A Theory of Rational Attitude Polarization*

Jean-Pierre Benoît
London Business School

Juan Dubra
Universidad de Montevideo

Revised January, 2016.

Abstract

Numerous experiments have demonstrated the possibility of *attitude polarization*. For instance, Lord, Ross & Lepper (1979) partitioned subjects into two groups, according to whether or not they believed the death penalty had a deterrent effect, and presented them with a mixed set of studies on the issue. Believers and skeptics both became more convinced of their initial views; that is, the population polarized. Many scholars have concluded that attitude polarization shows that people process information in a biased manner. We argue that not only is attitude polarization consistent with an unbiased evaluation of evidence, it is to be expected in many circumstances where it arises. At the same time, our theory identifies situations where the population should not polarize when given mixed evidence, as some experiments confirm.

Keywords: Attitude Polarization; Confirmation Bias; Bayesian Decision Making.
Journal of Economic Literature Classification Numbers: D11, D12, D82, D83

According to Gallup surveys, since the early 1990s around 68% of African Americans have held the view that the American justice system is biased against blacks. During the same time period, the percentage of whites who share this belief has dropped from 33% to 25%. Moving from beliefs to data, several studies have shown that police "stop and frisk" racial and ethnic minority members at higher rates than whites. What impact can these studies be expected to have, or to have had, on the views of blacks and whites on the American justice system?[1]

*We thank Gabriel Illanes and Oleg Rubanov for outstanding research assistance. We also thank Vijay Krishna, David Levine, Michael Mandler, Frederic Malherbe, Wolfgang Pesendorfer, Madan Pillutla, Debraj Ray, Jana Rodríguez-Hertz, Andrew Scott, and Stefan Thau for valuable comments.

[1]Gallup survey data can be found at http://www.gallup.com/poll/163610/gulf-grows-black-white-views-justice-system-bias.aspx. The Sentencing Project (2014) contains a discussion of this and related data. One study on police stops is Gelman, Fagan, and Kiss (2007).

More generally, how should we expect groups of people with differing opinions on an issue to react to the same piece of information? In a classic study, Lord, Ross and Lepper (1979) took two groups of subjects, one which believed in the deterrent effect of the death penalty and one which doubted it, and presented them with the same mixed evidence on the issue. Both groups became more convinced of their initial positions. Numerous, though by no means all, subsequent experiments, on a variety of issues, have also found that exposing people who disagree to the same mixed evidence may cause their initial attitudes to move further apart, or *polarize*.[2] Many scholars have concluded that these results provide evidence that people often process information in a biased manner, so as to support their pre-existing views. We argue that, on the contrary, this polarization of attitudes is often exactly what we should expect to find in a perfectly Bayesian population.

To begin, it is important to recognize that there are two aspects to attitude polarization, *pairwise polarization* and *population polarization*. Pairwise polarization occurs when the opinions of a particular pair of individuals move further apart after they receive a common piece of information. Population polarization occurs when this separating is systematic, so that the opinions of the population on the whole diverge. As we will see, population polarization is the relevant aspect to consider for the question of whether people update in a biased fashion.

Of course, population polarization cannot take place without pairwise polarization. Accordingly, the first part of our argument is that pairwise polarization is consistent with Bayesian updating and not particularly surprising, when viewed properly. This observation has been made by others as well, including Walley (1991), Seidenfeld and Wasserman (1993) and Andreoni and Mylovanov (2012) and we discuss their work in Section 2. We also provide a characterization of the conditions under which Bayesian pairwise polarization can arise, which has been missing from the literature.

Although pairwise polarization has been the focus of much research, the essential challenge posed by the attitude polarization literature does not, in fact, pertain to this aspect. On the contrary, a typical attitude polarization experiment deliberately crafts the information given to subjects to be ambiguous enough to legitimately have a positive or a negative impact on beliefs. In this way, some pairwise polarization arises quite naturally. It is population polarization that the attitude polarization literature takes as evidence of bias.

Consider Plous' (1991) well-cited nuclear deterrence experiment. Plous began by dividing his subjects into two groups, according to whether they entered the experiment with a belief

[2]Papers on attitude polarization include Darley and Gross (1983), Plous (1991), Miller, McHoskey, Bane, and Dowd (1993), Kuhn and Lao (1996), and Munro and Ditto (1997). Some experiments track both people's positive beliefs (e.g., do you believe capital punishment has a deterrent effect?) and normative opinions (e.g., are you in favour of capital punishment?). Throughout this paper, we only discuss movements in positive beliefs, as it is less clear how to evaluate changes in normative opinions.

that a strategy of nuclear deterrence made the United States safer or less safe. He then gave all subjects the same article to read, describing an actual incident where an erroneous alert caused the United States to enter a heightened state of readiness for nuclear war with the Soviet Union. The crisis lasted only three minutes, as officials quickly realized the alert was a false alarm. After reading the article, each groups' views on the safety of nuclear deterrence moved further in the direction of its initial inclinations.

How should unbiased subjects have reacted to the article? As Plous writes, "Given the fact that (a) the system malfunctioned and (b) the United States did not go to war despite the malfunction, the question naturally arises as to whether this breakdown indicates that we are safer or less safe than previously assumed." Plous deliberately chose the article so that some pairwise polarization was to be expected. This expectation comes from the fact that the evidence in the article is equivocal and its implications depend on beliefs about an ancillary consideration, to wit, whether it is more important for a system's safety that it have a well-functioning primary unit or that it have effective safeguards. It is not at all clear which one is more important, and a person could legitimately believe either one is, depending upon his previous information on the matter. A person who believed that the primary unit was more important would revise downwards his belief in the safety of nuclear deterrence, while a person who believed that safeguards were more important would revise upwards. A fortiori, the fact that two particular people polarized – an opponent of nuclear deterrence became more opposed while a proponent became more in favour – does not pose a challenge to unbiased reasoning. Results on pairwise polarization formalize this reasoning.

However, even if people can legitimately update in different directions, a challenge remains. Why would it be that, on the whole, the subjects who were initially in favour of nuclear deterrence responded positively to the evidence, while those who were initially opposed responded negatively? Put differently, why would it be that people who believed in the safety of nuclear deterrence also believed that safeguards were paramount, while people who were skeptical of nuclear deterrence also believed that primary units were crucial, rather than beliefs in these two dimensions being, say, uncorrelated? If these beliefs were uncorrelated, while there would be many instances of pairwise polarization, there would be just as many instances of pairs converging; overall these instances would cancel each other out and the population would not polarize. It is the fact that the population polarized which led Plous to conclude that people process information in a biased manner to support their initial beliefs. Bayesian explanations for pairwise polarization do not predict this population polarization, whereas, say, Rabin and Schrag's (1999) theory of confirmation bias does.

Is the conclusion of bias warranted? We now argue that it is not.

We are told that most of the subjects in the experiment knew of the false alarm incident before entering the experiment, though, presumably, they did not know all of the details provided in the article. (In a variant treatment, which also yielded population polarization,

subjects were provided with descriptions of near-miss incidents that were unfamiliar to them, rather than descriptions of an incident they were already aware of.) Which subjects would have entered the experiment with a favourable view of nuclear deterrence?

A reasonable presumption is that the subjects who entered with a favourable view, despite their knowledge of a previous malfunction that was caught by safeguards, are the ones that considered the reliability of safeguards to be more important than the reliability of the primary unit. These subjects would naturally tend to increase their belief that nuclear deterrence is safe after being given further evidence of properly functioning safeguards. On the other hand, subjects that considered a malfunction of the primary unit to be dispositive would have a negative view initially and would tend to revise downwards after being given further evidence about a shaky primary unit. Thus, population polarization is not only consistent with unbiased reasoning but even to be expected, at least in Plous' experiment.

In Lord, Ross and Lepper's (1979) capital punishment experiment, subjects were presented with a common piece of evidence that was "characteristic of research found in the current literature". Again, it is hardly surprising that it is those subjects for whom current evidence had previously led to a favourable conclusion on the efficacy of the death penalty that responded positively to additional similar evidence. (We return to this experiment in Section 4.5.

Darley and Gross (1983) is an influential study that uses a different methodology. We discuss how our model applies to it in sections 2.1 and 4.6. For now we note that, although this experiment is usually cited as providing **Do you want the word strong here or not?** strong evidence of biased reasoning, in fact it only finds polarization in 4 out of 8 instances.

Our general rationale for population polarization is as follows. Consider a group of people with differing opinions on an issue – the available information is equivocal and has induced positive views in some of them and negative views in others. Now suppose the group is exposed to an additional piece of information and that this information is similar in nature to the previous body of information. Those who previously considered this type of information to be positive are more likely to respond favourably than those who considered it to be negative, so that the population will polarize.

In contrast to a simple biased reasoning story, our population polarization argument implies that populations will not always polarize. Suppose the additional mixed information given to subjects in the experiment is novel in character. While some people may react positively to this information and others react negatively, or neutrally, there is no reason for their reactions to correlate with their initial positions, since these positions were formed on a completely different basis. Hence, there is now no reason to expect a particular pattern of belief changes at the level of the population. This no-population polarization prediction is consistent with some experimental findings. Previous Bayesian theories of pairwise polarization and bias theories of population polarization make no special distinction between

mixed information experiments where polarization occurs and those where it does not; in that regard, our theory better addresses the phenomenon of attitude polarization.

Our theory also makes some unexpected predictions, including that polarization will be especially pronounced among experts on an issue and among people with extreme beliefs. These predictions find some support in existing experimental results. To be more precise, the predictions find support under natural interpretations of those results as they are described in the experiments. However, those descriptions can be somewhat loose and other interpretations are also possible. This is perhaps unsurprising, as the experiments were not designed as tests of our theory. More surprisingly, many existing experiments do not provide good tests of a bias theory, either. Our results can serve as the basis for tests.

In Plous' experiment, in addition to asking his subjects for their views on nuclear deterrence, he explicitly asked them which was more significant for evaluating safety, the reliability of primary units or safeguards. Consistent with our reasoning, he found that those who believed that primary breakdowns were more significant revised downwards their belief in nuclear deterrence while those who felt that safeguards were more significant revised upwards.

However, Plous' reasoning on this finding is essentially the reverse of ours. Our logic can be summarized as: A belief that safeguards are important, combined with evidence that safeguards have worked in the past, has led some people to enter the experiment with a favourable view of a strategy of nuclear deterrence. These people tend to revise upwards when presented with additional evidence of safeguards working. Plous' logic is: Some people enter the experiment with a favourable view of a strategy of nuclear deterrence (for unspecified reasons). A desire to enhance that view leads them to believe that safeguards are important and to revise upwards.

In a similar vein, Plous found a strong correlation between an opposition to nuclear energy and a belief that the accident at the nuclear power plant in Chernobyl was relevant for the United States. For him, this is evidence that people assess the relevance of Chernobyl in a biased manner. Specifically, opponents of nuclear energy want to maintain this belief and so decide that Chernobyl is relevant, while proponents decide that it is not relevant. For us, the reverse is true or, at least, cannot be ruled out – people who feel that Chernobyl is relevant conclude that nuclear energy is not safe and are thus opponents at the time that Plous questions them; people who continue to favour nuclear energy are those that believe that Chernobyl is not relevant.

As we see, much evidence from attitude polarization experiments is consistent both with biased and unbiased reasoning. To help disentangle the two hypotheses, consider these implications of our model.

1. If the common evidence that people are presented with is novel in nature, the popu-

lation should not polarize. The reason is that supporters and opponents will not have been pre-sorted according to their reactions to this kind of evidence and so there is little reason for supporters to react more favourably than opponents (see Theorem 8).

Consistent with this prediction, Miller, McHoskey, Bane, and Dowd (1993) find no population polarization on the issue of the merits of affirmative action when subjects were presented with arguments that seemed unfamiliar to them (we provide greater detail in Section 1).

2. A population of people who have largely based their initial opinions on very similar evidence on the issue will be especially prone to polarization, as they will have been well sorted. In particular, this applies to experts that all have a good understanding of the current body of evidence on the issue but nevertheless disagree (see Theorem 4). This is consistent with Plous' finding that people who report high "issue involvement" polarize the most.

3. Groups with strong opinions polarize more (Theorem 5). For instance, the strongest believers in the deterrent effect of the death penalty will be the most likely to increase their belief and the strongest doubters will be the most likely to decrease their belief. This is consistent with Plous (1991) and with Miller, McHoskey, Bane, and Dowd (1993), who find that subjects with the strongest conviction are more likely to polarize. In addition, in many experiments, including Lord, Ross and Lepper, subjects are pre-selected to have strong convictions. On the other hand, Kuhn and Lao (1996) do not find that strength of opinion matters.

It is worth emphasizing the logic of attitude polarization experiments. They do not inquire as to why subjects' initial beliefs differ or whether or not these beliefs are rational or unbiased to begin with. Rather, the experiments implicitly accept that beliefs can legitimately differ and recognize that it is difficult to determine if beliefs have been rationally derived without knowing the information upon which they are based. Attitude polarization experiments circumvent this difficulty by examining how groups update their beliefs in response to a known piece of information. In this way, many of these experiments manage to examine beliefs about naturally occurring issues, rather than beliefs on artificial issues generated in the lab.

At the same time, these experiments are not about the persistence of disagreement or whether disagreement is common knowledge. Such issues are more or less orthogonal to the literature. At a theoretical level, while differences in beliefs that persist and are common knowledge could be explained by, for instance, positing that individuals start with different (inconsistent) priors or that rationality is not common knowledge – assumptions that would be neither here nor there for many psychologists – these assumptions would not explain why

or when populations polarize. At an empirical level, attitude polarization experiments have little to say about whether or not disagreement persists, as subjects are usually pre-selected from a larger pool precisely for their conflicting opinions. Even if beliefs in a population are converging, along the path to convergence it will always be possible to find subjects with conflicting beliefs.[3]

This paragraph is awfully repetitive, but we can keep it if you think it is better to have it. (I tried to think of a way t make it less repetitive but did not come up with anything, other than changing the word "addresses" to "explains" in the last sentence. As the prior literature stands, Bayesian theories address pairwise polarization but have little to say about population polarization, while population polarization is addressed by non-Bayesian theories. These previous theories either explain too little – pairwise polarization but not population polarization, which is the key finding of attitude polarization experiments – or too much – polarization whenever there is disagreement (and the signal is mixed). In contrast, our theory explains both the polarization found in some experiments and the absence of polarization in others.

