The role of source reliability in belief polarisation[*]

Leah Henderson · Alexander Gebharter

Abstract: Psychological studies show that the beliefs of two agents in a hypothesis can diverge even if both agents receive the same evidence. This phenomenon of belief polarisation is often explained by invoking biased assimilation of evidence, where the agents' prior views about the hypothesis affect the way they process the evidence. We suggest, using a Bayesian model, that even if such influence is excluded, belief polarisation can still arise by another mechanism. This alternative mechanism involves differential weighting of the evidence arising when agents have different initial views about the reliability of their sources of evidence. We provide a systematic exploration of the conditions for belief polarisation in Bayesian models which incorporate opinions about source reliability, and we discuss some implications of our findings for the psychological literature.

1 Introduction

Suppose two people, call them Alice and Bob, are members of a jury that has been appointed in order to decide on whether the defendant in a murder case is guilty. They must assess the hypothesis that the defendant is guilty. During the trial Alice and Bob are confronted with a number of pieces of evidence which tell either in favour or against this hypothesis. For example, they see a police report that a weapon such as was used in the murder was found in the defendant's house. This constitutes positive evidence for the hypothesis that the defendant is guilty. They also see forensic evidence which shows that the DNA traces left on the body do not match the DNA of the defendant. This is negative evidence which tells against the defendant's guilt. Suppose that Alice starts out more confident than Bob that the defendant is guilty, and after seeing the evidence, Alice becomes even more confident that the defendant is guilty, and Bob becomes even less sure that he is. This is a case of 'belief polarisation'. In belief polarisation, two individuals respond to the same evidence, but the result is not greater agreement, but more divergence in opinion.

There are experiments in psychology which arguably show that belief polarisation does occur. In some cases, belief polarisation has been seen on a single piece of evidence (Batson, 1975; Cook and Lewandowsky, 2016), but in a number of cases the evidence is of a mixed character (Lord et al., 1979; Plous, 1991). That is, part of the evidence supports the hypothesis in question, whereas part goes against it. In the most-cited study of this kind, people with differing prior views about the effectiveness of the death penalty as a crime deterrent were asked to read two fictional studies, one of which supported the idea that the death penalty is an effective crime deterrent, and the other which supported the idea that it is not (Lord et al., 1979). The study purported to show belief polarisation of the participants, though the experimental methodology has been subjected to later critique (Miller et al., 1993; Kuhn and Lao, 1996).

[*] This is a draft paper. Please do not cite or quote without permission. The final publication is forthcoming under the following bibliographical data: Henderson, L., & Gebharter, A. (in press). The role of source reliability in belief polarisation. Synthese.

In the psychological literature, polarisation has often been taken to arise as a result of some form of 'biased assimilation' of the evidence. This means that the way the evidence is taken up or processed is biased in some fashion by the prior views of the subject. However, in order to be precise about the notion of 'bias', it needs to be contrasted with some normative understanding of what would be an unbiased way to assimilate evidence. Clearly it is not wrong for prior opinions to play some role in belief updating. The question is whether they are, in cases of belief polarisation, playing too much of a role, or playing the wrong kind of role. In order to gain a notion of what 'unbiased' could mean, we can turn to normative models provided by Bayesianism. In Bayesian updating, prior opinion is combined with evidential information in a manner which is well-motivated by various normative arguments. Deviation from Bayesian updating may then potentially indicate that the agent has assimilated the evidence in a manner which has given too much weight to their prior beliefs.

An interesting question that then arises is whether belief polarisation actually has to be attributed to biased assimilation, or whether it can occur given the normatively correct method of updating specified by Bayesianism. Jern et al. (2014) have shown that belief polarisation can occur when two agents with different prior beliefs not just about the hypothesis in question but also about other factors update on the same evidence according to Bayesian conditionalisation. This suggests that belief polarisation should not necessarily be attributed to biased assimilation on the part of one or more of the parties involved. However, it raises the question of whether a more specific kind of explanation of belief polarisation might be possible, if we restrict attention to particular kinds of other factors which are involved in belief updating.

A plausible candidate for a more specific kind of explanation emerges when we consider the 'group polarisation' we see in society on a number of important topics. Group polarisation occurs when the beliefs held by members of subgroups in society diverge from those of other subgroups, despite exposure to the same evidence. For example, on a number of scientific issues, notably anthropogenic climate change, public opinion is sharply divided in the presence of shared evidence which is not disputed by experts. Also within the scientific community, different groups may respond to the same evidence in ways that lead to more extreme positions (Kahan et al., 2011; O'Connor and Weatherall, 2017). It is often striking that when a group polarises, individuals diverge not only in their attitudes towards specific propositions, but also on their views regarding the reliability of sources of information. We see, for example, those who hold certain opinions about climate change also tending to trust different news sources. It is natural to expect that agents have beliefs not only about specific hypotheses about which they may disagree, but also about the reliability of their sources of information. This prompts the question that we will address in this paper: could it be the case that belief polarisation can result from normal Bayesian updating of both attitudes about hypotheses and attitudes about reliabilities?

In order to explore this question, we examine simple Bayesian models which represent proper, non-biased assimilation of evidence and how it impacts our probabilities for hypotheses and for reliability of sources. We call these 'source reliability models'. These models are also found in the work of Bovens and Hartmann (2003),[1] where they have been extensively studied in relation to how they behave when the evidence from multiple sources agrees. We will focus rather on the cases where the evidence is mixed, as it is in the Lord et al. (1979) study of belief polarisation. Some of the key results are the following. We show that polarisation can arise in these models, under certain circumstances. This means that belief polarisation can under these conditions in principle be produced by normal Bayesian updating on hypotheses and reliabilities, without any

[1] See also (Merdes et al., 2020).

biased assimilation of evidence occuring. We find, however, that belief polarisation cannot arise simply because of a difference in pre-existing attitude about a hypothesis unless it is accompanied by different expectations regarding the reliability of the sources. On the other hand, a difference in prior expectations about the reliability of sources is sufficient to produce polarisation, even without any difference in initial attitude towards the hypothesis in question.

Before getting to the main results, we will first, in sections 2 and 3, explain what belief polarisation looks like in a Bayesian context, and then introduce source reliability models. Section 4 contains the main results. In section 5, we discuss the implications of these findings in relation to the existing literature on belief polarisation.

2 Belief polarisation in a Bayesian context

What does belief polarisation mean in a Bayesian context? In a Bayesian framework, an agent's degrees of belief are represented by probability distributions over random variables.[2] In the jury example, suppose the probability distributions for Alice and Bob are $p_A(\cdot)$ and $p_B(\cdot)$ respectively, where the probabilities may in each case be defined over a number of random variables. For example one of the random variables, H, could represent the hypothesis that the accused is guilty. This variable can then take two values: H meaning the accused is guilty, and ¬H meaning the accused is not guilty. Throughout this paper we will follow this convention, denoting random variables by italics, and the values of the variables in non-italic font. We may then consider receiving the value of an evidence variable E, such as a report of DNA testing. If this evidence variable also has two possible values, we will denote these as E and ¬E. Given the evidence, each of Alice and Bob update their prior probability for H to a posterior for H. Alice updates her prior $p_A(\mathrm{H})$ to the posterior $p_A(\mathrm{H}|E)$, and Bob updates his prior $p_B(\mathrm{H})$ to $p_B(\mathrm{H}|E)$. We denote the difference between Alice's posterior probability and her prior probability for H by $\Delta_A^{\mathrm{H}} = p_A(\mathrm{H}|E) - p_A(\mathrm{H})$ and the difference between Bob's posterior probability and his prior probability for H by $\Delta_B^{\mathrm{H}} = p_B(\mathrm{H}|E) - p_B(\mathrm{H})$.

In some cases, Alice and Bob's probabilities both move in the same direction—that is, Δ_A^{H} and Δ_B^{H} have the same sign. Following Jern et al. (2014), we call this *parallel updating*. In this case, Alice and Bob either both increase their probabilities (see Figure 1(a)), or they both decrease their probabilities, in the light of the evidence. Another possibility is *contrary updating*, where Alice and Bob update their probabilities in different directions. Alice may revise her probability to a lower value ($\Delta_A^{\mathrm{H}} < 0$), whilst Bob revises his to a higher value ($\Delta_B^{\mathrm{H}} > 0$), or vice versa. Contrary updating may be either *convergent*, where the beliefs of the two agents about the defendant's guilt come closer together as a result of the updating, or *divergent*, where the beliefs of the two agents move apart from one another. Convergent updating happens, for example, when Alice starts with a higher prior and revises her probability down after updating ($\Delta_A^{\mathrm{H}} < 0$), whilst Bob starts with a lower prior and revises his probability upwards ($\Delta_B^{\mathrm{H}} > 0$) (see Figure 1(b)). Divergent updating can happen when, for example, Alice starts with a higher prior and revises her probability upwards ($\Delta_A^{\mathrm{H}} > 0$), whilst Bob starts with a lower prior and revises his probability downwards ($\Delta_B^{\mathrm{H}} < 0$) (see Figure 1(c)). Belief polarisation then can be thought of as divergent contrary updating.

