When do populations polarize? An explanation.*

Jean-Pierre Benoît
London Business School
jpbenoit@london.edu

Juan Dubra
Universidad de Montevideo
dubraj@um.edu.uy

Revised April, 2018.

Abstract

Numerous experiments demonstrate *attitude polarization*. For instance, Lord, Ross & Lepper presented subjects with the same mixed evidence on the deterrent effect of the death penalty. Both believers and skeptics of its deterrent effect became more convinced of their views; that is, the population polarized. However, not all experiments find this attitude polarization. We propose a theory of rational updating that accounts for both the positive and negative experimental findings. This is in contrast to existing theories, which predict either too much or too little polarization.

Keywords: Attitude Polarization; Confirmation Bias; Bayesian Decision Making.
Journal of Economic Literature Classification Numbers: D11, D12, D81, D82, D83

In a classic study, Lord, Ross and Lepper (1979) took two groups of subjects, one of which believed in the deterrent effect of the death penalty and one of which doubted it, and presented them with the same mixed set of studies on the issue. Both groups became more convinced of their initial positions. Numerous, though by no means all, subsequent experiments, on a variety of issues, have also found that exposing groups of subjects who disagree to the same mixed evidence may cause their initial attitudes to move further apart

*We thank Gabriel Illanes and Oleg Rubanov for outstanding research assistance. We also thank Vijay Krishna, David Levine, Michael Mandler, Frederic Malherbe, Wolfgang Pesendorfer, Madan Pillutla, Debraj Ray, Jana Rodríguez-Hertz, Andrew Scott, and Stefan Thau for valuable comments. This is a substantially revised version of a paper that we circulated previously as "A theory of rational attitude polarization."

or *polarize*.[1] Many scholars have concluded that these results provide evidence that people process information in a biased manner, so as to support their pre-existing views. We argue that, on the contrary, this polarization of attitudes is exactly what we should expect to find in an unbiased Bayesian population in the context of the experiments that find polarization.[2]

There are two aspects to attitude polarization, which we term *pairwise polarization* and *population polarization*. Pairwise polarization occurs when the opinions of a particular pair of individuals move further apart after they receive a common piece of information. Population polarization occurs when this divergence is systematic, so that the opinions of the population on the whole diverge.

The economics literature has taken the view that there is, on the face of it, something puzzling about pairwise polarization and has examined the extent to which this polarization is consistent with Bayesian updating. The psychology literature, however, sees things quite differently. This literature finds nothing particularly intriguing about the polarization of two individuals. For the psychology literature, and we concur, it is population polarization, not pairwise polarization, that is the nub of attitude polarization.

Consider a report in BioNews (2015), describing a new finding on genome patterns that has led two teams of scientists to "opposing conclusions on the origins of Native Americans". The Berkeley team has concluded that the finding supports the theory that ancestral Native Americans entered from Siberia in a single wave and then split into two genetically distinct populations. The Harvard team has concluded that it supports the theory that there were two separate founding populations of the Americas. The fact that the same evidence has led to diverging theoretical conclusions is easily explained by the difficulty of interpreting complex data. Put differently, increased disagreement between two particular scientists, arising from the same piece of evidence, is hardly suggestive of biased reasoning. But, if it was also the case that all the scientists who initially favoured a single-wave theory interpreted the new finding as supporting a single-wave theory, while the scientists who initially favoured

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

[2] While we develop our ideas in a common priors rational setting, our reasoning is not restricted to rational agents. Full rationality merely provides a convenient benchmark of unbiased reasoning. For example, our theory can also be applied to unbiased subjects who are guilty of base rate neglect.

a two-group theory interpreted it as supporting a two-group theory, this would be suspect. Simply noting that a pair of scientists can legitimately update in different directions would not account for scientists systematically disagreeing in line with their prior beliefs. In fact, however, while the finding has reinforced the opinions of some scientists, it has reversed the opinions of others.[3]

A theory of rational pairwise polarization, which is what economists focusing on rational theories have explored, in and of itself neither explains population polarization when it occurs, nor predicts when it will occur. Indeed, for many psychologists, rather than pointing to the possibility of unbiased population polarization, the possibility of rational pairwise disagreement provides the mechanism that allows people to assimilate the evidence in a biased manner.

To illustrate this last point, take Plous' (1991) nuclear deterrence study. Plous begins by dividing his subjects into two groups, according to whether they entered the experiment with a belief that a strategy of nuclear deterrence makes the United States safer or less safe. He then gives all subjects the same article to read, describing an actual incident where an erroneous alert caused the United States to enter a heightened state of readiness for nuclear war with the Soviet Union. The crisis lasted only three minutes, as officials quickly realized the alert was a false alarm. After reading the article, the beliefs of subjects in each group move further in the direction of their initial inclinations.

How should unbiased subjects react to the article? As Plous writes, "Given the fact that (a) the system malfunctioned and (b) the United States did not go to war despite the malfunction, the question naturally arises as to whether this breakdown indicates that we are safer or less safe than previously assumed."

By the very nature of the article, some pairwise polarization is to be expected. The evidence in the article is equivocal – its implications depend on beliefs about an ancillary consideration, to wit, whether it is more important for a system's safety that it have a well-functioning primary unit or that it have effective safeguards. It is not at all clear which one is more important, and a person could legitimately believe either one is, depending upon his or her previous information on the matter. A person who believes that a well-functioning

[3]For instance, Professor David Reich, a geneticist at Harvard Medical School is quoted as saying "It's incredibly surprising. There's a strong working model in archaeology and genetics, of which I have been a proponent, that most Native Americans today extend from a single pulse of expansion south of the ice sheets – and that's wrong. We missed something very important in the original data."

primary unit is more important will revise downwards his belief in the safety of nuclear deterrence, while a person who believes sound safeguards matter more, will revise upwards. A fortiori, the fact that two particular subjects polarize – an opponent of nuclear deterrence becomes more opposed while a proponent becomes more in favour – is unproblematic.

However, even if people can legitimately update in different directions, a challenge remains. Why would it be that, on the whole, subjects who are in favour of nuclear deterrence respond positively to the evidence, while those who are opposed respond negatively? Put differently, why would it be that people who believe in the safety of nuclear deterrence also believe that safeguards are paramount, while people who are skeptical of nuclear deterrence also believe that primary units are crucial, rather than beliefs in these two dimensions being uncorrelated? If these beliefs were uncorrelated, while there would be many instances of pairwise polarization, there would be just as many instances of pairs converging; overall these instances would cancel each other out and the population would not polarize.

The fact that the *population* polarizes, not just isolated pairs, is what leads Plous to conclude that people process information in a biased manner to support their initial beliefs. He reasons that some people enter the experiment with a favourable view of nuclear deterrence, and a desire to enhance that view leads them to adopt the position that safeguards are dispositive. In this way, they justify revising upwards. Conversely for people who enter with an unfavourable view. In fact, Plous purposely chose the common evidence he presented to subjects to be mixed, in order to make such biased processing possible.

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

Plous tells us that most of the subjects in his experiment knew of the false alarm incident before entering the experiment, though, presumably, they did not know all of the details provided in the article. Which subjects would have entered the experiment with a favourable view of nuclear deterrence?

A reasonable presumption is that the subjects who entered with a favourable view, despite their knowledge of a previous malfunction that was caught by safeguards, are the ones that considered the reliability of safeguards to be more important than the reliability of the primary unit. These subjects would naturally tend to increase their belief that nuclear deterrence is safe after being given further evidence of properly functioning safeguards. On the other hand, subjects that considered a malfunction of the primary unit to be telling would have a negative view initially and would tend to revise downwards after being given

further evidence about a shaky primary unit. Population polarization is not only consistent with unbiased reasoning but even to be expected, at least in Plous' experiment.

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

The specific evidence that Lord, Ross, and Lepper provide to their subjects is two (supposed) studies, one that finds that murder rates tend to be lower in states that adopt the death penalty and one that finds that murder rates tend to be higher. Viewed as a single entity, the studies find that about half the time a state that adopts the death penalty subsequently has a lower murder rate and half the time a higher murder rate.

Why would some 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. 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 patterns in the murder rate. 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.

Darley and Gross (1983) is an influential study that uses a different methodology. We discuss how our model applies to it in sections 2.1 and 4.3. For now we note that, although this experiment is usually cited as providing strong evidence of biased reasoning, in fact it only finds polarization in 4 out of 8 instances.

Our general rationale for population polarization is as follows. Consider a group of people with differing opinions on an issue – the available evidence is mixed and has induced positive views in some people and negative views in others. Now suppose the group is exposed to some additional evidence and that this evidence is similar in nature to the previous body of evidence. Those who previously considered this type of information to be positive are

more likely to respond favourably than those who considered it to be negative, so that the population will polarize. While the basic intuition is simple enough, the complete argument is not quite so straightforward, as we will see.

As opposed to biased reasoning theories and previous Bayesian theories of pairwise polarization, our theory pays special attention to the interplay between prior information and new information. This allows us to make definite predictions on when we should expect evidence to cause the population to polarize and where polarization will be most marked. Current theories either explain too little – pairwise polarization but not population polarization – or too much – polarization whenever there is disagreement (and the new evidence is mixed).

We match the following findings in the experimental literature. (We discuss these findings in Section 1.3.)

1. Population polarization occurs when the new, common evidence presented to people is similar in nature to the previous information. If the common evidence that people are presented with is unfamiliar in nature, the population should not polarize. While some people may react positively to unfamiliar information and others react negatively, or neutrally, there is no reason for their reactions to correlate with their initial positions, since these positions were formed on a completely different basis. (See Theorem 2 and its corollary.)

2. A population of people that have largely based their initial opinions on very similar evidence on the issue will be especially prone to polarization, as they will have been especially 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 2.)

3. Groups with strong opinions polarize more. Thus, 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. (See Theorem 3.)

While, to some extent, our model was specifically designed to yield point 1, points 2 and 3 are predictions which the analysis yields. Outcomes in line with these predictions are often taken to be especially indicative of non-Bayesian thinking.

It is worth emphasizing the logic of attitude polarization experiments. These experiments recognize that field evidence is difficult to interpret, as it is hard to know if discrepancies in

people's beliefs reflect aspects of their reasoning processes or differences in the information they possess (for instance, different groups may read different newspapers). Moreover, the data on long run disagreement is mixed, at best. For instance, although partisan disagreement is often highlighted, Gerber and Green (1997) and Page and Shapiro (1992) examine long term survey data and conclude that attitudes of Democrats and Republicans in the United States move together.

Polarization experiments control the information that people receive in order to isolate a specific facet of reasoning. On the face of it, their focus is quite narrow. But if the experiments show, as they purport to, that people assimilate information in a biased way, then their implications are far-reaching, forcing us to re-evaluate our understanding of the way that people derive their beliefs. And although the experiments literally investigate one-shot belief updating, they have long run consequences. If people reason in a biased manner, giving them more and more information may not help resolve disputes – disagreements can be maintained forever.

