A Theory of Rational Attitude Polarization*

Jean-Pierre Benoît
London Business School

Juan Dubra
Universidad de Montevideo

Revised March, 2015.

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 set of studies on the issue. Believers and skeptics both become 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, some experiments do not find polarization, under the conditions in which our theory predicts the absence of polarization.

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%.[1] Moving from beliefs to data, several studies have confirmed what many people have long suspected – that police "stop and frisk" racial and ethnic minority members at higher rates than whites (See Gelman, Fagan, and Kiss (2007)). 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?

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

[1] http://www.gallup.com/poll/163610/gulf-grows-black-white-views-justice-system-bias.aspx. See The Sentencing Project (2014) for a discussion of this and related results.

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 not 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 process information in a biased manner to support their pre-existing views.

In this paper, we argue that, on the contrary, this polarization of attitudes is often exactly what we should expect to find in a perfectly rational unbiased population. Our argument consists of two parts. First of all, the mere fact that the opinions of two given people move further apart after they receive a common piece of information is perfectly consistent with Bayesian updating and not particularly surprising. This observation has been made by others as well, and we discuss their work in Section 3. Note, however, that the opinions of some individuals could diverge while the opinions of others converge, resulting in no polarization overall. The crux of the literature on attitude polarization is not about two people's opinions moving further apart but, rather, the systematic divergence of opinions. The second, more important, part of our argument is that when the mixed information that people are presented with is, in some sense, typical evidence, as is often the case in experiments, one should expect the views of people on either side of an issue to strengthen, so that the population on the whole polarizes. When the information is novel, one should not expect the population to polarize, as some experiments find.

Our reasoning is simple. Consider a group of people with differing opinions on an issue – the available information is equivocal and has induced favourable views in some of them and disfavourable 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 will be more likely to respond favourably than those who considered it to be negative, so that the population will polarize. On the other hand, suppose the mixed information is novel. While some people may react positively to it, and others react negatively or neutrally, there is no reason for their reactions to correlate with their initial positions, and no reason to expect a particular pattern of belief changes at the level of the population. Our model also predicts

[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 normative opinions (e.g., are you in favour of capital punishment?) and positive beliefs (e.g., do you believe capital punishment has a deterrent effect?). Throughout this paper, we only discuss movements in positive beliefs, as it is unclear how to evaluate changes in normative opinions.

other empirical findings that have been noted.

Although several authors have argued that Bayesian reasoning is consistent with *individuals* polarizing, no one previously has explained *populations* polarizing in a fully rational framework. Still, while we develop our ideas in a rational setting, our interest is not in rationality per se, but the extent to which attitude polarization is consistent with, and even predicted by, unbiased reasoning. Full rationality provides a convenient benchmark of unbiased reasoning.[3]

In the next section, we give a detailed informal exposition of our reasoning, including a simple numerical example in Section 1.1. In Section 2, we present a formal model that demonstrates group polarization; in Section 2.1 we provide conditions under which we would not expect groups to polarize; in Section 2.2 we characterize the conditions under which two individuals can polarize. In Section 3 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 The Basic Reasoning

Our starting point is an experiment in which subjects are twice asked to indicate their opinions on an issue. For instance, they may be asked to indicate the extent to which they believe that nuclear energy is safe, on a scale from -8 (completely safe) to 8 (not safe at all). Individuals arrive at the experiment with personal information from their own experience, and, depending on the particular experiment, before being asked to choose a number the first time, they may also be given differing pieces of information. In any case, before being asked their opinions a second time, all subjects are supplied with a common piece of information.

As an example, in a study by Darley and Gross (1983), subjects were presented with a description of a fourth-grade girl named Hannah. The subjects were partitioned into two groups, one of which was given information strongly suggesting that Hannah came from an upper class background and one of which was given information suggesting that she came from a lower class background – information that could potentially have a biasing effect on the way they processed subsequent information. At that point, the subjects were asked several questions about Hannah, among them to indicate at what grade level they believed she was actually functioning in mathematics, liberal arts, and reading. Subjects gave their answers on a scale that went from 0 (kindergarten) to 6.75 (sixth grade, nine months). Subjects who believed that Hannah came from a well-off family tended to rate her grade level in these three disciplines as slightly higher than those who believed she came from a poorer family.

[3] Thus, our theory can be applied to, say, unbiased subjects who are guilty of base rate neglect.

Next, subjects were provided with some specific evidence about Hannah's abilities. This evidence was the same for all the subjects. They were then again asked to rate her.[4] On average, the subjects who believed that Hannah was well-off and who had initially rated her higher, revised their estimates upwards, while the complementary group revised downwards. Notice that i) some subjects reacted positively to the additional information while others reacted negatively and ii) the difference in reaction was not distributed randomly among subjects.