A criterion for belief polarisation can be found in terms of likelihood ratios. The posterior probability for a hypothesis H given evidence E can be written as

$$p(\mathrm{H}|E) = \frac{h}{h + \bar{h}\, l}$$

[2] Or probability densities in the case of hypotheses concerning continuous random variables.

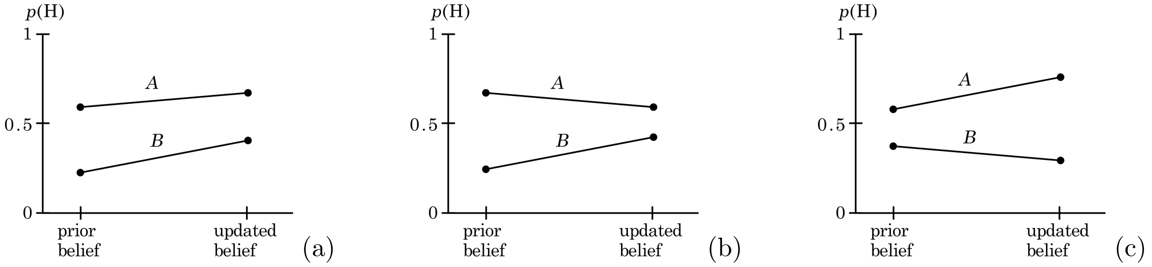


Figure 1: Different possibilities for two agents A and B to update their beliefs in a hypothesis H after collecting evidence: (a) parallel updating, (b) convergent updating, and (c) divergent updating. While both (b) and (c) are subspecies of contrary updating, only (c) constitutes a case of belief polarisation.

where $l = \frac{p(E|\neg \mathrm{H})}{p(E|\mathrm{H})}$ is the likelihood ratio. For convenience we use the notation h to represent the prior $p(\mathrm{H})$, and $\overline{h}$ denotes $1 - h$. It can be seen from this expression that the likelihood ratio tells us the direction of the belief update. If the likelihood ratio is greater than one, the posterior probability is lower than the prior, so the agent's probability for H goes down ($\Delta^{\mathrm{H}} < 0$). On the other hand, if the likelihood ratio is less than one, the agent's probability for H goes up ($\Delta^{\mathrm{H}} > 0$). If the likelihood ratio is exactly one, the prior probability and the posterior probability are equal and there is no change ($\Delta^{\mathrm{H}} = 0$). Thus, the condition for contrary updating is that Alice's likelihood ratio for H is greater than one, and Bob's is less than one (or vice versa). Let the prior for Alice on H be denoted by h_A, and the likelihood ratio for Alice on H be denoted by l_A, with similar notation for Bob. Then, divergent contrary updating, or belief polarisation, happens when Alice starts with a lower prior, and revises downwards ($\Delta^{\mathrm{H}}_{A} < 0$, likelihood ratio greater than one), whilst Bob starts with a higher prior and revises upwards ($\Delta^{\mathrm{H}}_{\mathrm{B}} > 0$, likelihood ratio less than one), or vice versa, switching roles for Alice and Bob. That is, belief polarisation occurs either when

$$l_B < 1 < l_A \text{ and } h_A \leq h_B$$

or when

$$l_A < 1 < l_B \text{ and } h_B \leq h_A$$

(Jern et al., 2014; Nielsen and Stewart, 2019).

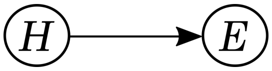
This is a general criterion which applies not only when E and H are the only variables under consideration, but also in the more typical situation where the agents have probabilistic opinions about other variables as well. In that case, the likelihoods for H would be determined in the usual way by marginalising out over the additional variables. As a simple example, if $p_A(\cdot)$ and $p_B(\cdot)$ are probability distributions concerning the variables E, H, and an additional variable V, then the likelihood for H would be calculated as $p(E|H) = \sum_V p(E|H,V)\, p(V)$. In realistic situations, people may have opinions about many variables. In cases where two parties have very different views about the relationships between other variables, it is not difficult to find situations where their opinions on a particular hypothesis may diverge from one another given the same evidence (an example is given in Jern et al. (2014) on p. 209). However, in many situations, the parties involved do not have completely divergent world-views, but rather have the same basic understanding of how the basic elements of the situation connect with one another. In such cases, one can ask whether it is still possible to have belief polarisation due to more limited divergences in prior belief between the two parties. Thus, in order to formulate precisely the question of when interesting belief polarisation can take place, we need a more precise characterisation of what should count as common ground between the agents, and where their prior opinions may differ.

Jern et al. (2014) suggest a way in which to spell out what it means for two agents to agree on the basic structure of a situation. This can be done in terms of Bayesian network models. It is well-known that joint probability distributions $p(\cdot)$ can be conveniently represented using Bayesian network models. A Bayesian network model consists of two elements: a graph and a parametrisation of that graph. In the graph, all the variables are represented by nodes. Some of the nodes are connected by arrows. Intuitively, an arrow from X to Y can be thought of as indicating that the variable X has a direct influence on the variable Y. If there is an arrow from X to Y, X is called a 'parent' of Y, and Y is called a 'child' of X. The graph must be 'acyclic', meaning that it is not possible to go in a cycle by following arrows. Thus, it is called a 'directed acyclic graph' (or 'DAG'). The DAG represents the probabilistic independencies between the variables in the joint probability distribution, given a specific precise condition.[3] However, there may be multiple probability distributions over the whole set of variables with the same set of independencies. Thus possession of a particular graph structure does not uniquely correspond to a given probability distribution. To fully specify a particular probability distribution, we add 'parameters' to the graph. These parameters specify conditional probability tables for all the nodes, given their parents. We also specify prior probabilities for the 'root' nodes, that is, those nodes with no parents.

The basic idea in Jern et al. (2014) is to say that two agents agree on the basic structure of a situation when they agree on the relevant variables in question, as well as on the Bayesian network structure and on the conditional probability tables that specify the relationships between the variables. The agents may still differ, however, in their prior beliefs regarding the values of the root nodes in the Bayesian network. In this paper, we follow Jern et al. in making these assumptions about the common ground that our agents share, and when we talk about whether belief polarisation is possible, it is under these conditions.

Suppose, as a simple example, the only variables included in the model are E and H, and the agents agree that the appropriate Bayesian network is the one depicted in Figure 2. Roughly this encodes the idea that the truth of the hypothesis affects the truth of the evidence, but not the other way around. In this simple case, this can be thought of simply as part of what it means for E to serve as evidence for H. We also assume that the two agents Alice and Bob agree on the conditional probability table (a) shown in Figure 2. However, they may have different priors: $p_A(\mathrm{H})$ may be different from $p_B(\mathrm{H})$. Since the likelihood ratio for H depends on $p(E|\mathrm{H})$ and $p(E|\neg\mathrm{H})$, Alice and

[3] The precise condition is the 'Markov condition'. More formally, a DAG is a graph $\langle \mathbf{V}, \mathbf{E} \rangle$, where $\mathbf{V}$ is a set of random variables V_i and $\mathbf{E}$ is an asymmetric binary relation on $\mathbf{V}$. We graphically represent $\langle V_i, V_j \rangle \in \mathbf{E}$ as $V_i \longrightarrow V_j$. $V_i, V_j \in \mathbf{V}$ are called adjacent if either $V_i \longrightarrow V_j$ or $V_j \longrightarrow V_i$. If each node V_i (with $i > 1$) in a tuple $\langle V_1, ..., V_n \rangle$ is adjacent to V_{i-1} and no V_i (with $1 < i < n$) appears more often than once in $\langle V_1, ..., V_n \rangle$, then $\langle V_1, ..., V_n \rangle$ is called a path between V_1 and V_n. A directed path from V_i to V_j is a path which takes the form $V_i \longrightarrow ... \longrightarrow V_j$. DAGs are acyclic, meaning that they do not feature a directed path $V_i \longrightarrow ... \longrightarrow V_i$. The set of a variable V_j's direct ancestors in the graph, i.e., the set of all V_i with $V_i \longrightarrow V_j$, is denoted by $\mathbf{Par}(V_j)$. Its elements are referred to as V_j's *parents*. The set of *decendants* of a variable V_j contains V_j itself as well as any node V_k such that there is a directed path from V_j to V_k. We denote the set of descendants of V_j by $\mathbf{Des}(V_j)$. Then the *Markov condition* is satisfied by a DAG $\langle \mathbf{V}, \mathbf{E} \rangle$ and a probability distribution $p(\cdot)$ over $\mathbf{V}$ iff every node in $\mathbf{V}$ is probabilistically independent of its non-descendants conditional on its parents, i.e., every node $V_j \in \mathbf{V}$ is independent of $\mathbf{V} \backslash \mathbf{Des}(V_j)$ conditional on $\mathbf{Par}(V_j)$. The Markov condition can also be interpreted as a principle characterising causal structures (Spirtes et al., 1993; Pearl, 2000). For a philosophical justification of its causal interpretation, see (Gebharter, 2017b; Schurz and Gebharter, 2016). For other realistic interpretations see, for example, (Schaffer, 2016; Gebharter, 2017a,c, 2019). In this paper, however, we refrain from any such realistic interpretation.



