Why More Evidence May Convince You Less: Evidence, Reliability and Bayesian Rationality

Edoardo Baccini (E.Baccini@lmu.de)
Munich Center for Mathematical Philosophy, LMU Munich, 80539 Munich (Germany)

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

Abstract

This paper considers the issue why more pieces of supporting evidence may be less persuasive than fewer, perhaps to the point of dissuading one's interlocutor. Challenging the view that such *more-is-less* effects result from failures to make correct inferences based on the evidence, this paper argues that they may be a rational response based on one's beliefs about a source's reliability. To do this, we present a Bayesian network framework for inference with potentially unreliable sources. While existing Bayesian models forbid *more-is-less* effects, our more general framework naturally accounts for them.

Keywords: Bayesian Epistemology, Reliability, Evidence

Introduction

If you want to convince someone that a particular claim holds, it may not be enough to provide evidence supporting your claim. Sometimes, your interlocutor, who might have initially believed you, may even come to disbelieve what you claim despite all the further evidence you may present. Dynamics of this type seem at play in a variety of contexts from interaction on social media to science communications (e.g. O'Connor and Weatherall (2019) and Stanovich (2021)). In all these situations, even reasoners who may have trusted some evidence provided, may eventually display resistance towards further evidence (Nyhan & Reifler, 2010; Sunstein, 2018). How can a larger body of evidence fail to persuade others? And why could it even dissuade them?

Explanations for this may be found in one's social environment or in the cognitive mechanisms driving belief change. One's social circle might determine which evidence an agent is exposed to (Nguyen, 2020), but even when agents have access to the relevant evidence, cognitive biases may hinder their ability to make sound inferences (Baccini et al., 2023; Nickerson, 1998). The fact that normative models of belief update also rule out that evidence can be so resisted reinforces the view that such resistance arises because of one's failure to adopt sound reasoning rules. According to Bayesian reasoning in particular, if evidence is trusted at face-value, access to more evidence always raises one's degree of belief in the supported hypothesis (Howson & Urbach, 2006).

Nevertheless, when evidence comes from others, one cannot believe it at face value, as one incurs the risk of trusting unreliable sources pointing away from the truth. How should one's beliefs change in these scenarios? Various Bayesian models have been proposed in this direction (Bovens & Hartmann, 2003; Harris et al., 2016; Jeffrey, 1990; Olsson, 2011; Pearl, 1990). Among these, the Bovens and Hartmann (2003) model has found wide application in the philosophy and psychology of reasoning and argumentation (Harris & Hahn, 2009; Henderson & Gebharter, 2021; Jarvstad & Hahn, 2011; Merdes et al., 2021). Even if this model factors in one's doubts about a source's reliability, it still excludes that more evidence from a single source may bring less confirmation.

Should one conclude that these *more-is-less* effects emerge only if reasoners fail to be rational? This paper argues that this conclusion is premature. We show that the absence of *more-is-less* effects in these ideal models is due to how reliability is represented. Indeed, they do not distinguish between one's beliefs about a source's *nature*, i.e., its telling truths, falsehoods or irrelevant facts, and one's beliefs about a source's *behaviour*, i.e., its tendency to give evidence supporting or countering a claim.

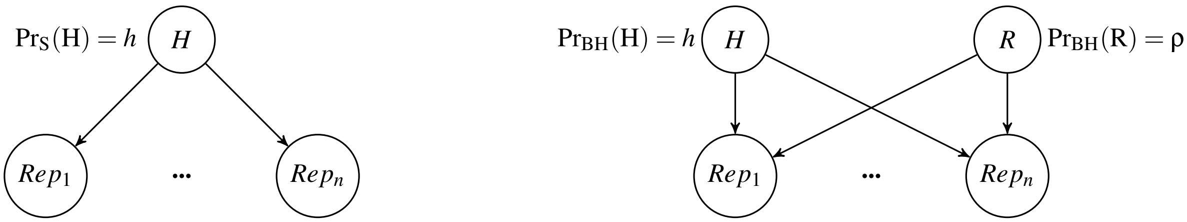
Building on the foregoing distinction, we introduce a novel Bayesian framework that generalises the model proposed by Bovens and Hartmann (2003). We then use this framework to show that *more-is-less* effects can arise through fully rational Bayesian updating. The underlying rationale is as follows. As additional evidence accumulates, a reasoner progressively learns about a source's behaviour. The inferred behavioural pattern can itself function as higher-order evidence about the source's nature. If this behaviour would be sufficiently unlikely for a truth-telling source, the reasoner may rationally reduce their confidence in the target hypothesis—even when earlier evidence had initially increased that confidence.

How More Evidence Convinces More

We consider the problem of belief update with new evidence within a Bayesian framework, where an agent's beliefs are represented via a joint probability distribution over a relevant set of random variables (Bovens & Hartmann, 2003; Lin, 2024; Sprenger & Hartmann, 2019). In particular, we will use Bayesian networks (Hartmann, 2021; Neapolitan, 2003).

