Why Bayesians Polarize: Towards A Unified Approach

David Young
School of Psychological Sciences, Birkbeck College, Malet Street, London, WC1E 7HX UK
Department of Psychology, University of Cambridge, Downing Place, Cambridge, CB2 3EB UK

Abstract

It is widely assumed that when belief polarization occurs – people updating their beliefs in opposing directions in response to the same evidence – people's reasoning must be in some way irrational. Yet, several authors have shown that Bayesian reasoning can, theoretically, cause polarization. I provide an accessible explanation for why Bayesians polarize, unifying some previous approaches. I elucidate the conditions under which Bayesian polarization can occur in two scenarios, providing non-technical explanations, with examples, and using Bayesian Networks to set up mathematical proofs. I show how these two scenarios can provide explanations for some prior documented cases of belief polarization in the literature.

Keywords: Bayesian; polarization; belief; updating; networks.

Introduction

There is confusion about the cognitive underpinnings of belief polarization, the phenomenon where two observers who observe the same piece of evidence regarding a particular hypothesis update their beliefs about the hypothesis in opposing directions, one becoming more confident it is true, one becoming more confident it is false (Bullock, 2009; Jern, Chang, & Kemp, 2014). In particular, there is debate about whether polarization can be caused by Bayesian reasoning, with the field historically supposing not.

For example, the most famous empirical example of belief polarization, Lord et al. (1979), is widely attributed to a form of procedurally-irrational reasoning called 'biased assimilation' (Baron, 2008; Lord & Taylor, 2009). Elsewhere, prominent authors claim Bayesianism requires people exposed to identical evidence to come closer together in their beliefs (Bartels, 2002; Kahan, 2016), making polarization *de facto* incompatible with Bayesian reasoning.

More recently it has been shown that there are circumstances where belief polarization can be Bayesian (Bullock, 2009; Cook & Lewandowsky, 2016; Jern et al., 2014; Mandelbaum, 2019; Olsson, 2013; Young et al., 2025). Articles which extemporize this perspective carry titles like "*Rational irrationality"* (Cook & Lewandowsky, 2016) and "*Belief polarization is not always irrational"* (Jern et al. 2014), underscoring the field's general view of belief polarization as the product of irrationality.

However, the literature establishing that polarization can be Bayesian has some limitations. Firstly, it contains internal conflicts – for instance, Bullock (2009) concludes that it is not possible for polarization over an "unchanging" hypothesis, whose state is fixed to occur, yet others, like Jern et al. (2014), contend it is. Secondly, the accessibility of this literature is in some cases limited due to its mathematical complexity. Thirdly, and most importantly, some accounts fall short of identifying sets of precise conditions which are *sufficient* for Bayesian polarization to occur – for example, Jern et al. (2014) identify three Bayesian Networks which permit polarization, but, as I will show, these networks must possess additional properties unstated by Jern et al. for polarization to be possible.

In this article I aim to communicate clearly and precisely why it is actually quite easy for Bayesians to polarize, by presenting two simple and broadly-applicable scenarios in which it can occur. I unify the approaches of Jern et al. (2014) and Benoît & Dubra (2019). To maximize accessibility, I present non-technical explanations of both scenarios, with examples, as well as mathematical proofs. I conclude by discussing how some prominent previous cases of belief polarization could fit into my framework. I hope the evidence I present will make it easier for others to adjudicate whether specific cases of belief polarization are compatible with Bayesian reasoning, or not.

Bayesian Polarization

Bayes' Rule tells us that when we observe some evidence, E, that bears upon some hypothesis, H, a true-or-false statement about the state of the world, we should update our confidence in H being true in proportion to the *likelihood ratio* of the evidence; this is how much likelier we think it is that we would observe the evidence if H were true than if it were false, $p(E|H)/p(E|\neg H)$.

Two individuals can polarize upon viewing the same evidence, then, when they have likelihood ratios which go in opposing direction, one calculating the likelihood ratio to be *above* 1 and so strengthening their in confidence H, the other calculating the likelihood ratio to be *below* 1, and weakening their confidence.