In Section 4.1, we give a simple numerical example of population polarization, which the reader can consult now if he or she is so inclined. In Section 1, we present a formal model of population polarization; in Section 1.2, we provide conditions under which we would not expect polarization. In Section 1.3, we characterize the conditions under which pairwise polarization occurs. In Section 2, we discuss the relationship of our work to the theoretical literature on attitude polarization and take a critical look at some experimental findings. The appendix examines some nuances of polarization and contains all proofs.

1 Formal Analysis

The essential elements of an attitude polarization study, as we see it, are the following. There is an issue of interest. Subjects have private information about the issue. They are provided with a common piece of evidence that, in some intuitive sense, bears directly on the issue. Subjects also have private information about an ancillary matter, which has little direct bearing on the issue but affects the interpretation of the evidence.[4]

The minimal setting that can capture these elements is one in which there is a proposition

[3]Miller, McHoskey, Bane, and Dowd (1993) is one of the few experiments that gives information about beliefs in the total pool of subjects. On the whole, they favored capital punishment to begin with – out of 337 participants, 251 were in favour and 86 were opposed. In Section 4.2, we show that population polarization is consistent with beliefs converging, in a common priors setting.

[4]For instance, the issue could be the safety of nuclear power, the evidence on the issue data on accidents and near-accidents in nuclear power plants and the ancillary matter the relative importance of primary units and safeguards.

about the issue that can take one of two values, say, it can be true or false, and there is an ancillary matter that can be in one of two states, say high or low. (In Section 1.1, we show how the model can be generalized beyond this 2×2 framework.) We make the stark assumption that the ancillary matter, in and of itself, has *no* direct bearing on the proposition; that is, information about the ancillary matter alone causes no revision in beliefs about the main issue.[5] Formally, the ancillary matter and the issue of concern are statistically independent in the prior.

The following is a straightforward Bayesian model (with common priors).

1. Nature chooses true or false for the proposition with probability $(a, 1-a)$ and, independently, high or low for the ancillary state with probability $(b, 1-b)$, where $1 > a, b > 0$. Thus, the prior over the possible states of nature is:

$$\begin{array}{ccc} & \text{Prior} & \\ & \text{True} & \text{False} \\ \text{High} & ab & (1-a)\,b \\ \text{Low} & a\,(1-b) & (1-a)\,(1-b) \end{array} \tag{1}$$

   We denote the state space by $\Omega = \{H, L\} \times \{T, F\}$.

2. Each individual receives a pair of private signals (s, σ).

   (a) The first element is a signal about the issue drawn from a finite sample space $\mathcal{S}$. The likelihood matrix for a signal $s \in \mathcal{S}$ is

$$\begin{array}{ccc} & \text{Likelihood of } s & \\ & \text{True} & \text{False} \\ \text{High} & p_s & q_s \\ \text{Low} & r_s & t_s \end{array} \tag{2}$$

   where $1 > p_s, q_s, r_s, t_s > 0$. Although we describe s as a single signal, it can be thought of as the sum total of the information the individual has about the issue.

   (b) The second element, σ, is a signal about the ancillary matter. The signal is drawn from a density $\pi_{\mathrm{H}}(\cdot)$ with support $[0, 1]$ when the ancillary state is high, and from the density $\pi_{\mathrm{L}}(\cdot)$ with support $[0, 1]$ when the ancillary state is low. We assume that $\frac{\pi_{\mathrm{H}}(\cdot)}{\pi_{\mathrm{L}}(\cdot)}$ is increasing in σ, so that the monotone likelihood ratio property is satisfied, and that $\lim_{\sigma \to 1} \frac{\pi_{\mathrm{H}}(\sigma)}{\pi_{\mathrm{L}}(\sigma)} = \infty$ and $\lim_{\sigma \to 0} \frac{\pi_{\mathrm{H}}(\sigma)}{\pi_{\mathrm{L}}(\sigma)} = 0$. The last

[5] Thus, just being told that safeguards are more important for safety than primary systems, without being given any information on the performance of nuclear power plants, says nothing about whether or not such plants are safe. Or, learning that a particular policy has been adopted because of political reasons unrelated to selection issues (as in Galiani *et al.* (2005), who discuss the privatizaton of water in Argentina) says nothing about the effectiveness of that policy.

two assumptions, as well as the assumption that the signal is drawn from $[0,1]$, rather than a finite sample space, are for ease of exposition. Note that, just as the ancillary matter by itself is unrelated to the truth of the proposition, we also assume that the signal about the ancillary matter is unrelated to the truth of the proposition.

Subject i , who has seen (s_i, σ_i), has **initial belief** about the truth of the proposition given by $P(T \mid s_i, \sigma_i)$.

3. All individuals observe a common signal $c \in \mathcal{C}$ with likelihood matrix:

Likelihood of c:

	True	False
High	p_c	q_c
Low	r_c	t_c

where $1 > p_c, q_c, r_c, t_c > 0$

Subject i's **updated belief** is $P(T \mid s_i, \sigma_i, c)$.

Definition 1 *Consider two individuals i and j who have received signals (s_i, σ_i) and (s_j, σ_j), respectively, and suppose that $P(T \mid s_i, \sigma_i) \geq P(T \mid s_j, \sigma_j)$. The pair **polarizes** if $P(T \mid s_i, \sigma_i, c) > P(T \mid s_i, \sigma_i)$ and $P(T \mid s_j, \sigma_j, c) < P(T \mid s_j, \sigma_j)$.*

The significance of the ancillary matter is that it can affect the interpretation of a signal. In the case of interest to us, a change in the ancillary state reverses the impact of a signal – for instance, if the state is high, the signal supports the proposition, while if the state is low, the signal goes against it. The condition for this to happen is that the signal be equivocal, as in the following definition.

Definition 2 *The signal c is **equivocal** if either i) $p_c > q_c$ and $r_c < t_c$ or ii) $p_c < q_c$ and $r_c > t_c$.*

We have the following theorem:

Theorem 1 *The signal c is equivocal if and only if either i) $P(T \mid \mathrm{H}, c, s) > P(T \mid \mathrm{H}, s)$ and $P(T \mid \mathrm{L}, c, s) < P(T \mid \mathrm{L}, s)$ for all $s \in \mathcal{S}$, or ii) $P(T \mid \mathrm{H}, c, s) < P(T \mid \mathrm{H}, s)$ and $P(T \mid \mathrm{L}, c, s) > P(T \mid \mathrm{L}, s)$ for all $s \in \mathcal{S}$. Moreover, if $p_c > q_c$ and $r_c < t_c$ then i) holds, while if $p_c < q_c$ and $r_c > t_c$ then ii) holds.*

All proofs are in the appendix.

- Without loss of generality, from now on we assume that when a signal $m = s, c$ is equivocal, $p_m > q_m$ and $r_m < t_m$. Thus, when the ancillary state is high, an equivocal signal increases the belief that the proposition is true; when the ancillary state is low, an equivocal signal decreases this belief.

The next result extends Theorem 1 to non-degenerate beliefs about the ancillary state.

Theorem 2 *Suppose c is equivocal. For all $s \in \mathcal{S}$, there exists an h_s such that $P(\mathrm{H} \mid s, \sigma) > h_s$ implies $P(T \mid c, s, \sigma) > P(T \mid s, \sigma)$ and $P(\mathrm{H} \mid s, \sigma) < h_s$ implies $P(T \mid c, s, \sigma) < P(T \mid s, \sigma)$.*

For any given signal about the issue, upon receiving an equivocal c, people with a large belief that the ancillary state is high revise upwards their beliefs that the proposition is true, while those with a small belief revise downwards. Although it may not always be obvious to the researcher what the ancillary matter is, in Plous (1991) it is pretty clear that the ancillary matter that renders near-misses equivocal is the relative importance of safeguards and the primary system. Specifically, a high state corresponds to safeguards being more important and a low state corresponds to primary units being more important. Plous provides somewhat of a direct test of Theorem 2, as he asks his subjects which is more important, the fact that safeguards worked or the fact that a breakdown occurred and, consistent with the theorem, he finds that those who feel that safeguards are more important revise upwards their beliefs that nuclear deterrence is safe while those who believe that breakdowns are more important revise downwards.

So far, we have analyzed how beliefs about the ancillary matter affect updating. However, the bulk of the work on attitude polarization is on how initial beliefs about the *issue* affect updating. We turn now to this question.

Subject i's previous information about the issue is summarized by s_i. If the equivocal common signal that the subject is given in the experiment is typical of existing information about the issue, as is explicitly the case in many experiments, we may expect that the subject's previous information was equivocal as well. The next result shows that, in that case, a person with a high initial belief in the truth of the issue revises upwards, while a person with low initial belief revises downwards.

Theorem 3 *Suppose that s and c are both equivocal. There exists a v_s such that $P(T \mid s, \sigma) > v_s$ implies $P(T \mid s, c, \sigma) > P(T \mid s, \sigma)$ and $P(T \mid s, \sigma) < v_s$ implies $P(T \mid s, c, \sigma) < P(T \mid s, \sigma)$.*

The reasoning behind this theorem is the following. If a person has observed an equivocal signal in the past, a large belief in the truth of the proposition indicates a large belief that the ancillary state is high (Lemma 2 in the appendix). In turn, a large belief that the ancillary state is high leads to an upward revision that the proposition is true following another equivocal c (Theorem 2). Theorem 3 combines these two results.

Theorem 3 concerns how individuals update. Theorem 9 in Section 1.3 extends this result to give precise conditions for pairwise polarization. Since our main interest is population polarization, we now move from individuals to the population. We begin with some definitions for polarization at the population level.

- Given $v \in (0,1)$ let P^v be the fraction of the population that initially believes the proposition to be true with probability greater than v and let P_v be the fraction that initially believes the proposition to be true with probability less than v. We think of the population as being "large", so that we identify the fraction of the population who have a particular belief with the probability of such a belief arising.

Definition 3 *Following a common signal c, the* ***population polarizes around v*** *if the fraction of those who initially believe the proposition to be true with probability greater than v that revises upwards is strictly greater than the fraction with initial belief less than v that revises upwards, and P^v, $P_v > 0$. Formally, a population, where individuals have observed potentially different signals s and σ, polarizes around v if the event $E_v = \{s, \sigma : P(T \mid s, \sigma) < v\}$ and its complement E_v^C have positive probability and*

$$P\left(s, \sigma : P(T \mid s, c, \sigma) > P(T \mid s, \sigma) \mid E_v^C\right) > P\left(s, \sigma : P(T \mid s, c, \sigma) > P(T \mid s, \sigma) \mid E_v\right).$$

Definition 4 *Following a common signal c, the* ***population polarizes completely around v*** *if everyone who initially believes the proposition to be true with probability greater than v revises upwards and everyone with belief less than v revises downwards, and P^v, $P_v > 0$. Formally,*

$$\begin{aligned} P(T) &> v \Rightarrow P(T \mid c) > P(T) \\ P(T) &< v \Rightarrow P(T \mid c) < P(T) \end{aligned}$$

and P^v, $P_v > 0$.

Definition 5 *Following a common signal c,* ***groups with the strongest opinions polarize completely*** *if there are $\overline{v}$ and $\underline{v} > 1 - \overline{v}$ such that everyone who initially believes the proposition to be true with probability greater than $\overline{v}$ revises upwards while everyone who believes the proposition to be false with probability greater than $\underline{v}$ revises downwards, and $P^{\overline{v}}, P_{1-\underline{v}} > 0$. Formally,*

$$\begin{aligned} P(T) &> \overline{v} \Rightarrow P(T \mid c) > P(T) \\ P(T) &< 1 - \underline{v} \Rightarrow P(T \mid c) < P(T) \end{aligned}$$

and $P^{\overline{v}}, P_{1-\underline{v}} > 0$

Definition 5 is especially important given that there is some evidence that polarization is more marked between sub-populations with the strongest opinions. Moreover, many experiments, including Lord, Ross and Lepper, pre-select people with strong opinions. When groups with the strongest opinions polarize completely, there will be a range of $\bar{w}$'s and $\underline{w}$'s such that most people who believe the proposition with probability greater than $\bar{w}$ increase their beliefs, while most people who disbelieve the proposition with probability greater than $\underline{w}$ increase their disbelief.

The following proposition follows immediately from definition 4.

Proposition 1 *If the population polarizes completely around some v^*, then the population polarizes around v, for all v with P^v, $P_v > 0$.*

Consider an issue on which various researchers have carried out studies. Each study provides a signal about the issue. Let $\overline{s}$ be the signal that is the composition of all these signals. The signal $\overline{s}$ represents the *body of knowledge* about the issue. We define an **expert** as someone who knows $\overline{s}$. Experts share the same knowledge about the issue but not necessarily about the ancillary matter.

As an example, experts on real business cycles have a thorough knowledge of the data on business cycles across time. However, these experts disagree about the economic theory that accounts for this data.

A stylized fact is that during a business cycle, wages move only a little while employment moves a lot. Although business cycle experts agree on this fact, they disagree on its import. To simplify a little, Neo-Keynesians take it as a sign that markets do not function smoothly – prices are sticky – while "freshwater" economists take it as evidence that markets function well, but the supply of labour is relatively flat. A future business cycle with similar movements can be expected to reinforce the opinions of (many of) those on both sides. The following result, which extends Theorem 3 to populations, formalizes this intuition

Theorem 4 *Suppose the body of knowledge about the issue and the common signal are both equivocal. Then, there is a v^* around which experts polarize completely. Formally, if $\overline{s}$ and c are equivocal, there is a v^* such that*

$$\begin{aligned} P(T \mid \overline{s}, \sigma) &> v^* \Rightarrow P(T \mid c, \overline{s}, \sigma) > P(T \mid \overline{s}, \sigma) \\ P(T \mid \overline{s}, \sigma) &< v^* \Rightarrow P(T \mid c, \overline{s}, \sigma) < P(T \mid \overline{s}, \sigma) \end{aligned}$$

and $P^{v^*} = P(\sigma : P(T \mid \overline{s}, \sigma) > v^*) > 0$, $P_{v^*} = P(\sigma : P(T \mid \overline{s}, \sigma) < v^*) > 0$.