The simplest possible scenario is one where an agent considers a hypothesis and receives multiple reports as to whether the hypothesis is true or not. The reports constitute evidence from which to infer the truth or falsehood of a hypothesis. In a Bayesian framework, the hypothesis is represented by a propositional random variable H that assumes value H ($\neg$H) when "The hypothesis is true (false) ". A report is represented by a propositional random variable *Rep* assuming value Rep

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



For $1 \leq i \leq n$:
$\Pr_S(\mathrm{Rep}_i|\mathrm{H}) = p$
$\Pr_S(\mathrm{Rep}_i|\neg\mathrm{H}) = q$

(a) Simple Bayesian conditioning

For $1 \leq i \leq n$:
$\Pr_{BH}(\mathrm{Rep}_i|\mathrm{H}\cdot\mathrm{R}) = 1$ $\quad \Pr_{BH}(\mathrm{Rep}_i|\neg\mathrm{H}\cdot\mathrm{R}) = 0$
$\Pr_{BH}(\mathrm{Rep}_i|\mathrm{H}\cdot\neg\mathrm{R}) = \alpha$ $\quad \Pr_{BH}(\mathrm{Rep}_i|\neg\mathrm{H}\cdot\neg\mathrm{R}) = \alpha$

(b) The Bovens and Hartmann (2003) model.

Figure 1: Bayesian network models of belief update with multiple reports.

($\neg$Rep) when "A positive (negative) report is received indicating that the hypothesis is true (false)". When considering multiple reports, we use labeled variables Rep_i where $i \in \mathbb{N}$. Throughout the paper, for any random variable X, for any value x for X and any probability distribution Pr, we use the compact notation $\Pr(x) := \Pr(X = x)$, which lifts to joint and conditional distributions. If $\Pr(x) = a$, then $\bar{a} := 1 - a$. Probability distributions are labelled by their associated model.

In the simplest case, the agent ignores the possibility that the source of reports may be unreliable. Assuming the agent takes each report as independent of any other report, we can represent their belief state with the Bayesian network in Figure 1(a), where a joint probability distribution $\Pr_S$ is specified over the variables $H, Rep_1, ..., Rep_n$. How should the agent update their belief in H based on n incoming reports? Bayesian epistemology demands that one's belief change by *conditioning* on new evidence. Thus,

$$\Pr_S^{(n)}(\mathrm{H}) := \Pr(\mathrm{H}|\mathrm{Rep}_1 \cdot \ldots \cdot \mathrm{Rep}_n) = \frac{h}{h + (\frac{q}{p})^n \cdot \bar{h}}.$$

We now ask: Can *more-is-less* effects arise in this model? Formally, we need to check whether two conditions hold:

(C1) For some $\Pr_S$, $\Pr_S^{(1)}(\mathrm{H}) > \Pr_S(\mathrm{H})$ and $\Pr_S^{(n)}(\mathrm{H}) < \Pr_S^1(\mathrm{H})$ for some $n \in \mathbb{N}$;

(C2) For some $\Pr_S$, $\Pr_S^{(1)}(\mathrm{H}) > \Pr_S(\mathrm{H})$ and $\Pr_S^{(n)}(\mathrm{H}) < \Pr_S(\mathrm{H})$ for some $n \in \mathbb{N}$.

Condition (C1) states that some prior probability distribution over the network is such that a single piece of evidence raises one's prior in H (confirms/supports H), which however decreases for some larger evidence set of size n. Condition (C2) states that some prior probability distribution is such that not only one's posterior after n pieces of evidence decreases below one posterior after a single piece of evidence, but it even decreases below one's initial prior in H (disconfirms H).

Neither condition is satisfied by the network in Fig. 1.[1]

[1] Proofs of all propositions are available at https://osf.io/vd6qs/overview?view_only=cf91638d74dd48e4a2d21d695f3924dd.

Proposition 1. *For all* $\Pr_S$, *for all* $n \geq 1$, *if* $\Pr_S^1(\mathrm{H}) > \Pr_S(\mathrm{H})$, *then* $\Pr_S^{(n+1)}(\mathrm{H}) > \Pr_S^{(n)}(\mathrm{H})$.

We now turn to the Bovens and Hartmann (2003) (henceforth, BH) model in Fig. 1(b). This model represents an agent who considers possible that the source of reports may be unreliable. In particular, two scenarios are possible: the source is reliable and always provides positive reports if H holds, or negative reports if H does not hold; the source is unreliable and provides positive reports with the same probability no matter whether H holds or not. This is modeled by a propositional variable R assuming value R ($\neg$R) when "The source is (not) reliable". A joint probability distribution $\Pr_{BH}$ is then specified such that $H \perp\!\!\!\perp R$, and such that the likelihood of each report depends on the values of H and R, as described in Figure 1(b). The likelihood of a positive report is 1 (0) if the source is reliable and H ($\neg$H) is the case. If the source is unreliable, the likelihood of a positive report is fixed by a *randomisation parameter* $\alpha \in (0, 1)$, no matter the actual value of H. Overall, either the source is always reporting the truth, or its reports are uncorrelated with the actual value of H.