Bayesian Networks can help us understand when this is possible. A Bayesian Network is a model of a reasoning problem which is parameterized so as to contain the information necessary for users to perform Bayesian inference to estimate, probabilistically, the states of some variables given observations of others (see Pearl, 1988; Pearl and Russell, 2002). Bayesian Networks contain information about so called parent-child relationships, where parent variables influence the states of child variables, depicted in the form of arrows (*directed edges*) drawn from parents to children. These arrows are intuitively suited to modelling

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

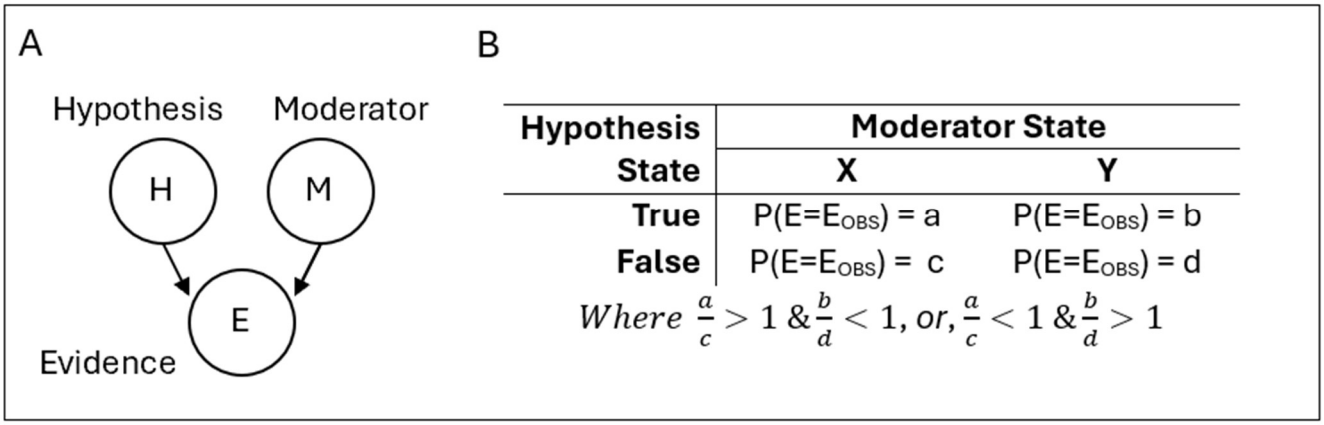


Figure 1: The structure of a Bayesian Network where belief polarization in response to a single piece of evidence can occur due to a **crossover** interaction. **A**. Diagram of the network in the form of a directed acyclic graph (DAG). **B.** The properties of the Conditional Probability Table; $P(E=E_{OBS})$ refers to the probability of the evidence node occupying its observed state, conditional upon the state of the Moderator variable and the state of the Hypothesis.

causal relationships but can also model purely correlational probabilistic dependencies. Conditional probability tables encode information about these relationships by specifying, for every child node, the probability of it occupying each of its different possible states for every possible combination of its parents' states. A large literature uses Bayesian Networks to model how human reasoners learn from evidence across a variety of contexts (Bhui & Gershman, 2020; Botvinik-Nezer et al., 2023; Fenton et al., 2013; Harris et al., 2013, 2016; Lagnado et al., 2013; Madsen, 2016, 2019; Young et al., 2023, 2025; Young & de-Wit, 2025).

Suppose we use the Bayesian Network in Figure 1a, to model a case where our interpretation of the evidence with respect to some hypothesis is affected by a Moderator, *M*. For simplicity's sake, I will assume all the nodes are binary, having only two possible states – *True* and *False* for *H*, *X* and *Y* for *M*, and since we observe *E* to be in one particular state, we do not need to worry about what its states are.

Crucially, the value of *M* affects our calculation of the likelihood ratio of *E* with respect to *H*, because both p(E|H) and p(E|¬H) have to be calculated as sums of the likelihood of *E* for both of *M*'s states, weighted by the probability of *M* occupying each state. That is, the likelihood ratio is equal to:

$$\frac{p(E|H,M=X)p(M=X)+p(E|H,M=Y)p(M=Y)}{p(E|\neg H,M=X)p(M=X)+p(E|\neg H,M=Y)p(M=Y)}$$

People with different priors for *M*, then, will have different likelihood ratios, and in certain cases, ones with opposing directions. Scenario #1 and Scenario #2 refer to broad types of reasoning problem which can be captured with a Bayesian Networks that has sufficient properties for this to occur.

Previous Approaches

My approach unifies two previous approaches, while adding some greater specificity and simplifying some elements. The Bayesian Network in Figure 1a is identical to one of the Networks Jern et al. (2014) identify as capable of causing belief polarization, which they call Network 'vi'. I label their node '*V*' as *M* and characterize it as a 'Moderator', which helps to illustrate its role. The Scenario #2 is a variation on this network that Jern et al. (2014) did not discuss. In contrast to Jern et al. (2014), I show that the properties of the conditional probability table for *E* are crucial for determining whether belief polarization is possible.