H	$p(\mathrm{E}\vert H)$
0	0.3
1	0.8

(a)

H	$p(\mathrm{E}\vert H)$
0	0.8
1	0.3

(b)

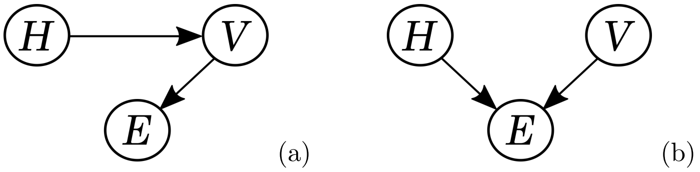
Figure 2: Simple Bayesian network model and two different example conditional probability tables. In the first column, a zero corresponds to H taking the value ¬H (i.e., H is not true), and a one corresponds to H taking the value H (i.e., H is true).

Bob have the same likelihood ratio as one another, and thus they agree on the direction of the update given evidence E. This means that contrary updating and particularly belief polarisation can never occur in such a model under the assumptions we are making. A simple case which would fit this structure is one where H is a variable representing whether or not a patient has a disease, and E is a variable with two values corresponding to a positive or negative test result. The conditional probability table in Figure 2(a) encodes the understanding that when the disease is present, it is more likely that the patient will present a positive test result than if the disease is not present.

Notice here that if the agents disagree about the conditional probability table, for example if Alice takes it to be the table in Figure 2(a) and Bob takes it to be the table in Figure 2(b), then this would express a very different understanding of the significance of the test. Alice takes a positive test result as evidence that the patient has the disease, whereas Bob thinks the test is now more likely to give a positive result if the disease is not present than if it is. Alice will increase her probability that the patient has the disease when she sees a positive test result (Alice's likelihood ratio is then less than one), whereas Bob decreases his probability that the patient has the disease when he sees a positive test result (his likelihood ratio is greater than one). This kind of fundamental disagreement over the conditional probability table is ruled out by our assumptions.

When the model includes three or more variables, whether or not belief polarisation can occur depends on the structure of the Bayesian network. Some graph structures never allow it, whereas some do, under certain parameter settings. Jern et al. (2014) provide a complete classification of all the three-variable networks in terms of whether they allow for belief polarisation or not. As a simple example, suppose the structure is that depicted in Figure 3(a). In this case, the likelihood ratio for H is $l = \frac{p(E|\mathrm{H})}{p(E|\neg\mathrm{H})} = \frac{\sum_V p(E|V)\,p(V|H)}{\sum_V p(E|V)\,p(V|\neg H)}$, which does not depend on the prior for H. Since H is the only root node, agents will always agree on the likelihood ratio, and so there can be no belief polarisation on H.

On the other hand, one of the structures which does allow for belief polarisation under certain conditions is Figure 3(b). In this case, the likelihood ratio is $l = \frac{p(E|\mathrm{H})}{p(E|\neg\mathrm{H})} = \frac{\sum_V p(E|V,\mathrm{H})\,p(V)}{\sum_V p(E|V,\neg\mathrm{H})\,p(V)}$. It is noteworthy that this likelihood ratio does not depend on the prior for the hypothesis $p(\mathrm{H})$, but does depend on the prior for the additional variable $p(V)$. The somewhat counterintuitive fact that the likelihood ratio depends on the prior for V but not for H will also be an important feature of the model we will present in section 3. Belief polarisation, then, is possible for some cases where Alice and Bob have different priors for V. Suppose, for example, that H represents whether the patient has a disease, and E is a test result as before. However, now there is a further variable which can influence the test result, namely the patient's blood sugar level, which we represent by the variable V (see Jern et al. (2014), pp. 208f). We assume that V, like H and E, is a binary variable: the patient's blood sugar level can be either high (V) or low (¬V), for example. Now let us suppose that a positive test result is more probable when the patient has the disease and has a high blood sugar level, but it is also quite likely when the patient does not have the disease but has a low blood



	p(H)	p(V)
Alice	0.8	0.8
Bob	0.2	0.2

H	V	$p(\mathrm{E}\|H,V)$
0	0	0.8
0	1	0.3
1	0	0.1
1	1	0.9

(c)

Figure 3: (a) A structure involving the three variables H, E, V which does not allow for belief polarisation; (b) a structure of the three variables which does allow belief polarisation for certain parameter settings, such as those shown in the table (c).

sugar level. Now suppose doctor Alice is fairly confident that the patient has the disease and also thinks that the patient has a high blood sugar level, then she becomes even more convinced that the patient has the disease when she sees a positive test result. If, on the other hand, doctor Bob is doubtful whether the patient has the disease, and also is inclined to think that the patient has a low blood sugar level, then he may become more convinced that the patient does not have the disease, given a positive test result. This situation could happen for example when the parameters are chosen as in the probability table shown in Figure 3(c). In this case, Alice's probability for H increases from her prior probability 0.8 to a posterior probability of 0.88, whilst Bob's probability for H decreases from his prior probability of 0.2 to a posterior probability of 0.08. Thus we see belief polarisation.

In summary then, belief polarisation in Bayesian models can be identified by looking at the likelihood ratios of each of the parties involved for a particular hypothesis. In general, these likelihood ratios depend on all the other opinions that the agent holds and which can be represented in a Bayesian network. There are some network structures which allow for belief polarisation, even in cases where the agents agree on the basic structure of the network. In these cases, the polarisation arises because of a difference in prior probabilities assigned to the root nodes of the network.

3 Source reliability models

In this paper, we are interested in the role of source reliability in belief polarisation. We therefore consider a special class of Bayesian network models, namely those which explicitly include nodes for not only hypotheses of interest, but also for the reliability of sources of evidence bearing on those hypotheses. In such a model, an agent is updating a joint assignment of degrees of belief for both hypotheses and source reliabilities. Several different Bayesian models of source reliability have been developed in the literature (Olsson, 2013; Bovens and Hartmann, 2003; Merdes et al., 2020). We focus here on the type of model presented by Bovens and Hartmann (2003). Bovens and Hartmann have explored a number of properties of these models, including the conditions for when agreement between multiple pieces of evidence boosts confirmation for a hypothesis. What we will

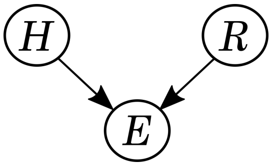
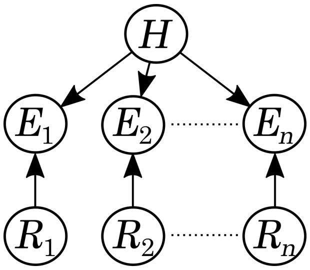


Figure 4: Simple source reliability model

do here is to examine the effects of mixed evidence, where different pieces of evidence give differing indications about whether the hypothesis in question is true. These cases are of interest, since as we have seen, psychological experiments show that belief polarisation can occur in situations where the evidence is mixed. The question then is whether belief polarisation can arise in simple Bayesian source reliability models, and if so, under what conditions.

As we have seen, one of the most basic models of evidence-collecting is that depicted in Figure 2. Here it is assumed that whether or not the hypothesis is true has bearing on whether or not certain evidence will be present. A source reliability model is a simple extension of this model, which takes the evidence to come in the form of a report from a source which has some reliability represented by a variable R. The basic structure is depicted in Figure 4. A basic assumption of this model is that two factors influence what evidence reports are received—first, the truth or falsity of the hypothesis itself, and second, the reliability of the source of the report. By using this structure, we make two further assumptions: i) the hypothesis H is independent of the reliability of the source R, and ii) H and R are dependent when conditioned on the evidence E. Assumption i) makes sense when the truth of the hypothesis does not influence the reliability of the source, nor does the reliability of the source influence the truth of the hypothesis. In our legal case introduced earlier, for instance, it is clear that whether or not the police report or the forensic report are reliable has no bearing on whether the defendant committed the crime. And also, whether or not the defendant committed the crime has no influence on the reliability of the reports. The reliability of the police report is determined by factors such as the integrity and disinterestedness of the police, for which it should be irrelevant whether or not the defendant committed the crime. Similarly, the reliability of the DNA test depends on the precise nature of the test that is carried out and what its false positive and false negative rates are. Again, the guilt of the defendant should have no influence on this. The independence assumption i) means that the prior probabilities for H and R should be assignable independently of each other. The model thus does not allow the possibility of biased influence of prior opinion, in the sense that it does not allow the agent's prior views of R to depend on her prior views of H, or vice versa. Assumption ii) is also a natural one. Suppose, for example, the evidence comes from an eyewitness. As soon as the agent is presented with evidence (E or ¬E), H and R should become relevant to each other. If Alice, for example, considers a particular witness highly reliable, then she will assign a higher probability to H after hearing this witness report E than she would had she not assigned such a high degree of reliability to this particular witness. And, vice versa, had Alice, for example, already assigned a high prior to H, then she should consider a witness less reliable after hearing the witness claim ¬E.