Although this theorem is stated for experts, it applies to any population that enters the experiment having seen more or less the same equivocal evidence on an issue. The assumption

of expertise provides one reason that individuals would have seen similar evidence on the issue.

From Theorem 4, there is a level of belief v^* such that everyone with belief in the truth of the proposition greater than v^* revises upwards and everyone with belief lower revises downwards. Of course, an experiment will be "noisy" so that we would not expect to find such a perfect separation in practice. Moreover, the level v^* need not correspond to the 'dividing line' in beliefs around which an experimenter checks for polarization. Nonetheless, from Proposition 1, the population polarizes around every v, so that polarization will be found regardless of the dividing line that is chosen.

As an example, suppose that the population polarizes completely around $v^* = 0.4$, but the experimenter, who is unaware of the value of v^* chooses a belief of 0.5 as the dividing line for polarization.[6] She will find that the population polarizes, as everyone who believes the proposition to be true with probability greater than 0.5 revises upwards while less than everyone with belief less than 0.5 revises upwards. Furthermore, focusing on people with the strongest beliefs, everyone who believes the proposition to be true with probability at least, say, 0.7 revises upwards while everyone who believes it to be false with probability at least 0.7 revises downwards. In general, experts with strong opinions will tend to exhibit a high degree of polarization. These results are in line with Plous' finding that subjects with high issue involvement and with strong convictions display a large degree of polarization, if we accept that "high issue involvement" suggests a good knowledge of the current body of evidence.

Theorem 4 concerns a population of subjects with a similar level of expertise. In most experiments, there will be subjects with varying degrees of expertise. While some subjects will be well acquainted with the literature, others will have only a brief knowledge of it. If the issue at hand is controversial, as is the case in most experiments, then even subjects with only a little knowledge will likely have seen equivocal evidence (and know that overall the evidence is equivocal enough for experts to disagree). The following theorem is for people who have all previously seen equivocal signals, though these signals may vary.

Theorem 5 *Suppose that each person's private signal about the issue is equivocal and that*

[6] In a typical experiment, subjects are not asked directly for a probability assessment but rather for a number that is, presumably, related to this assessment (see Section 1.1 for more on this) or, more informally, an adjective describing their beliefs. Consider an experiment in which subjects are asked to indicate the extent to which they believe a proposition by choosing an integer from -5 to 5. Although one might be tempted to associate the point 0 with a belief of 0.5, this is far from clear. For instance, consider the propostion that extraterrestials disguised as humans roam the earth. A person who thinks there is a 20% chance this is true could reasonably be described as someone with quite a strong agreement, say a 3 or 4. Arguably, the point 0 corresponds better to the average belief in the population or perhaps the prior, than to a belief of 0.5.

the common signal is equivocal. Then, groups with the strongest opinions polarize completely. Formally, there exist $\overline{v}$ *and* $\underline{v} > 1 - \overline{v}$ *such that for all* s *and* σ

$$\begin{aligned} P(T \mid s, \sigma) &> \overline{v} \Rightarrow P(T \mid c, s, \sigma) > P(T \mid s, \sigma) \\ P(T \mid s, \sigma) &< 1 - \underline{v} \Rightarrow P(T \mid c, s, \sigma) < P(T \mid s, \sigma) \end{aligned} \tag{3}$$

and $P^{\overline{v}} = P(s, \sigma : P(T \mid s, \sigma) > \overline{v})$, $P_{1-\underline{v}} = P(s, \sigma : P(T \mid s, \sigma) < 1 - \underline{v}) > 0$.

Thus if everyone's private signal is equivocal, then groups with the strongest opinions polarize. On their capital punishment experiment dealing with reported attitude change, Miller, McHoskey, Bane, and Dowd (1993) find the most polarization among subjects with the strongest beliefs. For their part, Lord, Ross, and Lepper (1979) find polarization in a group of subjects who have been pre-selected for their strong beliefs. On the other hand, Kuhn and Lao (1996) do not find an effect of strength of opinion.

It is easy to see that, in addition to groups with the strongest opinions polarizing, there are belief levels around which the population polarizes. In particular, the (entire) population polarizes around $\bar{v}$, as well as around $1 - \underline{v}$. However, in contrast to the results of Theorem 4, the population does not necessarily polarize around every v. It is possible to construct examples where the population does not polarize around every v if the various pieces of information on the issue are sufficiently dissimilar and the ancillary matter is sufficiently unimportant (see Section 4.4, for such an example). On the other hand, when all the signals have symmetric likelihood matrices – so that results are not being pushed in any particular direction – the population polarizes around every v.

Theorem 6 *Suppose that each person's private signal about the issue and the common signal are equivocal and have symmetric likelihood matrices. Then the population polarizes completely around the prior belief* $P(T) = a$. *Formally,*

$$\begin{aligned} P(T \mid s, \sigma) &> a \Rightarrow P(T \mid c, s, \sigma) > P(T \mid s, \sigma) \\ P(T \mid s, \sigma) &< a \Rightarrow P(T \mid c, s, \sigma) < P(T \mid s, \sigma) \end{aligned}$$

and $P^a = P(s, \sigma : P(T \mid s, \sigma) > a)$, $P_a = P(s, \sigma : P(T \mid s, \sigma) < a) > 0$.

From Proposition 1, Theorem 6 also yields that the population polarizes around every v.

1.1 What Do The Answers Mean?

Subjects in attitude polarization experiments are typically not asked for complete descriptions of their beliefs but, rather, for single numbers that somehow summarize their beliefs or how their beliefs change. For instance, in the case of Lord, Ross, and Lepper, subjects

indicate how much their views change by choosing a number on a scale from -8 to 8. What exactly does a subject's answer mean? Somehow, this question is rarely asked.

Our model restricts the main issue to taking one of two values, true or false. This allows us to skirt the issue of exactly how to interpret responses, as every change in a probability distribution over two values is a first order stochastic dominance (fosd) shift. A person whose beliefs shift up in an fosd sense should revise with a higher number under any reasonable interpretation of what her answer means, provided that her beliefs change sufficiently for her to indicate a change (in many experiments, a sizable fraction of subjects indicate no change). Conversely, a person who revises her response upward must have had an fosd shift up in her beliefs, since the alternative is an fosd shift down.

When an issue can assume several values, a change in beliefs that causes, say, the mean of beliefs to rise may cause the median to fall, making it difficult to evaluate single number responses. Any theoretical results that only demonstrate polarization of, say, mean beliefs will have restricted applicability. On the other hand, results that yield polarization in the sense that one group's beliefs have an fosd shift upward while another group's have an fosd shift downwards will be applicable to a wide range of experiments – when there is an fosd shift of beliefs in a certain direction, almost any reasonable point summary of these beliefs will move in the same direction. Our model can be modified to allow for a many-valued issue values and our results recast in terms of fosd shifts, at the cost of added complexity.[7]

There is an issue which can take values in $X = \{x_1, ..., x_n\} \subset \mathbf{R}$, with x_i increasing in i. We also generalize to allow the ancillary matter to take values in $A = \{a_1, ..., a_m\} \subset \mathbf{R}$, with a_i increasing in i. Nature chooses a state in $\Omega = X \times A$. The prior is independent, so that there are probability distribution functions g over X and h over A such that the probability of state (x, a) is $\pi(x, a) = g(x) h(a)$; we assume that both distributions have full support.

Individuals receive a signal σ which indicates which ancillary state has been chosen. That is, individuals observe a value of $\sigma \in (0, 1)$ that is drawn from the density f_a if the state is (x, a) for some x. Without loss of generality, we assume that states are ordered so that higher σ indicates a higher state. Formally, we assume the monotone likelihood ratio property: $f_{a'}(\sigma) / f_a(\sigma)$ is strictly increasing in σ for $a' > a$. As before, we also assume that extreme values of σ are "completely" informative about the state: $\lim_{\sigma \to 1} f_{a_m}(\sigma) / f_a(\sigma) = \infty$ for $a < a_m$ and $\lim_{\sigma \to 0} f_a(\sigma) / f_{a_1}(\sigma) = 0$ for $a > a_1$.

Each individual also observes a signal $s \in S$, whose probability distribution depends on both x and a. The probability of signal s being drawn in state $\omega \in \Omega$ is $p_\omega(s)$. The key property we will assume about a signal s is that it is **equivocal**, in the sense that s is "bad news" (indicates a low x) if a is low enough, while it is "good news" (indicates a high x) if a is high; we also assume that this transition is monotone as we increase a. Formally, we say

[7] The results of Baliga, Hanany, and Klibanoff (2013), which are discussed in Section 2, are for fosd shifts with ambiguity averse agents.

that s is equivocal if $p_{xa_m}(s)$ is strictly increasing in x, $p_{xa_1}(s)$ is strictly decreasing in x and $p_{x'a}(s)/p_{xa}(s)$ is strictly increasing in a for $x' > x$. As before, all individuals also observe a common signal $c \in C$.

As an example of a result in this more general setting, we generalize Theorem 5, which states that groups with strong opinions polarize. The statement simplifies somewhat, since the fosd ordering (which we denote by $\succeq$) is incomplete and the same distribution r^* can serve both as a "high" lower bar for people who believe x is large ($\overline{v}$ in Theorem 5), and as a "low" upper bar for people who believe x is small ($\underline{v}$ in Theorem 5).

Theorem 7 *Suppose that each person's private signal s about the issue is equivocal and that the common signal c is equivocal. Then, groups with the strongest opinions polarize completely. Formally, there is a distribution r^* over Ω such that the conditional marginal distributions $P(\cdot \mid s, \sigma)$ and $P(\cdot \mid c, s, \sigma)$ over X satisfy*

$$\begin{aligned} P(\cdot \mid s, \sigma) &\succ r^* \Rightarrow P(\cdot \mid c, s, \sigma) \succ P(\cdot \mid s, \sigma) \\ P(\cdot \mid s, \sigma) &\prec r^* \Rightarrow P(\cdot \mid c, s, \sigma) \prec P(\cdot \mid s, \sigma) \end{aligned}$$

and $P^{r^*} = P(s, \sigma : P(\cdot \mid s, \sigma) \succ r^*) > 0$, $P_{r^*} = P(s, \sigma : P(\cdot \mid s, \sigma) \prec r^*) > 0$.

Theorem 4 can also be adapted to this setting, though there are several ways to proceed and a discussion of the options would add little to our understanding of the phenomenon.

1.2 No Polarization

Miller, McHoskey, Bane, and Dowd (1993) carry out several experiments. In one capital punishment study the population of subjects polarizes, while in an affirmative action study the population does not polarize. More precisely, in the latter study, subjects whose attitudes polarize are counter-balanced by subjects whose attitudes depolarize. In both studies, the common information that subjects are given consists of two opposing essays.

What accounts for the different findings on the two studies? We quote from their paper, "Why did relatively more subjects in [the affirmative action] study report a depolarization of their attitudes? We have no convincing answer. Subjects may have been less familiar with detailed arguments about affirmative action relative to the capital punishment issue used in Experiments 1 and 2. A larger number of subjects were perhaps more informed by the essays in this study, and, as a result indicated a reversal of their position."

Miller et al. do not explain exactly why subjects would tend to polarize when presented with familiar arguments but instead be "informed" and revise upwards or downwards in a pattern inconsistent with biased assimilation when presented with novel arguments. Nevertheless, that is what is predicted by our model (under an appropriate interpretation of the quoted passage).

To see this, recall our argument that in a population of people that have (largely) derived their beliefs on nuclear deterrence from their knowledge of near-miss episodes, proponents of nuclear deterrence will tend to be people who believe that safeguards are critical and conversely for opponents. As a result, when the population is presented with further evidence of reliable backups, proponents will be more likely to revise upwards than opponents and the population will polarize. Now suppose that instead of being given evidence on primary systems and backups, this population is presented with the following information:

> i) Numerous experiments have found that people are very good at evaluating risks and rewards and will not take undue chances. A strategy of nuclear deterrence makes the United States safer because other countries will avoid actions that could provoke a nuclear reply.
>
> ii) Neurological research has shown that people react with the emotional part of their brain when confronted with extreme threats, making their actions unpredictable. Because of this, a strategy of nuclear deterrence is risky.

The combined impact of these two statements on an individual will depend on how much weight he or she places on experimental evidence as compared to neurological evidence. There is little reason for these weights to bear any particular relation to how important the individual believes primary units are relative to backups. Thus, while different individuals may respond differently to these two statements, there is little reason for these responses to correlate with their initial beliefs about nuclear deterrence and little reason to expect polarization at the population level. Information that is equivocal, but equivocal with respect to a dimension that is orthogonal to previous information, can cause some pairwise polarization but will not cause the population to polarize.

In order to formalize this reasoning, we need to introduce a second ancillary matter. Hence, in addition to an ancillary matter with states that take the values H or L, we introduce a second matter with states that take the values h or l. Nature chooses one of the states H or L with probabilities b and $1-b$ and, independently, one of the states h or l with probabilities d and $1-d$. Individuals enter the experiment having seen a signal about the issue and a signal $\sigma = (\sigma_1, \sigma_2)$, where σ_1 varies with states H, L and σ_2 varies with states h, l, and draws of σ_1 and σ_2 are independent. With respect to nuclear deterrence, H and L could correspond to whether backup units or primary units are more important, while h and l could correspond to whether experimental or neurological evidence is more compelling.

Definition 6 *Let s and c be two signals about the issue. These signals are* ***unrelated*** *to each other if their likelihoods depend upon different ancillary matters.*

If s and c are unrelated we can write their likelihood matrices as

$$\begin{array}{cccc} & T & F \\ Hh & p_s & q_s \\ Lh & p_s & q_s \\ Hl & r_s & t_s \\ Ll & r_s & t_s \end{array} \quad \text{and} \quad \begin{array}{cccc} & T & F \\ Hh & p_c & q_c \\ Lh & r_c & t_c \\ Hl & p_c & q_c \\ Ll & r_c & t_c \end{array}$$

The following theorem implies that a population will not polarize when people are presented with information that is unrelated to the previous information on which they based their opinions. Specifically, if the common signal is unrelated to previous information, then people with large beliefs in the proposition are just as likely to revise upwards as people with small beliefs.