To evaluate these findings, we start by considering two subjects, A and B, such that A's initial response was greater than B's. We will use the terms below to describe the different ways in which A and B's answers might move relative to each other after receiving additional common information. (For ease of exposition, we assume that the responses of both individuals change. A formal definition of polarizing is given in the next section.)

Definition 1 *Suppose that A's initial response is greater than B's. Following a common piece of information,*

A and B ***harmonize*** *if their responses both rise or both fall.*

A and B ***moderate*** *if A's response falls towards B's and B's response rises towards A's.*

A and B ***polarize*** *if A's response rises and B's response falls.*

For concreteness, suppose that A's initial response was 4.25, while B's was 3.75. Which type of response movement would seem, potentially at least, to be problematic? Trivially, neither harmonization nor moderation pose a problem. Indeed, if the common information was that a battery of reliable tests established Hannah to be operating at the level of a fifth grader, we would expect the responses to harmonize upwards towards 5, whereas if the information was that the tests established her level to be that of a beginning fourth-grader, we would expect the responses to moderate towards 4. Thus, at a somewhat intuitive level, the only potentially problematic change is that A and B polarize.

In fact, the additional information that Darley and Gross presented to their subjects was, by design, not so clear-cut. Rather, it was a video of Hannah taking an oral test in which she answered some difficult questions correctly but missed on some easy questions, was sometimes seen to be concentrating assiduously but was sometimes distracted. What conclusion should a subject draw from such mixed evidence? Consider the following three possibilities:

i Hannah's mixed performance is typical of an average fourth grade student.

[4] 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.

ii The fact that Hannah manages to answer some difficult questions and to concentrate when she wants to, is telling. The easy questions may just bore her. Hannah may not be the best student, but her level is certainly well above average (say, 4.5 or above).

iii The fact that Hannah misses some easy questions and cannot maintain her concentration is troublesome, the occasional correct answers on difficult questions notwithstanding. Hannah may not be the worst student, but her level is well below average (say, 3.5 or below).

All three of these conclusions strike us as defensible. To put it differently, while we may favour a particular one, none of the conclusions in and of themselves seem to be evidence of biased reasoning, especially given that the experimental subjects were not education experts, but merely Princeton undergraduates. Actually, we do not need to rely on our own intuition in this matter, as Darley and Gross ran a control where they asked subjects to rate Hannah based solely on the performance video, that is, to rate her without seeing any potentially biasing demographic information. Subjects' responses had a mean close to 4, but a significant fraction were at least 4.5 and a significant fraction were 3.5 or below.[5] That is to say, many of the necessarily unbiased subjects interpreted the performance video positively, as in ii), and many interpreted it negatively, as in iii). Hence, neither a rise nor a fall in response by either subject A or B would be problematic and, by extension, neither would polarization. For instance, A and B would unproblematically polarize if A reasons as in *ii*) above, while B reasons as in *iii*), and they reason this way independently of demographic information.

In general, moderation, harmonization and polarization are all consistent with rational unbiased reasoning. Rather than being surprising, this conclusion is almost tautological – when a person is presented with equivocal evidence, that is, evidence that can reasonably be interpreted as being either in favour or against a proposition, his beliefs can reasonably move either towards or away from accepting the proposition, or not move at all, and, by that very fact, the harmonization, moderation, and polarization of two individuals are all reasonable outcomes.[6] Actually, even evidence that is not mixed can lead to polarization, as we discuss later (see Theorem 8, Section 2.2, and Section 5.2). The conditions under which

[5] From Table 1 in their paper, the standard deviations for the 3 main dependent measures -Liberal arts, Reading, and Mathematics) - were .505, .581 and .238. If the scores were approximately normal, around 68% of scores would fall within one standard deviation of the mean. In the case of Liberal arts, the mean was 4, so 32% of scores would fall outside of the range 3.5-4.5. In the case of Reading, even more would fall outside that interval, but in mathematics, fewer would.

[6] Michael Mandler has made the argument to us that if moderation is possible, then, necessarily, polarization is also possible. Essentially, two people who have seen a moderating signal that may or may not be erroneous (say, pure noise), will polarize if they are later told that the signal turned out to be erroneous. This claim is also consistent with results in Baliga *et al.* (2013).

it is possible for two individuals to polarize has been the focus of much of the theoretical literature in economics to date, but it is not our main concern.

Even if people can legitimately update in different directions, a challenge remains. Why would it be the subjects who believed Hannah to be well-off and who initially rated her higher, that tended to revise upwards rather than downwards? More generally, why, when presented with the same information, would subjects on one side of an issue tend to update in an opposing direction to subjects on the other side? Moreover, why would this opposing updating be in directions that confirmed the subjects' initial predispositions? That is, why, or better yet, when, would there be polarization at the level of the *population*? This is the main query that we address in this paper.

Before answering this question, let us first consider an experiment by Plous (1991). Using a questionnaire, he divided subjects into two groups, according to whether they entered his experiment with a belief that a strategy of nuclear deterrence made the United States safer or less safe. He then gave all the subjects the same article to read, which described an 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, the subjects' views of nuclear deterrence moved further in the direction of their initial inclinations.

