The Myside Bias in Argument Evaluation: A Bayesian Model

Edoardo Baccini (e.baccini@rug.nl)
Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen
Groningen, 9747 AG (The Netherlands)

Stephan Hartmann (S.Hartmann@lmu.de)
Munich Center for Mathematical Philosophy, LMU Munich, 80539 Munich (Germany)

Abstract

The "myside bias" in evaluating arguments is an empirically well-confirmed phenomenon that consists of overweighting arguments that endorse one's beliefs or attack alternative beliefs while underweighting arguments that attack one's beliefs or defend alternative beliefs. This paper makes two contributions: First, it proposes a probabilistic model that adequately captures three salient features of myside bias in argument evaluation. Second, it provides a Bayesian justification of this model, thus showing that myside bias has a rational Bayesian explanation under certain conditions.

Keywords: Myside bias; Bayesianism; Argumentation

Introduction

Argumentation is central to our complex world. It often plays a critical role in political, social, and scientific communities and can play a key role in decision making and scientific research. Therefore, the understanding of argumentation has been the subject of numerous scientific investigations in different research areas (Hornikx & Hahn, 2012; Oaksford & Chater, 2020).

These investigations have shown that argumentative contexts are very complex and that it can be more difficult than expected to convince with one's arguments. This is especially true when the participants in a discussion have different prior assumptions about the topic under discussion. And indeed, research has shown that participants in a discussion are so influenced by their prior beliefs that they favor them over alternatives both in finding arguments and in evaluating other people's arguments (Perkins, 1985; Kuhn, 1991; Edwards & Smith, 1996; Nickerson, 1998; McKenzie, 2004; Taber & Lodge, 2006; Wolfe & Britt, 2008; Čavojová, Šrol, & Adamus, 2018; Stanovich, 2021).

In the literature, the influence of one's own prior beliefs in producing and evaluating arguments is commonly referred to as myside bias (Nickerson, 1998; Stanovich, West, & Toplak, 2013; Stanovich, 2021). Sometimes the same phenomenon is also referred to as confirmation bias (Stanovich, 2021). Here we adopt the terminology of Stanovich (2008b; 2013; 2021), and use the term myside bias exclusively to refer to the influence of one's prior beliefs in argumentation, while restricting confirmation bias to its original meaning of a bias in hypothesis testing. See Stanovich (2021) for a detailed discussion.

Myside bias has been studied in the context of a variety of other research topics in psychology, such as individual and group reasoning (Mercier, 2017, 2018), scientific thinking (Evans, 2002; Mercier & Heintz, 2014), intelligence and cognitive abilities (Stanovich & West, 2007, 2008b; Stanovich et al., 2013), human evolution (Mercier & Sperber, 2011, 2017; Peters, 2020) and political thinking (Taber & Lodge, 2006; Mercier & Landemore, 2012; Stanovich, 2021).

In the specific case of argument evaluation bias, research has shown the following two effects. On the one hand, discussants overestimate the strength of arguments that support their own prior beliefs or that attack opposing beliefs to their own; on the other hand, discussants underestimate arguments that either attack their own prior beliefs or that support opposing views to their own (Edwards & Smith, 1996; Nickerson, 1998; Toplak & Stanovich, 2003; Taber & Lodge, 2006; Stanovich & West, 2007, 2008b; Stanovich et al., 2013; Mercier, 2017; Stanovich, 2021).

For these reasons, it has been argued that myside-biased individuals risk to become overconfident in their beliefs and are less easily willing to revise them, regardless of their truth (Mercier, 2017; Mercier & Sperber, 2017; Stanovich, 2021). In addition, it has also been argued that myside bias contributes to undesirable social phenomena such as political polarization (Stanovich, 2021).

The purpose of this paper is twofold. First, it proposes a probabilistic model that adequately captures three important features of myside bias in argument evaluation. Second, it provides a Bayesian justification of this model, thus showing that myside bias has a rational Bayesian explanation under certain conditions. In doing so, this paper fills a gap in the literature, as despite the ever-growing literature on Bayesian approaches to reasoning and argumentation (for an overview, see Chater and Oaksford (2008), Zenker (2013), and Oaksford and Chater (2020)), there is still no systematic Bayesian model of myside bias in argument evaluation.

The Myside Bias

The myside bias in argument evaluation has been studied both in the context of formal argumentation, in which participants are asked to evaluate the conclusions of inferences that have a clear logical structure, and in the context of informal argumentation, in which subjects are asked to evaluate informal arguments that resemble real-world discussions (Čavojová et al., 2018). Overall, both lines of research show that the correspondence between arguers' prior beliefs and the content of

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

a conclusion or proposition influences how arguers judge the validity of the conclusion or the strength of the argument.