Theorem 8 *If signal c is unrelated to signal s, then, for any $\omega \in \Omega$,*

$$\begin{aligned} & P_\omega \left\{ \sigma : P\left(T \mid s, c, \sigma\right) > P\left(T \mid s, \sigma\right) \mid P_\omega \left(T \mid s, \sigma\right) > v \right\} \\ = \; & P_\omega \left\{ \sigma : P\left(T \mid s, c, \sigma\right) > P\left(T \mid s, \sigma\right) \mid P_\omega \left(T \mid s, \sigma\right) < v \right\}. \end{aligned} \tag{4}$$

whenever, $P^v = P_\omega \left(\sigma : P_\omega \left(T \mid s, \sigma\right) > v\right), P_v = P_\omega \left(\sigma : P_\omega \left(T \mid s, \sigma\right) < v\right) > 0.$

Theorem 8 is consistent with Miller et al (1993) analysis of their non-polarization finding, if we interpret their explanation that subjects were "less familiar" with the arguments to mean "the arguments were (largely) unrelated to how subjects had formed their initial views"

- While our basic framework as described in Section 1 has only one ancillary matter, other ancillary matters can easily be introduced. All our previous results carry through with the understanding that the common signal and the previous signals depend on the same ancillary matter.

1.3 Pairwise polarization

This paper is primarily concerned with the conditions under which populations polarize. Of course, a pre-condition for a population to polarize is that it is possible for two individuals polarize. The next theorem gives the conditions under which pairwise polarization can take place

First, we define a signal as unbalanced if the likelihood of the signal is always greater in one ancillary state than the other.

Definition 7 *The signal c is **unbalanced** if* $\min\left\{p_c, q_c\right\} > \max\left\{r_c, t_c\right\}$ *or* $\min\left\{r_c, t_c\right\} > \max\left\{p_c, q_c\right\}$.

Theorem 9 *A common signal c can cause pairwise polarization if and only if c is either equivocal or unbalanced . Formally, there exist initial beliefs $P(T \mid s_i, \sigma_i)$ and $P(T \mid s_j, \sigma_j)$ such that $P(T \mid s_i, \sigma_i) \geq P(T \mid s_j, \sigma_j)$, $P(T \mid s_i, \sigma_i, c) > P(T \mid s_i, \sigma_i)$ and $P(T \mid s_j, \sigma_j, c) < P(T \mid s_j, \sigma_j)$ if and only if c is either equivocal or unbalanced.*

While either an equivocal or an unbalanced signal can lead to pairwise polarization, unbalancedness does not naturally lead to population polarization (see the example in Section 4.2). Hence, the assumption that signals are unbalanced cannot be substituted for the assumption that they are equivocal in our previous theorems. Typical experiments on attitude polarization use common information that is equivocal.

2 Related literature

Walley (1991), Seidenfeld and Wasserman (1993), Andreoni and Mylovanov (2012), and Jern, Chang and Kemp (2014) argue that two individuals can polarize in a standard, rational setting, such as ours, if there is an ancillary matter (to put their result in our terms). Seidenfeld and Wasserman give sets of conditions for which individuals will polarize for all common signals c. Andreoni and Mylovanov provide a model where two individuals polarize after receiving one particular common signal c but they do not give a characterization of the properties that the likelihood of c must have in order for that to happen. Jern et al. provide examples of which Bayesian networks can generate polarization and which ones cannot. None of these papers address the question of when populations polarize.[8]

To grasp the distinction between pairwise polarization and population polarization at a technical level, suppose half the subjects enter with beliefs we will refer to as *type A beliefs,* and half with *type B beliefs*, as described by the following matrices:

$$\begin{array}{cc|cc} & \multicolumn{3}{c}{\text{Type A beliefs}} & & \multicolumn{3}{c}{\text{Type B beliefs}} \\ & T & F & & & T & F \\ H & \frac{1}{3}+a & \frac{1}{3}-a & \text{and} & H & \frac{1}{6}+b & \frac{1}{6}-b \\ L & \frac{1}{6} & \frac{1}{6} & & L & \frac{1}{3} & \frac{1}{3} \end{array}$$

The parameters a and b vary across individuals, with $-\frac{1}{6} \leq a, b \leq \frac{1}{6}$. Note that any subject for whom $a > 0$, and any subject for whom $b > 0$, has initial belief greater than $\frac{1}{2}$ in T and conversely.

[8] This is true both of Andreoni and Mylovanov's main model and their "More general environments" section. Andreoni and Mylovanov's principal concern is with the persistence of disagreement between individuals and when such disagreement can be common knowledge.

Suppose that subjects are presented with a common signal with likelihood matrix

	T	F
H	$\frac{1}{3}$	$\frac{1}{6}$
L	$\frac{1}{6}$	$\frac{1}{3}$

It is easily verified that all subjects with type A beliefs revise upwards while those with type B beliefs revise downwards. Thus, a type A subject with $a > 0$ and a type B subject with $b < 0$ polarize. This example, which captures the reasoning of the above papers, demonstrates that pairwise polarization is possible. The population will polarize if the distribution of a's is skewed towards $a > 0$, while the distribution of b's is skewed towards $b < 0$. However, as the example is presented, there is no particular reason to believe this to be the case and no indication of when it would be the case. If the distribution of a's is skewed towards $a < 0$, while the distribution of b's is skewed towards $b > 0$ the beliefs of the population will move closer together, rather than polarize. Under the neutral assumption that the distribution of a's and b' are the same, the movement in beliefs will be uncorrelated with initial beliefs. Thus, a demonstration of pairwise polarization says little about what happens as the population level.

Kondor (2012) shows that two individuals can polarize in a setting in which peoples' beliefs about the beliefs of others are important. Acemoglu, Chernozhukov, and Yildiz (2009) show that two individuals can persistently polarize if they disagree about the likelihoods of common signals. Glaeser and Sunstein (2013) show that two individuals with inconsistent beliefs can polarize.[9]

One of the clearest statements on polarization is found in Baliga, Hanany, and Klibanoff (2013), who are interested in the question of when two individuals can polarize. They let an issue take on many possible values and interpret a rise in a subject's response to indicate a first order stochastic dominance shift upwards in her beliefs and correspondingly for a fall in response. They first establish that, in a standard rational setting, if there is no ancillary matter (again, in our terms), then two individuals whose beliefs have common support cannot polarize. This result follows from Theorem 9, as assuming there is no ancillary matter is equivalent to setting $p_c = r_c$, $q_c = t_c$ and the theorem extends easily to issues that can take more than two values.[10] (Nevertheless, there is a sense in which polarization *in an fosd sense* can occur even without an ancillary state, as we show in Section 4.3 in the Appendix.)

Baliga et al. go on to argue that ambiguity aversion can explain polarization. Rabin and Schrag (1999) conclude that the literature on attitude polarization has shown that people

[9] All these papers largely interpret subjects' responses to reflect their mean beliefs. When issues can take on more than two values, so that changes in expected value are not isomorphic to fosd changes, individuals can polarize even when ancillary matters play no role, as the example in Section 4.3 shows (see also Baliga *et al.* (2013) and Dixit and Weibull (2007)).

[10] Of course, their result precedes our Theorem 9.

reason in a biased manner and develop a theory of confirmation bias. Fryer, Harms and Jackson (2013) show that two individuals can persistently polarize in a model in which agents are not fully rational. All three of these papers can be interpreted as showing population polarization as well as pairwise polarization, in non-standard settings. None of them make the distinction that we make between the types of information that should and should not produce polarization and, in fact, often predict polarization whenever there is disagreement.

Many experiments that find attitude polarization also find *biased assimilation* – subjects on either side of an issue both reporting that evidence that confirms their view is more credible than contrary evidence. As Lord, Ross and Lepper observe, this asymmetric assimilation in and of itself is not problematic, as it may be rational for a person to have greater confidence in a finding that confirms something she believes than a finding that disconfirms her belief. Gerber and Green (1999) show formally that biased assimilation can arise in a Bayesian model with normal signals, though their model does not allow for unbiased individuals to polarize. In a similar setting, Bullock (2009) shows that two unbiased individuals can polarize if they are estimating a parameter whose value is changing over time.

2.1 Further considerations on the literature

There is a considerable literature on attitude polarization and related phenomena. Unfortunately, it is easy for a casual reader to come away with an exaggerated impression of polarization findings. In a telling survey, Gerber and Green (1999) review the literature and conclude that the evidence for attitude polarization is mixed at best. One issue is that attitude polarization is more consistently found in experiments in which polarization is measured by asking subjects to choose a number indicating how their beliefs have changed than in experiments in which it is measured by having subjects choose a number indicating their initial beliefs and a number indicating their updated beliefs. Miller, McHoskey, Bane, and Dowd (1993), Munro and Ditto (1993) and Kuhn and Lao (1996), all find attitude polarization with the former type of question but not with the latter. It is not altogether clear what to make of this discrepancy.

Another difficulty in assessing the literature, is that a proper evaluation of experimental results often requires a close reading of the papers. In this section, we briefly consider three influential papers.

Darley and Gross (1983) provide subjects with descriptions of a fourth-grade girl. Half the subjects are given information strongly suggesting that the girl comes from an upper class background and half are given information suggesting that she comes from a lower class background – information that could potentially have a biasing effect on the way subjects process subsequent information. At that point the subjects are asked for their opinions of the girl's abilities on three academic subjects – liberal arts, reading, and mathematics –

and of her disposition on five traits – work habits, motivation, sociability, maturity, and cognitive skills. Subjects who believe that the girl comes from a well-off family tend to rate her slightly higher than those who believe she comes from a poorer family. Next, subjects are provided with some specific evidence about her abilities. This evidence is the same for all the subjects, who are then again asked to rate her.[11] The subjects beliefs polarize on four out of the eight questions, including the three academic subjects.

Although this experiment is typically touted as one that demonstrates polarization, this is hardly an overwhelming finding of polarization. Somewhat bizarrely, almost all the papers that cite Darley and Gross do not even mention the questions on which subjects do not polarize.[12] In fairness to Darley and Gross, they put their data through various tests to reach their conclusions of bias and it is beyond the scope of this paper to consider the merits of all their arguments. Nonetheless, at the very least, their conclusion that they have found evidence of polarization is open to doubt. We consider the paper in greater detail in Section 4.6.

Kunda (1987) gives subjects a scientific article claiming that women who are heavy drinkers of coffee are at high risk of developing fibrocystic disease, and asks them to indicate how convincing the article is. In one treatment, fibrocystic disease is characterized as a serious health risk and women who are heavy coffee drinkers rate the article as less convincing than women who are light drinkers of coffee (and than men). In a second treatment, the disease is described as common and innocuous and both groups of women rate the article as equally convincing. Note that in the first treatment, the article's claim is threatening to women who are heavy coffee drinkers, and only them, while in the second treatment the article's claim threatens neither group. Kunda's interpretation of her findings is that subjects engage in *motivated reasoning* and discount the article when it clashes with what they wanted to believe. However, when subjects are asked how likely they are to develop the disease in the next fifteen years, in both treatments women who are heavy coffee drinkers indicate about a 30% greater chance than light drinkers. That is, although heavy coffee drinkers in the serious health risk treatment describe the article as less convincing than in the innocuous risk treatment, they seem to be equally convinced in the two treatments. Kunda does not comment on this discrepancy (a chart is given without comment), but to us it makes the case for motivated reasoning here less than clear.

Nyhan and Reifler (2010) report on an extreme form of polarization, a so-called backfire

[11] Actually, in the experiment one group of subjects was given only demographic information, while another group was given both demographic information and additional common information. The two groups were presumed to be more or less identical a priori, and the results are universally interpreted to represent changes in responses following the additional information, while avoiding anchoring effects.

[12] Darley and Gross themselves explain away the negative findings. While one can debate the merits of their explanation, there is something a bit awkward when positive findings are taken as support of a hypothesis while negative ones are explained away – in a paper on hypothesis-confirming bias, no less.

effect. As they describe it, they give subjects articles to read that contain either a misleading statement by a politician or the misleading statement together with an independent correction and, rather than offsetting the misleading statement, the correction *backfires*, causing partisans to believe the statement even more.

In their first experiment, all subjects are given an article to read in which Bush justifies the United States invasion of Iraq in a manner that suggests that Iraq has weapons of mass destruction. For subjects in the correction condition, the article goes on to describe the Duelfer Report, which documents the absence of these weapons. However, "the correction backfired— conservatives who received a correction telling them that Iraq did not have WMD were *more* likely to believe that Iraq had WMD than those in the control condition."

It is worth looking at the actual "correction" that subjects are given and the question they are asked.

> **Correction**: While Bush was making campaign stops in Pennsylvania, the Central Intelligence Agency released a report that concludes that Saddam Hussein did not possess stockpiles of illicit weapons at the time of the U.S. invasion in March 2003, nor was any program to produce them under way at the time. The report, authored by Charles Duelfer, who advises the director of central intelligence on Iraqi weapons, says Saddam made a decision sometime in the 1990s to destroy known stockpiles of chemical weapons. Duelfer also said that inspectors destroyed the nuclear program sometime after 1991.
>
> **Question**: Immediately before the U.S. invasion, Iraq had an active weapons of mass destruction program, the ability to produce these weapons, and large stockpiles of WMD, but Saddam Hussein was able to hide or destroy these weapons right before U.S. forces arrived — Strongly disagree [1], Somewhat disagree [2], Neither agree nor disagree [3], Somewhat agree [4], Strongly agree [5]

To us, the so-called correction is far from a straightforward repudiation. First of all, it acknowledges that, at some point in time, Hussein did posses weapons of mass destruction, in the form of chemical weapons. It rather vaguely asserts that he made a decision to destroy stockpiles of chemical weapons, without asserting that he followed up on the decision. It goes on to say that inspectors destroyed the nuclear program sometime after 1991. But how difficult would it have been for Hussein to have hidden some weapons from the inspectors? The question asks if Iraq had "the ability to produce these weapons". Even if stockpiles of chemicals were destroyed, would that eliminate a country's ability to produce more?

All these issues muddy the interpretation of their findings. Some readers may think we are quibbling, but why not provide a more straightforward correction and question such as:

> **Correction**: In 2004, the Central Intelligence Agency released a report that concludes that Saddam Hussein did not possess stockpiles of illicit weapons at

the time of the U.S. invasion in March 2003, nor was any program to produce them under way at the time.