So far, we have only considered one evidence variable, but the agent may actually receive multiple pieces of evidence, which we represent by variables $E_1, E_2, ..., E_n$. If each of these pieces of evidence comes from an independent source, the structure of the Bayesian network is as depicted in Figure 5. Each source has a reliability R_i for $1 \leq i \leq n$. Following Bovens and Hartmann (2003), we



H	R_i	$p(\mathrm{E}_i\|H, R_i)$
0	0	a
0	1	0
1	0	a
1	1	1

Figure 5: Source reliability model with n pieces of evidence from n independent sources and conditional probability table.

make the following further assumptions about the conditional probabilities in the source reliability models. We take H, E_i, and R_i to be all binary variables. When the source is reliable ($R_i = 1$, which we denote R_i), the evidence that it produces perfectly discriminates between the truth and falsity of the hypothesis. That is, $p(\mathrm{E}_i|\mathrm{H}, \mathrm{R}_i) = 1$ and $p(\mathrm{E}_i|\neg\mathrm{H}, \mathrm{R}_i) = 0$. On the other hand, when the source is unreliable ($R_i = 0$, denoted $\neg\mathrm{R}_i$), it is indifferent to the truth or falsity of the hypothesis, and merely acts like a randomiser, giving a probability of positive evidence a_i regardless of whether the hypothesis is true or false. That is $p(\mathrm{E}_i|\mathrm{H}, \neg\mathrm{R}_i) = p(\mathrm{E}_i|\neg\mathrm{H}, \neg\mathrm{R}_i) = a_i$. We assume for simplicity that when each source is unreliable, it has the same randomisation parameter a, or chance of giving an incorrect report (i.e., we assume that $p(\mathrm{E}_i|\mathrm{H}, \neg\mathrm{R}_i) = p(\mathrm{E}_j|\mathrm{H}, \neg\mathrm{R}_j) = a$ for all i, j with $1 \leq i, j \leq n$). The conditional probability table for each evidence node is thus specified as in the table in Figure 5. As before, we will use the following notation for the priors: h will denote $p(\mathrm{H})$, ρ_i will denote $p(\mathrm{R}_i)$. We write $\overline{h}$ for $p(\neg\mathrm{H}) = 1 - h$, and similarly $\overline{\rho}_i$ denotes $p(\neg\mathrm{R}_i) = 1 - \rho_i$.

4 Results

In this section we present our findings regarding the source reliability models shown in Figures 4 and 5. We find that belief polarisation is not possible on the hypothesis in such models given only one piece of evidence. However, given multiple pieces of evidence, belief polarisation can arise due to differential weighting of the evidence produced by differences in the priors on reliability of the sources of evidence.

4.1 Simple model with one piece of evidence

4.1.1 Updating probability for hypothesis

First, let us consider a simple model where there is just one piece of evidence E (see Figure 4). In this case, there is no polarisation on H. This can be seen by considering the likelihood ratio for H, given a positive report E, which can be computed as:

$$
\begin{aligned}
l^{+} &= \frac{p(\mathrm{E}|\neg\mathrm{H})}{p(\mathrm{E}|\mathrm{H})} \\
&= \frac{\sum_R p(\mathrm{E}|\neg\mathrm{H}, R)\, p(R)}{\sum_R p(\mathrm{E}|\mathrm{H}, R)\, p(R)} \\
&= \frac{a\,\overline{\rho}}{a\,\overline{\rho} + \rho} \\
&= \frac{1}{1 + \frac{\rho}{a\,\overline{\rho}}} \qquad (1)
\end{aligned}
$$

It is clear that this expression is always less than one, since $\frac{\rho}{a\,\overline{\rho}}$ is positive. This means that the posterior probability for H always increases, given a piece of positive evidence E. Furthermore, we can see that the size of the update is governed by $\frac{\rho}{a\,\overline{\rho}}$ for a fixed h. Δ^{H} is greater when ρ is larger and/or when a is smaller. When ρ is larger, the agent has initially more trust in the reliability of the source, and thus is more responsive to what it says. The update is also larger when a is small, because this means that the chance that the positive report arises because the source is actually unreliable but erroneously gives a positive report is small.

Similar calculations show that when the evidence is negative ¬E, the likelihood ratio for H is:

$$
\begin{aligned}
l^{-} &= \frac{\overline{a}\,\overline{\rho} + \rho}{\overline{a}\,\overline{\rho}} \\
&= 1 + \frac{\rho}{\overline{a}\,\overline{\rho}} \qquad (2)
\end{aligned}
$$

Since $\frac{\rho}{\overline{a}\,\overline{\rho}}$ is positive, this is always greater than one. This means that the posterior probability for H always decreases, given a piece of negative evidence ¬E. Furthermore, the size of the update Δ^{H} for a given h is larger when the source is initially more trusted (high ρ). The update is also larger when a is large, because this means that the chance that the negative report arises because the source is actually unreliable but erroneously gives a negative report is small.

No matter what priors Alice and Bob start with then, either they both update to a higher posterior probability ($\Delta^{\mathrm{H}}_A > 0$ and $\Delta^{\mathrm{H}}_B > 0$), when the piece of evidence is positive, or they both update to a lower posterior ($\Delta^{\mathrm{H}}_A < 0$ and $\Delta^{\mathrm{H}}_B < 0$) when the piece of evidence is negative. Thus, there can be no belief polarisation with respect to the hypothesis given a single piece of evidence.

Since the likelihood ratio for H does not depend at all on the prior h for the hypothesis, the prior h has no effect on the direction of the update. If Alice and Bob have different priors for H, they will still update in exactly the same direction regardless of whether they have different priors for the reliability of the source, though the size of their update Δ^{H} depends on their individual priors h_A and h_B as well as on how reliable they consider the source to be.

4.1.2 Updating probability for reliability

It is also of interest to consider how the agents update their probabilities for the reliability of the source. Again, the update is determined by the likelihood ratio. For a positive piece of evidence E

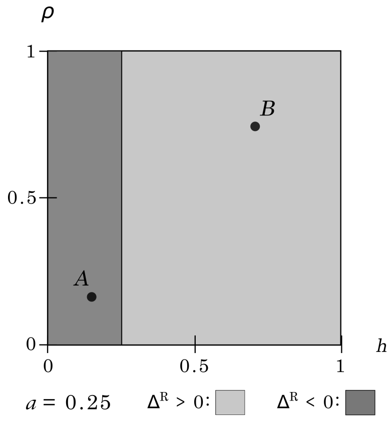


Figure 6: Each agent has a prior h and a prior ρ. All possible choices of priors are displayed on the $h - \rho$ plane. The directions of updates are shown, given one piece of positive evidence E. In the region where $h < a$, $\Delta^{\mathrm{R}} < 0$. In the region where $h > a$, on the other hand, $\Delta^{\mathrm{R}} > 0$. Belief polarisation for reliability can happen when, for example, Alice's prior is chosen as point A and Bob's as point B. In such a case, Alice starts from a lower prior ρ_A and her probability for R decreases. Bob starts from a higher prior ρ_B and his probability for R increases.

this is given by:

$$\begin{aligned} r^{+} &= \frac{p(\mathrm{E}|\neg\mathrm{R})}{p(\mathrm{E}|\mathrm{R})} \\ &= \frac{\sum_H p(\mathrm{E}|\neg\mathrm{R}, H)\, p(H)}{\sum_H p(\mathrm{E}|\mathrm{R}, H)\, p(H)} \\ &= \frac{a\, h + a\, \overline{h}}{h + 0\, \overline{h}} \\ &= \frac{a}{h} \end{aligned}$$

Thus, the likelihood ratio for the reliability does depend on the prior h, but it does not depend on the prior ρ. If $h < a$, the likelihood ratio for a positive piece of evidence will be greater than one, and hence the posterior for reliability is lower than the prior, $\Delta^{\mathrm{R}} < 0$. If $h > a$, on the other hand, $\Delta^{\mathrm{R}} > 0$. This is illustrated in Figure 6. This makes sense if we think that someone who initially sees the probability of H as low will take a positive report E as an indication that the source is unreliable. Whereas, if the probability of H is initially high, a positive report will reinforce the view that the source is reliable.