In contrast, our common prior Bayesian analysis implies that beliefs in the population will eventually converge. Nonetheless, as we show in Section 1.4, even as beliefs converge, a population may continue to display polarization. Thus, our immediate focus is on the short run, for which the data is the most compelling, but our model also highlights some subtleties in the interpretation of long run data.

1 Formal Analysis

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

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

[4]For instance, the issue of interest could be the safety of nuclear power, the evidence on the issue data on accidents and near-accidents in nuclear power plants, and the ancillary matter the relative importance of primary units and safeguards. Or the issue could be the effectiveness of capital punishment, the evidence on the issue how murder rates vary, and the ancillary matter whether capital punishment is adopted for selection reasons or for political reasons.

about the issue that takes one of two values, say, true or false, and there is an ancillary matter that can be in one of two states, say, high or low.[5] 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.[6] 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 $1 > a, b > 0$. We denote the state space by $\Omega = \{H, L\} \times \{T, F\}$.

2. Each member of a large population 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 realization $s \in \mathcal{S}$ is

		True	False
likelihood of s :	High	p_s	q_s
	Low	r_s	t_s

   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 continuous density $\pi_{\mathrm{H}}(\cdot)$ with support $[0, 1]$ when the ancillary state is high, and from the continuous 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

[5] Section 1.2.2 introduces an additional ancillary state. We could also move beyond a binary issue, at the cost of added complexity.

[6] 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 such plants are safe. Or, learning that a particular policy has been adopted because of political reasons unrelated to selection issues says nothing about the effectiveness of that policy.

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 an **initial belief** over Ω given by $P(\cdot \mid s_i, \sigma_i)$, so that, for example, we call $P(T \mid s_i, \sigma_i)$ subject i's initial belief in the truth of the proposition.

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

		True	False
likelihood of c :	High	p_c	q_c
	Low	r_c	t_c

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

Subject i's **updated belief** is $P(\cdot \mid s_i, \sigma_i, c)$.

A special case occurs when the ancillary state is superfluous with respect to signal $m = s, c$, so that $p_m = r_m$ and $q_m = t_m$. Our theory depends on the ancillary matter being relevant for some signals. We now define two ways in which a signal and the ancillary matter may interact.

- The signal $m = s, c$ is **equivocal** if either i) $p_m > q_m$ and $r_m < t_m$, or ii) $p_m < q_m$ and $r_m > t_m$.

Consider, for instance, condition i). When the ancillary state is high, the equivocal signal m is more likely to occur if the proposition is true; when the ancillary state is low the signal is more likely to occur if the proposition is false. Thus, the ancillary matter directly affects the interpretation of an equivocal signal: the signal is good news in the high state and bad news in the low state. Up to now, Bayesian theories of pairwise polarization have focused on this condition, although it had not been defined formally. (see for example Andreoni and Mylovanov, 2012).

- The signal $m = s, c$ is **unbalanced** if either i) $\min\{p_m, q_m\} > \max\{r_m, t_m\}$, or ii) $\min\{r_m, t_m\} > \max\{p_m, q_m\}$.

Consider condition i). Whether or not the proposition is true, the signal is more likely to occur in the high state than the low state. Thus, the signal unambiguously tends to indicate that the ancillary state is high. We will see the role unbalancedness plays in the next section. Unlike equivocal signals, unbalanced signals have not been identified and have not received any attention in the prior literature.

1.1 Pairwise Polarization

This paper is primarily concerned with the conditions under which populations polarize. Of course, a pre-condition for a population to polarize is that it is possible for two individuals to polarize. Accordingly, the first part of our argument is that pairwise polarization is consistent with unbiased Bayesian updating. Suppose that individual A has a greater initial belief in the truth of the proposition than individual B has. The pair **polarizes** upon seeing a piece of information if A's belief increases and B's decreases.

Baliga et al. (2013) establish that when the ancillary matter is superfluous, it is impossible for a pair of individuals to polarize (although Section 4.1 gives a caveat). The next theorem provides a characterization of the conditions under which pairwise polarization can take place. Although other papers, including Walley (1991), Seidenfeld and Wasserman (1993), and Jern et al. (2014), have pointed to the possibility of pairwise polarization, there has not been any characterization theorem.

Theorem 1 *A common signal c can cause a pair of individuals to polarize if and only if c is either equivocal or unbalanced. Formally, there exist initial beliefs $P(\cdot \mid s_i, \sigma_i)$ and $P(\cdot \mid s_j, \sigma_j)$ 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.*

The literature to date has emphasized that equivocal signals may lead to pairwise polarization. The intuition for this result is immediate. Suppose that c is equivocal, with, say, $p_c > q_c$ and $r_c < t_c$. An individual with a strong belief that the ancillary state is high will consider the signal c to be good news and revise upwards while the opposite is true for an individual with a strong belief that the ancillary state is low (see Lemma 1 in the next section).

The fact that an unbalanced signal may cause a pair to polarize is a bit more subtle. Suppose the signal c is unequivocal but unbalanced so that, say, *a)* $p_c > q_c$ and $r_c > t_c$

and b) $\min\{p_c, q_c\} > \max\{r_c, t_c\}$. From condition a), the signal itself tends to make people revise their beliefs in the truth of the proposition upwards in both H and L states. From condition b), they also revise upwards their belief that the ancillary state is high, regardless of whether the proposition is true. A person who has come to associate the high state with the proposition being false may end up updating his or her belief in the proposition downwards.

The next example illustrates how equivocal and unbalanced signals that lead to polarization can arise in a simple, natural setup.

Example 1 *Consider the proposition "capital punishment is an effective deterrent". Suppose that, absent capital punishment, the murder rate would be expected to rise in* $\frac{1}{10}$ *of jurisdictions and fall in* $\frac{9}{10}$ *of them. Suppose further that if capital punishment is effective, it will reverse a rise with probability* $\frac{1}{10}$ *in the locales where it is implemented; if it is ineffective it will have no effect. The ancillary matter is whether jurisdictions that adopt capital punishment do so because they are especially likely to face an increase in the murder rate or because of unrelated political considerations. In the former case, which we call ancillary state* H*, murder would otherwise have risen with probability* $\frac{8}{10}$ *in the jurisdictions where capital punishment is adopted; in the latter case, ancillary state* L*, murder would otherwise have risen with (the baseline) probability* $\frac{1}{10}$ *in the districts that adopt. Everyone's prior is that there is a 50% chance that capital punishment is effective and an independent 50% chance that jurisdictions adopt capital punishment for selection reasons, rather than political reasons.*

Perhaps due to different schooling, people come to believe more or less strongly that selection reasons rather than political reasons determine capital punishment policies. In terms of our model, each individual i *receives a signal* σ_i *about the ancillary matter. At this point, individuals differ in their beliefs that the ancillary state is* H *but share the belief that there is a* 50% *chance that capital punishment is effective* (T)*. Suppose that individual* A *believes the ancillary state is* H *with a* 17.1% *chance and individual* B *believes* H *with a* 1.3% *chance. Other individuals assign greater or smaller chances.*

Now a study is made of what happens to the murder rate in two jurisdictions following the adoption of capital punishment. Using the above numbers, we compute the following likelihood matrices for the chances that crime will increase or decrease in any particular

jurisdiction:

rate increases		T	F
	H	$\frac{72}{100}$	$\frac{8}{10}$
	L	$\frac{9}{100}$	$\frac{1}{10}$

and rate decreases

		T	F
	H	$\frac{28}{100}$	$\frac{2}{10}$
	L	$\frac{91}{100}$	$\frac{9}{10}$

Suppose the study finds that the murder rate has increased in one of the jurisdictions and decreased in the other. The likelihood matrix for this signal is

$$\textit{likelihood of } c_{id}: \quad \begin{array}{c|cc} & T & F \\ H & \frac{2016}{5000} & \frac{16}{50} \\ L & \frac{816}{5000} & \frac{9}{50} \end{array} \tag{1}$$

Everyone in the population is made aware of this study. The updated beliefs of A and B are given by the following matrices

$$\begin{array}{c|cc} & T & F \\ H & 16.9\% & 13.4\% \\ L & 33.2\% & 36.5\% \end{array} \quad \textit{and} \quad \begin{array}{c|cc} & T & F \\ H & 1.6\% & 1.3\% \\ L & 46.3\% & 50.8\% \end{array} \tag{2}$$

A's beliefs *B's beliefs*

Thus, A assigns a 50.1% chance to T, while B assigns a 47.9% chance.

Now, the population participates in an experiment in which they are presented with the results of a second study of two different locales. Consider two possibilities for this second study.

1) The second study also finds that the murder rate has gone up in one jurisdiction and down in the other. This signal, c_{id}, has the likelihood matrix in expression (1) and is equivocal, as in state H it is more likely to happen if the proposition is true, while in state L it is more likely to happen if the proposition is false. Subject A, who has a relatively high belief in H, increases her belief in T to 53.0%, while subject B, who has a relatively low belief in H, decreases his belief to 46.2% chance. Thus, these two individuals polarize.

2) Suppose instead that the second study finds that the murder rate has gone up in both jurisdictions. This signal, c_{ii}, has the likelihood matrix

$$\textit{likelihood of } c_{ii}: \quad \begin{array}{c|cc} & T & F \\ H & \frac{5184}{10000} & \frac{64}{100} \\ L & \frac{81}{10000} & \frac{1}{100} \end{array}$$

This signal is unequivocal, being unambiguously more likely when the proposition is false. An individual who is sure the ancillary state is H will increase his belief in F, as will an individual who is sure that the ancillary state is L. At the same time, this signal is unbalanced, being always more likely when the ancillary state is high than when it is low. As a result, the signal causes A and B to increase the probabilities they assign to H. From (2), both A and B believe that T is more likely than F in state H. For individual A, but not for B, this countervailing force is great enough to undo the negative tendency of the signal, and she updates positively. Following c_{ii}, individual A assigns a 50.2% chance to T and B assigns a 47.7% chance. Although the signal is unequivocal, A and B polarize.

This example shows how two Bayesian individuals can polarize. Moving beyond A and B to the entire population, simple calculations show the following.

3) When the second signal is c_{id}, everyone with an initial belief in T greater than 48.5% revises upwards and everyone with an initial belief smaller than 48.5% revises downwards. Thus, the population as a whole moves apart, not just individuals A and B.

4) When the second signal is c_{ii}, subjects with a belief in T greater than 51% and subjects with a belief smaller than 47% both revise their beliefs downwards. Although individuals A and B polarize, it is not generally true that people with relatively high beliefs and relatively low beliefs move in opposite directions.

In this example, although an equivocal and an unbalanced signal both lead to pairwise polarization, only the equivocal signal leads the population to polarize. This difference is not special to the example – only an equivocal signal can lead to population polarization. While an unequivocal, but unbalanced, signal can cause two particular individuals to polarize by shifting the weights they attach to the ancillary state, individuals with strong beliefs in the value of the ancillary state, be it H or L, will not have these weights shift sufficiently for this effect to have bite.