How should the subjects have reacted? 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." Put differently, the evidence provided by 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.

Plous himself explains why two particular subjects polarizing is not problematic. What about the population polarizing, is that evidence of bias, as he concludes?

We are told that most of the subjects 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 that also yielded population polarization, subjects were given descriptions of "near-miss" incidents that were unfamiliar to them, rather than descriptions of an incident they had already heard of. Which subjects would have entered the experiment with a favourable view of nuclear deterrence? Answering this question reveals the main mechanism at work in our theory.

A reasonable presumption is that the subjects with a favourable view, despite their knowledge of a previous malfunction that was caught by safeguards, were those who considered the reliability of the 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 about properly functioning safeguards. At the same time, those who considered a malfunction of the primary unit to be more dispositive than the quality of the safeguards 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. Section 1.1 provides a numerical illustration of this polarization.

In general, suppose the available information on an issue has induced positive views in some people and negative views in others. When exposed to an additional piece of information that is similar in nature, those who previously considered this type of information to be positive will be more likely to respond favourably than those who previously considered it to be negative, so that the population will polarize.

Before returning to Darley and Gross, which is not covered by this reasoning, consider Lord, Ross and Lepper's (1979) capital punishment experiment. There, 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 those for whom current evidence led to a favourable conclusion with regards to the efficacy of the death penalty responded positively to similar evidence.

The specific information that Lord, Ross, and Lepper provided their subjects was two (purported) studies, one that found that the murder rate tended to be lower in states following the adoption of the death penalty and one that found that the murder rate tended to be higher. Viewed as a single entity, the studies determined that about half the time, a state that adopted the death penalty subsequently had a lower murder rate and half the time a higher murder rate.

Why would some 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 one: some people believe that there is a selection issue, whereby states that adopt the death penalty are states with rising murder rates, and some do not. For people who believe there is a selection issue, 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.[7]

[7] Different beliefs in the importance of selection issues is only a possibility that we have provided to explain the differences in updating. The subjects themselves may have reasoned differently. In Section 5.4

People may be unsure of whether or not selection issues are relevant. Subjects with the strongest belief in the relevance of selection issues should be the ones who enter the experiment with the strongest belief that the death penalty is effective, given the mixed nature of experience with capital punishment to date. They will also be the ones who consider the additional evidence provided in the experiment to be the most favourable and the ones most likely to revise their beliefs that the death penalty deters crime upwards.

Returning to Plous' nuclear deterrence experiment, in it he asked his subjects which was more important, the fact that safeguards worked or the fact that a breakdown had occurred. Consistent with our reasoning, he found that those who felt that safeguards were more important revised upwards their belief in nuclear deterrence, while those who believed that breakdowns were more important revised downwards.

Plous' reasoning on this finding, however, 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 can 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-

we provide another possibility.

We note that Lord et al. 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 in this way is probably not very significant, as the design of the experiment primed them to.

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 7).

Consistent with this prediction, Miller, McHoskey, Bane, and Dowd (1993) found 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 2).

In Darley and Gross (1983), subjects were not in any way pre-sorted for their views on mixed performances. *Arguably* then, we should not expect groups to have interpreted a video of a mixed performance differently depending on whether the video featured a rich girl or a poor girl, so that we should not expect the population to have polarized. We discuss this experiment in some detail in Section 3.1. For now, we simply note that the findings in this study can be explained in many ways and that the case for polarization here is not as clear as it is usually made out to be. For instance, Darley and Gross asked their subjects questions about eight of Hannah's characteristics and did not find polarization on four of them.

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 (Theorems 4 and 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.

While we would like to have given predictions that clearly distinguish our model from a biased-reasoning hypothesis, this is difficult to do, as the broad category "biased reasoning" embraces several possibilities. Thus, while it would seem that, in contrast to 1), biased reasoners should evaluate novel information in a biased manner – accepting evidence they consider to be favorable to their views while rejecting unfavourable evidence – a contrary

argument could be made. For instance, biased reasoners may not have sufficient time in an experiment to come up with a satisfactory (to themselves) reason for rejecting only the unfavourable evidence, when all the evidence is novel. As to biased experts, they may be more emotionally invested and have a greater motivation to act in a biased manner, and hence be more prone to polarization. We do not immediately see a reason for a biased-reasoning hypothesis to also imply that any group of people who have seen similar information will be especially prone to polarize, but neither would we claim that no such reason can be produced.

1.1 A Simple Example

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

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 provided 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:

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

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 and that 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.

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,

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

2 Formal Analysis

The essential elements of an attitude polarization experiment, 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.[8]

The minimal setting that can capture these elements is one in which there is a proposition 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. 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.[9] 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 $\Omega = \{H, L\} \times \{T, F\}$.