For a negative piece of evidence ¬E, the likelihood ratio is:

$$\begin{aligned} r^- &= \frac{p(\neg\mathrm{E}|\neg\mathrm{R})}{p(\neg\mathrm{E}|\mathrm{R})} \\ &= \frac{\sum_H p(\neg\mathrm{E}|\neg\mathrm{R}, H)\, p(H)}{\sum_H p(\neg\mathrm{E}|\mathrm{R}, H)\, p(H)} \\ &= \frac{\overline{a}\, h + \overline{a}\, \overline{h}}{0\, h + 1\, \overline{h}} \\ &= \frac{\overline{a}}{\overline{h}} \end{aligned}$$

In this case, if $h < a$, the posterior probability for the reliability increases, and if $h > a$, the posterior probability for the reliability decreases. If someone thinks that the probability of H is very low, then a piece of negative evidence is taken as an indication that the source is reliable. Whereas, if someone thinks that the probability of H is high, then a piece of negative evidence is taken to indicate that the source is not so reliable.

Thus, in this model, even though there cannot be polarisation regarding the hypothesis itself, polarisation regarding the reliability of the source is possible. This could, for example, happen in the legal case if Alice has a lower prior for the reliability of a source of evidence than Bob, $\rho_A < \rho_B$, and she also thinks that the prior probability that the accused is guilty is low ($h_A < a$), while Bob considers it to be high ($h_B > a$). In that case, a positive piece of evidence will make Alice think the source is even less reliable ($\Delta_A^{\mathrm{R}} < 0$), whereas Bob will think it is more reliable ($\Delta_B^{\mathrm{R}} > 0$).

4.2 Multiple pieces of evidence from independent sources

In some of the canonical experiments on belief polarisation, the subjects were presented with mixed evidence—that is, multiple pieces of evidence where some of the evidence appears to be in favour of the hypothesis, whereas some goes against it. As we saw in section 3, we can represent this as a situation where the agents are presented with n pieces of evidence $E_1, ..., E_n$, all coming from different and independent sources.[4] For example, in a legal case, the jury might receive a police report, a report from forensic investigation, and eyewitness reports. In principle, each of these reports should be independent of the others. The graphical structure of a source reliability model representing these cases is depicted in Figure 5. This structure together with the Markov condition guarantees the independence of the variables $R_1, ..., R_n$ and, thus, the independence of the n sources. Again, we assume the probabilistic constraints in the table in Figure 5 in order to keep things simple and to guarantee that the variables $R_1, ..., R_n$ represent the reliability of the different sources assigned by the agents.

4.2.1 Updating probability for hypothesis

The direction of updating on the probability for the hypothesis H is determined, as we saw in section 2, by the likelihood ratio:

$$l = \frac{p(E_1, E_2, ..., E_n|\neg\mathrm{H})}{p(E_1, E_2, ..., E_n|\mathrm{H})}$$

[4] In this paper we do not consider the case of sequential pieces of evidence from the same source. Models which take account of this kind of updating can be found in Olsson (2013), Hahn et al. (2018) and Merdes et al. (2020).

On the basis of the DAG in Figure 5, the denominator $p(E_1, E_2, ..., E_n|\mathrm{H})$ is

$$p(E_1, E_2, ..., E_n|\mathrm{H}) = \prod_i \sum_{R_j} p(E_i|R_j, \mathrm{H})\, p(R_j)$$

and likewise for the numerator $p(E_1, E_2, ..., E_n|\neg\mathrm{H})$. Thus, the likelihood ratio for H factorises as

$$\begin{aligned} l &= \prod_i l_i \\ &= \prod_+ l_i^+ \prod_- l_i^- \end{aligned} \tag{3}$$

where l_i is the likelihood ratio for observation E_i. This is a product of the likelihood ratios for the positive pieces of evidence (for which $l_i = l_i^+ = \frac{p(\mathrm{E_i}|\neg\mathrm{H})}{p(\mathrm{E_i}|\mathrm{H})} = \frac{1}{1+\frac{\rho_i}{a\,\bar{\rho}_i}}$) and the likelihood ratios for the negative pieces of evidence (for which $l_i = l_i^- = \frac{p(\neg\mathrm{E_i}|\neg\mathrm{H})}{p(\neg\mathrm{E_i}|\mathrm{H})} = 1 + \frac{\rho_i}{\bar{a}\,\bar{\rho}_i}$).

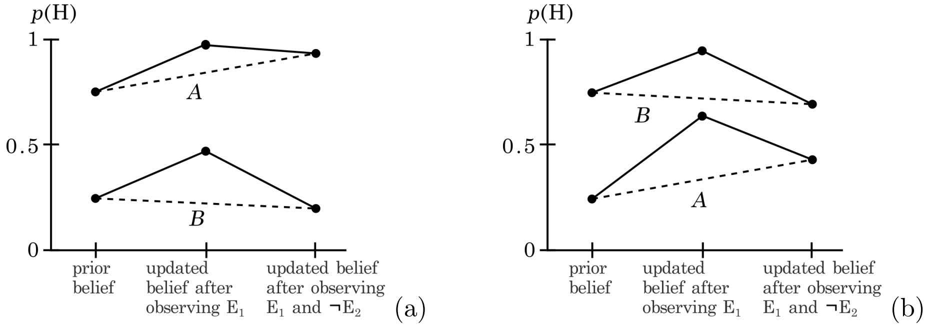
As in the model in subsection 4.1, the likelihood ratio does not depend at all on the prior h for the hypothesis H. It does, however, depend on the priors for the reliabilities of the sources, namely ρ_i. If Alice and Bob assign different priors for the reliability of sources, they may update to different extents on each piece of evidence. In some situations this may give rise to belief polarisation. Suppose, for example, that Alice and Bob receive two pieces of evidence. Alice starts with a higher prior for H than Bob. She also assigns a higher prior to the reliability of the source of the first piece of evidence than to the source of the second piece of evidence. Bob, on the other hand, assigns a higher prior to the reliability of the source of the second piece of evidence than the first. Suppose Alice and Bob now receive a positive piece of evidence from the first source and a negative piece of evidence from the second. Then because Alice initially trusts the first source more than the second, she is more responsive to the positive evidence than the negative. The overall effect of updating on the mixed evidence is that Alice's probability increases, $\Delta_A^{\mathrm{H}} = p_A(\mathrm{H}|\mathrm{E_1}, \neg\mathrm{E_2}) - p_A(\mathrm{H}) > 0$. Bob, on the other hand, is more responsive to the negative evidence than to the positive evidence, and so his probability decreases, $\Delta_B^{\mathrm{H}} = p_B(\mathrm{H}|\mathrm{E_1}, \neg\mathrm{E_2}) - p_B(\mathrm{H}) < 0$. Thus there is belief polarisation. Notice that given the same reliability priors, if Alice had started with a lower prior for H than Bob, the same updating can result in convergence of their posterior probabilities. Both cases are illustrated in Figure 7.

In general, the probability update Δ^{H} depends on the reliability priors for the different sources, the value of the parameter a, and the relative number of pieces of positive and negative evidence. The probability update Δ^{H} is negative when the overall contribution of the negative evidence outweighs the overall contribution of the positive evidence in making the overall likelihood ratio greater than one. The condition for this is:

$$\prod_i^- \left(1 + \frac{\rho_i}{\bar{a}\,\bar{\rho}_i}\right) > \prod_i^+ \left(1 + \frac{\rho_i}{a\bar{\rho}_i}\right)$$

Here, the product on the left hand side is over all negative pieces of evidence and the product on the right is over all positive pieces of evidence. For the special case where we have just two pieces of evidence, one positive and one negative, the likelihood ratio is

$$l^{+-} = \left(\frac{1}{1 + \frac{\rho_1}{a\,\bar{\rho}_1}}\right)\left(1 + \frac{\rho_2}{\bar{a}\,\bar{\rho}_2}\right)$$



	p(H)	$p(R_1)$	$p(R_2)$
Alice	0.75	0.6	0.4
Bob	0.25	0.4	0.6

(c)

	p(H)	$p(R_1)$	$p(R_2)$
Alice	0.25	0.6	0.4
Bob	0.75	0.4	0.6

(d)

H	R_i	$p(E_i \mid H, R_i)$
0	0	0.4
0	1	0
1	0	0.4
1	1	1

(e)

Figure 7: Assume Alice and Bob both observe E_1 and $\neg E_2$ and that they both agree on the conditional probabilities in (e). Assume further that the only difference between them is that Alice starts with a higher prior for H and assigns a higher reliability to the first source and a lower reliability to the second source than Bob does as in (c). Then this results in belief polarisation (a). If Alice and Bob switch their priors for H while keeping all the other probabilities the same—i.e., if they would now have the priors given in (d)—this would result in convergence (b).