Although it is necessary for population polarization that the common signal be equivocal, it is not sufficient; the common signal must also be "similar" to previous information that subjects have seen. The next section explores population polarization.

1.2 Population Polarization

Although there are numerous psychology papers on polarization, the literature has not always been careful in defining what is meant by the term. The basic idea is that people react to new

information in line with their prior beliefs, so that people who have a high belief in the truth of the proposition are more likely to revise upwards than people with a low initial belief. Consider a population in which some people revise upwards and some revise downwards, upon seeing a common signal. We say that the **population polarizes** if the proportion of people who revise upwards is a (non-constant) increasing function of their initial beliefs.

This is a relatively weak notion of polarization. A strong notion is that a **population polarizes completely around** v if, upon seeing a common signal, *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 both these sets are non-empty); the **population polarizes completely** if there is such a v. Of course, if a population polarizes completely, it polarizes.

If a population does not polarize completely, we might look at how (sub)groups of the population behave. We say that **groups with the strongest opinions polarize completely** if there is a $\overline{v}$ and a $\underline{v} > 1 - \overline{v}$ 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. This notion plays a special role as there is some evidence that polarization is more marked between sub-populations with the strongest opinions (see Section 1.3).

In Example 1, the equivocal signal c_{id} causes the population to polarize completely. The next example shows that an equivocal signal that leads to some pairwise polarization need not lead the population to polarize. Thus, an understanding of pairwise polarization is not sufficient to understand population dynamics.

Example 2 *Each of the four states in* $\Omega = \{H, L\} \times \{T, F\}$ *has a prior probability of* $\frac{1}{4}$. *Individuals first observe a signal* h *or* l *about the ancillary state that is correct with probability* $\frac{2}{3}$, *so that* $P(h \mid H) = P(l \mid L) = \frac{2}{3}$.[7] *They then observe one of two signals about the issue,* t *or* f, *with* $P(t \mid T) = P(f \mid F) = \frac{2}{3}$. *This partitions the population into four groups, with initial beliefs: i)* $P(T \mid h, t) = .67$, *ii)* $P(T \mid h, f) = .33$, *iii)* $P(T \mid l, t) = .67$, *and iv)* $P(T \mid l, f) = .33$. *For concreteness, suppose the actual state of the world is* (H, T). *With a large population, group i) makes up* $\frac{4}{9}$ *of the population, group ii) makes up* $\frac{2}{9}$ *of the population, group iii) makes up* $\frac{2}{9}$ *of the population, and group iv) makes up* $\frac{1}{9}$ *of the*

[7] To keep the example simple, we deviate from the model and use a binary signal about the ancillary matter. This feature is not essential.

population.

Now, everyone is presented with a common equivocal signal c with likelihood matrix

$$\begin{array}{cc} T & F \\ H \quad \frac{1}{3} & \frac{1}{6} \\ L \quad \frac{1}{6} & \frac{1}{3} \end{array}$$

Posteriors for the four groups are i) $P(T \mid c, h, t) = .71$*, ii)* $P(T \mid c, h, f) = .38$*, iii)* $P(T \mid c, l, t) = .61$*, and iv)* $P(T \mid c, l, f) = .29$*. Although an individual from group i) and an individual from group iv) polarize, the population does not polarize. In fact, there is no difference at all in how groups with high beliefs and low beliefs update: two thirds of people with a high initial belief in T,* .67*, revise upwards and two thirds of people with a low individual belief,* .33*, revise upwards. Put differently, the level of initial beliefs is unrelated to whether these beliefs are revised up or down.*

To understand what goes wrong in the previous example, we now lay out the basic mechanism for population polarization. We can think of the population polarizing upon seeing an equivocal signal as a two-step process. The first step is that individuals with a large belief that the ancillary state is high revise in the opposite direction than individuals with a small belief that the state is high. That is the content of the following lemma.

Lemma 1 *Let P be a belief over* Ω *that assigns strictly positive probability to every state. If c is equivocal there exists an h such that,*

$$\begin{aligned} P(H \mid \sigma) \quad &> \quad h \Rightarrow (p_c - q_c)\left(P(T \mid c, \sigma) - P(T \mid \sigma)\right) > 0 \\ P(H \mid s, \sigma) \quad &< \quad h \Rightarrow (p_c - q_c)\left(P(T \mid c, \sigma) - P(T \mid \sigma)\right) < 0. \end{aligned}$$

For concreteness, suppose that $p_c > q_c$ and $r_c < t_c$. Then agents with a large belief in H revise upwards upon seeing c, while agents with a small belief revise downwards. For the population to polarize, rather than just some individuals, a second step is needed. It must be that, systematically, agents with a large belief that the ancillary state is high also have a large initial belief in the truth of the proposition and conversely. But why would this be the case? For many psychologists, biased reasoning is the answer: people with a large belief in the proposition "decide" that the ancillary state is (probably) high in order to maintain their belief.

However, there is no need to invoke bias. There is a Bayesian reason the equivocal signal leads to population polarization in Example 1 but not in Example 2. In Example 1, the information on the issue that subjects have seen previous to the experiment is equivocal in the same way as the common information that is given to them (in fact, the signals are identical). Because of this, people who enter the experiment with a relatively large belief in T are those who have a relatively large belief in H. When presented with another equivocal signal for which $p > q$ they revise upwards. Conversely for people with a large belief in F. When the equivocal common signal and previous information are similar, a correlation is created between beliefs in H and T, which is necessary for the population to polarize.

In contrast, in Example 2 the previous information on which subjects based their beliefs has little connection to the common signal. Although some individuals react positively and some react negatively to the equivocal common signal, these differences are unrelated to their initial beliefs. Beliefs in the proposition and the ancillary matter are uncorrelated and pairwise polarization does not lead to population polarization.

For the population to polarize upon seeing the equivocal signal c with $p_c > q_c$, it must be that the previous information is more indicative of T when the ancillary state is H than when it is L. This will be true of a previous signal s if s is also equivocal with $p_s > q_s$. But it will also be true under the weaker condition that $\frac{p_s}{q_s} > \frac{r_s}{t_s}$. A parallel condition holds if c is equivocal with $p_c < q_c$, leading to the following definition.

- Signals s and c are **similar** if $(p_s t_s - q_s r_s)(p_c t_c - q_c r_c) > 0$

1.2.1 Theorems

In an attitude polarization experiment, subjects enter an experiment with previous information about an issue and are given some additional "mixed" evidence on it. In many of these experiments, this common piece of evidence is explicitly chosen to be typical of pre-existing information about the issue. In our terms, subjects are given an *equivocal* signal that is *similar* to previous signals that subjects have seen.

Consider an issue on which various researchers have carried out studies. Each study provides a signal about the issue. Let $\bar{s}$ be the signal that is the composition of all these signals. The signal $\bar{s}$ represents the *body of knowledge* about the issue. We define an **expert** as someone who knows $\bar{s}$. 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 formalizes this intuition.

Theorem 2 *Consider a population of experts who have all observed a signal $\bar{s}$ and then observe a common signal c, with $p_c \neq q_c$, $r_c \neq t_c$. The population polarizes completely if and only if c is equivocal and $\bar{s}$ is similar to c. Formally, there is a v such that*

$$P(T \mid \bar{s}, \sigma) > v \Rightarrow P(T \mid c, \bar{s}, \sigma) > P(T \mid \bar{s}, \sigma) \tag{3}$$

$$P(T \mid \bar{s}, \sigma) < v \Rightarrow P(T \mid c, \bar{s}, \sigma) < P(T \mid \bar{s}, \sigma) \tag{4}$$

and the antecedents are non-empty, if and only if c is equivocal and $\bar{s}$ is similar to c.

If c and $\bar{s}$ are not similar, or if c is not equivocal, then the population does not polarize – either i) everyone revises in the same direction, or ii) there are some groups with low beliefs and some groups with high beliefs who all revise in the same direction, or iii) high belief groups and low belief groups move towards each other.

Although this theorem is stated for experts, it applies to any population that enters the experiment having seen more or less the same equivocal evidence on an issue. The assumption of expertise provides one reason that individuals would all have seen the same evidence on the issue.

From the theorem, there is a level of belief v such that everyone with a belief in the truth of the proposition greater than v revises upwards and everyone with a belief lower revises downwards. Of course, an experiment will be "noisy" so that we would not expect to find such a perfect separation in practice. Moreover, the level v need not correspond to the 'dividing line' in beliefs around which an experimenter checks for polarization. In practice, there will be a range of $\widetilde{v}$'s for which most people with belief greater than $\widetilde{v}$ revise upwards and most people with belief smaller than $\widetilde{v}$ revise downwards.

Theorem 2 concerns a population of subjects with congruent levels of expertise. In many situations, subjects will have disparate degrees of expertise. While some subjects will be well acquainted with the literature, others will have only a superficial 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 be aware of the general tenor of the existing evidence. The following theorem says that in a population in which the common signal is similar to previous evidence that people have seen, which may vary from individual to individual, groups with the strongest opinions polarize.

Theorem 3 *Suppose each individual i has observed a signal s_i and everyone then observes a common signal c with $p_c \neq q_c$, $r_c \neq t_c$. Groups with the strongest opinions polarize completely if and only if c is equivocal and for all i, s_i is similar to c. Formally, if c is equivocal and for all i, s_i is similar to c there exist $\overline{v}$ and $\underline{v} > 1 - \overline{v}$ such that*

$$P(T \mid s_i, \sigma) > \overline{v} \Rightarrow P(T \mid c, s_i, \sigma) > P(T \mid s_i, \sigma) \tag{5}$$

$$P(T \mid s_i, \sigma) < 1 - \underline{v} \Rightarrow P(T \mid c, s_i, \sigma) < P(T \mid s_i, \sigma) \tag{6}$$

and the antecedents are non-empty. Conversely, if there exist $\overline{v}$ and $\underline{v} > 1 - \overline{v}$ such that (5) and (6) hold non-trivially, c must be equivocal, and for every s_i such that the antecedents in (5) and (6) hold, s_i must be similar to c.

Suppose that $p_c > q_c$. If everyone has previously seen evidence that is similar to c, then the groups with the strongest belief in T will be those with the strongest belief in H. Individuals in these groups will all respond positively to the equivocal signal. Conversely for groups with strong beliefs in F.

When groups with the strongest opinions polarize, there will be a range of $\bar{w}$'s and $\underline{w}$'s such that most people who believe the proposition with probability greater than $\bar{w}$ increase their beliefs, while most people who disbelieve the proposition with probability greater than $\underline{w}$ increase their disbelief. However, if the various prior pieces of information that individuals have seen on the issue are sufficiently variegated and the ancillary matter is sufficiently unimportant, the population as a whole may not polarize: It is possible to construct examples where the fraction of the population that revises upwards is not a monotonically increasing function of initial beliefs, even if all prior signals are similar to a common equivocal signal (see Section 4.2). On the other hand, when all the signals have likelihood matrices that

are symmetric along both diagonals – so that results are not being pushed in any particular direction – the population polarizes completely.