As for the study of formal arguments, early research on belief bias showed that the prior credibility of the conclusions of a deductive inference influences the arguer's judgment about the validity of the inference (Evans, Newstead, & Byrne, 1993; Evans, 1989, 2002, 2007). Further research has shown that the credibility of an argument's conclusion in light of one's prior beliefs is a predictor of whether an actor judges a conclusion to be valid or invalid. For example, Čavojová et al. (2018) found that participants had difficulty accepting the conclusions of logically valid conclusions about abortion whose content conflicted with their prior beliefs about abortion. At the same time, participants had difficulty rejecting invalid conclusions whose conclusions they agreed with (Čavojová et al., 2018).

Much of the experimental research on myside bias is conducted in an informal argumentation framework (Čavojová et al., 2018), and myside bias in argument evaluation has been documented in a substantial number of experiments (Nickerson, 1998; Edwards & Smith, 1996; Taber & Lodge, 2006; Stanovich & West, 2007, 2008a, 2008b; Stanovich et al., 2013). For example, in a commonly used experimental paradigm, participants are first asked to express their opinion on a particular topic, such as abortion or public policy; participants are then asked to rate the strength of arguments from a set of arguments that include arguments both for and against the participants' positions on an issue (Edwards & Smith, 1996; Taber & Lodge, 2006; Stanovich & West, 2007, 2008b; Stanovich et al., 2013).

Overall, three salient features of myside bias in the evaluation of arguments stand out:

1. Reasoners overweight arguments that favour their own prior beliefs and disfavour views opposite to their own (McKenzie, 2004; Stanovich & West, 2007, 2008b; Stanovich et al., 2013; Stanovich, 2021). Simultaneously, an argument that attacks the reasoner's prior opinion or confirm opposite views is generally rated as a weak argument (Nickerson, 1998).

2. Reasoners that are neutral to the topic that is being discussed tend not to show a myside bias in evaluation tasks (Taber & Lodge, 2006).

3. The myside bias occurs in various gradations: Proponents who believe more firmly in their point of view tend to exhibit a stronger bias than proponents who hold a milder opinion (Stanovich & West, 2008a). In other words, two arguers who are on the same side of an issue may differ in the extent to which they believe the side of the issue and thus show stronger or weaker bias depending on how strong their prior beliefs are.

In summary, myside bias can be interpreted as a difference of opinion about the extent to which an argument confirms (or refutes) a belief (or its opposite) between an agent who is neutral toward the belief versus the case in which the agent endorses either the truth or falsity of the belief.

A Bayesian Model

To provide a Bayesian justification of myside bias, we introduce binary propositional variables A and B (in italic script) which have the values A and ¬A, and B and ¬B (in roman script), respectively, with a prior probability distribution P defined over them. In the present context, B is the *target proposition* and A is an *argument* in support of B. A and B are contingent propositions and we assume that $P(\mathrm{A}), P(\mathrm{B}) \in (0,1)$. P represents the subjective probability function of an agent and $P(\mathrm{A})$ measures how strongly they believe in A. See Sprenger and Hartmann (2019) for a philosophical justification of the Bayesian framework.

Next, we are interested in the posterior probability of B after learning A. According to Bayes theorem, it is given by $P^*(\mathrm{B}) = P(\mathrm{B}|\mathrm{A})$ which can also be written as

$$P^*(\mathrm{B}) = \frac{P(\mathrm{B})}{P(\mathrm{B}) + x \cdot P(\neg\mathrm{B})}. \tag{1}$$

Here the likelihood ratio x is given by

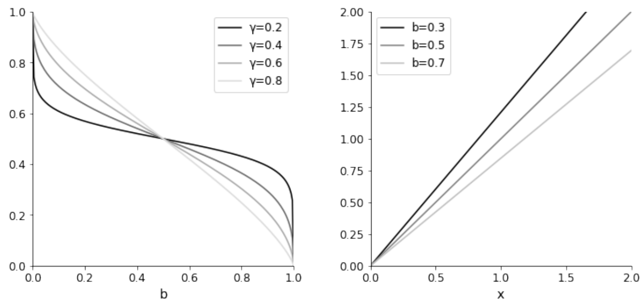
$$x := \frac{P(\mathrm{A}|\neg\mathrm{B})}{P(\mathrm{A}|\mathrm{B})}, \tag{2}$$

where we follow the convention used in Bovens and Hartmann (2003). Then the following proposition holds:

Proposition 1. *Let A and B be two binary propositional variables with a prior probability distribution P and a posterior distribution P^* defined over them. Then eq. (1) implies that (i) $P^*(\mathrm{B}) > P(\mathrm{B})$ iff $0 \leq x < 1$, (ii) $P^*(\mathrm{B}) = P(\mathrm{B})$ iff $x = 1$, and (iii) $P^*(\mathrm{B}) < P(\mathrm{B})$ iff $x > 1$.*

This proposition directly relates confirmation and disconfirmation of one's own beliefs to the likelihood ratio x: if $x < 1$, then the agent's degree of beliefs in B increases, and therefore the argument A confirms the target belief B; if $x > 1$, then the agent's degree of belief in B decreases, and therefore the argument A disconfirms the target belief B ("A attacks B"); if $x = 1$, then learning A does not make any difference for the agent's degree of belief in B, which means that the argument A is not relevant for the truth or falsity of B.

The likelihood ratio x is also referred to as the *diagnosticity* of an argument A relative to a belief B. Here the term "diagnosticity" refers to the fact that the likelihood ratio measures how much A specifically supports the truth of B against its falsity. For instance, consider a case in which an argument A is more likely to be true if B is true than if ¬B is true, i.e. when $P(\mathrm{A}|\mathrm{B}) > P(\mathrm{A}|\neg\mathrm{B})$: then $x < 1$ and, by Proposition 1, the argument A will increase the degree of belief in the target proposition B. The more likely it is that an argument A is true if B is true, compared to the case where ¬B is true, the smaller x is and the higher the confirmation of B is.



(a) x' as a function of b for $x = 1/2$ and different values of γ.

(b) x' as a function of x for $\gamma = 1$ and different values of b.

Figure 1: The perceived likelihood ratio x'.

The Perceived Likelihood Ratio

The central idea of the proposed model is that the myside bias affects the way an agent judges the diagnosticity of an argument A relative to B. This, in turn, affects the way the agent updates the strength of their belief in B based on the argument A. See also Nickerson (1998).

Therefore, we model the myside bias as a distortion of the (pure) likelihood ratio x of A relative to B, via a perceived likelihood ratio function x'. If an agent assigns a high degree of belief to B, then using x' yields more confirmation than using the (pure) likelihood ratio x, provided that x is confirmatory (i.e., $x < 1$); on the other hand, if x is disconfirmatory (i.e., $x > 1$), then x' yields less confirmation than x. More specifically, we propose the following functional form:

Definition 1. *An agent considers the propositions A (= the argument) and B (= the target belief) with a probability distribution P defined over the corresponding propositional variables. x is the (pure) likelihood ratio defined in eq. (2) and $b := P(\mathrm{B})$ is the agent's prior degree of belief in B. Then the agent's perceived likelihood ratio x' is given by*

$$x'(x,b) = 2x \cdot \frac{\bar{b}^{\gamma}}{b^{\gamma} + \bar{b}^{\gamma}}, \tag{3}$$

with $\bar{b} := 1 - b$ and $0 < \gamma < 1$.

Note that the perceived likelihood ratio x' is a function of the prior probability of the target proposition, as opposed to the pure likelihood ratio x, which is considered independent of the prior probability. Fig. 1 (a) shows the perceived likelihood ratio x' as a function of the agent's prior degree of belief b. We see that $x' < x$ if $b > 1/2$ and $x' > x$ if $b < 1/2$. Furthermore, the parameter γ, which determines the convexity of the function, characterizes different ways in which the agent's prior belief b can distort the (pure) likelihood ratio. We will see below that γ has to be in the open interval $(0,1)$. Then the distortion is much stronger for values of b close to the extremes (i.e., 0 and 1), than for middling values of b.

Fig. 1 (b) plots the perceived likelihood ratio x' as a function of the (pure) likelihood ratio x for fixed values of b and γ. In this case, x' is a linear function of x, where $x' > x$ if $b > 1/2$. Similarly, $x' < x$ if $b < 1/2$. We summarize our findings in two propositions:

Proposition 2. *The perceived likelihood ratio $x'(x,b)$ has the following features: (i) If $b > 1/2$, then $x' < x$, (ii) if $b = 1/2$, then $x' = x$, and (iii) if $b < 1/2$, then $x' > x$.*

Proposition 3. *The perceived likelihood ratio $x'(x,b)$ is strictly monotonically decreasing in b.*

Propositions 2 and 3 demonstrate that the perceived likelihood ratio $x'(x,b)$ is adequate to represent the myside bias, as it incorporates the three salient features of myside biased identified above. In particular, Proposition 2 shows that if the agent is more convinced of B than of ¬B, any argument will be perceived as more confirmatory or less disconfirmatory compared to the evaluation of a neutral observer (who uses the pure likelihood ratio x). On the other hand, if the agent is prone to believe that the target proposition is false, they will tend to perceive arguments as less confirmatory or more disconfirmatory than a neutral observer. Furthermore, an agent who is indifferent between B and ¬B will not be biased towards either of the two sides. This is consistent with the first two salient features of myside bias.