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

[9] 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 reading a paper that argues that a particular policy was adopted because of political reasons unrelated to selection (as in Galiani *et al.* (2005), who discuss the privatizaton of water in Argentina) says nothing about the effectiveness of that policy.

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 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:

$$\begin{array}{ccc} & \text{Likelihood of } c\text{:} & \\ & \text{True} & \text{False} \\ \text{High} & p_c & q_c \\ \text{Low} & r_c & t_c \end{array}$$

   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 2 *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 individuals **polarize** 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 3 *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. The bulk of the work on attitude polarization is on how initial beliefs about the *issue* affect updating.

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.

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 an individual updates. We now move from individuals to the population. We begin with some definitions for polarization at the population level. (Formal statements are given in the theorems, as the definitions are used).

- 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 4 *Following a common signal, 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, $P_v > 0$.*

Definition 5 *Following a common signal, 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, $P_v > 0$.*

Definition 6 *Following a common signal, the* ***population polarizes everywhere*** *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, for all v with P^v, $P_v > 0$.*

Definition 7 *Following a common signal,* ***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* $P^{\overline{v}}, P_{1-\underline{v}} > 0$.

Definition 7 is especially important given that there is some evidence that polarization is more marked between sub-populations with the strongest opinions and many experiments pre-select people with strong opinions. If groups with the strongest opinions polarize completely, there will be a range of $\overline{w}$'s and $\underline{w}$'s such that *most* people who believe the proposition with probability greater than $\overline{w}$ 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 definitions 5 and 6.

Proposition 1 *If the population polarizes completely around some* v*, then the population polarizes everywhere.*

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 the theorem is stated for experts, it applies to any population who enter 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. 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.

The level v^* in Theorem 4 need not correspond to the 'dividing line' around which an experimenter checks for polarization. Nonetheless, from Proposition 1, the population polarizes everywhere, 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 her dividing line.[10] 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. Moreover, 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.

Consider what happens as a population of experts receives more and more equivocal information. Although Theorem 4 indicates that the population polarizes at each step, this does not imply that more and more equivocal information inexorably leads to more disagreement or that enough information will not resolve an issue.

To understand this, think again about a finding of a lower murder rate in 50% of jurisdictions with capital punishment and a higher rate in 50%. As we have noted, this could indicate that the death penalty is effective but selection is important or that the death penalty is ineffective and selection plays no role. However, in the first instance there is no particular reason to expect precisely a 50/50 outcome, whereas if the death penalty is ineffective and selection plays no role, 50/50 is exactly what one would expect if murder rates fluctuate

[10] In a typical experiment, a subject is not asked directly for a probability assessment but rather for a number that is, presumably, related to this assessment (see Section 3 for more on this.) 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.

randomly. As a consequence, repeated 50/50 findings can eventually lead everyone to agree that capital punishment is ineffective, even if at each step there are extremists who polarize. We provide a specific illustration of this in Section 5.1 in the appendix.

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 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*

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

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

If everyone's private signal is equivocal, then groups with the strongest opinions polarize.[11] 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) pre-screen their subjects and select only those with 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 $\overline{v}$ and the population also polarizes 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 everywhere if the various pieces of information on the issue are sufficiently dissimilar and the ancillary matter is sufficiently unimportant (see Section 5.3, for an example). On the other hand, when all the signals are equivocal and have symmetric likelihood matrices – so that results are not being pushed in any particular direction – the population polarizes everywhere.

[11] In fact, as shown following the proof of Theorem 5 in the appendix, there is a group of people with belief in the proposition greater than the prior who all revise upwards and a group with belief less than the prior who all revise downwards.

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 everywhere.

2.1 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 moderate, or "depolarize", as Miller *et al.* put it. 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 randomly when presented with novel arguments. Nevertheless, that is what is predicted by our model.

To see this, recall that in a population of people that have (largely) derived their beliefs on nuclear deterrence from their knowledge of near-miss episodes and their views on the primary/safeguards question, proponents of nuclear deterrence will tend to be people who believe that safeguards are most important 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 this type of evidence, the population is presented with the following,

> 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 places on each of them. There is little reason for these weights to bear any particular relation to how important the individual believes primary units are compared to backups. Thus, while different individuals may respond differently to these two arguments, 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, will not cause population polarization.