Question: Immediately before the U.S. invasion, Iraq had an active weapons of mass destruction program and large stockpiles of WMD – Strongly disagree, Somewhat disagree, Neither agree nor disagree.

In fact, Nyhan and Reifler run a follow-up study in which this is precisely the correction and question that they use. And with this formulation they do not find a backfire effect. However, their reason for this alternate formulation is not to test their original finding and they do not conclude that the original backfire effect was spurious. Rather, they provide several explanations for the different finding. One explanation starts with the observation that the follow-up experiment took place a year later and in the intervening year the belief that Iraq had weapons of mass destruction had fallen among Republicans. Notice that this observation itself belies the notion that polarization is inevitable. Another explanation acknowledges that the different result may be related to the "minor wording changes." These do not strike us as minor changes, but our intent is not to enter in a debate here. The authors report the two different findings, as well as another, and they make a case for their interpretation. What is unfortunate is that others who refer to them typically quote the first experiment without even mentioning the follow-up.

We do not doubt that there is a real phenomenon here – indeed, that is why we have written this paper – but it is important to do a proper assessment of experimental results.

3 Conclusion

Our results show that unbiased Bayesian reasoning will often lead populations to polarize. To some extent, this should come as no surprise. After all, the differences in opinions between different schools of thought – be it Neo-Keynesians versus freshwater economists, communists versus fascists, republicans versus democrats, or Freudians versus Jungians – do not result from access to different information on the issues they discuss, but from differences in how they interpret the information. It is hardly surprising when members of the different schools continue to interpret evidence in different ways. Essentially, the schools of thought correspond to the ancillary matters that play a crucial role in our analysis.

Nonetheless, if reasoning is unbiased there are limitations to the polarization that should take place. In keeping with this prediction, some experiments do not find polarization. In the political sphere, an analysis of Gallup poll surveys across 36 years by Gerber and Green (1999) shows that the approval ratings of United States presidents by Democrats, Republicans and Independents move up and down together with a very high correlation in the way in which partisan groups update their assessments. Where there are persistent differences in political

beliefs, it is often not clear what these differences show about how people reason. After all, many political questions concern issues where fundamentals are changing over time, where evidence is hard to come by, where even partisans are often ill-informed, and where factual discussions are confounded with discussions about values – hardly an ideal setting for a convergence of beliefs (see Bullock (2009) for a further discussion).

Returning to the question we began with, what effect should we expect evidence of racial disparities in police stop and frisk rates to have on different groups' views of the American justice system? Surveys show that many white Americans see disparate treatment by the police as a rational response to differences in crime rates where many black Americans see a discriminatory police force. The evidence on stop and frisks is consistent with both viewpoints. Indeed, while scholars are quick to cite opinion polls showing disparities in beliefs between different racial groups in the United States, most of these disparities have few implications for Bayesian reasoning.[13] Different racial groups in the United States have markedly different experiences and the same evidence interpreted in light of different experiences may yield varying conclusions. This does not mean that there is no evidence that should lead members of different groups to react similarly. Gelman, Fagan, and Kiss (2007) find that, not only were blacks and Hispanics in New York city stopped by police more often than whites in the late 1990s, they were also stopped more often than whites relative to their respective crime rates and that stops of blacks and Hispanics were less likely to lead to arrests. While this data is not devoid of all ambiguity, it is more likely to lead to a harmonization of beliefs than simple data on overall stop rates.

We have shown not just that it is possible to concoct some Bayesian model in which groups polarize, but that Bayesian polarization can arise quite naturally. This does not mean that biased reasoning never occurs. However, a finding of attitude polarization is a long way from a demonstration of biased reasoning. Our results can be used to design experiments that test for bias.

Many scholars have asked what can be done to reduce persistent disagreements among various groups. Our model suggests that, rather than provide people with yet more direct evidence on the issue at hand, it would often be better to give them information on an ancillary matter that is only indirectly related to the issue, in order to first make their beliefs on the ancillary matter converge. Our reasoning is not far from Pascal's: "When we wish to correct with advantage and to show another that he errs, we must notice from what side he views the matter, for on that side it is usually true, and admit that truth to him, but reveal to him the side on which it is false." (Pensées, translated by W. F. Trotter.)

[13] There may be implications for whether or not different beliefs are common knowledge and whether or not rationality is common knowledge, but common knowledge assumptions are quite strong.

4 Appendix

4.1 A Simple Example

In this section, we provide a numerical example that illustrates population polarization, using the question of whether nuclear deterrence makes a country safer.

Suppose a nuclear deterrence system consists of two components, a primary unit and a backup, each of which can be either reliable, r, or (relatively) unreliable, u. Let (r, u) denote that the primary system is reliable and the backup unreliable, and so forth for the other three possibilities. The safety of the system depends not only on the reliability of its components, but also on which component is critical for systems of this sort. If primary units are critical, then a system is safe if and only if its primary unit is reliable (say if the primary unit fails too often, sooner or later the backup will fail to catch it, so the primary unit must be reliable). Call this, condition $\mathcal{P}$. If, on the other hand, backups are critical, then a system is safe provided its backup unit is reliable (perhaps initial mistakes are inevitable but it is easier to catch an error than prevent one, so a reliable backup is all that is needed). Call this, condition $\mathcal{B}$. People are uncertain which one of $\mathcal{P}$ and $\mathcal{B}$ holds. An individual's belief on the matter comes from his information about the determinants of safety for systems of this type.

Let $\mathcal{T}$ indicate that it is true that nuclear deterrence makes a country safer and $\mathcal{F}$ that it is false. It is convenient to describe the world as being in one of four possible states, as indicated by the following matrix:

$$\begin{array}{cc} \mathcal{T} & \mathcal{F} \\ \mathcal{B} \quad (r,r),(u,r) & (u,u),(r,u) \\ \mathcal{P} \quad (r,r),(r,u) & (u,r),(u,u) \end{array}$$

The matrix shows that the state can be $\mathcal{BT}$ in one of two possible ways: backups are critical and both components are reliable, or backups are critical and only backups are reliable. The states $\mathcal{BF}$, $\mathcal{PT}$, and $\mathcal{PF}$ are established in similar fashion. Suppose that, a priori, each component is reliable with a 50% chance, backups are critical with a 50% chance, and all these probabilities are independent. Then each state has a $\frac{1}{4}$ probability.

Independent signals emanate about the reliability of the two components. Specifically, if a component is reliable the signal $\hat{r}$ is issued with probability $\frac{2}{3}$ and the signal $\hat{u}$ with probability $\frac{1}{3}$; if a component is unreliable, the signal $\hat{u}$ is issued with probability $\frac{2}{3}$ and $\hat{r}$ with probability $\frac{1}{3}$. The pair $(\hat{r}, \hat{r})$ can be thought of as a positive signal about the safety of nuclear deterrence, the pair $(\hat{u}, \hat{u})$ as a negative signal, and the pairs $(\hat{u}, \hat{r})$ and $(\hat{r}, \hat{u})$ as equivocal signals, where the first element of each pair emanates from the primary unit and the second from the backup. For example, $(\hat{u}, \hat{r})$, an unreliable primary unit and a reliable

backup, is equivocal because it points to a safe system if backups are important, but an unsafe system if primary units are more relevant.

A near-miss incident corresponds to the signal $(\hat{u}, \hat{r})$. In the state $\mathcal{BT}$, the probability of receiving signal $(\widehat{u}, \widehat{r})$ is given by

$$P\left(\widehat{u}, \widehat{r} \mid \mathcal{BT}\right) = P\left(\widehat{u}, \widehat{r} \mid \mathcal{B}, u, r\right) P\left(\mathcal{B}, u, r \mid \mathcal{BT}\right) + P\left(\widehat{u}, \widehat{r} \mid \mathcal{B}, r, r\right) P\left(\mathcal{B}, r, r \mid \mathcal{BT}\right) = \frac{1}{3}.$$

Similar calculations for the other states show the likelihood matrix for the signal $(\hat{u}, \hat{r})$ to be

Likelihood of $(\widehat{u},\widehat{r})$

	$\mathcal{T}$	$\mathcal{F}$
$\mathcal{B}$	$\frac{1}{3}$	$\frac{1}{6}$
$\mathcal{P}$	$\frac{1}{6}$	$\frac{1}{3}$

In addition to the information about the reliability of the primary and secondary unit, each person also receives a signal about whether the state is $\mathcal{B}$ or $\mathcal{P}$. Let person i's information be a draw $\sigma_i \in (0, 1)$, where higher values are more likely if $\mathcal{B}$ holds, independently of other parameters. (For instance, if the state is $\mathcal{BT}$ or $\mathcal{BF}$ the individual samples σ from a density $\pi_{\mathcal{B}}(\sigma) = 2\sigma$, while in states $\mathcal{PT}$ or $\mathcal{PF}$ he samples from $\pi_{\mathcal{P}}(\sigma) = 2(1-\sigma)$.)

Consider a population of subjects who have derived their beliefs on nuclear deterrence from their knowledge of a single near-miss incident in the past, evaluated in light of their views about what is critical to the safety of systems. Those who believe that nuclear deterrence is probably safe will be those who believe that backups are likely to be critical; those who believe that nuclear is probably not safe will be those who believe that primary units are likely to be critical. That is,

$$\begin{aligned} P\left(\mathcal{T} \mid (\hat{u}, \hat{r}), \sigma\right) &> \frac{1}{2} \Leftrightarrow P\left(\mathcal{B} \mid (\hat{u}, \hat{r}), \sigma\right) > \frac{1}{2} \\ P\left(\mathcal{F} \mid (\hat{u}, \hat{r}), \sigma\right) &> \frac{1}{2} \Leftrightarrow P\left(\mathcal{P} \mid (\hat{u}, \hat{r}), \sigma\right) > \frac{1}{2}. \end{aligned}$$

Now suppose the subjects are all told about another near-miss incident; that is, they are given further evidence that the primary unit is relatively unreliable but the backup is reliable. This signal is positive for subjects who believe that backups are critical; these are also the subjects who have an initially positive view of nuclear deterrence. Similarly on the negative side. Hence, the population polarizes – those subjects who believe that nuclear is probably safe and those who believe it is probably not safe both become more convinced of their views. That is,

$$\begin{aligned} P\left(\mathcal{T} \mid (\hat{u}, \hat{r}), \sigma\right) &> \frac{1}{2} \Rightarrow P\left(\mathcal{T} \mid (\hat{u}, \hat{r}), (\hat{u}, \hat{r}), \sigma\right) > P\left(\mathcal{T} \mid (\hat{u}, \hat{r}), \sigma\right) \\ P\left(\mathcal{F} \mid (\hat{u}, \hat{r}), \sigma\right) &> \frac{1}{2} \Rightarrow P\left(\mathcal{F} \mid (\hat{u}, \hat{r}), (\hat{u}, \hat{r}), \sigma\right) > P\left(\mathcal{F} \mid (\hat{u}, \hat{r}), \sigma\right). \end{aligned}$$

4.2 Convergence with Polarization

In this section, we present an example that illustrates that polarization may take place even as there is growing agreement in a population, as noted in footnote 3. The example also shows that a signal that is unequivocal and unbalanced may cause pairwise polarization.

Consider the issue of capital punishment. Let i be a finding that the murder rate has increased in a jurisdiction with capital punishment and d a finding that the rate has decreased. Suppose that i and d have the following likelihood matrices

	T	F
H	$\frac{4}{5}$	$\frac{9}{10}$
L	$\frac{1}{10}$	$\frac{1}{2}$

i

	T	F
H	$\frac{1}{5}$	$\frac{1}{10}$
L	$\frac{9}{10}$	$\frac{1}{2}$

d (5)

where H corresponds to selection issues being important and L to these issues being irrelevant.[14] Suppose the prior over the four states is uniform.

Let $\mathcal{C} = \mathcal{S}$ be the set of unordered draws from two jurisdictions with capital punishment. Thus, $\mathcal{C}$ consists of three signals, c_{ii}, c_{dd}, and c_{id}, where, for instance, the signal c_{id}, indicates that the murder rate has increased in one jurisdiction and decreased in one. Their likelihoods are

	T	F
H	$\frac{16}{25}$	$\frac{81}{100}$
L	$\frac{1}{100}$	$\frac{1}{4}$

c_{ii}

	T	F
H	$\frac{1}{25}$	$\frac{1}{100}$
L	$\frac{81}{100}$	$\frac{1}{4}$

c_{dd}

	T	F
H	$\frac{8}{25}$	$\frac{18}{100}$
L	$\frac{18}{100}$	$\frac{1}{2}$

c_{id}

Note that c_{id} is an equivocal signal.

Say the existing body of knowledge is $\bar{s} = c_{id}$. Consider a population of experts, who have all seen this signal. The experts all agree upon the experience that jurisdictions have had with capital punishment to date but they disagree about the importance of selection issues.

Now suppose they are presented with information from two additional jurisdictions and that this signal is again c_{id}. The population polarizes completely around an initial belief of (about) 0.55 that the proposition is true. That is, everyone with an initial belief in the proposition greater than 0.55 revises upwards upon seeing an additional c_{id}, while everyone with an initial belief smaller than 0.55 revises downward.

Let us consider what happens as the population is given more and more common information. We can model this process as more and more conditionally independent draws from $\mathcal{C}$. Suppose the actual state of the world is LF, where the modal draw is c_{id}. First consider the

[14] In this example, when selection issues are important jurisdictions that adopt capital punishment have such sharply rising murder rates that, even if the punishment is an effective deterrent, there is still a large chance of $\frac{4}{5}$ that the murder rate increases. This feature is unimportant for our immediate purposes but allows the example to also be used to demonstate the effect of an unbalanced signal.

unlikely possibility that every draw happens to be this equivocal signal. Take a person with initial belief of 0.62 that capital punishment is effective (that is, $P(T \mid \bar{s}, \sigma_i) = 0.62$). As we know, after seeing one more instance of c_{id}, she revises upward. For the next six iterations, her belief continually increases, reaching 0.96. However, at the seventh additional draw, her belief decreases and continues to decrease from then on. The reason for the downturn is that, while c_{id} is equivocal, it is most likely to occur in the state LF. Eventually the effect of this fact dominates and she revises downwards.