In addition, Proposition 3 shows that, all things being equal, the perceived likelihood ratio decreases as the strength of belief in B increases, and increases as the strength of belief in B decreases. Intuitively, this means that myside bias gets more pronounced as the degree of belief in the target proposition increases, in accordance with the third salient feature of myside bias.

Predictions of the Model

An agent who commits the myside bias does not update with the (pure) likelihood ratio x, but with the perceived likelihood ratio $x'(x,b)$ provided in Definition 1. Using Bayes theorem with x' instead of x, the posterior degree of belief in the target proposition B, after updating on the argument A, is then given by

$$P^{**}(\mathrm{B}) = \frac{b}{b + x'(x,b) \cdot \bar{b}}. \tag{4}$$

From eqs. (1) and (4) and Proposition 2 we then obtain:

Proposition 4. *The following claims hold: (i) If $b > 1/2$, then $P^{**}(\mathrm{B}) > P^{*}(\mathrm{B})$; (ii) if $b = 1/2$, then $P^{**}(\mathrm{B}) = P^{*}(\mathrm{B})$, and (iii) if $b < 1/2$, then $P^{**}(\mathrm{B}) < P^{*}(\mathrm{B})$.*

Therefore, our model predicts that agents' posterior degrees of belief will be more extreme than those of an agent who uses eq. (1) to calculate his posterior degree of belief (unless they are indifferent to the target statement). More specifically, agents who rate ¬B as more likely than B will have a lower posterior degree of belief than that obtained using eq. (1). Conversely, agents who believe B more strongly than ¬B will have a higher posterior degree of belief than an agent who uses eq. (1). This prediction is consistent with recent findings presented in Bains and Petkowski (2021).

Another interesting consequence of the proposed model is that the new updating rule (i.e., eq. 4) is non-commutative, i.e.

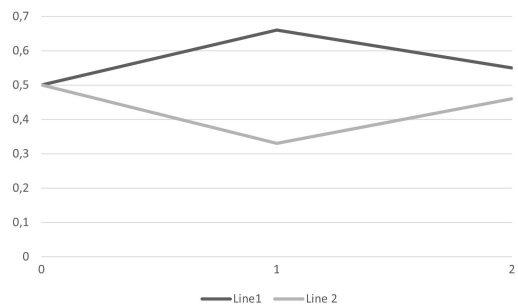


Figure 2: The result of the update on two arguments, A_1 (with likelihood ratio x_1) and A_2 (with likelihood ratio x_2) for $b = 1/2$ and $\gamma = 1/2$. Line 1: first update on A_1, then on A_2. Line 2: first update on A_2, then on A_1.

the result of an update on two or more arguments depends on the order in which the update takes place.

Fig. 2 illustrates this point. Here we consider an agent who is initially indifferent between B and ¬B (and therefore sets $b = 1/2$). Then the agent is presented with two arguments, A_1 (with the pure likelihood ratio x_1) and A_2 (with the pure likelihood ratio x_2), such that $x_1 < 1 < x_2$, i.e. the first argument (A_1) is confirmatory and the second argument (A_2) is disconfirmatory. If the agent first updates on A_1, then their degree of belief in B will increase; in turn, this will determine an underweighting of the disconfirmatory strength of A_2, since $P^{**}(B) > 1/2$, by Proposition 2. However, if the agent first updates on A_2, then $P^{**}(B) < 1/2$ and her second update (on A_1) uses the perceived likelihood ratio $x'_1 > x_1$, again by Proposition 2. Thus, the agent assigns a higher posterior degree of belief to B if they first update on the stronger argument A_1, followed by a second update on the weaker argument A_2.

Mathematically, the reason for the non-commutability of the new updating rule is the fact that the perceived likelihood ratio is a function of the prior probability. (While updating by Bayes theorem is commutative, it is well known that updating by Jeffrey conditionalization is non-commutative, however for a different reason.) It is worth noting that likelihood ratios that depend on the prior probability of the hypothesis being tested, while unusual, are not uncommon in the literature. See, e.g., chapter 5 of Bovens and Hartmann (2003) for a discussion.

We summarize our findings on the non-commutativity of my-side biased updating in the following proposition.