I also draw on Benoît & Dubra (2019), who show Bayesian polarization can occur when an "ancillary matter" affects how the evidence should be interpreted. I capture "ancillary matters" using the variable *M*, providing a firm computational definition. Importantly, the condition which I will establish must be fulfilled for a Bayesian Network to allow belief polarization in Scenario #1 is identical, besides labelling, to one of the conditions Benoît and Dubra identify as necessary for Bayesian belief polarization – namely, using their nomenclature, that a signal is "equivocal" (see p. 1678-1679). My contribution is to provide a substantially simpler proof of this, to offer a different and perhaps clearer way of understanding the point by situating it within the Bayesian Network framework, and to differentiate this kind of polarization from Scenario #2 cases, where observers receive two pieces of conflicting evidence, which is a distinction Benoît and Dubra do not make.

Scenario #1

Non-Technical Explanation

Scenario #1 occurs when the relationship between the evidence we observe, *E,* and the hypothesis we are interested in, *H*, is affected by some kind of Moderator variable, *M*, which changes whether *E* is more likely to occur when *H* is true or when *H* is false. If we think *E* is more likely to be observed when *H* is true, we will regard it as positive evidence, but if we think *E* is more likely to be observed when *H* is false, we will regard it as *negative* evidence. Thus, if two people have different beliefs about *M*, one can regard the evidence as positive evidence while the other regards it as negative, causing polarization. I provide three examples of

polarization occurring in this way, each taken from a different broad class of cases where Scenario #1 can arise.

In the first class of cases, 'Taste' cases, we seek to make a valence judgment about whether something is good or bad after observing that it possesses a particular characteristic. The Moderator simply affects whether we regard the characteristic as indicating goodness or badness. For example, if we learn that our favorite musician is about to release a 'country' album, whether or not we believe the album will be good or bad will be affected by our belief about whether country music is good or bad. Thus, a person who thinks country music is good is likely to increase their confidence the album will be good, but a person who thinks country music is bad will increase their confidence the album will be bad. This is the most basic way a stimulus can be "polarizing" – people simply have different beliefs about whether its characteristics make it good or bad.

In the second class of Scenario #1 cases, 'Trade-Off' cases, we again seek to make a valence judgment about something after learning about a characteristic that it possesses. This characteristic has both good and bad implications, and the Moderator is whether we judge the good to outweigh the bad, or vice versa. For example, suppose we learn that a zoo located very close to where we live is extremely cheap. This has the good implication of making it more affordable, but also the bad implication that the zoo may not be able to afford adequate security measures to prevent dangerous animals from escaping, potentially endangering us. Two people considering whether the proximity of the zoo is a good or bad thing can therefore polarize upon learning about its cheap cost – someone who values the greater affordability but assumes the security risks are minimal will increase their confidence it is good, but someone who has little interest in going anyway but worries an under-funded zoo will shirk on security will increase their confidence it is bad.

In the third class of cases, 'Knowledge' cases, a difference in the *knowledge* possessed by two people can lead to polarization when some piece of background knowledge is critical to determining the direction of the evidence. For example, suppose you are at a dinner party with two friends, one of whom hates fish. You learn the host is planning to serve 'Bombay duck'. You, assuming Bombay duck to be a dish containing duck, increase your confidence that your fish-hating friend will enjoy their meal. But your other friend, knowing that Bombay duck actually contains fish, increases their confidence that the friend will dislike their meal.

Technical Explanation

The defining feature of Scenario #1 is that a Bayesian Network of the problem structure consists of an interaction between the hypothesis-under-consideration H (with states 'True' or 'False') and a Moderator M (with states X and Y), which influences the state of the evidence variable, E, with a *crossover interaction*. By *crossover interaction*, I mean that the state of M changes whether the probability of E being in its observed state is greater when H is true or when it is false; this means that the state of H with the highest *likelihood* given E's state differs depending on the state of M, and thus E can provide positive or negative evidence for H being true depending on our prior for M. Figure 1 shows the structure of this Bayesian Network and defines the properties of its conditional probability table.

To prove that Bayesian Networks with these properties allow for belief polarization to occur, I will show that the direction of updating can be either positive or negative depending on the observer's prior for M.

