asian, Hispanic, Native American, White, Other race, or Prefer not to answer. Individuals could mark multiple choices.

Religion
Individuals could choose: Buddhist, Hindu, Christian, Jewish, Muslim, Not religious, or Prefer not to answer

Educational Attainment
We asked what was the highest level of education that the participant achieved. Answer choices and corresponding indicator variables were Some High School, High School Graduate, College with no degree, Bachelor's Degree, Graduate degrees (Master, PhD, etc.), Other and Prefer not to answer.

College - We coded the variable College as equal to 1 if particpants selected Bachelor's degree or Graduate degrees. We coded the variable as 0 if participants did not select Bachelor's degree or Graduate degree and did not select Prefer not to answer.

Employment Status
Employment status choices include: Employed, A Student, Unemployed and seeking work, Not formally employed and not seeking formal employment, Retired, Other, or Prefer not the answer.

Political affiliation
Indicators were constructed for Democrat, Republican, Independent, and Prefer not to answer

Wages and Annual Income
Hourly wage categories include: \$0.00-\$2.00, \$2.01-\$4.00, \$4.01-\$7.00, \$7.01-\$10.00, \$10.01-\$15.00, \$15.01-\$20.00, \$20.01-\$30.00, \$30.01-\$50.00, \$50.01 or more and Prefer not to answer.

For the purposes of calculating the mean in Table 1, we used the midpoint of each category (e.g. \$1.00 for the first category, \$3.00 for the next category, etc.). We coded the \$50.00 or more category as equal to \$60.00.

Approximate annual income categories include: \$0.00-\$5,000, \$5,001-\$10,000,\$10,001-\$20,000, \$20,001-\$30,000, \$30,001-\$40,000, \$40,001-\$60,000, \$60,001-\$80,000, \$80,001 or more and

Prefer not to answer.

For the purposes of calculating the mean in Table 1, we used the midpoint of each catogory (e.g. \$2,500 for the first category, \$7,500 for the next category etc.). We coded the \$80,000 or more category as equal to \$100,000.

Location
We first defined the boundaries of the continental United States using the following boundary lines: Western boundary = -124.8 degrees; Eastern boundary = -66.9 degrees; Northern boundary = 49.4 degrees; Southern boundary = 24.4 degrees. Then we split it into quadrants along the East-West line = -95.85 degrees and the North-South line = 36.9 degrees.

We used the location coordinates that were recorded based on the participant's IT address to assign each participant to a quadrant: Northwest, Northeast, Southwest and Southeast.

We also collected location type indicator variables: Urban, Suburban, Rural, and Prefer not to answer

Appendix Table 1: Summary of Belief Divergence Results

Lord, Ross, and Lepper (1979)	Experimental subjects were provided with evidence for and against the detterant effect of the death penalty. Subjects of all beliefs report that the article matching their baseline is more convincing, and students became more confident in their original position.
Darley and Gross (1983)	Subjects were asked to rate a student's academic ability and performance after seeing different videos of the student's playground either a poor-looking inner-city neighborhood or a wealthier-looking suburban neighborhood. Subjects gave lower grades in the inner-city treatment. Subjects who also viewed a video of the child answering a variety of quiz questions (some correctly, some incorrectly, sometimes paying attention, sometimes not) before rating the child displayed even greater divergence.
Plous (1991)	Subjects with varying opinions on nuclear energy and deterrence were provided with articles on the Three Mile Island disaster and a narrowly-averted accidental missile launch. Subjects of all viewpoints expressed increased confidence in their original viewpoints after reading the articles.
Munro and Ditto (1997)	Subjects with high and low levels of prejudice towards homosexuals were presented with two fictional studies on the empirical prevalence of a homosexual stereotype. Follow-up interviews revealed evidence of both biased assimilation and attitude polarization.
Russo, Meloy, and Medvec (1998)	Experimenters sequentially provided subjects with information on two fictional brands. In later stages, once participants have formed preferences, neutral information causes subjects to identify more strongly with their preferred brand.
McHoskey (2002)	Students were randomly selected to review either information supporting the claim that Lee Harvey Oswald acted alone in assassinating John F. Kennedy and or information pointing to a larger conspiracy. Students with extreme opinions intensify their positions when presenting with information supporting their beliefs and relax their beliefs to a lesser degree when confronted with contradictory evidence.
Kahan et al (2007)	Subjects were surveyed on their beliefs about the safety of nanotechnology after half were randomly provided with factual information about risks and benefits. Those who were exposed to information displayed greater polarization than those who were not.
Nyhan and Reifler (2010), Nyhan, Reifler, and Ubel (2013)	These studies detail a series of five experiments in which participants are asked to assess the validity of a false or misleading statement by a politician. In each case, the additional information leads the most-committed members of the targeted subgroup to intensify their misperceptions, rather than weakening them.

Appendix Table 2: Regressions of Interpretations on Priors - Question by Question

	Summary 1	Summary 2	Summary 3	Summary 4	Summary 5	Summary 6
	(1)	(2)	(3)	(4)	(5)	(6)
Climate Change						
Prior Belief	0.147***	0.117***	0.084***	0.100***	0.032	0.079***
	(0.025)	(0.024)	(0.028)	(0.027)	(0.026)	(0.024)
Constant	5.210***	5.587***	-5.663***	-5.311***	-0.850***	0.200
	(0.131)	(0.124)	(0.148)	(0.138)	(0.135)	(0.127)
Observations	608	608	608	608	608	608
Death Penalty						
Prior Belief	0.070***	0.084***	0.127***	0.102***	0.051***	0.036
	(0.025)	(0.028)	(0.030)	(0.027)	(0.016)	(0.030)
Constant	5.569***	3.535***	-3.600***	-5.132***	-0.118	-1.823***
	(0.123)	(0.136)	(0.148)	(0.131)	(0.079)	(0.147)
Observations	608	608	608	608	608	608

Notes: This table presents estimates of the influence of prior beliefs on interpretation of each individual summary. Column (1) refers to summary 1, column(2) refers to summary 2 etc. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Appendix Table 3: Main Results with Controls

	Climate Change		Death Penalty	
	(1)	(2)	(3)	(4)
Pro Abstracts				
Prior Belief	0.069**	-0.066	0.080***	0.108
	(0.028)	(0.061)	(0.025)	(0.073)
Constant	6.885***	-2.093	9.294***	14.480***
	(1.495)	(4.279)	(2.605)	(4.555)
Observations	982	208	982	208
Con Abstracts				
Prior Belief	0.060***	0.149***	0.054**	0.064
	(0.022)	(0.050)	(0.024)	(0.067)
Constant	-5.423**	3.739	-2.024	0.598
	(2.124)	(3.594)	(2.060)	(6.650)
Observations	1,473	312	1,473	312
Unclear Abstracts				
Prior Belief	0.030	0.151	0.044*	0.042
	(0.033)	(0.097)	(0.024)	(0.078)
Constant	11.468***	1.789	-1.254	-10.993*
	(2.676)	(6.193)	(2.344)	(6.455)
Observations	491	104	491	104

Notes: This table presents estimates of the influence of prior beliefs on interpretation of the summaries by category of summary, controlling for a full set of demographic variables. The Data Appendix contains a description of all demographics and their definitions. 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. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively. All standard errors are clustered at the individual level.