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 detterence, 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 8 *Signal s is* ***unrelated*** *to signal c if their likelihoods depend upon different ancillary matters. Thus, if s and c are unrelated we can write their likelihood matrices as, say,*

$$\begin{array}{ccc} & T & F \\ Hh & w_s & x_s \\ Lh & w_s & x_s \\ Hl & y_s & z_s \\ Ll & y_s & z_s \end{array} \quad \text{and} \quad \begin{array}{ccc} & T & F \\ Hh & p_c & q_c \\ Lh & r_c & t_c \\ Hl & p_c & q_c \\ Ll & r_c & t_c \end{array}$$

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. It says that 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 7 *If signal c is unrelated to signal s, then, for any $\omega \in \Omega$,*

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

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

Theorem 7 is consistent with Miller et al (1993). Munro and Ditto (1997) investigate movements in subjects' beliefs on stereotypes pertaining to homosexuals. They divide subjects into groups according to their level of prejudice towards homosexuals, measured by the HATH questionnaire, and present them with (supposed) scientific studies on this issue. The level of prejudice has a statistically significant correlation with beliefs on the stereotypes, but this correlation is low. In addition, we speculate that the scientific information in the experiment is novel in that most individuals form views on homosexual stereotypes without knowing that there are studies on the issue.[12] As a result, we do not predict polarization here. The results in the study are mixed with respect to this prediction – polarization is not found on Experiment 1 but is found on Experiment 2.[13]

We note that, while our basic framework has only one ancillary matter, a second 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.

2.2 Individuals polarizing

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 two individuals can polarize.

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

Definition 9 *The signal* c *is* ***unbalanced*** *if* $\min\{p_c, q_c\} > \max\{r_c, t_c\}$ *or* $\min\{r_c, t_c\} > \max\{p_c, q_c\}$.

Theorem 8 *A common signal* c *can cause two individuals to polarize 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 two individuals to polarize, unbalancedness does not naturally lead to population polarization (see the example in

[12] The situation s is different with capital punishment, where even someone who has never read a study on the effect of capital punishment likely knows that i) some jurisdictions have the death penalty while others do not and ii) the resulting evidence on its effectiveness is mixed (given that there is no general consensus.)

[13] As we do throughout this paper, we ignore the portions of their experiment that relate to non-factual questions, specifically, general feelings about homosexuals.

Section 5.1). 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 mixed, or equivocal.

2.3 What do responses mean?

In a standard attitude polarization experiment, subjects are not asked for the distribution of their beliefs but, rather, for a single number that somehow summarizes this distribution. In the case of Darley and Gross, subjects are asked to place Hannah somewhere on a 27-point scale from kindergarten through sixth grade. That is, a subject with a presumably non-degenerate probability distribution over this scale is asked to name a single point. In the case of Lord, Ross, and Lepper, subjects are asked how much their views change on a scale from -8 to 8. What exactly does a subject's answer indicate? Somehow, this question is rarely asked. Perhaps the most obvious possibility (to an economist) is that a subject's response represents the mean of her beliefs, but there are countless other possibilities, including his median beliefs. Actually, for attitude polarization purposes, it is not necessary to decide exactly what a subject's response indicates, only what a change in response means.

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

If we move to an issue that can take one of n values, then a change in beliefs that causes, say, the mean of beliefs to rise may cause the median to fall, making it difficult to interpret the findings of an experiment. Any theoretical results that only demonstrate polarization of mean beliefs will have limited applicability. On the other hand, any results that show 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 quite strong – when there is an fosd shift of beliefs in a certain direction, almost any reasonable point summary of beliefs will move in the same direction.

Our model can be modified to allow for n values and the results recast in terms of fosd shifts, at the cost of added complexity. Thus, let the main issue take one of n values $X = \{x_1, ..., x_n\}$, so that the state space is $\Omega = \{H, L\} \times X$. Let the probability of signal s in state ω be $f_\omega(s)$, and say that signal s is equivocal if $f_{\mathrm{H}x_i}(s)$ is strictly increasing in x_i, and $f_{\mathrm{L}x_i}(s)$ is strictly decreasing in x_i. As an example of a theorem in this framework, let

$r = (r_1, ..., r_n)$ be any probability distribution over the issue. Then, experts whose beliefs fosd r are more likely to have their beliefs shift up than experts for whom r fosd their belief.

3 Related literature

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. A constant response, then, represents either no change in beliefs or a change that is not ordered by fosd. They show that if there is no ancillary matter (to put their result in our terms), then two individuals whose beliefs have common support cannot polarize. This result follows from Theorem 8, 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.[14] (Nevertheless, there is a sense in which an fosd shift can occur even without an ancillary state, as we show in Section 5.2 in the Appendix.) They go on to relate ambiguity aversion to polarization.

Andreoni and Mylovanov (2012) emphasize that two individuals can polarize if there is an ancillary matter (in our terms). They are not particularly concerned with the question of when populations polarize but, rather, are interested in the persistence of disagreement between individuals and when such disagreement can be common knowledge. Kondor (2012) shows that two individuals can polarize in a setting in which peoples' beliefs about the beliefs of others are important. Glaeser and Sunstein (2013) show that two individuals with inconsistent beliefs can polarize.[15] Acemoglu, Chernozhukov, and Yildiz (2009) show that two individuals can persistently polarize if they disagree about the likelihoods of common signals.

Rabin and Shrag (1999) conclude that the literature on attitude polarization has shown that people 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.

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, Lepper, and Ross observe, biased assimilation by

[14] Of course, their result precedes our Theorem 8.