To make the equations less cluttered, I let $p(M{=}X) = m$, where $0 \geq m \geq 1$ (I prevent $m = 0$ and $m = 1$ as in these cases, the observer would have no need to consider M as a variable, knowing its state for certain, and so it is not the case that they would be using the Scenario #1 network). As Figure 1b shows, our critical property is that either $a/c > 1$ and $b/d < 1$, or $a/c < 1$ and $b/d > 1$. I begin by assuming $a/c > 1$ and $b/d < 1$. The likelihood ratio governing how much a person updates in response to the evidence is:

$$LR = \frac{am + b(1-m)}{cm + d(1-m)} \tag{1.1}$$

Positive updating occurs when $LR > 1$. We can therefore deduce the conditions under which the value of m will lead to positive updating as follows:

$$\frac{am + b(1-m)}{cm + d(1-m)} > 1 \tag{1.2}$$

$$am + b(1-m) > cm + d(1-m) \tag{1.3}$$

$$(a-c)m > (d-b)(1-m) \tag{1.4}$$

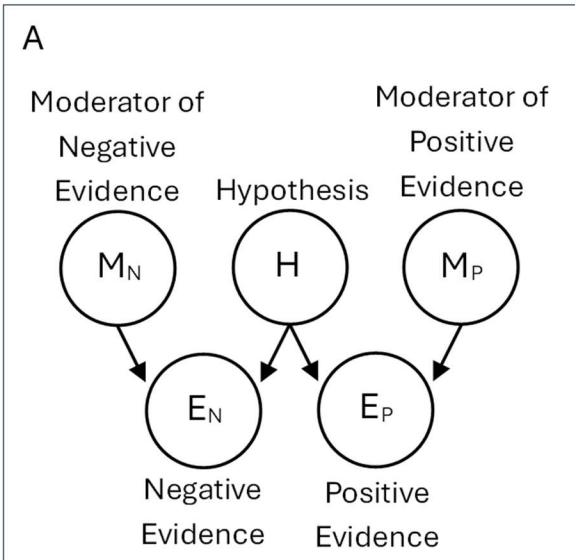
$$\frac{m}{1-m} > \frac{d-b}{a-c} \tag{1.5}$$

Since $a/c > 1$, $a - c$ is positive, and since $b/d < 1$, $d - b$ is positive too. This means $(d - b)/(a - c)$ is positive and non-infinite, and since $0 \geq m \geq 1$, $m/(1 - m)$ must also have a positive non-infinite value. Therefore certain values of m will satisfy this inequality, leading to positive updating.

But other values of m will lead to negative updating for the same values of a, b, c, and d. Negative updating occurs when $LR < 1$. Therefore suppose the inequality was flipped from the start in *Eq.* 1.2; since the inequality does not change from *Eq* 1.2 to *Eq* 1.5, we know that following the same logic will result in the condition for negative updating being:

$$\frac{m}{1-m} < \frac{d-b}{a-c} \tag{1.6}$$

Again, since the only restriction on the values of both quantities in this equation is that they are positive and non-infinite, it will always be possible for certain values of m to satisfy this inequality. Therefore, depending on a person's prior for M, it will be possible or them to update in either direction. In the special case where $d - b = a - c$, positive updating will occur when $m/(1 - m) > 1$, i.e., when $m > 0.5$, whereas negative updating will occur when $m < 0.5$. In other cases, the critical value of m will be different. Effectively, (d



B

Hypothesis State	Moderator of Positive Evidence State	
	X	Y
True	P(E=POS) = a	P(E=POS) = b
False	P(E=POS) = c	P(E=POS) = d

Hypothesis State	Moderator of Negative Evidence State	
	X	Y
True	P(E=NEG) = c	P(E=NEG) = d
False	P(E=NEG) = a	P(E=NEG) = b

$$Where \frac{a}{c} > \frac{b}{d}, or, \frac{a}{c} < \frac{b}{d}$$

Figure 2: The structure of a Bayesian Network where belief polarization in response to two pieces of evidence, one positive and one negative, can occur due to an **attenuating** interaction. **A**. Diagram of the network in the form of a directed acyclic graph (DAG). **B.** The properties of the Conditional Probability Table; P(E=POS) and P(E=NEG) refer, respectively, to the probability, for each evidence node, of it occupying a positive and negative state, conditional upon the state of the Moderator variable and the state of the Hypothesis.

- b)/(a - c) acts as a threshold which determines the direction of updating depending on whether m/(1-m) is greater or lesser.