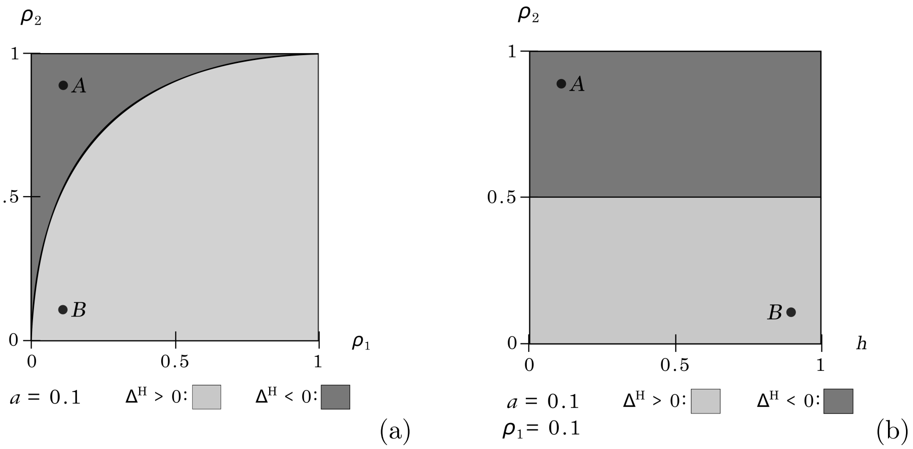


Figure 8: Direction of updates to probability for H after receiving conflicting evidence E_1 and $\neg E_2$. (a) Direction of updates to probability for H shown on the $\rho_1 - \rho_2$ plane. These do not depend on the prior for H. Belief polarisation happens if Alice starts with a lower prior for H, $h_A < h_B$, and her reliability priors fall in the $\Delta^H < 0$ region, whereas Bob's reliability priors fall in the $\Delta^H > 0$ region. An example is shown where Alice has priors at A and Bob has priors at B. (b) Direction of updates to probability for H shown on the $h - \rho_2$ plane for the case where $\rho_1 = 0.1$. A and B denote choices of priors for Alice and Bob which would give rise to belief polarisation on H.

which is greater than one when:

$$\left(1 + \frac{\rho_2}{\overline{a}\,\overline{\rho}_2}\right) > \left(1 + \frac{\rho_1}{a\overline{\rho}_1}\right)$$

This occurs when:

$$\frac{\rho_2}{\overline{\rho}_2} > \frac{\overline{a}\,\rho_1}{a\,\overline{\rho}_1}$$

For any given a, there is thus a region on the $\rho_1 - \rho_2$ plane where it is possible that the size of the update on the negative evidence is greater than the size of the update on the positive evidence. In these cases, the probability for H decreases given both pieces of evidence ($\Delta^H < 0$).

Suppose now that $h_A < h_B$. Polarisation will occur exactly when Alice's priors for reliability fall in the region where Δ^H is negative, and when Bob's priors for reliability fall in the region where Δ^H is positive. Such a case is illustrated in Figure 8.

The effect of varying a on the regions is shown in Figure 9. We see that the region where $\Delta^H < 0$ grows for higher a. This is because, as we have seen, for higher a, a positive piece of evidence has less effect on the probability update and a negative piece of evidence has more. An interesting case is where $a = 0.5$. In this case, the probability of getting positive evidence when the hypothesis is false is equal to the probability of getting negative evidence when the hypothesis is true. In this balanced situation, the probability update Δ^H is positive on exactly half of the $\rho_1 - \rho_2$ plane. Since Δ^H_A is negative in half of Alice's parameter space of reliability priors, and Δ^H_B is positive in half of Bob's parameter space, belief polarisation will occur in one quarter of the total parameter space of Alice and Bob's reliability priors. This is the maximum proportion of the parameter space on which belief polarisation can occur. Moving a away from 0.5 reduces the size of the region where

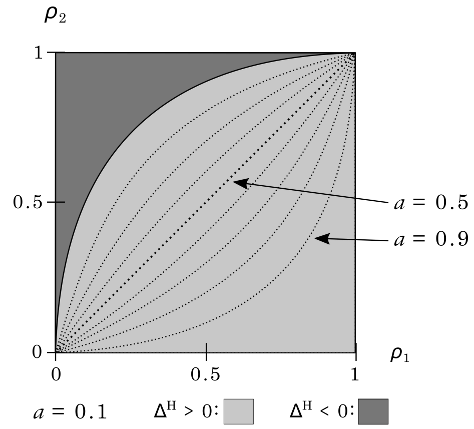


Figure 9: Effect of varying a on Δ^{H} after receiving conflicting evidence E_1 and $\neg\mathrm{E}_2$ shown on the $\rho_1 - \rho_2$ plane. Each dotted line represents an increase of a by 0.1, starting from $a = 0.1$ and up to $a = 0.9$. The higher a becomes, the smaller the region where $\Delta^{\mathrm{H}} > 0$ becomes. If $a = 0.5$, then the regions where $\Delta^{\mathrm{H}} > 0$ and $\Delta^{\mathrm{H}} < 0$ each make up exactly half of the $\rho_1 - \rho_2$ plane. Under the assumption that each possible combination of h, ρ_1, and ρ_2 is equally probable for both Alice and Bob, this is the situation where belief polarization is as likely as it can get.

belief polarisation occurs, since it produces an imbalance between the size of the $\Delta^{\mathrm{H}} < 0$ and $\Delta^{\mathrm{H}} > 0$ regions. Similarly, as we increase the amount of evidence beyond two pieces of evidence, an imbalance between the number of pieces of positive and negative evidence will also produce such an imbalance between the size of the $\Delta^{\mathrm{H}} < 0$ and $\Delta^{\mathrm{H}} > 0$ regions, and hence reduce the region in which belief polarisation occurs. In the case where the amount of evidence is increased, but the evidence remains balanced (i.e., equal numbers of pieces of positive and negative evidence), the proportion of the parameter space where belief polarisation occurs still cannot be increased beyond the maximum of one quarter.[5]

4.2.2 Updating probability for reliability

We will now look at the direction of update of the posterior probabilities for the reliabilities when there are multiple pieces of evidence. We compute the likelihood ratio for one of the reliabilities, say R_1:

$$r = \frac{p(E_1, E_2, ..., E_n|\neg\mathrm{R}_1)}{p(E_1, E_2, ..., E_n|\mathrm{R}_1)} \tag{4}$$

[5] Thus, in this model, we do not see the ‘information overload’ effect which has been found in other models, such as the model considered in Pothos et al. (2021), where considering more evidence leads to a greater chance of polarisation.

On the basis of the DAG in Figure 5, the denominator $p(E_1, E_2, ..., E_n|\mathrm{R}_1)$ can be computed as:

$$\begin{aligned}
p(E_1, E_2, ..., E_n|\mathrm{R}_1) &= \sum_{H,R_2,R_3,...,R_n} p(H)\, p(E_1|H, \mathrm{R}_1)\, p(E_2|H, R_2)\, p(R_2) \dots p(E_n|H, R_n)\, p(R_n) \\
&= \sum_H p(H)\, p(E_1|H, \mathrm{R}_1) \prod_{j=2}^{n} p(E_j|H) \\
&= h\, p(E_1|\mathrm{H}, \mathrm{R}_1) \prod_{j=2}^{n} p(E_j|\mathrm{H}) + \overline{h}\, p(E_1|\neg\mathrm{H}, \mathrm{R}_1) \prod_{j=2}^{n} p(E_j|\neg\mathrm{H}) \qquad (5)
\end{aligned}$$

A similar calculation gives the numerator:

$$p(E_1, E_2, ..., E_n|\neg\mathrm{R}_1) = h\, p(E_1|\mathrm{H}, \neg\mathrm{R}_1) \prod_{j=2}^{n} p(E_j|\mathrm{H}) + \overline{h}\, p(E_1|\neg\mathrm{H}, \neg\mathrm{R}_1) \prod_{j=2}^{n} p(E_j|\neg\mathrm{H}) \qquad (6)$$

Substituting into Equation 4 the expressions given by Equation 5 and Equation 6 thus gives the likelihood ratio:

$$\begin{aligned}
r &= \frac{h\, p(E_1|\mathrm{H}, \neg\mathrm{R}_1) \prod_{j=2}^{n} p(E_j|\mathrm{H}) + \overline{h}\, p(E_1|\neg\mathrm{H}, \neg\mathrm{R}_1) \prod_{j=2}^{n} p(E_j|\neg\mathrm{H})}{h\, p(E_1|\mathrm{H}, \mathrm{R}_1) \prod_{j=2}^{n} p(E_j|\mathrm{H}) + \overline{h}\, p(E_1|\neg\mathrm{H}, \mathrm{R}_1) \prod_{j=2}^{n} p(E_j|\neg\mathrm{H})} \\
&= \frac{h\, p(E_1|\mathrm{H}, \neg\mathrm{R}_1) + \overline{h}\, p(E_1|\neg\mathrm{H}, \neg\mathrm{R}_1) \prod_{j=2}^{n} l_j}{h\, p(E_1|\mathrm{H}, \mathrm{R}_1) + \overline{h}\, p(E_1|\neg\mathrm{H}, \mathrm{R}_1) \prod_{j=2}^{n} l_j} \qquad (7)
\end{aligned}$$