Proposition 5. *An agent considers the propositions* A_1, A_2 *and* B *with a prior probability distribution P defined over them. The corresponding likelihoods are* x_1 *and* x_2*, respectively. Let* $P^{\dagger}(B)$ *be the posterior probability of B after updating first on* A_1 *and then on* A_2*, and let* $P^{\ddagger}(B)$ *be the posterior probability of B after updating first on* A_2 *and then on* A_1*. Then* $P^{\ddagger}(B) > P^{\dagger}(B)$ *iff* $x_1 > x_2$.

Hence, it is epistemically advantageous for a myside-biased agent to first update on the stronger argument, i.e. on the argument with the smaller (pure) likelihood ratio.

Our model also predicts that reasoners are easily persuaded of their own position and harder to change. For instance, the stronger an agent's prior degree of belief becomes, the stronger contrasting arguments needed to be in order to sway the reasoner. In contrast with this view, Mercier (2017, 2020) and Mercier and Sperber (2017) argue that myside bias does not directly affect an agent's evaluation of external arguments, and that reasoners are able to accept good arguments even when they challenge their own view. Within this framework, one would not expect to observe differences in argument evaluation between reasoners differing in prior degrees of beliefs. This contrasts with our prediction that argument evaluation changes as a function of an arguers' prior degree of belief.

While the correctness of one or the other of these predictions remains an open question, our model has the advantage of more intuitively explain harmful group-level phenomena, such as polarization in peer groups and communication difficulties between polarized groups as an effect of one-sided exchanges and evaluations of arguments (Stanovich, 2021).

Discussion

So far, we have presented a model that is consistent with the three salient features of myside bias. The model is Bayesian because it models the bias in a Bayesian way: The agent assigns a prior probability to B, is then presented with an argument A, and updates B accordingly. This requires specifying a likelihood ratio, and our model identifies an appropriate choice, viz. $x'(x,b)$. However, this choice must be justified. Otherwise, the model would be a purely ad hoc solution. So how can the choice and the proposed functional form of $x'(x,b)$ be justified?

To address this question, we introduce a new propositional variable E and argue that the agent does not only learn A, but also E. In the present context, the appropriate posterior probability of B is therefore $P^{***}(B) = P(B|A,E)$. We will then see that, under certain conditions, $P^{***}(B) = P^{**}(B)$.

The new propositional variable E has the values E: "The target belief coheres with the background beliefs" and ¬E: "The target belief does not cohere with the background beliefs" We take E to be supporting evidence for B. That is, it is rational to assign a higher degree of belief to a proposition that fits well to one's background beliefs than to a proposition that does not. Hence, it is rational that $P(B|E) > P(B|\neg E)$.

It has already been suggested that the link between an agent's beliefs and their background beliefs justifies their putative bias in evaluating arguments (Evans & Over, 1996; Evans, 2002). For example, Evans (2002) argues that a broadly coherent system of beliefs is necessary to make sense of the world, and that this justifies an individual's biased attitude toward her or his own view and toward alternatives. Our proposal is in line with this research.

Before proceeding, it is important to note that the agent considers proposition E on the basis of the argument A put forward. Considerations of coherence with background be-

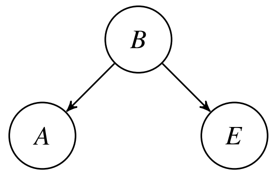


Figure 3: The Bayesian network for the myside bias.

liefs also play a role, of course, in an agent's determination of the prior probability of B. Here, however, the focus is on the following question: does B cohere with the agent's background beliefs *in light of* A?

Next, we note that $A \perp\!\!\!\perp E|B$. That is, once we know that B, learning E will not change the degree of belief an agent assigns to A: The truth (or falsity) of the argument only depends on the target belief. This plausible assumption then suggests the Bayesian network represented in Fig. 3. Note that there are arcs from B to A and from B to E, indicating that the corresponding propositional variables are directly probabilistically dependent on each other. For an introduction to the theory of Bayesian networks, see Neapolitan (2003).

To complete the Bayesian network, we have to specify the prior probability of the root node B, i.e.

$$P(\mathrm{B}) = b, \tag{5}$$

and the conditional probabilities of the child nodes (i.e. A and E) given the values of their parent (i.e. B). These likelihoods are given by

$$\begin{aligned} P(\mathrm{A}|\mathrm{B}) &:= p_1 \quad , \quad P(\mathrm{A}|\neg\mathrm{B}) := q_1 \\ P(\mathrm{E}|\mathrm{B}) &:= p_2 \quad , \quad P(\mathrm{E}|\neg\mathrm{B}) := q_2. \end{aligned} \tag{6}$$

With this we can calculate the posterior probability of B after learning A and E.