The Obverse Case

What about when $a/c < 1$ and $b/d > 1$? Now we obtain the opposite end results: when *Eq*. 1.5 is satisfied, *negative* updating occurs; when *Eq*. 1.6 is satisfied, *positive* updating occurs. This is because, if we proceed from *Eq.* 1.2 as before to establish the conditions under which positive updating occurs, when we reach *Eq*. 1.4 and try to proceed further by dividing both sides by $a - c$, this quantity is negative, which requires us to flip the inequality, meaning we end up with *Eq*. 1.6. Similarly, if we began with the inequality flipped in *Eq*. 1.2 to establish the conditions under which negative updating occurs, we would end up having to flip the inequality at the equivalent point, leading to *Eq.* 1.5. So, updating can still occur in either direction, but the rule governing *which* direction is reversed.

The Null Case

We can also see that polarization will not be possible for this structure of Bayesian Network *unless* there is a crossover interaction in the conditional probability table as defined. There cannot be a crossover interaction if either $a/c = 1$ or $b/d = 1$. If $a/c = 1$, $a - c = 0$, meaning $(d - b)/(a - c)$ is undefined and so neither *Eq*. 1.5 or *Eq*. 1.6 can be satisfied. If $b/d = 1$, $b = d$, meaning $(d - b)/(a - c) = 0$ and *Eq* 1.6 cannot be satisfied. In either case, it is not possible for there to be values of m which can satisfy both inequalities. The other way in which a crossover interaction does not occur is if both a/c and $b/d > 1$, or both a/c and $b/d < 1$. In both cases, either $a - c$ or $b - d$ is negative, meaning $(d - b)/(a - c)$ has a negative value, and so *Eq*. 1.6 cannot be satisfied.

Thus, when there is no crossover interaction, for any given set of parameters a, b, c, and d, updating will only be possible in one direction.

Scenario #2

Non-Technical Explanation

In Scenario #2, people receive *two* pieces of evidence for the same hypothesis, one positive and one negative. For each piece of evidence, there is a moderating factor which affects its *strength*. Two people can then polarize when one's beliefs about the moderating factors mean they regard the positive evidence as stronger than the negative evidence, leading to a positive net update, but the other sees the negative evidence as stronger than the positive, leading to a negative net update.

As a specific example, suppose two people read two newspaper articles about whether a politician is guilty of a scandal, one arguing for their guilt, one arguing for their innocence. How strongly they update in response to each article will be influenced by their past experience of the newspapers' reliability in commenting on political matters. If one person has typically found the newspaper arguing for the politician's guilt to be more reliable than the other, they will become more convinced of their guilt; but if the other person has typically found the *other* newspaper to be more reliable, they will become more convinced of the politician's

innocence. Thus, the two people polarize due to their differing perceptions of the newspapers' reliabilities.

Technical Explanation

The defining features of Scenario #2 are that a Bayesian Network of the problem structure contains two pieces of evidence, each of which is the product of an interaction between the hypothesis-under-consideration *H* and a moderating factor *M,* where each interaction creates an *attenuating interaction* in the conditional probability table. By *attenuating interaction*, I mean that for one state of *M*, the degree to which *E* is likelier given *H* than ¬*H* is reduced, altering the strength of the evidence as our belief in *M* occupying that state increases (this also includes cases where the state of *M reverses* whether *E* is likelier given *H* or ¬*H*, as the strength of the evidence in its nominal direction is reduced). This allows some people to weight the positive evidence more than the negative, but for others to do the opposite, polarizing. The structure of the network and the properties its conditional probability tables must have are shown in Figure 2. We can see from Figure 2a that the structure of the Bayesian Network is effectively two of the Bayesian Networks from Figure 1a joined together at the *H* node; notably, Jern et al. (2014) did not propose networks with this structure.

To prove that Bayesian Networks with these properties can allow for polarization, I show that the direction of updating depends on the relative strength of the observer's priors for the Moderator variables. To make the equations less cluttered, I let p($M_P = X$) = m_p, and p($M_N = X$) = m_N, neither of which can equal 0 or 1, as before.

The likelihood ratio for the positive evidence, LR_P, is:

$$LR_P = \frac{m_P a + (1 - m_P)b}{m_P c + (1 - m_P)d} \tag{2.1}$$

Similarly the likelihood ratio for the negative evidence, LR_N, is:

$$LR_N = \frac{m_N c + (1 - m_N)d}{m_N a + (1 - m_N)b} \tag{2.2}$$

As Figure 2a shows, it can either be the case that $a/c > b/d$ or $a/c < b/d$. I will start with the case where $a/c > b/d$. This gives a condition to which we will return shortly:

$$ad - bc > 0 \tag{2.3}$$

To get positive updating, $LR_P \times LR_N > 1$, and therefore:

$$\frac{m_P a + (1 - m_P)b}{m_P c + (1 - m_P)d} > \frac{m_N a + (1 - m_N)b}{m_N c + (1 - m_N)d} \tag{2.4}$$