Typically, additional draws will not consist of unbroken strings of one increase/one decrease, although, in the limit, the data will show that the murder rate has risen half the time (in the state LF). For i.i.d. draws, we have the following:

1. *Eventually (almost) everyone agrees that the proposition is false and the ancillary state is low.* Formally, let c^∞ be a sequence of iid draws from $\mathcal{C}$, and c^t the first t draws. For any σ, $P\{c^\infty : \lim_{t\to\infty} P(LF \mid c^t, \bar{s}, \sigma) = 1\} = 1$.

2. *Eventual harmonization.* Initially, two given experts may polarize. Eventually, however, they will harmonize. Formally, for any σ,σ', $c \in \mathcal{C}$,

$$\lim_{t\to\infty} P\left\{c^t : P\left(T \mid c, c^t, \bar{s}, \sigma\right) < P\left(T \mid c^t, \bar{s}, \sigma\right) \text{ and } P\left(T \mid c, c^t, \bar{s}, \sigma'\right) < P\left(T \mid c^t, \bar{s}, \sigma'\right)\right\} = 1.$$

3. *While more and more people revise downwards upon seeing an equivocal signal, there are always extremists who revise upwards.* Formally, for all t and c^t, there exist v_t and h_t such that $P(T \mid \bar{s}, \sigma) > v_t \Rightarrow P(T \mid c_{id}, c^t, \bar{s}, \sigma) > P(T \mid c^t, \bar{s}, \sigma)$ and $P(H \mid \bar{s}, \sigma) > h_t \Rightarrow P(T \mid c_{id}, c^t, \bar{s}, \sigma) > P(T \mid c^t, \bar{s}, \sigma)$.

Although the population always polarizes upon seeing an equivocal signal, as evidence accumulates more and more people become convinced that the proposition is false and more and more people harmonize.

The signal c_{ii} is unequivocal. In both ancillary states H and L, the signal causes a downward revision that the proposition is true – that is, for all s, $P(T \mid H, c_{ii}, s) < P(T \mid H, s)$ and $P(T \mid L, c_{ii}, s) < P(T \mid L, s)$. However, the signal c_{ii} is also unbalanced, being always more likely in ancillary state H than L, and it can lead an individual who is uncertain of the ancillary state to revise upwards. For instance, an expert who initially believes the ancillary state is high with probability .52 revises upwards. The reason he revises upwards is that c_{ii} increases his belief that the state is high, and when the is high, his initial belief in the proposition is relatively large. This expert has an initial belief of .46 that the proposition is true. At the same time, an expert with initial belief of .38 that the population is true revises downwards, so that the unequivocal c_{ii} causes these two individuals to polarize. However, everyone with initial belief greater than .53 also revises downwards and the population does not polarize upon seeing c_{ii}.

4.3 Polarization without an ancillary state

The following example shows that even without an ancillary state, an experiment could find that beliefs polarize in an fosd sense depending on the exact question that is asked.

Consider the issue of how safe nuclear energy is. Suppose its safety can be described as a parameter that takes on the values $1, 2, 3$, or 4 (say, 1 means there is more than a 3% chance of an accident, 2 means a $1-3\%$ chance, etc...), and that, a priori, all four values are equally likely. Individuals receive private information that is one of four signals with likelihoods:

$S_A \downarrow \Theta \rightarrow$	**1**	**2**	**3**	**4**
s_1	$\frac{3}{4}$	$\frac{1}{4}$	0	0
s_2	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
s_3	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$
s_4	0	0	$\frac{1}{4}$	$\frac{3}{4}$

Likelihoods

Suppose that person I sees signal s_2 and II sees signal s_3. Their updated beliefs are

	1	**2**	**3**	**4**
$I : p(\cdot \mid s_2)$	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
$II : p(\cdot \mid s_3)$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$

Posteriors (6)

so that II's beliefs fosd I's. Now I and II are shown the common signal c with likelihoods

	1	**2**	**3**	**4**
c	0	1	1	0

Likelihoods

In this setting, Baliga et al. have shown that fosd polarization of two individuals cannot occur. This no-polarization also follows from Theorem 9, extended to issues that take on more than one value. Indeed, posterior beliefs are

	1	**2**	**3**	**4**
$I : p(\cdot \mid s_2,c)$	0	$\frac{2}{3}$	$\frac{1}{3}$	0
$II : p(\cdot \mid s_3,c)$	0	$\frac{1}{3}$	$\frac{2}{3}$	0

Posteriors (7)

and there is no polarization in an fosd sense. In fact, for both I and II beliefs have neither risen nor fallen in an fosd sense.

Suppose, however, that the experimenter does not ask subjects for their beliefs over the four point scale. Instead, the experimenter asks for their beliefs that nuclear energy is "safe".

Say that both subjects agree that nuclear energy is safe if it rates a 3 or 4. The beliefs of the subjects before and after the common signal are

	Posteriors after signals	
	$Dangerous$	$Safe$
$I : s_2$	$\frac{5}{8}$	$\frac{3}{8}$
$II : s_3$	$\frac{3}{8}$	$\frac{5}{8}$
$I : s_2, c$	$\frac{2}{3}$	$\frac{1}{3}$
$II : s_3, c$	$\frac{1}{3}$	$\frac{2}{3}$

Before the common signal, II's beliefs fosd I's. Following c, II's beliefs shift up and I's shift down, so there is polarization in an fosd sense. This example is in the spirit of BHK's assumptions which guarantee no polarization. As they write, the key to their result is that "conditional on the parameter, all individuals agree on the distribution over signals and their independence". Here too, conditional on the underlying parameters, all individuals have this agreement. However, while the experimenter has asked a natural enough question, it is (perhaps inevitably) only a function of the underlying parameters and that function does not have the same properties.

Note also that the initial question (where there is no polarization in an fosd sense) shows that polarization in an expected value sense does not require an ancillary state (or a "miscalibrated" question). From equation (6), $E(\theta \mid s_2) = 2.37$ and $E(\theta \mid s_3) = 2.62$, while from equation (7) $E(\theta \mid s_2, M) = 2.33$ and $E(\theta \mid s_3, M) = 2.67$.

4.4 Polarization, but not everywhere

The following example shows that the population may not polarize around every v even if all signals are equivocal.

Suppose the prior is uniform ($a = b = \frac{1}{2}$) and that the ancillary signal is heavily concentrated around $\sigma's$ such that $\frac{\pi_H(\sigma)}{\pi_L(\sigma)} \in [0.9, 1.1]$. Then the bulk of the ancillary signals are not very informative about the ancillary state. Let $\mathcal{S} = \{s_1, s_2, s_3\}$, where, for $\varepsilon \approx 0$, the likelihood of each signal in each state is

$$\begin{array}{cc} s_1 & \\ \frac{3}{7}+\varepsilon & \frac{3}{7}-\varepsilon \\ \frac{2}{7}+\varepsilon & \frac{4}{7}-\varepsilon \end{array}, \quad \begin{array}{cc} s_2 & \\ \frac{4}{7}-\varepsilon & \frac{2}{7}+\varepsilon \\ \frac{3}{7}-\varepsilon & \frac{3}{7}+\varepsilon \end{array} \quad \text{and} \quad \begin{array}{cc} s_3 & \\ 0 & \frac{2}{7} \\ \frac{2}{7} & 0 \end{array}$$

and let c have likelihood matrix

$$\begin{array}{cc} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{array}$$

Suppose that, as it happens, the actual state of the world is (H, T) and consider a large group of subjects that have all seen one signal about the issue. Then, $\frac{3}{7}$ of the subjects

have seen s_1 and $\frac{4}{7}$ have seen s_2. Consistent with Theorem 5, everyone who believes the proposition is true with probability at least .59 revises upwards and everyone who believes it is false with probability at least .59 revises downwards.
However, for $\sigma's$ such that $\frac{\pi_H(\sigma)}{\pi_L(\sigma)} \in [0.9, 1.1]$, which form the bulk of σ's, $P(T \mid s_1, \sigma) < \frac{1}{2} < P(T \mid s_2, \sigma)$. We also have $P(T \mid c, s_1, \sigma) > P(T \mid s_1, \sigma)$ if and only if $\frac{\pi_H(\sigma)}{\pi_L(\sigma)} > 0.94$, while $P(T \mid c, s_2, \sigma) > P(T \mid s_2, \sigma)$ if and only if $\frac{\pi_H(\sigma)}{\pi_L(\sigma)} > 1.0$. Hence, for $v \approx \frac{1}{2}$, the proportion of people with belief greater than v that revises upwards is *smaller* than the proportion with belief less than v that revises upwards.

There are three particular features of this counter-example:

1. Although there is an ancillary state, its importance is minimal as almost all the subjects have very similar beliefs about it.

2. Although the private signals the subjects have seen are equivocal, they are not very equivocal. For instance, the signal s_1 is essentially negative for the proposition – it is more or less neutral in state H, and it is bad news in state L. By the same token, signal s_2 is essentially positive.

3. Although the private signals are equivocal, they are also quite different from the common signal. For instance, in contrast to s_1 and s_2, the signal c is neither good news nor bad news for the proposition.

While these three points are each important separately, Theorem 6 addresses 2) and 3) together, by considering only *symmetric* signals.

4.5 Lord, Ross, and Lepper revisited

Lord, Ross and Lepper (1979) find that views on capital punishment move further apart after subjects view a common piece of evidence. The specific evidence that Lord, Ross, and Lepper provide to their subjects is two (purported) studies, one that finds that murder rates tend to be lower in states following the adoption of the death penalty and one that finds that murder rates tend to be higher. Viewed as a single entity, the studies determine that about half the time a state that adopts the death penalty subsequently has a lower murder rate and half the time a higher murder rate.

Why would some (unbiased) people consider this type of data to be evidence in favour of the death penalty and others evidence against? It is not crucial that we, as analysts, know the reason why but let us propose two.[15]

[15]Rabin and Schrag (1999), commenting on this experiment in their footnote 8, write that polarization may happen if people are "predisposed" to making different interpretations of ambiguous evidence, in a

1. Some people believe that states that adopt the death penalty are states with rising murder rates. For these people, the fact that murder rates drop in half the states is evidence that the death penalty has a deterrent effect. Indeed, even evidence that the murder rate increased in all states would not be strong evidence against the death penalty. Other people believe that states adopt the death penalty according to the politics of the state, politics that are unrelated to current murder rates. For such people, the studies provide evidence that the death penalty is not effective, as murder rates seem to rise or fall independently of its adoption.

2. Footnote 2 in LRL reads "Subjects were asked... whether they thought the researchers had favored or opposed the death penalty... Analyses...showed only that subjects believed the researcher's attitudes to coincide with their stated results..." That is, subjects believed that researchers who found evidence of a deterrent effect also favoured the death penalty and correspondingly for researchers who did not find a deterrent effect. What are we to make of this? Is it that subjects believed that the researchers became convinced by their own research? That is a possibility, although opposition to the death penalty depends not just on its deterrent effect. Moreover, the statement that the researchers *had* favoured the death penalty suggests that their attitudes preceded their findings. But then how can it be that researchers' beliefs always coincide with their findings? They could be faking their findings, or consciously or subconsciously making research decisions that influence their findings, or perhaps only publishing research that coincides with their views of the death penalty. With an ancillary matter of whether researchers who are to the left politically or to the right are more honest and forthcoming, a 50/50 finding easily leads to polarization in our model. This ancillary matter is in keeping with the persuasion literature, which notes the importance of source credibility in shaping beliefs.[16]

4.6 Hannah revisited

Recall Darley and Gross (1983)'s experiment discussed in Section 2.1. Half the subjects were given information indicating that a girl named Hannah came from an upper class background and half information indicating that she came from a lower class background. At this point, they were asked to evaluate Hannah in eight domains. The subjects were then shown a

way that "departs from common-priors Bayesian information processing". (They do not provide an explicit model). In contrast, our model uses common priors Bayesaian processing.

[16] Lord et al. also asked their subjects to evaluate the studies presented. Subjects tended to give (implausible) methodological critiques of the studies that went against their inital views. However, as the authors note, the fact that subjects answered with methodlogical critiques is probably not very significant, as the design of the experiment primed them to.

video of her engaged in various tasks, and were again asked to evaluate her. The responses of the two groups of subjects polarized in four out of the eight domains. Although we do not consider this to be a strong finding of polarization, some might argue that it is still a finding of polarization. Either way, the experiment does not provide a test of our theory.

To see this, note that the different groups of subjects are effectively asked about two different girls, a rich one and a poor one, and the same behaviour could well have different implications for children from different demographics. For instance, subjects could believe that a child that attends a rich school will perform well on national tests provided that she is able to concentrate moderately well while a child that attends a poor school will perform well only if she has exceptional concentration skills. Then, evidence that Hannah concentrates moderately well would be good news for *rich Hannah* but bad news for *poor Hannah.* Thus, a finding of attitude polarization would be consistent with our model, with the child's background being the ancillary matter. On the other hand, a strong finding of polarization is not particularly predicted by our model, as people were not pre-sorted according to their beliefs. Hence, the results of this experiment say little about our theory.

In fact, even without the benefit of our model, and even if we are to consider only the domains where polarization is found, we are not persuaded the experiment would demonstrate biased reasoning. The strongest findings of attitude polarization are on the three academic subjects Darley and Gross ask about. Let us be a bit more precise about these findings. When given only demographic information about Hannah, subjects initially rated rich Hannah as slightly better than poor Hannah on the three subjects, though in two out of three cases the difference was not statistically significant. A fair summary is that, overall, the two Hannah's were initially rated more or less equally. To quote from the paper, initial "estimations of the child's ability level tended to cluster closely around the one concrete fact they had at their disposal: the child's grade in school." , though in two out of three cases the difference was not statistically significant. A fair summary is that, overall, the two Hannah's were initially rated more or less equally. To quote from the paper, initial "estimations of the child's ability level tended to cluster closely around the one concrete fact they had at their disposal: the child's grade in school."

As Darley and Gross realize, it is a bit odd that the two Hannah's were rated almost equally, given the advantages that wealthy schools confer upon their students (and which we might well expect Princeton University subjects to be aware of) and given that many studies have shown positive correlations between social class and school performance. Darley and Gross provide a possible explanation for this: "Base-rate information... represents probabilistic statements about a class of individuals, which may not be applicable to every member of the class. Thus, regardless of what an individual perceives the actual base rates to be, rating any one member of the class requires a higher standard of evidence."