Theorem 4 *Suppose that each person's private signal s_i about the issue and the common signal c have symmetric likelihood matrices. The population polarizes completely if and only if c is equivocal and every s_i is similar to c. In particular, the population polarizes completely around the prior belief $P(T) = a$. Formally, for $a = P(T)$*

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

and the antecedents are non-empty, if and only if c is equivocal and similar to each s_i.

1.2.2 Unfamiliar Evidence

Previous theories, both bias theories and rational theories, have emphasized the role played by the equivocal nature of the common evidence. Our theory adds that this evidence must be similar to previous evidence. Hence, the population will not polarize when the common evidence is equivocal but is affected by a different ancillary matter than the previous information. This will be the case, in particular, when the new evidence is of an unfamiliar nature.

Recall, for example, our argument that in a population of people that have (largely) derived their beliefs on nuclear deterrence from their knowledge of near-miss episodes, proponents of nuclear deterrence will tend to be people who believe that safeguards are critical and conversely for opponents. As a result, when the population is presented with further evidence of reliable backups, proponents will be more likely to revise upwards than opponents and the population will polarize.

Now suppose that instead of being given evidence on a narrow miss, or some other evidence related to primary systems and backups, this population is presented with the following information:

> i) Numerous experiments have found that people are very good at evaluating risks and rewards and will not take undue chances. A strategy of nuclear deterrence makes the United States safer because other countries will avoid actions that could provoke a nuclear reply.

ii) Neurological research has shown that people react with the emotional part of their brain when confronted with extreme threats, making their actions unpredictable. Because of this, a strategy of nuclear deterrence is risky.

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

In order to formalize this reasoning, we need to expand our model slightly. In addition to an ancillary matter with states that take the values H or L, suppose there is a second matter with states that take the values h or l.[8] Nature chooses one of the states H or L with probabilities b and $1-b$ and, independently, one of the states h or l with probabilities d and $1-d$. Individuals enter the experiment having seen a signal about the issue and a signal $\sigma = (\sigma_1, \sigma_2)$, where σ_1 varies with states H, L and σ_2 varies with states h, l, and draws of σ_1 and σ_2 are independent. (With respect to nuclear deterrence, H and L could correspond to whether backup units or primary units are more important, while h and l could correspond to whether experimental or neurological evidence is more compelling.)

Let s be the previous information a subject has seen. We say that c is **unfamiliar** if c varies with a different ancillary matter than s. That is, if c is unfamiliar, then we can write the likelihood matrices of s and c as

$$\text{likelihood of } s: \begin{array}{c|cc} & T & F \\ Hh & p_s & q_s \\ Lh & p_s & q_s \\ Hl & r_s & t_s \\ Ll & r_s & t_s \end{array} \quad ; \text{ and likelihood of } c: \begin{array}{c|cc} & T & F \\ Hh & p_c & q_c \\ Lh & r_c & t_c \\ Hl & p_c & q_c \\ Ll & r_c & t_c \end{array}$$

The next result, which follows from Theorem 2, shows that unfamiliar evidence does not lead to population polarization.

[8] All our previous results can be adapted to this setting.

Corollary 1 (to Theorem 2) *A population of experts presented with unfamiliar evidence does not polarize.*

When presented with mixed but unfamiliar evidence, there may be pairs of subjects that polarize but they will be counterbalanced by subjects whose opinions move closer together. Similarly, groups with the strongest opinions will not polarize when presented with unfamiliar evidence, and the population will not polarize even if the signal is symmetric.

1.3 Empirical Support

In this section, we consider some empirical support for our theory in existing experiments. The strength of this evidence should not be overstated, as the experiments were not designed as tests of our theory.

Lemma 1 says that when $p_c > q_c$, people with a large belief that the ancillary state is H should revise upwards and conversely. 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. 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 the lemma, as he asks his subjects which is more important, the fact that safeguards worked or the fact that a breakdown occurred and, as predicted by the lemma, 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.

Theorem 2, on experts, is in line with Plous' finding that subjects with high issue involvement display a large degree of polarization, if we accept that "high issue involvement" suggests a good knowledge of the current body of evidence.

Theorem 3 says that groups with the strongest opinions polarize. On their capital punishment experiment dealing with reported attitude change, Miller et al. (1993) find the most polarization among subjects with the strongest beliefs. Plous (1991) reports that subjects with strong convictions polarize the most. Moreover, many experiments, including Lord, Ross and Lepper (1979), pre-select people with strong opinions. On the other hand, Kuhn and Lao (1996) do not find an effect of strength of opinion.

According to Corollary 1, subjects should not polarize when presented with unfamiliar evidence. Miller et al. (1993) carry out several studies. In a capital punishment study they find population polarization but no such polarization on an affirmative action study. More precisely, on the affirmative action study subjects whose attitudes polarize are offset by subjects whose attitudes depolarize. 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" when presented with unfamiliar arguments and revise upwards or downwards in a pattern inconsistent with biased assimilation. Nevertheless, that is what is predicted by our model. Munro and Ditto (1997) present subjects with equivocal and (arguably) unfamiliar information on stereotypes pertaining to homosexuals. They find no population polarization in their Experiment 1 but polarization in their Experiment 2.

1.4 Longer term implications

Suppose that individuals repeatedly receive (conditionally) independent signals. Standard results imply that, in our common prior Bayesian setup, the beliefs of the population will eventually converge. Despite this convergence, the data may continually display polarization. We demonstrate this possibility using the capital punishment example 1 from Section 1.1.

Recall that in the example, there is a signal space on the issue that consists of two possible signals: a finding that the murder rate has increased and a finding that it has gone down. Call this signal space $\mathcal{C}$. Initially, everyone has seen a study consisting of two draws from $\mathcal{C}$, one showing an increase in the murder rate and one showing a decrease, as well as private signals on the ancillary matter.

We extend the example through time by considering what happens as individuals receive more and more independent draws from $\mathcal{C}$. For concreteness, suppose the actual state of the world is HT. In the limit, i.i.d. draws will show that the murder rate has risen 72% of the time.

Let c^∞ be an infinite sequence of i.i.d draws from $\mathcal{C}$ and c^t be the first t draws. We have the following:

1. *Eventually almost everyone agrees that the proposition is true and the ancillary state is high.* Formally, for any σ, $P\{c^\infty : \lim_{t\to\infty} P(HT \mid c^t, \bar{s}, \sigma) = 1\} = 1$.

2. *While more and more people revise upwards, at any point in time an equivocal signal causes groups with strong opinions to polarize completely.* Formally, for all t, there exist $\overline{v}_t$ and $\underline{v}_t > 1 - \overline{v}_t$ such that

$$\begin{aligned} P\left(T \mid c^t, \overline{s}, \sigma\right) &> \overline{v}_t \Rightarrow P\left(T \mid c_{id}, c^t, \overline{s}, \sigma\right) > P\left(T \mid c^t, \overline{s}, \sigma\right) \\ P\left(T \mid c^t, \overline{s}, \sigma\right) &< 1 - \underline{v} \Rightarrow P\left(T \mid c_{id}, c^t, \overline{s}, \sigma\right) < P\left(T \mid c^t, \overline{s}, \sigma\right) \end{aligned}$$

3. *An equivocal signal causes the population to "almost" polarize completely infinitely often (although the bulk of the population revises in the same direction.)* Formally, for all ε and t', there exists a $t \geq t'$, a set of (sequences of) signals $D \subset \{c_i, c_d\}^\infty$ with $\Pr(D) \geq 1 - \varepsilon$, and a v such that for all $c^\infty \in D$

$$\begin{aligned} P\left(T \mid c^t, \overline{s}, \sigma\right) &> v + \varepsilon \Rightarrow P\left(T \mid c_{id}, c^t, \overline{s}, \sigma\right) > P\left(T \mid c^t, \overline{s}, \sigma\right) \\ P\left(T \mid c^t, \overline{s}, \sigma\right) &< v - \varepsilon \Rightarrow P\left(T \mid c_{id}, c^t, \overline{s}, \sigma\right) < P\left(T \mid c^t, \overline{s}, \sigma\right) \end{aligned}$$

2 Related Literature

Walley (1991), Seidenfeld and Wasserman (1993), Andreoni and Mylovanov (2012), and Jern, Chang and Kemp (2014) argue that two individuals can polarize in a standard, rational setting, such as ours, if there is an ancillary matter (to put their result in our terms). Seidenfeld and Wasserman give conditions for which, for a given set of prior beliefs, and a distribution over signals, the signals are such that the maximum of beliefs increases while the minimum of beliefs decreases. Andreoni and Mylovanov provide a model where two individuals polarize after receiving a common signal c but they do not give a characterization of the properties that the likelihood of c must have in order for that to happen. Jern et al. provide examples of which Bayesian networks can generate polarization and which ones cannot. None of these papers address the question of when populations polarize.[9]

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

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

One of the clearest statements on polarization is found in Baliga, Hanany, and Klibanoff (2013), who are interested in the question of when two individuals can polarize. They let an issue take on many possible values and interpret a rise in a subject's response to indicate a first order stochastic dominance shift upwards in her beliefs and correspondingly for a fall in response. They first establish that, in a standard rational setting, if there is no ancillary matter (again, in our terms), then two individuals whose beliefs have common support cannot polarize. (Nevertheless, there is a sense in which polarization *in an fosd sense* can occur even without an ancillary state, as we show in Section 4.1 in the Appendix.) They go on to argue that ambiguity aversion can explain pairwise polarization.

Rabin and Schrag (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. Loh and Phelan (2017) provide conditions for when long run polarization can occur, and when it cannot, when individuals do not store the full distribution over the multidimensional state space, but only the marginals over each dimension. All these papers can be interpreted as showing population polarization as well as pairwise polarization, in non-standard settings. None of them make the distinction that we make between the types of information that should and should not produce polarization and, in fact, often predict polarization whenever there is disagreement. The non-standard settings lead to the possibility of disagreement in the long run and, except for Baliga et al. (2013), these papers address this question.

Many experiments that find attitude polarization also find *biased assimilation* – subjects on either side of an issue both reporting that evidence that confirms their view is more credible than contrary evidence. As Lord, Ross and Lepper observe, this asymmetric assimilation in and of itself is not problematic, as it may be rational for a person to have greater confidence in a finding that confirms something she believes than a finding that disconfirms her belief. Gerber and Green (1999) show formally that biased assimilation can arise in a

Bayesian model with normal signals, though their model does not allow for unbiased individuals to polarize. In a similar setting, Bullock (2009) shows that two unbiased individuals can polarize if they are estimating a parameter whose value is changing over time.

At a broader level, our paper is related to the literatures on confirmation bias and cognitive dissonance, which provide bias explanations for attitude polarization.