[15] All three papers interpret individuals' responses to reflect their mean beliefs. In fact, when issues can take on more than two values, so that changes in expected value are not isomorphic to fosd changes, individual polarization can arise even in a very narrow setting, as the example in Section 5.2 shows (see also Baliga *et al.* (2013) and Dixit and Weibull (2007)).

itself is not problematic, as it is rational for a person to have greater confidence in a finding that confirms something she believes than a disconfirming finding. 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.

3.1 Hannah revisited

Our theory does not particularly predict population polarization in the experiment of Darley and Gross (1983), as people were not pre-sorted according to their beliefs. Nonetheless, a close examination of this experiment shows that its findings do not conflict with our predictions.

First of all, we note that subjects were asked for their opinions of Hannah's level on three academic subjects, liberal arts, reading, and mathematics, and on five traits, work habits, motivation, sociability, maturity, and cognitive skills. Although the experiment is typically described as one that finds polarization, of these eight categories, polarization was only found on four, hardly an overwhelming finding of polarization.[16]

Leaving aside the negative findings, the strongest positive findings of attitude polarization were on the three academic subjects. Let us be a bit more precise about these results. When given only demographic information, 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 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.

[16] 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.

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 they are shown a video of Hannah that clearly establishes one thing: 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 polarization by modelling Darley and Gross' own words, although not in the way they themselves would choose to model them. Moreover, it is not hard to come up with other ways to obtain unbiased polarization in this experiment, as it is by no means clear that the same behaviour should have the same implications for children from different backgrounds.

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 the conclusion that they have found evidence of biased reasoning is open to doubt, even without appealing to our model.

3.2 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 a distorted impression of this area, as many papers underplay negative findings and provide only a superficial analysis of the experiments they discuss. (This is especially true of papers by non-psychologists, who tend to have narrow goals when invoking the literature.) For instance, papers that quote Darley and Gross typically do not mention the questions on which Darley and Gross do not find polarization, or consider alternate explanations for the positive findings.

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 none with the latter. It is not altogether clear what to make of this discrepancy. Another difficulty is that a proper evaluation of experimental results often requires a close reading

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

of the papers. We have already seen this with Darley and Gross; in this section, we briefly consider two other influential papers.

Kunda (1987) gave subjects a scientific article claiming that women who were heavy drinkers of coffee were at high risk of developing fibrocystic disease, and asked them to indicate how convincing the article was. In one treatment, fibrocystic disease was characterized as a serious health risk and women who were heavy coffee drinkers rated the article as less convincing than women who were light drinkers of coffee (and than men). In a second treatment, the disease was described as common and innocuous and both groups of women rated the article as equally convincing. Note that in the first treatment, the article's claim was threatening to women who were heavy coffee drinkers, and only them, while in the second treatment the article's claim threatened neither group. Kunda's interpretation of her findings is that subjects engaged in *motivated reasoning* and discounted the article when it clashed with what they wanted to believe. However, when subjects were asked how likely they were to develop the disease in the next fifteen years, in both treatments women who were heavy coffee drinkers indicated about a 30% greater chance than light drinkers. That is, although heavy coffee drinkers in the serious health risk treatment described the article as less convincing than in the innocuous risk treatment, they seem to have been 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 (2012) report on an extreme form of polarization, a so-called backfire effect. As they describe it, subjects were given articles to read that contained 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 *backfired*, causing partisans to believe the statement even more.

In their first experiment, all subjects were given an article to read in which Bush justified the United States invasion of Iraq in a manner that suggested that Iraq had weapons of mass destruction. For subjects in the correction condition, the article went on to describe the Duelfer Report, which documented 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 were given and the question they were 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 ran a follow-up study in which this is precisely the correction and question that they used. And with this formulation they did not find a backfire effect. However, their reason for this alternate formulation was 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 recognize the negative findings as well as the positive ones.

4 Conclusion

Unbiased Bayesian reasoning can lead to population polarization. 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 closely together with a very high correlation in the way in which partisan groups update their assessments. Moving to Latin America, free market reforms in the 1990's did not have the large impact on growth that many had promised. Commentators on the left took this as evidence against the supposed benefits of free markets while commentators on the right concluded that the reforms were not extensive enough to produce the desired results. Despite new evidence, old disagreements persisted – which is easily explainable as the product of unbiased reasoning. Nonetheless, while the differences between the left and the right in Latin America run deep, this does not mean that there is never any convergence on any issue. For instance, where once they disagreed, left wing and right wing parties in Uruguay now concur that government debts should be paid.[18]

Where there are persistent differences in political beliefs, it is often not clear what these differences show about how people reason. Many political questions concern issues where

[18] See Garcé and Yaffe (2004) p139 for a speech in 1989 against paying debts by Astori, who later paid debts as minister of economics and later vice president of the country, during the governments of the left wing coalition that came to power in Uruguay in 2005.