Let us put some numbers to this notion of base rates and a higher standard of evidence.

Suppose that subjects think that, nationwide, a fourth grade student attending a school with poor resources is likely to be operating at a level of 3.5, while a student attending a wealthy school is likely to be operating at a level of 4.5. However, there is a 35% chance that any child is exceptional, that is, exceptionally bad or exceptionally good, and subjects require 75% certitude to make a judgement of an individual member of a demographic class.[17] Since the 75% standard has not been met, initially everyone reports that Hannah is operating at a level of 4. Now subjects are shown a video of Hannah, answering questions among other things. By design, the video clearly establishes one thing about Hannah: she is not exceptional. The required standard of evidence is now met and subjects' responses polarize to 3.5 and 4.5, the levels for the two types of schools. We have obtained unbiased population polarization by modelling Darley and Gross' own words, although not in the way they themselves would choose to model them.

4.7 Proofs

Proof of Theorem 1. Suppose c is equivocal, and assume $p_c > q_c$ and $r_c < t_c$. This holds if and only if

$$\begin{aligned} P(T \mid \mathrm{H}, c, s) &= \frac{p_c p_s ab}{p_c p_s ab + q_c q_s (1-a) b} = \frac{p_s ab}{p_s ab + \frac{q_c}{p_c} q_s (1-a) b} \\ &> \frac{p_s ab}{p_s ab + q_s (1-a) b} = P(T \mid \mathrm{H}, s) \end{aligned} \tag{8}$$

and similarly $P(T \mid \mathrm{L}, c, s) < P(T \mid \mathrm{L}, s)$. The proof that $p_c < q_c$ and $r_c > t_c$ if and only if $ii)$ holds is omitted. ■

Recall the sign function is defined by $sgn(x) = -1$ if $x < 0$, 0 if $x = 0$, and 1 if $x > 0$.

Lemma 1 *Suppose that c is equivocal and let B be a belief over Ω that assigns strictly positive probability to every state. There exists $\sigma_B \in (0,1)$ such that $sgn[B(T \mid c, \sigma) - B(T \mid \sigma)] = sgn[\sigma - \sigma_B]$ for all σ.*

Proof. We have that $B(T \mid c, \sigma) - B(T \mid \sigma)$ has the same sign as

$$\frac{p_c B(T\mathrm{H} \mid \sigma) + r_c B(T\mathrm{L} \mid \sigma)}{q_c B(F\mathrm{H} \mid \sigma) + t_c B(F\mathrm{L} \mid \sigma)} - \frac{B(T\mathrm{H} \mid \sigma) + B(T\mathrm{L} \mid \sigma)}{B(F\mathrm{H} \mid \sigma) + B(F\mathrm{L} \mid \sigma)}$$

which, letting $g = \frac{\pi_{\mathrm{H}}(\sigma)}{\pi_{\mathrm{L}}(\sigma)}$ can be written as

$$[pB(\mathrm{TH})\, g + rB(\mathrm{TL})]\,[B(\mathrm{FH})\, g + B(\mathrm{FL})] - [B(\mathrm{TH})\, g + B(\mathrm{TL})]\,[qB(\mathrm{FH})\, g + tB(\mathrm{FL})] \tag{9}$$

[17] See Benoît and Dubra (2004) for an example of a model where such a decision making rule arises in a utility-maximizing setting.

Define

$$\begin{aligned} f(\sigma) &\equiv B(\text{FH})\,B(\text{TH})\,\frac{\pi_{\text{H}}(\sigma)}{\pi_{\text{L}}(\sigma)}(p-q)+ \\ &\quad B(\text{TH})\,B(\text{FL})\,p - B(\text{FH})\,B(\text{TL})\,q - B(\text{TH})\,B(\text{FL})\,t + B(\text{TL})\,B(\text{FH})\,r \end{aligned}$$

and note that $f(\sigma)$ is increasing in σ. Expression(9) can be written as

$$M(\sigma) \equiv \frac{\pi_{\text{H}}(\sigma)}{\pi_{\text{L}}(\sigma)} f(\sigma) - B(T\text{L})\,B(F\text{L})\,(t-r),$$

so that $B(T \mid c,\sigma) - B(T \mid \sigma)$ has the same sign as $M(\sigma)$.

As $\sigma \to 0$, $f(\sigma)$ converges to a constant and $\frac{\pi_{\text{H}}(\sigma)}{\pi_{\text{L}}(\sigma)}$ converges to 0; hence $M(\sigma)$ converges to $-B(T\text{L})\,B(F\text{L})\,(t-r) < 0$. As $\sigma \to 1$, $\frac{\pi_{\text{H}}(\sigma)}{\pi_{\text{L}}(\sigma)} f(\sigma) \to \infty$, so that $M(\infty) > 0$. Since $\frac{\pi_{\text{H}}(\sigma)}{\pi_{\text{L}}(\sigma)}$ and $f(\sigma)$ are increasing and continuous, $M(\sigma)$ is also increasing and continuous and there exists a unique $\sigma_B \in (0,1)$ such that $M(\sigma_B) = 0$. Then, $sgn\left[B(T \mid c,\sigma) - B(T \mid \sigma)\right] = sgn\left[M(\sigma) - M(\sigma_B)\right] = sgn\left[\sigma - \sigma_B\right]$. ■

Proof of Theorem 2. Let $B = P(\cdot \mid s)$ and set $h_s = P(\text{H} \mid s, \sigma_B)$ for σ_B as in Lemma 1. Since $P(\text{H} \mid s,\sigma)$ is strictly increasing in σ, we obtain that

$$\begin{aligned} sgn\left[P(\text{H} \mid s,\sigma) - h_s\right] &= sgn\left[P(\text{H} \mid s,\sigma) - P(\text{H} \mid s,\sigma_B)\right] \\ &= sgn\left[\sigma - \sigma_B\right] = sgn\left[P(T \mid c,s,\sigma) - P(T \mid s,\sigma)\right] \end{aligned}$$

as was to be shown. ■

Lemma 2 *Suppose s is equivocal. Then $P(T \mid s,\sigma') > P(T \mid s,\sigma)$ implies $P(\text{H} \mid s,\sigma') > P(\text{H} \mid s,\sigma)$ and $P(T \mid s,\sigma') < P(T \mid s,\sigma)$ implies $P(\text{H} \mid s,\sigma') < P(\text{H} \mid s,\sigma)$.*

Proof. Note first that

$$\begin{aligned} P(T \mid s,\sigma) &= \frac{abp_s\pi_{\text{H}}(\sigma) + a(1-b)\,r_s\pi_{\text{L}}(\sigma)}{abp_s\pi_{\text{H}}(\sigma) + (1-a)\,bq_s\pi_{\text{H}}(\sigma) + a(1-b)\,r_s\pi_{\text{L}}(\sigma) + (1-a)(1-b)\,t_s\pi_{\text{L}}(\sigma)} \\ &= \frac{abp_s + a(1-b)\,r_s\frac{\pi_{\text{L}}(\sigma)}{\pi_{\text{H}}(\sigma)}}{abp_s + (1-a)\,bq_s + (ar_s + (1-a)\,t_s)(1-b)\frac{\pi_{\text{L}}(\sigma)}{\pi_{\text{H}}(\sigma)}}. \end{aligned}$$

We have

$$\frac{dP(T \mid s,\sigma)}{d\frac{\pi_{\text{L}}}{\pi_{\text{H}}}} = \frac{ab(q_sr_s - p_st_s)(1-a)(1-b)}{\left(abp_s + (1-a)\,bq_s + (ar_s + (1-a)\,t_s)(1-b)\frac{\pi_{\text{L}}(\sigma)}{\pi_{\text{H}}(\sigma)}\right)^2} < 0.$$

Since $\frac{\pi_{\text{L}}(\sigma)}{\pi_{\text{H}}(\sigma)}$ is strictly decreasing in σ, we have that $P(T \mid s,\sigma)$ is strictly increasing in σ. But then,

$$P(\text{H} \mid s,\sigma) = \frac{abp_s + (1-a)\,bq_s}{abp_s + (1-a)\,bq_s + a(1-b)\,r_s\frac{\pi_{\text{L}}(\sigma)}{\pi_{\text{H}}(\sigma)} + (1-a)(1-b)\,t_s\frac{\pi_{\text{L}}(\sigma)}{\pi_{\text{H}}(\sigma)}}$$

ensures $sgn\left[P\left(\mathrm{H}\mid s,\sigma'\right)-P\left(\mathrm{H}\mid s,\sigma\right)\right]=sgn\left[\sigma'-\sigma\right]=sgn\left[P\left(T\mid s,\sigma'\right)-P\left(T\mid s,\sigma\right)\right]$ as was to be shown. ■

Proof of Theorem 3. Let $B=P\left(\cdot\mid s\right)$ in Lemma 1, and let σ_B be such that $sgn\left[P\left(T\mid c,s,\sigma\right)-P\left(T\mid s,\sigma\right)\right]=sgn\left[\sigma-\sigma_B\right]$. Define $v_s=P\left(T\mid s,\sigma_B\right)$. Then by Lemma 2 we have the second equality below, and by Lemma 1, the fourth

$$\begin{aligned} sgn\left[P\left(T\mid s,\sigma\right)-v_s\right] &= sgn\left[P\left(T\mid s,\sigma\right)-P\left(T\mid s,\sigma_B\right)\right]=sgn\left[P\left(\mathrm{H}\mid s,\sigma\right)-P\left(\mathrm{H}\mid s,\sigma_B\right)\right] \\ &= sgn\left[\sigma-\sigma_B\right]=sgn\left[P\left(T\mid c,s,\sigma\right)-P\left(T\mid s,\sigma\right)\right]. \end{aligned}$$

■

Proof of Theorem 4. The v^* around which experts polarize completely is given by $v^*=v_{\bar{s}}$ in Theorem 3. Note that because $v_{\bar{s}}=P\left(T\mid\bar{s},\sigma_B\right)$ for $\sigma_B\in(0,1)$ from Lemma 1, we have that $P^{v_{\bar{s}}},P_{v_{\bar{s}}}>0$. ■

Proof of Theorem 5. For each s compute $\sigma_B\in(0,1)$ from Lemma 1 with $B=P\left(\cdot\mid s\right)$ and define $v_s=P\left(T\mid s,\sigma_B\right)$. Note that because for each s we have $\sigma_B\in(0,1)$, there is a positive mass of signals σ such that $P\left(T\mid s,\sigma\right)>P\left(T\mid s,\sigma_B\right)=v_s$. We obtain that for $\overline{v}=\max_{s\in\mathcal{S}}\left\{v_s\right\}$, $P^{\overline{v}}>0$. Similarly, for $1-\underline{v}=\min_{s\in\mathcal{S}}\left\{v_s\right\}\leq\overline{v}$ we obtain $P_{1-\underline{v}}<1$. By Theorem 3

$$P\left(T\mid s,\sigma\right)>\overline{v}\Rightarrow P\left(T\mid s,\sigma\right)>v_s\Rightarrow P\left(T\mid c,s,\sigma\right)>P\left(T\mid s,\sigma\right)$$

which establishes (3). Similarly, $P\left(T\mid s,\sigma\right)<1-\underline{v}\Rightarrow P\left(T\mid c,s,\sigma\right)<P\left(T\mid s,\sigma\right)$ as was to be shown. ■

Proof of Proposition 6. If s and c are symmetric, $P\left(T\mid s,\sigma,c\right)>P\left(T\mid s,\sigma\right)$ if and only if

$$\begin{aligned} \frac{pp_sab\pi_{\mathrm{H}}\left(\sigma\right)+qq_sa\left(1-b\right)\pi_{\mathrm{L}}\left(\sigma\right)}{qq_sb\pi_{\mathrm{H}}\left(\sigma\right)\left(1-a\right)+pp_s\left(1-b\right)\left(1-a\right)\pi_{\mathrm{L}}\left(\sigma\right)} &> \frac{p_sab\pi_{\mathrm{H}}\left(\sigma\right)+q_sa\left(1-b\right)\pi_{\mathrm{L}}\left(\sigma\right)}{q_s\left(1-a\right)b\pi_{\mathrm{H}}\left(\sigma\right)+p_s\left(1-b\right)\left(1-a\right)\pi_{\mathrm{L}}\left(\sigma\right)}\Leftrightarrow \\ \frac{pp_sb\pi_{\mathrm{H}}\left(\sigma\right)+qq_s\left(1-b\right)\pi_{\mathrm{L}}\left(\sigma\right)}{qq_sb\pi_{\mathrm{H}}\left(\sigma\right)+pp_s\left(1-b\right)\pi_{\mathrm{L}}\left(\sigma\right)} &> \frac{p_sb\pi_{\mathrm{H}}\left(\sigma\right)+q_s\left(1-b\right)\pi_{\mathrm{L}}\left(\sigma\right)}{q_sb\pi_{\mathrm{H}}\left(\sigma\right)+p_s\left(1-b\right)\pi_{\mathrm{L}}\left(\sigma\right)}\Leftrightarrow \\ b\pi_{\mathrm{H}}\left(\sigma\right) &> \left(1-b\right)\pi_{\mathrm{L}}\left(\sigma\right). \end{aligned} \tag{10}$$

We have

$$P\left(T\mid s,\sigma\right)=\frac{abp_s\pi_{\mathrm{H}}\left(\sigma\right)+a\left(1-b\right)q_s\pi_{\mathrm{L}}\left(\sigma\right)}{abp_s\pi_{\mathrm{H}}\left(\sigma\right)+a\left(1-b\right)q_s\pi_{\mathrm{L}}\left(\sigma\right)+\left(1-a\right)bq_s\pi_{\mathrm{H}}\left(\sigma\right)+\left(1-a\right)p_s\left(1-b\right)\pi_{\mathrm{L}}\left(\sigma\right)}$$

Letting $y=\frac{b\pi_{\mathrm{H}}(\sigma)}{(1-b)\pi_{\mathrm{L}}(\sigma)}$, we obtain

$$\begin{aligned} P\left(T\mid s,\sigma\right) &> a\Leftrightarrow\frac{1}{1+\frac{1-a}{a}\frac{q_sy+p_s}{p_sy+q_s}}>a\Leftrightarrow \\ \frac{q_sy+p_s}{p_sy+q_s} &\Leftrightarrow \frac{b\pi_{\mathrm{H}}\left(\sigma\right)}{\left(1-b\right)\pi_{\mathrm{L}}\left(\sigma\right)}>1 \end{aligned}$$