Proposition 6. *An agent considers the propositions* A, B *and* E *with a prior probability distribution P defined in eqs. (5) and (6). The corresponding propositional variables satisfy the conditional independencies encoded in the Bayesian network in Fig. 3. Then*

$$P(\mathrm{B}|\mathrm{A},\mathrm{E}) = \frac{b}{b + x'' \cdot \bar{b}}.$$

with $x'' = x \cdot x_E$ *and* $x := q_1/p_1$ *and* $x_E := q_2/p_2$.

To establish that $P^{***}(\mathrm{B}) = P(\mathrm{B}|\mathrm{A},\mathrm{E}) = P^{**}(\mathrm{B})$, we need to show that

$$x_E := q_2/p_2 = \frac{2\,\bar{b}^\gamma}{b^\gamma + \bar{b}^\gamma}. \tag{7}$$

This obtains if one sets

$$\begin{aligned} p_2 &:= \tfrac{1}{2} \cdot \left(\bar{b}^\gamma + b^\gamma\right) \\ q_2 &:= \bar{b}^\gamma. \end{aligned} \tag{8}$$

We will now argue that this is a good choice. And indeed, eqs. (8) are plausible. First, we mentioned already that a prior dependence of the likelihoods has already been used in other contexts. Second, an agent who believes B more strongly than ¬B (i.e. who assigns $b > 1/2$) expects the target belief to cohere more with their background beliefs under the assumption that B is true than if it is false. Likewise, an agent who believes ¬B more strongly than B (i.e. who assigns $b < 1/2$) expects the target belief to cohere less with their background beliefs under the assumption that B is true than if it is false. It is easy to see that eqs. (8) account for this. See also Fig. 4. Third, q_2 is a decreasing function of b. That is, the probability that the target belief coheres with the background beliefs decreases with the prior probability of the target belief, under the assumption that it is false. Fourth, p_2 has a maximum at $b = 1/2$ if $\gamma < 1$. See also Fig. 4. This is plausible as a proposition with a middling prior probability is most "flexible" and one would expect it to easily fit into a system of background beliefs. This is not to be expected with a proposition of whose truth or falsity one is much more convinced. In this case (i.e. for $b \approx 0$ or $b \approx 1$), the chance should be $1/2$ that the target belief coheres with the background beliefs. For $\gamma > 1$, p_2 has a minimum at $b = 1/2$. As this is not plausible (given the above considerations), we restrict the range of γ to the open interval $(0, 1)$ (see Definition 1).

The crucial idea of the present proposal is that an agent, who holds a belief B and who is confronted with an argument A for or against B does not only update their strength of belief on A but also investigates, prompted by the argument A, whether B fits to the agent's background beliefs. This will lead to an increase or decrease of the agent's strength of belief in B–the myside bias–which then, under these assumptions, turns out to be a rational response.

In closing this section, let us shortly comment on the notion of coherence that is used here. "Coherence" is a notoriously vague term that plays a key role in the coherence theory of justification in epistemology (see, e.g., BonJour (1985)). It refers to the property of an information set to "hang together well" which is often taken to be a sign of its truth. Witness reports in murder cases are good illustrations of this. But while we have a good intuitive sense of which information sets are coherent and which not (and which of two information sets is more coherent), it is notoriously hard to make precise what coherence means and to substantiate the claim that coherence is, under certain conditions, truth conducive (or at least probability conducive) in the sense that a more coherent set is, given certain conditions, more likely to be true (or has a higher posterior probability). These questions have been addressed in the literature in formal epistemology. See, e.g., Bovens and Hartmann (2003); Douven and Meijs (2007); Olsson (2005) and it will be interesting to relate the qualitative proposal made in this paper to that literature. This will allow for a more fundamental derivation of the perceived likelihood ratio proposed in this paper. We leave this task for another occasion.

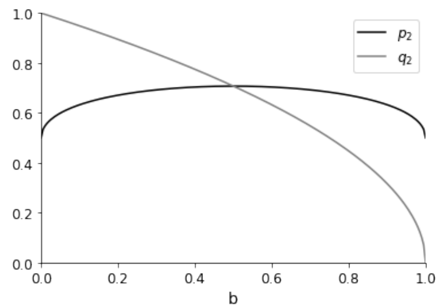


Figure 4: The likelihoods p_2 and q_2 as a function of b for $\gamma = 1/2$.

Conclusion

In this paper, we proposed a novel probabilistic model of myside bias in argument evaluation, in which myside bias is represented as a bias in the diagnosticity of an argument relative to a belief. Moreover, we have shown that our model can be derived from Bayesian assumptions if we assume that an agent takes into account the coherence of the belief under consideration with her background beliefs.