2.1 Further considerations on the literature

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

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

Darley and Gross (1983) provide subjects with descriptions of a fourth-grade girl. Half the subjects are given information strongly suggesting that the girl comes from an upper class background and half are given information suggesting that she comes from a lower class background – information that could potentially have a biasing effect on the way subjects process subsequent information. At that point the subjects are asked for their opinions of the girl's abilities on three academic subjects – liberal arts, reading, and mathematics – and of her disposition on five traits – work habits, motivation, sociability, maturity, and cognitive skills. Subjects who believe that the girl comes from a well-off family tend to rate her slightly higher than those who believe she comes from a poorer family. Next, subjects are provided with some specific evidence about her abilities. This evidence is the same for

all the subjects, who are then again asked to rate her.[10] The subjects beliefs polarize on four out of the eight questions, including the three academic subjects.

Although this experiment is typically touted as one that demonstrates polarization, this is hardly an overwhelming finding of polarization. Somewhat bizarrely, almost all the papers that cite Darley and Gross do not even mention the questions on which subjects do not polarize.[11] In fairness to Darley and Gross, they put their data through various tests to reach their conclusions of bias and it is beyond the scope of this paper to consider the merits of all their arguments. Nonetheless, at the very least, their conclusion that they have found evidence of polarization is open to doubt. We consider the paper in greater detail in Section 4.3.

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

[10] 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.

[11] 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.

it makes the case for motivated reasoning here less than clear.

Nyhan and Reifler (2010) report on an extreme form of polarization, a so-called backfire effect. As they describe it, they give subjects articles to read that contain either a misleading statement by a politician or the misleading statement together with an independent correction and, rather than offsetting the misleading statement, the correction *backfires*, causing partisans to believe the statement even more.

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

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

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

To us, the so-called correction is far from a straightforward repudiation. First of all, it acknowledges that, at some point in time, Hussein did posses weapons of mass destruction, in the form of chemical weapons. It rather vaguely asserts that he made a decision to destroy stockpiles of chemical weapons, without asserting that he followed up on the decision. It

goes on to say that inspectors destroyed the nuclear program sometime after 1991. But how difficult would it have been for Hussein to have hidden some weapons from the inspectors? The question asks if Iraq had "the ability to produce these weapons". Even if stockpiles of chemicals were destroyed, would that eliminate a country's ability to produce more?

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

> **Correction**: In 2004, the Central Intelligence Agency released a report that concludes that Saddam Hussein did not possess stockpiles of illicit weapons at the time of the U.S. invasion in March 2003, nor was any program to produce them under way at the time.
>
> **Question**: Immediately before the U.S. invasion, Iraq had an active weapons of mass destruction program and large stockpiles of WMD – Strongly disagree, Somewhat disagree, Neither agree nor disagree.

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

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

3 Conclusion

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

Although we have presented our theory as a positive description of reasoning processes, another interpretation is that we have presented a benchmark model of rationality. Our theory shows that existing findings on attitude polarization do not, by themselves, point to non-Bayesian reasoning.

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

4 Appendix

4.1 Polarization without an ancillary state

Our model is *fully* rational – not only do subjects understand the signalling structure perfectly, the experimenter also (implicitly) understands the structure and asks questions in line with the structure. To see this implications of this, consider the issue of how safe nuclear energy is. Let us say its safety can be described by a parameter that takes on the values 1, 2, 3, or 4 (for instance, 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 consists of one of four signals with likelihoods:

$$\begin{array}{rccccc} & S_A \downarrow \Theta \rightarrow & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ & s_1 & \frac{3}{4} & \frac{1}{4} & 0 & 0 \\ \text{likelihoods of signals } s_1, ..., s_4: & s_2 & \frac{1}{8} & \frac{1}{2} & \frac{1}{4} & \frac{1}{8} \\ & s_3 & \frac{1}{8} & \frac{1}{4} & \frac{1}{2} & \frac{1}{8} \\ & s_4 & 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{array}$$

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

$$\begin{array}{rccccc} & & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ \text{posteriors after } s_2, s_3: & I: p(\cdot \mid s_2) & \frac{1}{8} & \frac{1}{2} & \frac{1}{4} & \frac{1}{8} \\ & II: p(\cdot \mid s_3) & \frac{1}{8} & \frac{1}{4} & \frac{1}{2} & \frac{1}{8} \end{array} \tag{7}$$

However, the experimenter does not ask subjects for their beliefs over the four point scale. Instead, the experimenter asks them 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, and dangerous if it rates a 1 or 2. Subject I believes nuclear energy is safe with probability $\frac{1}{4} + \frac{1}{8} = \frac{3}{8}$; subject II believes it is safe with probability $\frac{1}{2} + \frac{1}{8} = \frac{5}{8}$.

Now I and II are shown the common signal c with likelihoods

$$\begin{array}{rccccc} & & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ \text{likelihood of } c: & c & 0 & 1 & 1 & 0 \end{array}$$

Posterior beliefs are

$$\begin{array}{rccccc} & & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ \text{posteriors after signal } c: & I: p(\cdot \mid s_2,c) & 0 & \frac{2}{3} & \frac{1}{3} & 0 \\ & II: p(\cdot \mid s_3,c) & 0 & \frac{1}{3} & \frac{2}{3} & 0 \end{array}$$

Subject I's belief in the safety of nuclear energy decreases to $\frac{1}{3}$ while II's belief increases to $\frac{2}{3}$. Thus, the pair polarize.

Note that there is no ancillary matter here. Or, equivalently, there may be an ancillary matter that is superfluous and does not affect the likelihoods. In any case, the common signal is neither equivocal not unbalanced. Nevertheless, polarization can arise here because the experimenter asks a question that is not properly aligned with the signalling structure.

This example also applies to Baliga et al.'s (2013) result on no pairwise 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, this question is, perhaps inevitably, only a function of the underlying parameters. As a consequence, conditional on "safe" or "unsafe", individuals disagree on the likelihood of the "c" signal and there is polarization. This could also be seen as a foundation for the assumption in Acemoglu et al. (2009) that individuals disagree on the likelihoods of signals.

4.2 Only similar signals

The following example shows that the population may not polarize even if all previous signals are similar to the common signal.

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 3, 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, the population does not polarize. That is, the fraction of those revising upwards is not an increasing function of initial beliefs. To see this note that for $\sigma's$ such that $\frac{\pi_H(\sigma)}{\pi_L(\sigma)} \in [0.9, 1.1]$, which form the bulk of σ's, $P(T \mid s_1, \sigma) < v = \frac{1}{2} < P(T \mid s_2, \sigma)$. So that, for polarization, the chances that individuals who observed s_2 increase their beliefs should be *larger* than for those who observed s_1. But $P(T \mid c, s_1, \sigma) > P(T \mid s_1, \sigma)$ 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 $\frac{\pi_H(\sigma)}{\pi_L(\sigma)} > 1.0$. That is, the

likelihood of σ signals which lead to increases in the belief of T when subjects have observed s_2 is *smaller* than the likelihood of σ signals which lead to increases in the belief of T when subjects have observed s_1.

There are three particular features of this counter-example:

1. Although there is an ancillary matter, its importance is minimal as the large bulk of subjects have very similar beliefs about the ancillary state.

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 in itself is neither good news nor bad news for the proposition.

While these three points are each important separately, Theorem 4 addresses 2) and 3) together, by considering only *symmetric* signals.

4.3 Hannah revisited

Recall Darley and Gross (1983)'s experiment discussed in Section 2.1. Half the subjects were given information indicating that a girl named Hannah came from an upper class background and half information indicating that she came from a lower class background. At this point, they were asked to evaluate Hannah in eight domains. The subjects were then shown a video of her engaged in various tasks, and were again asked to evaluate her. The responses of the two groups of subjects polarized in four out of the eight domains. Although we do not consider this to be a strong finding of polarization, some might argue that it is still a finding of polarization. Either way, the experiment does not provide a test of our theory.

To see this, note that the different groups of subjects are effectively asked about two different girls, a rich one and a poor one, and the same behaviour could well have different implications for children from different demographics. For instance, subjects could believe that a child that attends a rich school will perform well on national tests provided that she is able to concentrate moderately well while a child that attends a poor school will

perform well only if she has exceptional concentration skills. Then, evidence that Hannah concentrates moderately well would be good news for *rich Hannah* but bad news for *poor Hannah.* Thus, a finding of attitude polarization would be consistent with our model, with the child's background being the ancillary matter. On the other hand, a strong finding of polarization is not particularly predicted by our model, as people were not pre-sorted according to their beliefs. Hence, the results of this experiment say little about our theory.

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

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

Let us put some numbers to this notion of base rates and a higher standard of evidence. Suppose that subjects think that, nationwide, a fourth grade student attending a school with poor resources is likely to be operating at a level of 3.5, while a student attending a wealthy school is likely to be operating at a level of 4.5. However, there is a 35% chance that any child is exceptional, that is, exceptionally bad or exceptionally good, and subjects require 75% certitude to make a judgement of an individual member of a demographic class.[12] Since the

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

75% standard has not been met, initially everyone reports that Hannah is operating at a level of 4. Now subjects are shown a video of Hannah, answering questions among other things. By design, the video clearly establishes one thing about Hannah: she is not exceptional. The required standard of evidence is now met and subjects' responses polarize to 3.5 and 4.5, the levels for the two types of schools. We have obtained unbiased population polarization by modelling Darley and Gross' own words, although not in the way they themselves would choose to model them.

4.4 Proofs

Proof of Theorem 1. Write j and i's initial beliefs as

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

For i, we have

$$\begin{aligned} P(T \mid c, s_i, \sigma_i) - P(T \mid s_i, \sigma_i) &= \frac{p_c\overline{a} + r_c\overline{c}}{p_c\overline{a} + q_c\overline{b} + r_c\overline{c} + t_c\overline{d}} - \frac{\overline{a} + \overline{c}}{\overline{a} + \overline{b} + \overline{c} + \overline{d}} > 0 \Leftrightarrow \\ 0 &< \frac{\overline{a}\overline{b}p_c - \overline{a}\overline{b}q_c + \overline{a}\overline{d}p_c - \overline{b}\overline{c}q_c + \overline{b}\overline{c}r_c - \overline{a}\overline{d}t_c + \overline{c}\overline{d}r_c - \overline{c}\overline{d}t_c}{\left(\overline{a}p_c + \overline{b}q_c + \overline{c}r_c + \overline{d}t_c\right)\left(\overline{a} + \overline{b} + \overline{c} + \overline{d}\right)} \Leftrightarrow \\ 0 &< \overline{a}\overline{b}(p_c - q_c) + \overline{a}\overline{d}(p_c - t_c) + \overline{b}\overline{c}(r_c - q_c) + \overline{c}\overline{d}(r_c - t_c). \quad (8) \end{aligned}$$

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 (8) 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 (8) 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 (8), 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 (8) 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. ■

We will use repeatedly that for P a belief over Ω that assigns strictly positive probability to every state $P(T \mid c, \sigma) - P(T \mid \sigma)$ has the same sign as

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

which, letting $g(\sigma) = \frac{\pi_{\mathrm{H}}(\sigma)}{\pi_{\mathrm{L}}(\sigma)}$ has the same sign as

$$M(g) \equiv [p_c P(\mathrm{TH})\, g + r_c P(\mathrm{TL})]\,[P(\mathrm{FH})\, g + P(\mathrm{FL})] - [P(\mathrm{TH})\, g + P(\mathrm{TL})]\,[q_c P(\mathrm{FH})\, g + t_c P(\mathrm{FL})]\,. \tag{9}$$

This is a parabola in g, and the coefficient in g^2 is $P(\text{FH})\,P(\text{TH})\,(p_c - q_c)$, while the intercept is $P(\text{TL})\,P(\text{FL})\,(r_c - t_c)$.