Consider the special case where $n = 2$ and suppose the first piece of evidence is positive, E_1, and the second piece of evidence is negative, $\neg\mathrm{E}_2$. Then the likelihood ratio

$$r^{+-} = \frac{p(\mathrm{E}_1, \neg\mathrm{E}_2|\neg\mathrm{R}_1)}{p(\mathrm{E}_1, \neg\mathrm{E}_2|\mathrm{R}_1)}$$

is given by substituting $l_2 = l^-$, given by the expression in Equation 2, into Equation 7:

$$r^{+-} = \frac{a}{h}\left(h + \overline{h}\,(1 + \frac{\rho_2}{\overline{a}\,\overline{\rho}_2})\right)$$

Thus the updating of the probability for R_1 depends on the prior h and the prior ρ_2, as well as a. Again, there is a region of parameter space where $\Delta^{\mathrm{R}_1} > 0$ and a region where $\Delta^{\mathrm{R}_1} < 0$. These regions are shown in Figure 10(a) for the case where $\rho_1 = 0.1$ and $a = 0.1$. There can be cases of polarisation on R_1 where Alice has priors in the $\Delta^{\mathrm{R}_1} < 0$ region and Bob has priors in the $\Delta^{\mathrm{R}_1} > 0$ region. In Figure 10(b) we show how these regions intersect with the regions where $\Delta^{\mathrm{H}} > 0$ and $\Delta^{\mathrm{H}} < 0$. The parameter space is then divided into four regions. It is then possible that Alice and Bob's probabilities for both the hypothesis and the reliability update in the same direction (when they both have priors in the same region). But it is also possible for certain choices of priors that Alice and Bob polarise on the hypothesis, or on the reliability, or both.

5 Discussion and relation to other work

How do our results relate to work in experimental psychology on belief polarisation? One of the most well-known studies of belief polarisation looked at how people updated their beliefs about

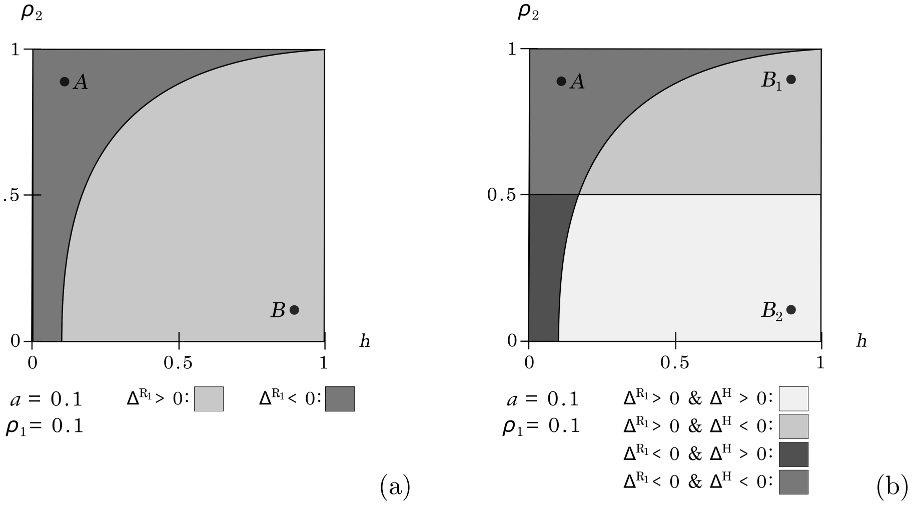


Figure 10: Dependence of update directions on priors after receiving conflicting evidence E_1 and $\neg\mathrm{E}_2$, shown on the $h-\rho_2$ plane for the case where $a = 0.1$ and $\rho_1 = 0.1$. (a) Direction of updates to probability for R_1. The plane is divided into a region where $\Delta^{\mathrm{R}_1} > 0$ and a region where $\Delta^{\mathrm{R}_1} < 0$. Belief polarisation for reliability can occur when Alice starts with a lower prior for R_1, $\rho_{1A} < \rho_{2B}$, and her reliability priors fall in the $\Delta^{\mathrm{R}_1} < 0$ region, whereas Bob's reliability priors fall in the $\Delta^{\mathrm{R}_1} > 0$ region. An example is shown where Alice has priors at A and Bob has priors at B. (b) Direction of updates to both H and R_1. Depending on how priors are chosen, Alice and Bob can polarise on H, on R_1, on both, or neither. For example, if Alice chooses A and Bob chooses B_1, there is belief polarisation on R_1, but not on H. On the other hand, if Alice chooses a prior at A and Bob chooses a prior at B_2, there is belief polarisation on both R_1 and H.

the effectiveness of the death penalty as a crime deterrent after seeing mixed evidence (Lord et al., 1979). Participants were asked to read about two fictional studies, one of which supported the idea that the death penalty is an effective crime deterrent, and the other which supported the idea that it is not. It was observed that supporters of the death penalty who already believed it to have a deterrent effect became more convinced that it was an effective crime deterrent after seeing the studies, whereas opponents who initially did not believe in the deterrent effect became more convinced that it was not an effective deterrent. The same evidence thus led to belief updates in opposite directions after seeing both studies. It was also found that the participants did respond to the individual studies in the sense that they all shifted their attitude in favour of deterrent efficacy when presented with the prodeterrent study and shifted against it when presented with the antideterrent study. However, the amount by which the opinions shifted differed between proponents and opponents. Proponents revised their opinion more than opponents after reading the prodeterrent study, and less than opponents after reading the antideterrent study.

A conclusion that has often been drawn from such experimental studies of belief polarisation is that it results from biased assimilation of the evidence presented. The key idea is that, as Lord et al. (1979) put it, ‘people tend to interpret subsequent evidence so as to maintain their initial beliefs’ (p. 1099). There have been a number of proposals concerning what the exact mechanism can be that drives this differential weighting of evidence, with some favouring more affective and others favouring more cognitive explanations. If what the agent is doing is simply discounting evidence that disagrees with their prior views on H, then this would seem to amount to a rather irrational form of dogmatism (Kelly, 2008). It may then be a case of motivated reasoning or confirmation bias (Taber and Lodge, 2006; Taber et al., 2009). Various other processes have been suggested which may not be so blatantly a case of irrational bias. It may be, for example, that prior beliefs influence how evidence is to be interpreted (Fryer et al., 2013). Or it may be that people have the tendency to scrutinise evidence which disagrees with their prior views to a greater extent than evidence that agrees with it (Lord et al., 1979; Munro and Ditto, 1997; McHoskey, 1995). Another proposal is that real human agents have bounded memories, and it may make sense to forget reasons and evidence which does not fit into a coherent picture (Singer et al., 2019). For all these theories, some kind of biased influence of the initial belief in the hypothesis on the way evidence is handled or processed is postulated.

However, as we saw in section 2, for some belief networks, differences in the priors assigned to a hypothesis H have no impact on the likelihood ratio and thus are not responsible for belief polarisation on H. Belief polarisation can nonetheless be produced by differences in priors for other variables in the agents' belief networks. Jern et al. (2014) have used this point to suggest an alternative explanation for the results in Lord et al. (1979). They propose that the Lord et al. (1979) results could also be produced, for example, by a simple network with the structure shown in Figure 3(b) (Jern et al., 2014, pp. 211f). In this model, H is a variable representing the hypothesis that the death penalty is an effective crime deterrent, E is a study which may either support the idea that the death penalty is an effective crime deterrent (E) or support the idea that it is not (¬E), and V is a variable representing the view that the consensus expert opinion supports the effectiveness of the death penalty. Jern et al. show that if Alice and Bob have different priors for V as well as H, that a pattern of updating like that observed in the Lord et al. study can be seen in such a model. Jern et al. do not claim that this is necessarily the mechanism which is at work in this experiment. Their point is simply that alternative explanations are available, which do not involve any biased evaluation of the evidence. Whether or not such an alternative explanation is the correct one depends on whether the beliefs of the subjects really are governed by a specific extra belief like V.