Expanding and simplifying, we find:

$$\begin{gathered} adm_P(1 - m_N) + bcm_N(1 - m_P) > \\ adm_N(1 - m_P) + bcm_P(1 - m_N) \end{gathered} \tag{2.5}$$

$$(ad - bc)m_P(1 - m_N) > (ad - bc)m_N(1 - m_P) \tag{2.6}$$

Now we can simplify this further by dividing both sides by $ad - bc$, which *Eq*. 2.3 shows is positive. Therefore we do not need to reverse the inequality in this case, giving:

$$m_P(1 - m_N) > m_N(1 - m_P) \tag{2.7}$$

$$m_P - m_P m_N > m_N - m_P m_N \tag{2.8}$$

$$m_P > m_N \tag{2.9}$$

So positive updating occurs when $m_P > m_N$ .As for negative updating, we can see that this will occur when $m_P < m_N$. This is because to establish the conditions under which negative updating occurs, we would start with $LR_P \times LR_N < 1$, giving:

$$\frac{m_P a + (1 - m_P)b}{m_P c + (1 - m_P)d} < \frac{m_N a + (1 - m_N)b}{m_N c + (1 - m_N)d} \tag{2.10}$$

Which is just *Eq*. 2.4 with the signed reversed. Following the same logic as before then, since the sign does not flip between *Eq*. 2.4 and *Eq*. 2.9, we know we will end up with:

$$m_P < m_N \tag{2.11}$$

Therefore, what determines the direction of updating is very simple: is the person's prior for p($M_P = X$) greater than p($M_N = X$), or vice versa? These cases lead to positive and negative updating respectively.

The Obverse Case

We began the preceding proof by assuming that $a/c > b/d$, but we can also have an attenuating interaction when $a/c < b/d$ Like with Scenario #1, flipping over this inequality just means that the conditions for positive and negative updating are reversed. With *a*, *b*, *c*, and *d* parameters that have this property, positive updating occurs when *Eq*. 2.11 is satisfied, whereas negative updating occurs when *Eq*. 2.9 is satisfied.

To see why this is the case, we need to first establish that, since $a/c < b/d$, *Eq*. 2.3 is not true, rather:

$$ad - bc < 0 \tag{2.12}$$

Thus, to understand when positive updating takes place, we can follow from *Eq*. 2.4 to *Eq*. 2.6 as before, but when we divide both sides by $ad - bc$, must flip the inequality since this is a negative quantity, giving not *Eq*. 2.7, but rather:

$$m_P(1 - m_N) < m_N(1 - m_P) \tag{2.13}$$

Which simplifies to *Eq*. 2.11, $m_P < m_N$. Similarly, if we began with *Eq*. 2.10 to understand when negative updating occurs, we would have to flip the inequality at the same juncture, leading us to *Eq*. 2.9, $m_P > m_N$.

The Null Case

If there is no attenuating interaction, polarize cannot occur. In this case, $a/c = b/d$, meaning:

$$ad - bc = 0 \tag{2.14}$$

This means we cannot update either positively or negatively, because we cannot move from *Eq*. 2.6 to *Eq*. 2.7,

as dividing by $ad - bc$ would require dividing by 0, which we cannot do.

More Complicated Cases

To simplify matters, in the preceding cases both Moderators M_N and M_P shared a set of common parameters, and their conditional probability tables were mirror images. In the Supplementary Materials, I show that in cases where they have different parameters, polarization can still occur as long as both Moderators have attenuating interactions (https://osf.io/rhbq4/?view_only=d77e48ce8ad5496f99cbc4eaeea5ad09).

Discussion

There are two easy-to-imagine scenarios where sufficient conditions are in place for Bayesians to polarize. These Bayesians agree about the structure of the problem they are reasoning about, and the conditional probabilities which describe the relationships between the variables involved, but they update in opposing directions – polarize – due to differences in their prior beliefs about moderating variables which affect how the evidence is interpreted.