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.[19] 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 demonstrating that biased reasoning took place. Interestingly enough, the impact that repeated draws of the same evidence has on beliefs can change over time, without that being a "psychological" phenomenon (see Section 5.1).

The logic of an attitude polarization experiment is as follows. Subjects with varying beliefs on an issue are gathered. The experiment does not inquire as to why subjects' beliefs differ or whether or not these beliefs are rational to begin with. Rather, it implicitly accepts that beliefs can legitimately differ and recognizes that it is difficult to determine if beliefs have been rationally derived without knowing the information upon which they are based. The experiment investigates the way in which people update their beliefs in response to a known piece of information, in particular, the direction in which they update. We have shown

[19] 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 extremely (implausibly?) strong. Moreover, while people who are reasoning in a more or less Bayesian fashion can be expected to draw conclusions that are more or less Bayesian, "small" departures from common knowledge assumptions easily lead to very different conclusions.

that, appearances notwithstanding, this direction is typically in accord with an unbiased application of Bayes' rule. We have not addressed the larger question of just how rational people are.

To push the point, consider subjects whose initial beliefs are simply wrong. Perhaps they stem from an unfounded fear of spiders, a distorted view of how the world works, or baseless prejudices. The question addressed by an attitude polarization study is, given these erroneous views, do subjects commit an additional error and update erroneously, or do they behave consistently with the erroneous views they have. Our model suggests that they do not commit an additional error, or, more precisely, that attitude polarization does not show they do. (At the same time, the fact that a person has beliefs that are largely wrong does not indicate that he or she is irrational. These beliefs may have been rationally derived from the information available to the person.)

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." (Pensees, translated by W. F. Trotter.)

5 Appendix

5.1 Convergence with Polarization

In this section, we present an example to illustrate that polarization may take place even as there is growing agreement in a population, as discussed in Section 2. The example also illustrates the effect of an unbalanced signal.

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

$$\begin{array}{c|c|c|} & \mathrm{T} & \mathrm{F} \\ \hline \mathrm{H} & \frac{4}{5} & \frac{9}{10} \\ \hline \mathrm{L} & \frac{1}{10} & \frac{1}{2} \\ \hline \end{array} \qquad \begin{array}{c|c|c|} & \mathrm{T} & \mathrm{F} \\ \hline \mathrm{H} & \frac{1}{5} & \frac{1}{10} \\ \hline \mathrm{L} & \frac{9}{10} & \frac{1}{2} \\ \hline \end{array} \tag{5}$$

i d

where H corresponds to selection issues being important and L to these issues being irrele-

vant.[20] 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 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*

[20] 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.

is low. Formally, let c^{∞} be a sequence of iid draws from $\mathcal{C}$, and c^t the first t draws. For any σ, $P\left\{c^{\infty} : \lim_{t\to\infty} P\left(LF \mid c^t, \bar{s}, \sigma\right) = 1\right\} = 1$.

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 $P\left(T \mid \bar{s}, \sigma\right) > v_t \Rightarrow P\left(T \mid c_{id}, c^t, \bar{s}, \sigma\right) > P\left(T \mid c^t, \bar{s}, \sigma\right)$ and $P\left(H \mid \bar{s}, \sigma\right) > h_t \Rightarrow P\left(T \mid c_{id}, c^t, \bar{s}, \sigma\right) > P\left(T \mid c^t, \bar{s}, \sigma\right)$.

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, $P\left(T \mid H, c_{ii}, s\right) < P\left(T \mid H, s\right)$ and $P\left(T \mid L, c_{ii}, s\right) < P\left(T \mid L, s\right)$. 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}.

5.2 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 8, 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 "mis-calibrated" 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 $E(\theta \mid s_3, M) = 2.67$.

5.3 Polarization, but not everywhere

The following example shows that the population may not polarize everywhere 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.

5.4 Lord, Ross, and Lepper revisited

We have argued that different opinions as to the importance of selection issues could have caused groups in Lord, Lepper and Ross' capital punishment study to polarize. In this section, we offer another possibility, based on evidence from the experiment. Footnote 2 in the paper reads, in part, "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. In fact, if the true data is predominantly in one direction or the other, then only researchers on one side of the issue need be guilty of distortions. 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.

5.5 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

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

Define

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

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

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

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 $\frac{\pi_{\mathrm{H}}(\sigma)}{\pi_{\mathrm{L}}(\sigma)}$ converges to 0; hence $M(\sigma)$ converges to $-B(T\mathrm{L})\,B(F\mathrm{L})\,(t-r) < 0$. As $\sigma \to 1$, $\frac{\pi_{\mathrm{H}}(\sigma)}{\pi_{\mathrm{L}}(\sigma)} f(\sigma) \to \infty$, so that $M(\infty) > 0$. Since $\frac{\pi_{\mathrm{H}}(\sigma)}{\pi_{\mathrm{L}}(\sigma)}$ 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, $sgn\,[B(T \mid c, \sigma) - B(T \mid \sigma)] = sgn\,[M(\sigma) - M(\sigma_B)] = sgn\,[\sigma - \sigma_B]$. ■