Also, for any P, we have

$$P(\text{H} \mid \sigma) = \frac{\pi_{\text{H}}(\sigma)\,(P(T\text{H}) + P(F\text{H}))}{\pi_{\text{H}}(\sigma)\,(P(T\text{H}) + P(F\text{H})) + \pi_{\text{L}}(\sigma)\,(P(T\text{L}) + P(F\text{L}))} = \frac{1}{1 + \frac{\pi_{\text{L}}(\sigma)}{\pi_{\text{H}}(\sigma)} \frac{P(T\text{L}) + P(F\text{L})}{P(T\text{H}) + P(F\text{H})}}$$

which is strictly increasing in σ.

Proof of Lemma 1. Assume without loss of generality that $p_c > q_c$ (the case of $p_c < q_c$ is symmetric and omitted). Then the intercept in the parabola (9) is $P(\text{TL})\,P(\text{FL})\,(r_c - t_c) < 0$ and the coefficient on g^2 is greater than 0. So among $g > 0$, there exists a unique g, and hence a unique $\sigma_P \in (0,1)$, such that $M(g(\sigma_P)) = 0$, and $sgn\left[P(T \mid c, \sigma) - P(T \mid \sigma)\right] = sgn\left[M(g(\sigma)) - M(g(\sigma_P))\right] = sgn\left[\sigma - \sigma_P\right]$. The proof is complete by setting $h_s = P(\text{H} \mid s, \sigma_P)$, because $P(\text{H} \mid s, \sigma)$ is strictly increasing in σ. ■

Lemma 2 *Let P (in general $P(\cdot) = P(\cdot \mid s)$) be a belief over Ω that assigns strictly positive probability to every state. If c is equivocal there exists $\sigma_P \in (0,1)$ such that $sgn\left[P(T \mid c, \sigma) - P(T \mid \sigma)\right] = sgn\left[(\sigma - \sigma_P)(p_c - q_c)\right]$ for all σ.*

If c is generic ($p_c \neq q_c$ and $r_c \neq t_c$) and not equivocal there exist $\overline{\sigma}_P \in (0,1)$ and $0 < \underline{\sigma}_P \leq \overline{\sigma}_P$ such that for all $\sigma' > \overline{\sigma}_P$ and $\sigma < \underline{\sigma}_P$,

$$(P(T \mid c, \sigma) - P(T \mid \sigma))\,(P(T \mid c, \sigma') - P(T \mid \sigma')) > 0. \tag{10}$$

In addition, if individuals with extreme beliefs revise upwards (downwards) their belief in T for P, they will revise upwards (downwards) for any other belief B over Ω that assigns strictly positive probability to every state. Formally for $\sigma' > \max\{\overline{\sigma}_P, \overline{\sigma}_B\}$ and $\sigma < \min\{\underline{\sigma}_P, \underline{\sigma}_B\}$

$$(P(T \mid c, \sigma') - P(T \mid \sigma'))\,(B(T \mid c, \sigma') - P(T \mid \sigma')) > 0 \tag{11}$$

$$(P(T \mid c, \sigma) - P(T \mid \sigma))\,(B(T \mid c, \sigma) - P(T \mid \sigma)) > 0. \tag{12}$$

Proof. Assume without loss of generality that $p_c > q_c$ (the case of $p_c < q_c$ is symmetric and omitted). Sufficiency follows from Lemma 1 by letting σ_P be such that $P(\text{H} \mid \sigma_P) = h$ because $P(\text{H} \mid \sigma)$ is strictly increasing and continuous in σ.

If c is not equivocal, the intercept of the parabola in (9) is $P(\text{TL})\,P(\text{FL})\,(r_c - t_c) > 0$. Because we have $g > 0$, we focus on the existence of positive roots. There are either:
I) no real positive roots, in which case beliefs in T increase after c for all σ, $P(T \mid c, \sigma) > P(T \mid \sigma)$. For this case set $\underline{\sigma} = \overline{\sigma} = \frac{1}{2}$ to obtain (10).

II) or two positive real roots $g_1 \equiv g(\underline{\sigma})$ and $g_2 \equiv g(\overline{\sigma}) \geq g(\underline{\sigma})$, such that:
II.a) for $\sigma' > \overline{\sigma}$, $g(\sigma') > g(\overline{\sigma})$ and therefore $M(g(\sigma')) > 0$ which implies $P(T \mid c, \sigma') > P(T \mid \sigma')$.
II.b) for $\sigma < \underline{\sigma}$, $g(\sigma) < g(\underline{\sigma})$ and therefore $M(g(\sigma)) > 0$ which implies $P(T \mid c, \sigma) > P(T \mid \sigma)$.
So II.a and II.b establish (10) as was to be shown.

To establish that individuals with extreme beliefs revise in the same direction, regardless of their initial beliefs, notice that what determines whether they will revise up or down their beliefs in T depends only on the sign of $p-q$: if $p-q$ is positive, people with extreme beliefs will revise upwards, regardless of whether their initial belief was P or B; if $p-q<0$, they will revise downwards. ■

Lemma 3 *For all s, σ and $\sigma' \neq \sigma$, $(\sigma'-\sigma)(P(\mathrm{H} \mid s, \sigma') - P(\mathrm{H} \mid s, \sigma)) > 0$. Additionally, suppose s is such that $p_s t_s > q_s r_s$. Then for $\sigma' \neq \sigma$ we have*

$$(P(T \mid s, \sigma') - P(T \mid s, \sigma))(P(\mathrm{H} \mid s, \sigma') - P(\mathrm{H} \mid s, \sigma)) > 0.$$

If $p_s t_s < q_s r_s$ then $(P(T \mid s, \sigma') - P(T \mid s, \sigma))(P(\mathrm{H} \mid s, \sigma') - P(\mathrm{H} \mid s, \sigma)) < 0$.

Proof. Assume $p_s t_s > q_s r_s$, the case of $p_s t_s < q_s r_s$ is analogous and omitted. 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. ■

Theorem 2 is a consequence of the following.

Theorem 5 *Suppose individuals observe a signal s (the body of knowledge) and then a common signal c; suppose c is such that $p_c \neq q_c$, $r_c \neq t_c$. Then, there is a v around which experts polarize completely if and only if c is equivocal and s is similar to c. Formally, there is a v such that*

$$P(T \mid s, \sigma) > v \Rightarrow P(T \mid c, s, \sigma) > P(T \mid s, \sigma) \tag{13}$$

$$P(T \mid s, \sigma) < v \Rightarrow P(T \mid c, s, \sigma) < P(T \mid s, \sigma) \tag{14}$$

and $P^v = P(\sigma : P(T \mid s, \sigma) > v) > 0$, $P_v = P(\sigma : P(T \mid s, \sigma) < v) > 0$ if and only if if c is equivocal and s is similar to c.

Moreover:

(i) if c is equivocal and s is not similar to c with $p_s t_s \neq q_s r_s$, there is moderation: there is a v such that

$$P(T \mid s, \sigma) > v \Rightarrow P(T \mid c, s, \sigma) < P(T \mid s, \sigma)$$

$$P(T \mid s, \sigma) < v \Rightarrow P(T \mid c, s, \sigma) > P(T \mid s, \sigma)$$

and $P^v, P_v > 0$.

(ii) if c is not equivocal and $p_c \neq q_c$, $r_c \neq t_c$ people with extreme beliefs harmonize: there are $\overline{v}$ and $\underline{v}$ such that for all σ_l and σ_h, with $P(T \mid s, \sigma_h) > \overline{v} \geq \underline{v} > P(T \mid s, \sigma_l)$ we obtain

$$(P(T \mid c, s, \sigma_l) - P(T \mid s, \sigma_l))(P(T \mid c, s, \sigma_h) - P(T \mid s, \sigma_h)) > 0.$$

(iii) if c is not equivocal and $p_c = q_c$, $r_c = t_c$ all individuals update in the same direction after c: for all σ and σ',

$$sgn[P(T \mid c, s, \sigma) - P(T \mid s, \sigma)] = sgn[P(T \mid c, s, \sigma') - P(T \mid s, \sigma')].$$

Proof. Without loss of generality, assume throughout that $p_c > q_c$.

Part 1, sufficiency. Signal c is equivocal and s is weakly similar. Since s is weakly similar to c, $p_s t_s > q_s r_s$.

Let P in Lemma 2 be $P(\cdot) = P(\cdot \mid s)$. Then, set $v = P(T \mid \sigma_P)$ for $\sigma_P \in (0,1)$ as in Lemma 2.

Then,

a) by Lemma 3, with $\sigma' = \sigma_P$, $(P(T \mid s, \sigma_P) - P(T \mid s, \sigma))(\sigma_P - \sigma) > 0$, so that

$$P(T \mid s, \sigma) > P(T \mid s, \sigma_P) = v \Leftrightarrow \sigma > \sigma_P \tag{15}$$

and $P^v = P(\sigma : P(T \mid s, \sigma) > v) = P(\sigma : \sigma > \sigma_P) > 0$ and $P_v = P(\sigma : P(T \mid s, \sigma) < v) = P(\sigma : \sigma < \sigma_P) > 0$.

b) by Lemma 2, and $p_c > q_c$,

$$\sigma > \sigma_P \Leftrightarrow P(T \mid c, s, \sigma) > P(T \mid s, \sigma). \tag{16}$$

Combining (15) and (16) we obtain

$$P(T \mid s, \sigma) > v \Leftrightarrow P(T \mid c, s, \sigma) > P(T \mid s, \sigma)$$

as was to be shown.

Part 2, Necessity; s not weakly similar; case (i). Continue to assume that c is equivocal and that $p_c > q_c$, but suppose s is not similar to c. If we had $p_s t_s = q_s r_s$ then $P(T \mid s, \sigma)$ is constant in σ, and there is no v such that $P(T \mid s, \sigma) > v$ for some σ while $P(T \mid s, \sigma') < v$ for some other σ', so polarization cannot obtain. Assume then $p_s t_s < q_s r_s$. In that case, if for some v, σ and σ' we have $P(T \mid s, \sigma) > v$ and $P(T \mid s, \sigma') < v$, by Lemma 3 we know $\sigma' > \sigma$. Two cases arise:

a) if $\sigma' > \sigma_P$, with σ_P from Lemma 2, we have $P(T \mid c, s, \sigma') > P(T \mid s, \sigma')$, violating (14).

b) if $\sigma_P \geq \sigma' > \sigma$, $P(T \mid c, s, \sigma) < P(T \mid s, \sigma)$, violating (13).