The two scenarios suggested here can provide theoretical Bayesian explanations for a number of well-known cases of belief polarization in the literature. For example, to explain why Lord et al. (1979) found that providing mixed evidence for-and-against the effectiveness of the death penalty as a crime deterrent led both people who initially doubted its effectiveness and those who initially believed in its effectiveness to report that it had strengthened their views, we can view this as an example of Scenario #2. The moderating factor here could have been the reliability of the sources, as evidence from less-reliable sources should naturally attenuate the strength of the source's evidence. Whichever side participants initially took, they likely had previous experience that people who took the same view of the death penalty as them tended to be more reliable than people in the opposing camp, as the attitudinally-congruent source group's claims would have, in past experiences, appeared more plausible in light of their pre-existing views (see Collins et al., 2018). Since pro-death penalty claims are more likely to be made by people who support the death penalty, and vice versa, people can probabilistically infer that the evidence which conflicts with their view is more likely to come from an unreliable source, leading people to weight the attitudinally-consistent information more heavily, creating polarization. Indeed, this explanation is quite similar to one suggested by Jern et al. (2014), where participants might down-weight the attitudinally-inconsistent claims due to perceived bias of the source.

Furthermore Young et al. (2025) have empirically demonstrated that when people are exposed to conflicting claims made by two source groups, and have reason to regard one group as less reliable (in this case because their members were more inter-dependent), this can cause belief polarization between people who disagree about *which* group is less reliable. This is a case of Scenario #2 polarization.

An example of Scenario #1 polarization can be found in Benoît and Dubra (2019)'s explanation of the polarization observed by Plous (1991), who found that telling people about an incident where a nuclear power station narrowly avoided going into meltdown caused people who initially thought nuclear power unsafe to view it as less safe, but for those who initially thought it safe, to view it as even more safe. They point out that the 'near miss' has both negative implications – dangerous events can and do occur – and positive ones – nuclear plants have robust and reliable safety protocols. Thus, people who disagree about whether the positives outweigh the negatives can update in opposing directions – this then is Scenario #1 example of the 'Trade-Off' class. Moreover, they point out that how people evaluate this trade-off is likely to have influenced their view of similar near-miss incidents in the past, leading those who prioritize the negative implications to doubt nuclear safety, and those who prioritize the positive implications to have faith in nuclear safety, which explains why it was the former who updated negatively and the latter who updated positively.

Of course, I have no empirical evidence that the Scenario #1 and #2 Bayesian networks explain these cases; I can merely point out the logical plausibility. One piece of evidence, however, comes from Cook and Lewandowsky (2016), who proposed a Bayesian network that could explain polarization over climate change evidence, which has the form of a Scenario #1 network; their US participants perceived a crossover interaction in the conditional probability table, and polarization occurred. Future studies could systematically test whether polarization does occur when people perceive the reasoning problem they face to have the form of a Scenario #1 or #2 network and the conditions I have identified are satisfied.

My two scenarios, and the example cases of them given, can help researchers determine if a Bayesian explanation for a given example of polarization might be plausible. The computational specificity afforded here offers greater clarity on this matter than exists currently in the literature – I go beyond Jern et al. (2014) by identifying specific conditions that must be satisfied regarding people's priors and the conditional probability tables, and I go beyond Mandelbaum's (2019) attempt to differentiate between Bayes-compatible and Bayes-incompatible cases of polarization by providing computational rather than purely verbal arguments. However, my framework cannot be used as a foolproof guide, as the explanations for Bayesian polarization offered here are not exhaustive – there are other scenarios where polarization may be possible. Indeed, Jern et al. (2014) show that belief polarization can occur in two other kinds of network, which future work can investigate. Therefore, it should be noted that a case of polarization which cannot be modelled as arising from either Scenario #1 or #2 might still be Bayesian.

My hope is that future work will build on the framework established here to provide a clear, exhaustive, and empirically-validated explanation of why Bayesians polarize.

Acknowledgments

Many thanks to four anonymous reviewers for their interesting and constructive comments.

While working on this manuscript the author was funded by an ESRC DTP Studentship for his PhD (https://doi.org/10.17863/CAM.106767), a postdoc position funded by an award from the Templeton World Charity Foundation to Lee de-Wit and Jens Madsen (https://doi.org/10.54224/31453), and finally a second postdoc position funded by the Horizon Europe project as part of the Perycles project.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

The author acknowledges support by the European Union under the Horizon Europe project Perycles (Participatory Democracy that Scales).

Funded by the European Union

References

Baron, J. (2008). *Thinking and Deciding* (Fourth). Cambridge University Press.