The proposed model is specific enough to be empirically tested. This should be done, first, to evaluate whether the proposed model is quantitatively adequate to describe myside bias in argument evaluation. Second, it is important to empirically investigate the relationships between the influence of coherence and other epistemic motives on myside bias and, more generally, on argument evaluation. Finally, it will be interesting to compare the proposed Bayesian explanation with other explanations for the myside bias and develop criteria for evaluating them.

Proofs

Proposition 2

Definition 1 implies that

$$\frac{x'}{x} = \frac{2}{1 + (\bar{b}/b)^{\gamma}}. \tag{9}$$

Next, we note that $(\bar{b}/b)^{\gamma} < 1$ for $b > 1/2$, $(\bar{b}/b)^{\gamma} = 1$ for $b = 1/2$ and $(\bar{b}/b)^{\gamma} > 1$ for $b < 1/2$. Hence, from eq. (9), $x'/x < 1$ if $b > 1/2$, $x'/x = 1$ if $b = 1/2$ and $x'/x > 1$ if $b < 1/2$. From this, the proposition follows. □

Proposition 3

We differentiate $x'(x, b)$ with respect to b and obtain:

$$\frac{\partial x'}{\partial b} = -2\gamma x \cdot \frac{(b\bar{b})^{\gamma-1}}{(b^{\gamma} + \bar{b}^{\gamma})^2} < 0$$

Hence, $x'(x, b)$ is a strictly monotonically decreasing function of b. □

Proposition 4

It is easy to see from eqs. (1) and (4) that (i) $P^{**}(\mathrm{B}) > P^{*}(\mathrm{B})$ iff $x' < x$, (ii) $P^{**}(\mathrm{B}) = P^{*}(\mathrm{B})$ iff $x' = x$, and (iii) $P^{**}(\mathrm{B}) < P^{*}(\mathrm{B})$ iff $x' < x$. Using Proposition 2 then completes the proof. □

Proposition 5

We begin with some notation. We denote the probability of B after first updating on A_1 by b' and the probability of B after first updating on A_2 by c'. Likewise, we denote the probability that results after first updating on A_1 and then on A_2 by b'' and the probability that results after first updating on A_2 and then on A_1 by c''. These are given by

$$\begin{aligned}
b' &= \frac{b}{b + \bar{b}\, x'(x_1, b)} =: \frac{b}{N_1} \\
b'' &= \frac{b'}{b' + \overline{b'}\, x'(x_2, b')} =: \frac{b'}{N_2} \\
c' &= \frac{b}{b + \bar{b}\, x'(x_2, b)} =: \frac{b}{N_3} \\
c'' &= \frac{c'}{c' + \overline{c'}\, x'(x_1, c')} =: \frac{c'}{N_4}
\end{aligned}$$

Next, we calculate $\Delta := c'' - b''$:

$$\begin{aligned}
\Delta &= \frac{1}{N_2 N_4} \cdot \left(c' \left(b' + \overline{b'}\, x'(x_2, b')\right) - b' \left(c' + \overline{c'}\, x'(x_1, c')\right).\right) \\
&= \frac{1}{N_2 N_4} \cdot \left(c'\, \overline{b'}\, x'(x_2, b')) - b'\, \overline{c'}\, x'(x_1, c')\right)) \\
&= \frac{b\bar{b}}{N_1 N_2 N_3 N_4} \cdot \left(x'(x_1, b)\, x'(x_2, b') - x'(x_2, b)\, x'(x_1, c')\right)
\end{aligned}$$

Plugging in the expressions for the various x', one obtains after some algebra that

$$\Delta = K \cdot \left(\left(\frac{\overline{b'}\, c'}{b'\, \overline{c'}}\right)^{\gamma} - 1\right), \tag{10}$$

where K is a positive constant. Hence, $\Delta > 0$ iff $\overline{b'}\, c' > b'\, \overline{c'}$. This holds iff $c' > b'$ which in turn holds iff $x_1 > x_2$. This completes the proof. □

Proposition 6

We first note that

$$P(\mathrm{B}|\mathrm{A}, \mathrm{E}) = \frac{P(\mathrm{A}, \mathrm{B}, \mathrm{E})}{P(\mathrm{A}, \mathrm{E})}.$$

Next we apply the product rule from the theory of Bayesian networks (see, e.g., Hartmann (2021)) and obtain:

$$\begin{aligned}
P(\mathrm{B}|\mathrm{A}, \mathrm{E}) &= \frac{P(\mathrm{B})\, P(\mathrm{A}|\mathrm{B})\, P(\mathrm{E}|\mathrm{B})}{\sum_B P(B)\, P(\mathrm{A}|B)\, P(\mathrm{E}|B)} \\
&= \frac{b\, p_1\, p_2}{b\, p_1\, p_2 + \bar{b}\, q_1\, q_2} \\
&= \frac{b}{b + \bar{b}\, (q_1/p_1)\, (q_2/p_2)}
\end{aligned}$$

From this, the proposition follows. □

References

Bains, W., & Petkowski, J. J. (2021). Astrobiologists are rational but not Bayesian. *International Journal of Astrobiology*, *20*, 312–318.