That establishes that there exists no such v. To establish moderation (case i), since s has $p_s t_s \neq q_s r_s$ and is not similar to c, $p_s t_s < q_s r_s$ and note that if v is in the range of $P(T \mid s, \sigma)$ for $\sigma \in [0, 1]$, there is a unique σ_v such that $P(T \mid s, \sigma) > v \Leftrightarrow \sigma < \sigma_v$ (beliefs in T are decreasing in σ by Lemma 3). Also, by Lemma 2 there is a σ_P such that $sgn[P(T \mid c, \sigma) - P(T \mid \sigma)] = sgn(\sigma - \sigma_P)$.

If $\sigma_P \geq \sigma_v$, nobody with high initial beliefs increases their beliefs, while some with low initial beliefs do increase them. This is so, since high initial beliefs imply low σ ($P(T \mid s, \sigma) > v \Leftrightarrow \sigma < \sigma_v$), which implies that (because $\sigma \leq \sigma_v \leq \sigma_P$) beliefs get revised downward ($sgn[P(T \mid c, \sigma) - P(T \mid \sigma)] = sgn(\sigma - \sigma_P)$). Also, for all $\sigma > \sigma_P \geq \sigma_v$, initial beliefs are low, and they get revised upwards.

If $\sigma_P < \sigma_v$ in this case, all those with beliefs less than v increase their belief in T (as $P(T \mid s, \sigma) < v$ implies $\sigma > \sigma_v > \sigma_P$, which ensures $P(T \mid c, \sigma) > P(T \mid \sigma)$) while some with high beliefs decrease (since $\sigma < \sigma_P$ implies both a high initial belief in T, and a decrease after observing c), so again there is moderation.

Part 3, Necessity; c not equivocal; case (ii). Signal c is not equivocal. Recall we had assumed $p_c > q_c$ and because c is not equivocal, and generic, we know $r_c > t_c$.

To obtain a contradiction, assume there is one such v. In order for the antecedents (3) and (4) to be nonempty, we must have $p_s t_s \neq r_s q_s$, and then $P(T \mid s, \cdot)$ is a (strictly) monotone function by Lemma 3.

Suppose $p_s t_s > q_s r_s$. In that case, the set $\{\sigma : P(T \mid s, \sigma) > v\}$ is of the form $\{\sigma : \sigma > \sigma^*\}$ for σ^* such that $P(T \mid s, \sigma^*) = v$ and $\{\sigma : P(T \mid s, \sigma) > v\} = \{\sigma : \sigma < \sigma^*\}$. Suppose instead $p_s t_s < q_s r_s$, so that $\{\sigma : P(T \mid s, \sigma) > v\} = \{\sigma : \sigma < \sigma^*\}$ and $\{\sigma : P(T \mid s, \sigma) < v\} = \{\sigma : \sigma > \sigma^*\}$. In either case, extreme values of σ ensure extreme beliefs in T, and then Lemma 2 ensures that for all $\sigma' > \max\{\sigma^*, \overline{\sigma}_P\}$ and $\sigma < \min\{\sigma^*, \underline{\sigma}_P\}$,

$$(P(T \mid c, s, \sigma) - P(T \mid s, \sigma))(P(T \mid c, s, \sigma') - P(T \mid s, \sigma')) > 0 \tag{17}$$

which violates (13) and (14), since individuals who observe extreme values of σ update in the same direction after observing c. This proves necessity.

To establish (ii), we define $\underline{v} \equiv P(T \mid s, \underline{\sigma}_P)$ and $\overline{v} \equiv P(T \mid s, \overline{\sigma}_P)$ for $\underline{\sigma}_P$ and $\overline{\sigma}_P$ from Lemma 2 if $p_s t_s > q_s r_s$. This ensures that

$$\begin{aligned} P(T \mid s, \sigma_h) &> P(T \mid s, \overline{\sigma}_P) = \overline{v} \Leftrightarrow \sigma_h > \overline{\sigma}_P \\ P(T \mid s, \sigma_l) &< P(T \mid s, \underline{\sigma}_P) = \underline{v} \Leftrightarrow \sigma_l < \underline{\sigma}_P \end{aligned}$$

and by Lemma 2, for all $\sigma' > \overline{\sigma}_P$ and $\sigma < \underline{\sigma}_P$ (in particular, for $\sigma_h = \sigma'$ and $\sigma_l = \sigma$), equation (17) holds, as was to be shown. If $p_s t_s < q_s r_s$, define $\overline{v} \equiv P(T \mid s, \underline{\sigma}_P)$ and $\underline{v} \equiv P(T \mid s, \overline{\sigma}_P)$ for $\underline{\sigma}_P$ and $\overline{\sigma}_P$ from Lemma 2. Then,

$$\begin{aligned} P(T \mid s, \sigma_h) &> P(T \mid s, \underline{\sigma}_P) = \overline{v} \Leftrightarrow \sigma_h < \underline{\sigma}_P \\ P(T \mid s, \sigma_l) &< P(T \mid s, \overline{\sigma}_P) = \underline{v} \Leftrightarrow \sigma_l > \overline{\sigma}_P \end{aligned}$$

and by Lemma 2, for all $\sigma' > \overline{\sigma}_P$ and $\sigma < \underline{\sigma}_P$ (in particular, for $\sigma_l = \sigma'$ and $\sigma_h = \sigma$), (17) holds, establishing (ii) and completing the proof.

Part 4, Necessity, c not equivocal with $p_c = q_c$ and $r_c = t_c$. In this case, all subjects update in the same direction after observing c, since from equation (9) the intercept and the coefficient on g^2 vanish. ■

Proof of Theorem 3. Assume without loss of generality that $p_c > q_c$; the case of $p_c < q_c$ is symmetric and omitted.

Sufficiency. For each s compute $\sigma_P \in (0,1)$ from Lemma 2 with $P = P(\cdot \mid s)$ and define $v_s = P(T \mid s, \sigma_P)$. Note that because for each s we have $\sigma_P \in (0,1)$, there is a positive mass of signals σ such that $P(T \mid s, \sigma) > P(T \mid s, \sigma_P) = v_s$. We obtain that for $\overline{v} = \max_{s \in \mathcal{S}} \{v_s\}$, the antecedent in (5) holds with positive probability. Similarly, for $1 - \underline{v} = \min_{s \in \mathcal{S}} \{v_s\} \leq \overline{v}$. As in the proof of Theorem 5

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

which establishes (5). 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.

Necessity. Suppose c is equivocal (continue to assume $p_c > q_c$), but some s is not similar. If $p_s t_s = q_s r_s$, we have that $P(T \mid s, \sigma)$ is constant in σ. If $P(T \mid s, \sigma) > \overline{v}$, we will have that for small σ, $P(T \mid c, s, \sigma) < P(T \mid s, \sigma)$ contradicting (5), while $P(T \mid s, \sigma) < 1 - \underline{v}$ will contradict (6) for large σ.

Suppose $p_s t_s < q_s r_s$ for some s such that the antecedent in (6) holds. Then it will continue to hold for all higher σ (because $P(T \mid s, \sigma)$ is decreasing by Lemma 3) but $P(T \mid c, s_j, \sigma) > P(T \mid s_j, \sigma)$ for some high enough σ, by Lemma (2).

Suppose now c is not equivocal. Suppose also that for some s the antecedent in (5) holds, so that $P(T \mid s, \sigma) > \overline{v}$ for some σ. If $p_s t_s = q_s r_s$, then for all σ we also have $P(T \mid s, \sigma) > \overline{v}$, but for small enough σ we will have $P(T \mid c, s, \sigma) < P(T \mid s, \sigma)$ contradicting (5). A similar argument establishes that a violation also occurs if the antecedent in s the antecedent in (6) holds but again $p_s t_s = q_s r_s$. Assume therefore $p_s t_s \neq q_s r_s$ for any s such that the antecedents in (5) or (6) holds.

If c is not equivocal, and the antecedents in (5) and (6) hold non-trivially pick any $\overline{s}$ such that for some σ, $P(T \mid \overline{s}, \sigma) > \overline{v}$; and pick any $\underline{s}$ such that for some σ, $P(T \mid \underline{s}, \sigma) < 1 - \underline{v}$. The rest of the proof consists in using the facts that:

a) $P(T \mid \overline{s}, \sigma)$ and $P(T \mid \underline{s}, \sigma)$ update in the same direction for extreme values of σ;

b) for both $\overline{s}$ and $\underline{s}$ there are extreme values of σ such that the antecedents in (5) and (6) continue to hold.

Therefore, for some pair of extreme values of σ, both antecedents will hold, but only one of the conclusions of (5) or (6) will hold.

Note that if $P(T \mid \overline{s}, \sigma) > \overline{v}$ holds for some σ, it will hold for all higher σ if $p_{\overline{s}} t_{\overline{s}} > q_{\overline{s}} r_{\overline{s}}$, or for all lower σ if $p_{\overline{s}} t_{\overline{s}} < q_{\overline{s}} r_{\overline{s}}$ (and conversely for $\underline{s}$). This establishes (b).

In Lemma 2 let $P = P(\cdot \mid \overline{s})$ and $B = P(\cdot \mid \underline{s})$, and we know that for all high enough σ' and low enough σ, P revises in the same direction (equation 10); from equations (11) and (12) we know that $P(\cdot \mid \overline{s}, \sigma')$ and $P(\cdot \mid \underline{s}, \sigma')$ revise in the same direction, as do $P(\cdot \mid \overline{s}, \sigma)$ and $P(\cdot \mid \underline{s}, \sigma)$. So the antecedent in (5) will hold for $\overline{s}$ and all high enough σ' or low enough σ, and the antecedent will also hold in (6) for $\underline{s}$ and all high enough σ' or low enough σ; but only one of the consequent statements can hold as $sgn(P(T \mid c, s', \widehat{\sigma}) - P(T \mid s', \widehat{\sigma})) = sgn(P(T \mid c, s, \widetilde{\sigma}) - P(T \mid s, \widetilde{\sigma}))$ for $s, s' \in \{\overline{s}, \underline{s}\}$ and $\widetilde{\sigma}, \widehat{\sigma} \in \{\sigma, \sigma'\}$. ■