Bartels, L. M. (2002). Beyond the Running Tally: Partisan Bias in Political Perceptions. *Political Behavior*, *24*(2), 117–150. http://www.jstor.org/stable/1558352

Benoît, J., & Dubra, J. (2019). Apparent Bias: What Does Attitude Polarization Show? *International Economic Review*, *60*(4), 1675–1703. https://doi.org/10.1111/iere.12400

Bhui, R., & Gershman, S. J. (2020). Paradoxical effects of persuasive messages. *Decision*, *7*(4), 239–258. https://doi.org/10.1037/dec0000123

Botvinik-Nezer, R., Jones, M., & Wager, T. D. (2023). A belief systems analysis of fraud beliefs following the 2020 US election. *Nature Human Behaviour*, *7*(7), 1106–1119. https://doi.org/10.1038/s41562-023-01570-4

Bullock, J. G. (2009). Partisan Bias and the Bayesian Ideal in the Study of Public Opinion. *The Journal of Politics*, *71*(3), 1109–1124. https://doi.org/10.1017/S0022381609090914

Collins, P. J., Hahn, U., von Gerber, Y., & Olsson, E. J. (2018). The Bi-directional Relationship between Source Characteristics and Message Content. *Frontiers in Psychology*, *9*, 18. https://doi.org/10.3389/fpsyg.2018.00018

Cook, J., & Lewandowsky, S. (2016). Rational Irrationality: Modeling Climate Change Belief Polarization Using Bayesian Networks. *Topics in Cognitive Science*, *8*(1), 160–179. https://doi.org/10.1111/tops.12186

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A General Structure for Legal Arguments About Evidence Using Bayesian Networks. *Cognitive Science*, *37*(1), 61–102. https://doi.org/10.1111/cogs.12004

Harris, A. J. L., Corner, A., & Hahn, U. (2013). James is polite and punctual (and useless): A Bayesian formalisation of faint praise. *Thinking & Reasoning*, *19*(3–4), 414–429. https://doi.org/10.1080/13546783.2013.801367

Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach. *Cognitive Science*, *40*(6), 1496–1533. https://doi.org/10.1111/cogs.12276

Jern, A., Chang, K. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224. https://doi.org/10.1037/a0035941

Kahan, D. M. (2016). The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It. In R. A. Scott & S. M. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences* (1st ed., pp. 1–16). Wiley. https://doi.org/10.1002/9781118900772.etrds0417

Lagnado, D. A., Fenton, N., & Neil, M. (2013). Legal idioms: A framework for evidential reasoning. *Argument & Computation*, *4*(1), 46–63. https://doi.org/10.1080/19462166.2012.682656

Lord, C. G., & Taylor, C. A. (2009). Biased Assimilation: Effects of Assumptions and Expectations on the Interpretation of New Evidence: Biased Assimilation. *Social and Personality Psychology Compass*, *3*(5), 827–841. https://doi.org/10.1111/j.1751-9004.2009.00203.x

Madsen, J. K. (2016). Trump supported it?! A Bayesian source credibility model applied to appeals to specific American presidential candidates' opinions. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 165–170). Cognitive Science Society.

Madsen, J. K. (2019). Voter Reasoning Bias When Evaluating Statements from Female and Male Political Candidates. *Politics & Gender*, *15*(02), 310–335. https://doi.org/10.1017/S1743923X18000302

Mandelbaum, E. (2019). Troubles with Bayesianism: An introduction to the psychological immune system. *Mind & Language*, *34*(2), 141–157. https://doi.org/10.1111/mila.12205

Olsson, E. J. (2013). A Bayesian Simulation Model of Group Deliberation and Polarization. In F. Zenker (Ed.), *Bayesian Argumentation* (Vol. 362, pp. 113–133). Springer Netherlands. https://doi.org/10.1007/978-94-007-5357-0_6

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman. https://doi.org/10.1016/B978-0-08-051489-5.50002-3

Plous, S. (1991). Biases in the Assimilation of Technological Breakdowns: Do Accidents Make Us Safer? *Journal of Applied Social Psychology*, *21*(13), 1058–1082. https://doi.org/10.1111/j.1559-1816.1991.tb00459.x

Young, D. J., & de-Wit, L. H. (2025). The Bias-and-Expertise Model: A Bayesian Network Model of Political

Source Characteristics. *Cognitive Science*, *49*(11), e70141. https://doi.org/10.1111/cogs.70141

Young, D. J., Madsen, J. K., & de-Wit, L. H. (2025). Belief polarization can be caused by disagreements over source independence: Computational modelling, experimental evidence, and applicability to real-world politics. *Cognition*, *259*, 106126. https://doi.org/10.1016/j.cognition.2025.106126

Young, D. J., Madsen, J. K., & Yousefi, S. (2023). The Epistemic Weight of Silence. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*. https://escholarship.org/uc/item/57h4j672