In this paper, we have examined whether taking into account opinions about the reliability of the source of evidence can also provide alternative explanations of the belief polarisation phenomenon. Such an explanation has already been suggested in Cook and Lewandowsky (2016) in a model which also includes worldview as a variable. We find that indeed a source reliability model such as depicted in Figure 5 can also reproduce the pattern of weighting of evidence seen in the Lord et al. (1979) experiment. Whereas the alternative explanations invoked by Jern et al. (2014) rely on specific extra beliefs which subjects then may or may not be entertaining, it is perhaps plausible that in fact we always do have some beliefs about the reliability of our sources which we are updating in tandem with our views about the hypotheses in question. Thus, the type of explanation we offer potentially has a more generic character.

However, it is actually not so clear that the mechanism modeled in Figure 5 actually represents a plausible alternative explanation of the Lord et al. (1979) set-up. This is because the explanation relies, as we have seen, on subjects assigning different prior reliabilities to the two sources. Notice that if $\rho_1 = \rho_2$, the condition for the likelihood ratio (Equation 3) no longer depends on the priors at all, and thus belief polarisation is not possible. However, in the Lord et al. study there is no particular reason why the participants should set their priors differently for the reliabilities of the two studies, given that they are presented in exactly the same way. In the experiment, participants were simply given cards which present the results of the studies. It seems natural then to expect that they should assign the same prior for reliability to each of the studies, and if this is the case, then the model would predict no belief polarisation. In the experiment, participants were asked to assess the reliability of the studies, but on the basis of what the studies themselves said. Thus, what was examined here was not a prior probability for the reliability of the study, but a posterior which already depends on the content of the study itself.

Even if the model does not provide a convincing explanation of the Lord et al. (1979) results, we still think that the mechanism which it elucidates may well be at work in real-life contexts. In many real-life settings, agents do have prior views about the reliability of their sources. They may, for example, trust one news source more than another. In the legal case, some jurors might have greater initial trust in the police than others. Some may have greater trust in forensic investigations than others, or in the reliability of eye-witnesses. The experimental set-up of Lord et al. may actually represent a rather unusual situation, since it is arranged in such a way that agents have no independent way to form prior opinions about the reliability of their sources. It is even possible that such an unusual set-up effectively forces people to assess the reliability of their information differently from how they normally would, making use of their own prior views about the hypothesis since that is all they have access to.

What our model shows is that even small initial differences in how reliable we take our sources to be can in certain circumstances be amplified into divergence of opinion on crucial hypotheses, even if there is initially no difference of opinion on these hypotheses. The reliability priors determine whether or not a subject updates in a positive or negative direction given mixed evidence. However, there is no systematic connection between having higher prior for the hypothesis and having reliability priors which produce positive updates, or vice versa—if this is the case, then there will be belief polarisation. But it is also possible to have a low prior for the hypothesis, and reliability priors which induce a positive update, in which case, there will be convergence—as we see in Figure 10(b), all different combinations are possible. Thus, in this model the correlation between having a high prior for H and having reliability priors that lead to positive update can be accidental, rather than driven by any kind of bias. The model predicts then that we should not expect always to see polarisation. Rather whether polarisation occurs depends on the subjects happening to have a certain constellation of prior opinions. In fact, follow-up experiments to the Lord et al. (1979) study have shown that indeed belief polarisation only occurs in a certain subset of subjects, and then rather infrequently Kuhn and Lao (1996). This kind of result is what a model like ours would predict. On the other hand, the natural expectation if the effect is caused by some kind of consistent biased assimilation is that it should occur more of the time. An exception of course would hold if there were some reason to expect that the bias manifests itself in some people but not others. Jern et al. (2014) have set up an experiment to test whether in a particular case, belief polarisation can be explained by a normative Bayesian model rather than by biased assimilation. To make such a comparison it is necessary to carefully compare the effects of certain manipulations of prior beliefs on the proportions of subjects responding in a certain way (Jern et al., 2014, pp. 215–218). In principle it may be possible to do similar experiments to study the effects of subjects' prior beliefs about source reliability, as suggested by simple source reliability models.

6 Conclusion

In this paper, we have considered the question of whether beliefs about reliability of sources of information may play a role in driving belief polarisation. We have found that in a simple Bayesian model in which agents update not only their opinions about hypotheses but also about source reliability, belief polarisation can occur on mixed evidence. In this model, the amount by which an agent's opinion changes when it is updated on a piece of evidence depends on how reliable she takes the source of the evidence to be. Thus, if an agent initially has more trust in the reliability of a particular source than another agent, she may update more strongly on evidence from that source. When two agents are presented with mixed evidence, consisting of some evidence in favour of and some evidence against a certain hypothesis, their differential updates due to differences in prior opinions about reliability of sources may produce belief polarisation. This kind of mechanism for producing belief polarisation differs from mechanisms invoked in many of the standard explanations in that it does not rely in any way on the agents involved being influenced by their prior views on the hypothesis in any biased or undue way.

Acknowledgments

Leah Henderson acknowledges support from a Veni grant (275-20-058) from the Dutch Organisation for Scientific Research (NWO).

References

Batson, C. D. (1975). Rational processing or rationalization? The effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology 32*, 176–184. 1

Bovens, L. and S. Hartmann (2003). *Bayesian epistemology.* Oxford: Oxford University Press. 1, 3, 3

Cook, J. and S. Lewandowsky (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science 8*, 160–179. 1, 5

Fryer, R. G., P. Harms, and M. O. Jackson (2013). Updating beliefs with ambiguous evidence: Implications for polarization. *NBER working papers*. 5

Gebharter, A. (2017a). Causal exclusion and causal Bayes nets. *Philosophy and Phenomenological Research 95*(2), 353–375. 3

Gebharter, A. (2017b). *Causal nets, interventionism, and mechanisms.* Philosophical foundations and applications. Cham: Springer. 3

Gebharter, A. (2017c). Uncovering constitutive relevance relations in mechanisms. *Philosophical Studies 174*(11), 2645–2666. 3

Gebharter, A. (2019). A causal Bayes net analysis of Glennan's mechanistic account of higher-level causation (and some consequences). *British Journal for the Philosophy of Science*. 3

Hahn, U., C. Merdes, and M. von Sydow (2018). How good is your evidence and how would you know? *Topics in Cognitive Science 10*, 660–678. 4

Jern, A., K. K. Chang, and C. Kemp (2014). Belief polarization is not always irrational. *Psychological Review 121*(2), 206–224. 1, 2, 2, 2, 5

Kahan, D. M., H. Jenkins-Smith, and D. Braman (2011). Cultural cognition of scientific consensus. *Journal of Risk Research 14*(2), 147–174. 1

Kelly, T. (2008). Disagreement, dogmatism, and belief polarization. *Journal of Philosophy 105*(10), 611–633. 5

Kuhn, D. and J. Lao (1996). Effects of evidence on attitudes: Is polarization the norm? *Psychological Science 7*(2), 115–120. 1, 5

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. 1, 5

McHoskey, J. W. (1995). Case closed? On the John F. Kennedy assassination: Biased assimilation of evidence and attitude polarization. *Basic and Applied Social Psychology 17*(3), 395–409. 5

Merdes, C., M. von Sydow, and U. Hahn (2020). Formal models of source reliability. *Synthese*. 1, 3, 4

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. 1

Munro, G. D. and P. H. Ditto (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin 23*(6), 636–653. 5

Nielsen, M. and R. T. Stewart (2019). Persistent disagreement and polarization in a Bayesian setting. *British Journal for the Philosophy of Science*. 2

O'Connor, C. and J. O. Weatherall (2017). Scientific polarization. *European Journal for Philosophy of Science 8*(3), 855–875. 1

Olsson, E. (2013). A Bayesian simulation model of group deliberation and polarization. In F. Zenker (Ed.), *Bayesian argumentation: the practical side of probability*. Springer. 3, 4

Pearl, J. (2000). *Causality* (1 ed.). Cambridge: Cambridge University Press. 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. 1

Pothos, E. M., S. Lewandowsky, I. Basieva, A. Barque-Duran, K. Tapper, and A. Krennikov (2021). Information overload for (bounded) rational agents. *Proceedings of the Royal Society of London. Series B. 288*, 2020957. 5

Schaffer, J. (2016). Grounding in the image of causation. *Philosophical Studies 173*(1), 49–100. 3

Schurz, G. and A. Gebharter (2016). Causality as a theoretical concept: Explanatory warrant and empirical content of the theory of causal nets. *Synthese 193*(4), 1073–1103. 3

Singer, D. J., A. Bramson, P. Grim, B. Holman, J. Jung, K. Kovaka, A. Ranginani, and W. J. Berger (2019). Rational social and political polarization. *Philosophical Studies 176*, 2243–2267. 5

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, prediction, and search* (1 ed.). Dordrecht: Springer. 3

Taber, C. S., D. Cann, and S. Kucsova (2009). The motivated processing of political arguments. *Political Behavior 31*, 137–155. 5

Taber, C. S. and M. Lodge (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science 50*, 755–769. 5