BonJour, L. (Ed.). (1985). *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

Bovens, L., & Hartmann, S. (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Chater, N., & Oaksford, M. (Eds.). (2008). *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford: Oxford University Press.

Douven, I., & Meijs, W. (2007). Measuring coherence. *Synthese*, *156(3)*, 405–425.

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology*, *71*, 5–24.

Evans, J. S. (1989). *Bias in Human Reasoning: Causes and Consequences*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Evans, J. S. (2002). The influence of prior belief on scientific thinking. In P. Carruthers, S. Stich, & M. Siegal (Eds.), *The Cognitive Basis of Science.* Cambridge University Press.

Evans, J. S. (2007). *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*. New York, NY: Psychology Press.

Evans, J. S., Newstead, S. E., & Byrne, R. M. J. (1993). *Human Reasoning: The Psychology of Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Evans, J. S., & Over, D. E. (1996). *Rationality and Reasoning*. Oxford, UK: Psychology/Erlbaum (Uk) Taylor & Fr.

Hartmann, S. (2021). Bayes nets and rationality. In M. Knauff & W. Spohn (Eds.), *The Handbook of Rationality* (pp. 253–264). Boston, MA: MIT Press.

Hornikx, J., & Hahn, U. (2012). Reasoning and argumentation: Towards an integrated psychology of argumentation. *Thinking & Reasoning*, *18*, 225–243.

Kuhn, D. (1991). *The Skills of Argument*. Cambridge University Press.

McKenzie, C. R. M. (2004). Hypothesis testing and evaluation. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making.* Oxford: Blackwell Publishing.

Mercier, H. (2017). Confirmation bias—Myside bias. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing Phenomena in Thinking, Judgment and Memory.* New York, NY: Routledge/Taylor & Francis Group.

Mercier, H. (2018). A related proposal: An interactionist perspective on reason. *Behavioral and Brain Sciences*, *41*, e53.

Mercier, H. (2020). *Not Born Yesterday*. Princeton, NJ: Princeton University Press.

Mercier, H., & Heintz, C. (2014). Scientists' argumentative reasoning. *Topoi*, *33*, 513–524.

Mercier, H., & Landemore, H. (2012). Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology*, *33*, 243–258.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*, 57–74.

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason: A New Theory of Human Understanding*. Cambridge: Harvard University Press.

Neapolitan, R. E. (2003). *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, *71*, 305–330.

Olsson, E. J. (Ed.). (2005). *Against Coherence: Truth, Probability, and Justification*. Oxford: Oxford University Press.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology*, *77*, 562–571.

Peters, U. (2020). What is the function of confirmation bias? *Erkenntnis*, 1–26. doi: 10.1007/s10670-020-00252-1

Sprenger, J., & Hartmann, S. (2019). *Bayesian Philosophy of Science*. Oxford: Oxford University Press.

Stanovich, K. E. (2021). *The Bias That Divides Us: The Science and Politics of Myside Thinking*. Cambridge: MIT Press.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*, 225–247.

Stanovich, K. E., & West, R. F. (2008a). On the failure of cognitive ability to predict myside and one-sided thinking biases. *Thinking & Reasoning*, *14*, 129–167.

Stanovich, K. E., & West, R. F. (2008b). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*, 672–695.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*, 259–264.

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

Toplak, M. E., & Stanovich, K. E. (2003). Associations between myside bias on an informal reasoning task and amount of post-secondary education. *Applied Cognitive Psychology*, *17*, 851–860.

Wolfe, C. R., & Britt, M. A. (2008). The locus of the myside bias in written argumentation. *Thinking & Reasoning*, *14*, 1–27.

Zenker, F. (Ed.). (2013). *Bayesian Argumentation: The Practical Side of Probability*. Dordrecht: Springer.

Čavojová, V., Šrol, J., & Adamus, M. (2018). My point is valid, yours is not: myside bias in reasoning about abortion. *Journal of Cognitive Psychology*, *30*, 656–669.