Hence,

$$\begin{aligned} P(T \mid s,\sigma) &> a \Rightarrow b\pi_{\mathrm{H}}(\sigma) > (1-b)\,\pi_{\mathrm{L}}(\sigma) \\ &\Rightarrow P(T \mid s,\sigma,c) > P(T \mid s,\sigma) \end{aligned}$$

and similarly for $P(T \mid s,\sigma) < a$. ■

Proof of Theorem 8. The prior over the eight states is

$$\begin{array}{cll} & T & F \\ Hh & abd & (1-a)\,bd \\ Lh & a(1-b)\,d & (1-a)(1-b)\,d \\ Hl & ab(1-d) & (1-a)\,b(1-d) \\ Ll & a(1-b)(1-d) & (1-a)(1-b)(1-d) \end{array} \tag{11}$$

It is easy to check that we can write an agent's posteriors as,

	posterior after s and σ proportional to T	posterior after s and σ proportional to F		posterior after s,c and σ proportional to T	posterior after s,c and σ proportional to F
Hh	$afgw$	$(1-a)\,fgx$		$afgwp$	$(1-a)\,fgxq$
Lh	$a(1-f)\,gw$	$(1-a)(1-f)\,gx$	&	$a(1-f)\,gwr$	$(1-a)(1-f)\,gxt$
Hl	$af(1-g)\,y$	$(1-a)\,f(1-g)\,z$		$af(1-g)\,yp$	$(1-a)\,f(1-g)\,zq$
Ll	$a(1-f)(1-g)\,y$	$(1-a)(1-f)(1-g)\,z$		$a(1-f)(1-g)\,yr$	$(1-a)(1-f)(1-g)\,zt$

for some f and g. We have,

$$\begin{aligned} \frac{P(T \mid s,\sigma)}{1-P(T \mid s,\sigma)} &= \frac{a}{1-a}\frac{fgw+(1-f)\,gw+f(1-g)\,y+(1-f)(1-g)\,y}{fgx+(1-f)\,gx+f(1-g)\,z+(1-f)(1-g)\,z} > \frac{v}{1-v} \Leftrightarrow \\ \frac{1-a}{a}\frac{v}{1-v} &< \frac{gw+(1-g)\,y}{gx+(1-g)\,z} \end{aligned}$$

Since, $P > v \Leftrightarrow \frac{P}{1-P} > \frac{v}{1-v}$, we have that $sgn\left[P(T \mid s,\sigma)-v\right]$ depends on g but not on f.

Similarly

$$\begin{aligned} P(T \mid s,c,\sigma) &> P(T \mid s,\sigma) \Leftrightarrow \\ \frac{fgpw+(1-f)\,grw+f(1-g)\,py+(1-f)(1-g)\,ry}{fgqx+(1-f)\,gtx+f(1-g)\,qz+(1-f)(1-g)\,tz} &> \frac{gw+(1-g)\,y}{gx+(1-g)\,z} \Leftrightarrow \\ \frac{fp+(1-f)\,r}{fq+(1-f)\,t}\frac{gw+(1-g)\,y}{gx+(1-g)\,z} &> \frac{gw+(1-g)\,y}{gx+(1-g)\,z} \Leftrightarrow \frac{fp+(1-f)\,r}{fq+(1-f)\,t} > 1 \end{aligned}$$

so $sgn\left[P(T \mid s,c,\sigma)-P(T \mid s,\sigma)\right]$ depends on f but not g.

Therefore, conditioning on $sgn\left[P_\omega(T \mid s,\sigma)-v\right]$ does not affect the probability that $P(T \mid s,c,\sigma) > P(T \mid s,\sigma)$, which establishes the desired result. ■

Proof of Theorem 9. Write j and i's initial beliefs as

	True	False		True	False
High	$\widetilde{a}$	$\widetilde{b}$	High	$\overline{a}$	$\overline{b}$
Low	$\widetilde{c}$	$\widetilde{d}$	Low	$\overline{c}$	$\overline{d}$
	j's beliefs			i's beliefs	

For i, we have

$$\begin{aligned} P(T \mid c, s_i, \sigma_i) - P(T \mid s_i, \sigma_i) &= \frac{p_c\overline{a} + r_c\overline{c}}{p_c\overline{a} + q_c\overline{b} + r_c\overline{c} + t_c\overline{d}} - \frac{\overline{a} + \overline{c}}{\overline{a} + \overline{b} + \overline{c} + \overline{d}} > 0 \Leftrightarrow \\ 0 &< \frac{\overline{a}\overline{b}p_c - \overline{a}\overline{b}q_c + \overline{a}\overline{d}p_c - \overline{b}\overline{c}q_c + \overline{b}\overline{c}r_c - \overline{a}\overline{d}t_c + \overline{c}\overline{d}r_c - \overline{c}\overline{d}t_c}{\left(\overline{a}p_c + \overline{b}q_c + \overline{c}r_c + \overline{d}t_c\right)\left(\overline{a} + \overline{b} + \overline{c} + \overline{d}\right)} \Leftrightarrow \\ 0 &< \overline{a}\overline{b}(p_c - q_c) + \overline{a}\overline{d}(p_c - t_c) + \overline{b}\overline{c}(r_c - q_c) + \overline{c}\overline{d}(r_c - t_c) \quad (12) \end{aligned}$$

and similarly for j. First suppose that c is equivocal. For $\varepsilon \approx 0$, set $\overline{b} = \overline{a} = \frac{1}{2} - \varepsilon$, $\overline{c} = \overline{d} = \varepsilon$, $\widetilde{b} = \widetilde{a} = \varepsilon$ and $\widetilde{c} = \widetilde{d} = \frac{1}{2} - \varepsilon$. Then $P(T \mid s_i, \sigma_i) = \overline{a} + \overline{c} = \frac{1}{2} = P(T \mid s_j, \sigma_j)$. The right hand side of expression (12) becomes

$$\overline{a}^2(p_c - q_c) + \overline{a}\left(\frac{1}{2} - \overline{a}\right)(p_c - t_c + r_c - q_c) + \left(\frac{1}{2} - \overline{a}\right)^2(r_c - t_c)$$

which is greater than 0 for $\varepsilon \approx 0$, so that i revises upwards. Writing expression (12) for j, the right hand side is less than 0 for $\varepsilon \approx 0$, so that j revises downwards.

Suppose now that c is unbalanced with $\min\{p_c, q_c\} > \max\{r_c, t_c\}$ (the case $\min\{r_c, t_c\} > \max\{p_c, q_c\}$ is analogous and omitted). For $\varepsilon \approx 0$, set $\overline{a} = \overline{d} = \frac{1}{2} - \varepsilon$, $\overline{b} = \overline{c} = \varepsilon$, $\widetilde{a} = \widetilde{d} = \varepsilon$ and $\widetilde{c} = \widetilde{b} = \frac{1}{2} - \varepsilon$. A similar argument to the one above shows that i revises upwards and j revises downwards.

To show the converse, we argue by contradiction. Assume that c is neither equivocal nor unbalanced and suppose that for some initial beliefs, i and j polarize. We must then have that of the four terms in brackets in (12), some are strictly positive and some are strictly negative.

a) Suppose $p_c > q_c$, so that we must find which of the other three bracketed terms in (12) is negative.

- If $t_c > r_c$ the signal is equivocal, contradicting our assumption. So assume $r_c \geq t_c$.
- If $t_c > p_c$, we have $r_c \geq t_c > p_c > q_c$, so that $\min\{r_c, t_c\} > \max\{p_c, q_c\}$, and c is equivocal. So assume $p_c \geq t_c$.
- If $q_c > r_c$ we obtain $p_c > q_c > r_c \geq t_c$, so that the signal is unbalanced, contradicting the assumption.

b) Suppose $p_c = q_c$. Of the three remaining bracketed terms, one must be positive and one negative.

- If $p_c > t_c$, if either of the final two terms is negative ($p_c = q_c > r_c$ or $t_c > r_c$), then $\min\{p_c, q_c\} > \max\{r_c, t_c\}$ so again the signal is unbalanced.
- If $p_c = t_c$, the two remaining brackets are $(r_c - p_c)$, so they have the same sign and polarization is not possible.
- If $p_c < t_c$, if either of the final two terms is positive ($p_c = q_c < r_c$ or $t_c < r_c$), then $\max\{p_c, q_c\} < \min\{r_c, t_c\}$ so again the signal is unbalanced, contradicting our assumption.

The case $p_c < q_c$ is analogous. ■

References

Acemoglu, D., V. Chernozhukov and M. Yildiz (2009) "Fragility of Asymptotic Agreement under Bayesian Learning," mimeo.

Andreoni, J. and T. Mylovanov (2012) "Diverging Opinions," American Economic Journal: Microeconomics, 4(1): 209–232

Baliga, S., E. Hanany and P. Klibanoff (2013), "Polarization and Ambiguity," *American Economic Review* **103(7)**, 3071–83.

Benoît, J.-P. and J. Dubra (2004), "Why do Good Cops Defend Bad Cops?", International Economic Review.

Bullock, John G. (2009), "Partisan Bias and the Bayesian Ideal in the Study of Public Opinion," *The Journal of Politics*, **(71) 3,** 1109-1124.

Cason, C. and C. Plott (2014), "Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing," *Journal of Political Economy* **122(6)**, pp. 1235-70.

Darley, J.M. and P.H. Gross (1983), "A Hypothesis-Confirming Bias in Labeling Effects," *Journal of Personality and Social Psychology* **44(1)**, 20-33.

Dixit, A. and J. Weibull, (2007), "Political Polarization", Proceedings of the National Academy of Sciences, 104, 7351–7356.

Fryer, R., P. Harms and M. Jackson (2013), "Updating Beliefs with Ambiguous Evidence: Implications for Polarization," NBER WP 19114.

Galiani, S., P. Gertler and E. Schargrodsky (2005), "Water for Life: The Impact of the Privatization of Water Services on Child Mortality," *Journal of Political Economy,* **113(1)**], 83-120.

Garcé, A. y J. Yaffé (2004) *La Era Progresista*, Ed. Fin de Siglo, Montevideo Uruguay.

Gelman, A., J. Fagan, and A. Kiss (2007),An Analysis ofthe New York City Police Department's "Stop and Frisk" Policy in the Context of Claims of Racial Bias, Journal of the American Statstical Association, **(102)** 476, 813-823.

Gerber and Green (1999) , "Misperceptions About Perceptual Bias," *American Review of Political Science*, **2,**189-210.

Glaeser, E.L. and C.R. Sunstein (2013) "Why does balanced news produce unbalanced views?" NBER WP 18975.

Gneezy, U. and J. List (206) "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments," *Econometrica* **74(5)**, 1365-84.

Jern, A., K. K. Chang and C. Kemp (2014) "Belief polarization is not always irrational," *Psychological Review* **121(2)**, 206-24.

Kondor, P. (2012), "The More We Know about the Fundamental, the Less We Agree on the Price," Review of Economic Studies (2012) 79, 1175–1207

Kuhn, D., and J. Lao (1996) , "Effects of Evidence on Attitudes: Is Polarization the Norm?," *Psychological Science*, **7**(2), 115-120.

Levitt, S.D., J. List, and S. Sadoff (2011), "Checkmate: Exploring Backward Induction among Chess Players," *American Economic Review* **101(2)** 975-90.

List, J. (2004), "Neoclassical Theory Versus Prospect Theory: Evidence from the Marketplace," *Econometrica*, **72(2)**, 615-25.

List, J. (2006), "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions," *Journal of Political Economy*, **114(1)**, 1-37.

List, J. (2007), "On the Interpretation of Giving in Dictator Games," *Journal of Political Economy*, **115(3)**,482-94.

Lord, CG., Lepper, M.R, & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. Journal of Personality and Social Psychology, 47, 1231-1243.

Lord, C.G. L. Ross and M.R. 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.

Miller, A. G., J. W. McHoskey, C. M. Bane, and T. 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*, **64(4)**, 561–574.

Munro and Ditto (1997), Biased Assimilation, Attitude Polarization, and Affect in Reactions to Stereotype-Relevant Scientific Information *Personality and Social Psychology Bulletin.* **(23)6,** 636-653.

Nyhan, B., and J. Reifler (2010), When Corrections Fail: The Persistance of Political Misperceptions," *Political Behavior* **32(2)**: 303–330.

Palacios-Huerta, I. and R. Serrano (2006), "Rejecting Small Gambles under Expected Utility," *Economics Letters* **91**, pp. 250-9.
Palacios-Huerta, I. and O. Volij (2009), "Field Centipedes," *American Economic Review* **99(4)**, pp. 1619–35.
Plott, C. and K. Zeiler (2005), "The Willingness to Pay–Willingness to Accept Gap, the "Endowment Effect," Subject Misconceptions, and Experimental Procedures for Eliciting Valuations," *American Economic Review* **95(3)**, pp. 530-45.
Plott, C. and K. Zeiler (2007), "Exchange Asymmetries Incorrectly Interpreted as Evidence of Endowment Effect Theory and Prospect Theory?," *American Economic Review* **97(4)**, pp. 1449-66.
Plous, S. (1991), "Biases in the Assimilation of Technological Breakdowns: Do Accidents Make Us Safer?," *Journal of Applied Social Psychology,* **21(13)**, 1058-82.
Pascal, B., Pensees, translatd by W.F. Trotter
Rabin, M. and J. Schrag (1999), "First Impressions Matter: A Model Of Confirmatory Bias," *The Quarterly Journal of Economics* **114(1)**.
Rubinstein, A. (2003), "Economics and Psychology? The case of Hyperbolic Discounting," *International Economic Review* **44(4)**.
Seidenfeld, T. and L. Wasserman (1993), "Dilation for Sets of Probabilities," *Annals of Statistics* **21(3)**, pp. 1139-54.
Stoop, J., Noussair, C., van Soest, D. "From the Lab to the Field: Cooperation among Fishermen," *Journal of Political Economy* **120(6)**, pp. 1027-56.
The Sentencing Project (2014), Race and Punishment: Racial Perceptions of Crime and Support for Punitive Policies.
Walley, P. (1991). Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London.