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

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

as was to be shown. ■

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

Proof. Note first that

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

We have

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

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

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

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

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

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

■

Proof of Theorem 4. The v^* around which experts polarize completely is given by $v^* = v_{\bar{s}}$ in Theorem 3. Note that because $v_{\bar{s}} = P(T \mid \bar{s}, \sigma_B)$ 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 $B = P(\cdot \mid s)$ and define $v_s = P(T \mid s, \sigma_B)$. Note that because for each s we have $\sigma_B \in (0,1)$, there is a positive mass of signals σ such that $P(T \mid s, \sigma) > P(T \mid s, \sigma_B) = v_s$. We obtain that for $\overline{v} = \max_{s \in \mathcal{S}} \{v_s\}$, $P^{\overline{v}} > 0$. Similarly, for $1 - \underline{v} = \min_{s \in \mathcal{S}} \{v_s\} \leq \overline{v}$ we obtain $P_{1-\underline{v}} < 1$. By Theorem 3

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

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

Footnote 11 claims that a group with beliefs greater than the prior revises up. To see this, just let $v = \max_{s \in \mathcal{S}} \{v_s\}$ for v_s as in the proof of Theorem 5, and $\overline{v} = \max\{v, a\}$ and note that since signals are equivocal, there is a positive mass of individuals with beliefs greater than $\overline{v}$, all of whom revise upward.

A similar argument shows there is also a group with beliefs less than the prior all of which revise downwards.

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

$$\begin{aligned}
\frac{pp_s ab\pi_{\mathrm{H}}(\sigma) + qq_s a(1-b)\pi_{\mathrm{L}}(\sigma)}{qq_s b\pi_{\mathrm{H}}(\sigma)(1-a) + pp_s(1-b)(1-a)\pi_{\mathrm{L}}(\sigma)} &> \frac{p_s ab\pi_{\mathrm{H}}(\sigma) + q_s a(1-b)\pi_{\mathrm{L}}(\sigma)}{q_s(1-a)b\pi_{\mathrm{H}}(\sigma) + p_s(1-b)(1-a)\pi_{\mathrm{L}}(\sigma)} \Leftrightarrow \\
\frac{pp_s b\pi_{\mathrm{H}}(\sigma) + qq_s(1-b)\pi_{\mathrm{L}}(\sigma)}{qq_s b\pi_{\mathrm{H}}(\sigma) + pp_s(1-b)\pi_{\mathrm{L}}(\sigma)} &> \frac{p_s b\pi_{\mathrm{H}}(\sigma) + q_s(1-b)\pi_{\mathrm{L}}(\sigma)}{q_s b\pi_{\mathrm{H}}(\sigma) + p_s(1-b)\pi_{\mathrm{L}}(\sigma)} \Leftrightarrow \\
b\pi_{\mathrm{H}}(\sigma) &> (1-b)\pi_{\mathrm{L}}(\sigma). \qquad (10)
\end{aligned}$$

We have

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

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

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

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 7. The prior over the eight states is

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

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

$$\begin{array}{lcc} & \text{posterior after } s \text{ and } \sigma \text{ proportional to} & \\ & T & F \\ Hh & afgw & (1-a)\,fgx \\ Lh & a(1-f)\,gw & (1-a)(1-f)\,gx \\ Hl & af(1-g)\,y & (1-a)\,f(1-g)\,z \\ Ll & a(1-f)(1-g)\,y & (1-a)(1-f)(1-g)\,z \end{array} \quad \& \quad \begin{array}{lcc} & \text{posterior after } s,c \text{ and } \sigma \text{ proportional to} & \\ & T & F \\ Hh & afgwp & (1-a)\,fgxq \\ Lh & a(1-f)\,gwr & (1-a)(1-f)\,gxt \\ Hl & af(1-g)\,yp & (1-a)\,f(1-g)\,zq \\ Ll & a(1-f)(1-g)\,yr & (1-a)(1-f)(1-g)\,zt \end{array}$$

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 8. Write j and i's initial beliefs as

$$\begin{array}{lcc} & \text{True} & \text{False} \\ \text{High} & \widetilde{a} & \widetilde{b} \\ \text{Low} & \widetilde{c} & \widetilde{d} \\ & j\text{'s beliefs} & \end{array} \qquad \begin{array}{lcc} & \text{True} & \text{False} \\ \text{High} & \overline{a} & \overline{b} \\ \text{Low} & \overline{c} & \overline{d} \\ & i\text{'s beliefs} & \end{array}$$

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 (2013) "Diverging Opinions," American Economic Journal: Microeconomics 2012, 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.

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

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?, Pschological Science, **7(**2), 115-120.

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

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

The Sentencing Project (2014), Race and Punishment: Racial Perceptions of Crime and Support for Punitive Policies.