Proof of Theorem 4. If s and c are symmetric, we have $p_s = t_s$, $q_s = r_s$, $p_c = t_c$ and $q_c = r_c$. We have $P(T \mid s, \sigma, c) > P(T \mid s, \sigma)$ if and only if

$$\begin{aligned}
\frac{p_c p_s ab\pi_{\mathrm{H}}(\sigma) + q_c q_s a(1-b)\pi_{\mathrm{L}}(\sigma)}{q_c q_s b\pi_{\mathrm{H}}(\sigma)(1-a) + p_c p_s(1-b)(1-a)\pi_{\mathrm{L}}(\sigma)} &> \frac{p_s ab\pi_{\mathrm{H}}(\sigma) + q_s a(1-b)\pi_{\mathrm{L}}(\sigma)}{q_s(1-a)b\pi_{\mathrm{H}}(\sigma) + p_s(1-b)(1-a)\pi_{\mathrm{L}}(\sigma)} \Leftrightarrow \\
\frac{p_c p_s b\pi_{\mathrm{H}}(\sigma) + q_c q_s(1-b)\pi_{\mathrm{L}}(\sigma)}{q_c q_s b\pi_{\mathrm{H}}(\sigma) + p_c p_s(1-b)\pi_{\mathrm{L}}(\sigma)} &> \frac{p_s b\pi_{\mathrm{H}}(\sigma) + q_s(1-b)\pi_{\mathrm{L}}(\sigma)}{q_s b\pi_{\mathrm{H}}(\sigma) + p_s(1-b)\pi_{\mathrm{L}}(\sigma)} \Leftrightarrow \\
(p_c - q_c)(b\pi_{\mathrm{H}}(\sigma) - (1-b)\pi_{\mathrm{L}}(\sigma)) &> 0. \qquad (18)
\end{aligned}$$

Also,

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

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

$$P(T \mid s, \sigma) > a \Leftrightarrow \frac{1}{1 + \frac{1-a}{a}\frac{q_s y + p_s}{p_s y + q_s}} > a \Leftrightarrow 1 > \frac{q_s y + p_s}{p_s y + q_s}. \qquad (19)$$

Sufficiency. Assume without loss of generality that $p_c > q_c$, and since every s is similar to c, $p_s > q_s$ (the case of $p_c < q_c$ is analogous and omitted). From equation (19),

$$P(T \mid s, \sigma) > a \Leftrightarrow p_s y + q_s > q_s y + p_s \Leftrightarrow y = \frac{b\pi_{\mathrm{H}}(\sigma)}{(1-b)\pi_{\mathrm{L}}(\sigma)} > 1 \overset{(18)}{\Leftrightarrow} P(T \mid s, \sigma, c) > P(T \mid s, \sigma)$$

as was to be shown. That the antecedents hold non-trivially follows from the fact that $P(T \mid s, \sigma) > a \Leftrightarrow \frac{\pi_{\mathrm{H}}(\sigma)}{\pi_{\mathrm{L}}(\sigma)} > \frac{1-b}{b}$ and that $\frac{\pi_{\mathrm{H}}(\sigma)}{\pi_{\mathrm{L}}(\sigma)}$ ranges from 0 to ∞.

Necessity. If $p_s = q_s$, the initial beliefs are constant in σ, and both antecedents in the theorem cannot hold non-trivially. Assume then $p_s \neq q_s$. Similarly, if $p_c = q_c$, by equation (18) there is no updating after c, so both conclusions in the theorem fail to hold. Assume then $p_c \neq q_c$.

If c is equivocal, assume still $p_c > q_c$ without loss of generality, but suppose some s is not similar to c, so that $p_s < q_s$. Then, by equation (19),

$$P(T \mid s, \sigma) > a \Leftrightarrow b\pi_{\mathrm{H}}(\sigma) < (1-b)\pi_{\mathrm{L}}(\sigma) \Leftrightarrow P(T \mid s, \sigma, c) < P(T \mid s, \sigma),$$

so we obtain moderation for that particular s. Since, being symmetric and with $p_c \neq q_c$, c must be equivocal, the proof is complete. ■

Proof of Corollary 1. For

$$\begin{aligned}\Pr(c \mid hT, \sigma_1) &= \frac{\Pr(c, hT, \sigma_1)}{\Pr(hT, \sigma_1)} = \frac{p_c P(\sigma_1 \mid HhT) P(HhT) + r_c P(\sigma_1 \mid LhT) P(LhT)}{\Pr(hT, \sigma_1)} \\ &= \frac{p_c \pi_H(\sigma_1) abd + r_c \pi_L(\sigma_1) a(1-b) d}{\pi_H(\sigma_1) abd + r_c \pi_L(\sigma_1) a(1-b) d} = \frac{p_c \pi_H(\sigma_1) b + r_c \pi_L(\sigma_1)(1-b)}{\pi_H(\sigma_1) b + \pi_L(\sigma_1)(1-b)} \\ &= \Pr(c \mid lT, \sigma_1)\end{aligned}$$

and any fixed distribution π of σ_1 define $\overline{p}_c$ to be the probability with which a subject will observe signal c (without having observed yet σ_1) in state hT :

$$\overline{p}_c = E[\Pr(c \mid hT, \sigma_1)] = \int \Pr(c \mid hT) \pi(\sigma_1) d\sigma_1.$$

Notice that because $\Pr(c \mid hT, \sigma_1) = \Pr(c \mid lT, \sigma_1)$, we obtain that for $\overline{r}_c$ the probability with which a subject will observe signal c (without having observed yet σ_1) in state lT, $\overline{p}_c = \overline{r}_c$. Similarly when we define $\overline{q}_c$ to be the probability with which a subject will observe signal c (without having observed yet σ_1) in state hF, and $\overline{t}_c$ that in state lF, we obtain $\overline{q}_c = \overline{t}_c$.

Then, if $\overline{p}_c > \overline{q}_c$, we obtain $\overline{r}_c > \overline{t}_c$, so that signal c, is not equivocal with respect to states $\{h, l\} \times \{T, F\}$. Then, by Theorem 2 with states h and l playing the role of H and L, and $\overline{p}_c, \overline{q}_c, \overline{r}_c$ and $\overline{t}_c$ playing the role of p_c, q_c, r_c and t_c (as if the agents had not observed σ_1 yet) there is no v such that

$$\begin{aligned}P(T \mid s, \sigma_2) &> v \Rightarrow P(T \mid c, s, \sigma_2) > P(T \mid s, \sigma_2) \\ P(T \mid s, \sigma_2) &< v \Rightarrow P(T \mid c, s, \sigma_2) < P(T \mid s, \sigma_2).\end{aligned}$$

Similarly, if $\overline{p}_c = \overline{q}_c$, by Part (iii) of Theorem 5, again there is no polarization (this result is of course trivial: when $p_c = q_c = r_c = t_c$, there is no updating after c).

That establishes that there is no complete polarization, but in fact one can derive a stronger conclusion. It is easy to check that we can write an agent's posteriors after observing

s and σ as follows (for some $f, g \in (0,1)$, where f is a function of σ_2 and g of σ_1):

	posterior after s and σ proportional to			posterior after s,c and σ proportional to	
	T	F		T	F
Hh	$afgp_s$	$(1-a)fgq_s$		$afgp_sp_c$	$(1-a)fgq_sq_c$
Lh	$a(1-f)gp_s$	$(1-a)(1-f)gq_s$	&	$a(1-f)gp_sr_c$	$(1-a)(1-f)gq_st_c$
Hl	$af(1-g)r_s$	$(1-a)f(1-g)t_s$		$af(1-g)r_sp_c$	$(1-a)f(1-g)t_sq_c$
Ll	$a(1-f)(1-g)r_s$	$(1-a)(1-f)(1-g)t_s$		$a(1-f)(1-g)r_sr_c$	$(1-a)(1-f)(1-g)t_s$

We have,

$$\begin{aligned}\frac{P(T \mid s,\sigma)}{1-P(T \mid s,\sigma)} &= \frac{a}{1-a}\frac{fgp_s+(1-f)gp_s+f(1-g)r_s+(1-f)(1-g)r_s}{fgq_s+(1-f)gq_s+f(1-g)t_s+(1-f)(1-g)t_s} > \frac{v}{1-v} \Leftrightarrow \\ \frac{1-a}{a}\frac{v}{1-v} &< \frac{gp_s+(1-g)r_s}{gq_s+(1-g)t_s}\end{aligned}$$

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

Similarly

$$\begin{aligned} P(T \mid s,c,\sigma) &> P(T \mid s,\sigma) \Leftrightarrow \\ \frac{fgp_cp_s+(1-f)gr_cp_s+f(1-g)p_cr_s+(1-f)(1-g)r_cr_s}{fgq_cq_s+(1-f)gt_cq_s+f(1-g)q_ct_s+(1-f)(1-g)t_ct_s} &> \frac{gp_s+(1-g)r_s}{gq_s+(1-g)t_s} \Leftrightarrow \\ \frac{fp_c+(1-f)r_c}{fq_c+(1-f)t_c}\frac{gp_s+(1-g)r_s}{gq_s+(1-g)t_s} &> \frac{gp_s+(1-g)r_s}{gq_s+(1-g)t_s} \Leftrightarrow \\ \frac{fp_c+(1-f)r_c}{fq_c+(1-f)t_c} &> 1 \end{aligned}$$

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

Suppose for simplicity that $p_c > q_c$. Then for any v, the population with $P(T \mid s,\sigma) > v$ will increase their belief in T iff $\sigma_1 > \overline{\sigma}_1$ for some cutoff $\overline{\sigma}_1$; but the same is true for those who have a σ_2 such that $P(T \mid s,\sigma) < v$, so that the proportion who increase their belief is constant in their initial beliefs. This establishes no polarization. ■

References

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

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

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

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

BioNews (20015), Scientists disagree over Native American origins, https://www.bionews.org.uk/page_9

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.

Gerber A, Green DP. 1997. Rational learning and partisan attitudes. Am. J. Polit. Sci. 42:794–818

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

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

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

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

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

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

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

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

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

List, J. (2007), "On the Interpretation of Giving in Dictator Games," *Journal of Political*

Economy, **115(3)**,482-94.

Loh, I. and G. Phelan (2017), "Dimensionality and Disagreement: Asymptotic Belief Divergence in Response to Common Information," WP 2016-18 Williams College.

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

Lord, C.G. L. Ross and M.R. Lepper (1979), "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence," *Journal of Personality and Social Psychology*, **37(11)**, 2098-2109.

Miller, A. G., J. W. McHoskey, C. M. Bane, and T. G. Dowd (1993), "The Attitude Polarization Phenomenon: Role of Response Measure, Attitude Extremity, and Behavioral Consequences of Reported Attitude Change," *Journal of Personality and Social Psychology*, **64(4)**, 561–574.

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

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

Page BI, Shapiro RY. 1992. The Rational Public: Fifty Years of Trends in Americans'Policy Preferences. Chicago: Univ. Chicago Press

Plous, S. (1991), "Biases in the Assimilation of Technological Breakdowns: Do Accidents Make Us Safer?," *Journal of Applied Social Psychology,* **21(13)**, 1058-82.

Pascal, B., Pensees, translatd by W.F. Trotter

Rabin, M. and J. Schrag (1999), "First Impressions Matter: A Model Of Confirmatory Bias," *The Quarterly Journal of Economics* **114(1)**.

Seidenfeld, T. and L. Wasserman (1993), "Dilation for Sets of Probabilities," *Annals of Statistics* **21(3)**, pp. 1139-54.

Walley, P. (1991). Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London.