Given n positive reports, one's posterior belief in H is:

$$\Pr_{BH}^{(n)}(\mathrm{H}) := \Pr_{BH}(\mathrm{H}|\mathrm{Rep}_1 \cdot \ldots \cdot \mathrm{Rep}_n) = \frac{h}{h + \bar{h}\frac{\bar{\rho}\cdot\alpha^n}{\rho + \bar{\rho}\cdot\alpha^n}}$$

A closer look reveals that *more-is-less* effects are impossible in the BH model: Any distribution $\Pr_{BH}$ violates (C1) and (C2). More positive reports always bring more confirmation.

Proposition 2. *For all* $\Pr_{BH}$, *for all* $n \geq 1$, $\Pr_{BH}^{(n+1)}(\mathrm{H}) > \Pr_{BH}^{(n)}(\mathrm{H})$.

A General Bayesian Framework

Consider a scenario, where an agent interrogates a source about H. As usual, upon a single interrogation, the source produces a single report, which could either be positive or negative according to some fixed likelihood. Imagine now that an agent interrogates the same source multiple times, by which the source in question provides a larger set of reports. This set may contain a mix of positive and negative reports.

What factors affect the ratio of positive to negative reports within the set generated by the source?

First, whether H *in fact* holds or not will generally matter as to the ratio of positive to negative reports in the set of reports obtained from a source. For instance, under the assumption that H holds, a truth-telling source will provide more positive than negative reports, where negative reports are reporting mistakes. Conversely, if the hypothesis would not hold, a truth-telling source would provide more negative than positive reports – where the positive reports now count as mistakes.

Second, the nature of a source, i.e., whether the source is telling truths, falsehoods or simply randomising as in the BH model, will itself contribute to determining the ratio of positive to negative reports in the generated set of reports. For example, assuming that the hypothesis holds, a false-telling source will generate a set of reports containing more negative than positive reports, while, as pointed out above, a truth-telling source would do the opposite.

Note that even if an agent knows a source's nature and the actual value of H, they may still not know the likelihood of receiving a positive (or a negative) report from the source in question. In other words, the actual value of H and the actual nature of the source jointly *underdetermine* the likelihood of obtaining a positive (or a negative) report from the source in question. In addition, the converse also holds: Knowing the exact likelihood of obtaining a positive or a negative report from a source will generally underdetermine the actual value of H and the nature of the source. For example, while one expects a false-telling source to produce negative reports more often than positive ones if H holds, one may however remain uncertain about the exact value of that likelihood. In turn, the likelihood of obtaining a positive report from a given source, if known, may be consistent with the possibility that the source is a truth-teller providing information for (resp. against) the fact that H holds and the possibility that the source is a false-teller reporting against (resp. for) the fact that H does not hold.

We now construct a Bayesian network model encoding the intuitions above. Beyond an agent's beliefs about H, we introduce two levels of uncertainty in regard to the reporting source. First, we consider the reasoner's beliefs about a source's *nature*, i.e., whether the source is expected to report truthfully, to report falsehoods, or to randomise in the sense of the BH model. Second, we consider one's beliefs about a source's *behaviour*, i.e., beliefs about the different tendencies of producing positive or negative reports that a source could have under different scenarios.

The Bayesian network is constructed as in Figure 2, where N_S is a variable representing a source's *nature*, which ranges over a subset V_{N_S} of the set of value $\{tt, ft, rand\}$, representing that the source is a truth-teller, a false-teller or a randomiser, respectively. A random variable B_S is defined to represent beliefs about source *behaviour* and which ranges over a finite set of rational numbers representing the frequency with which a source would produce positive reports.

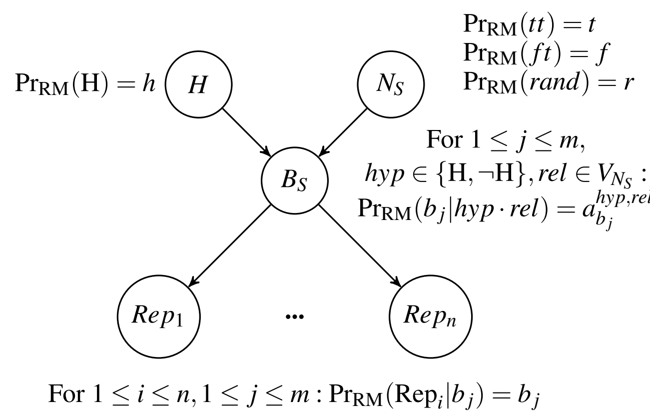


Figure 2: General reliability model.

Definition 1. *A reliability model RM is a Bayesian network* $\langle G, \Pr_{\mathrm{RM}} \rangle$ *where:*

- *G is a directed acyclic graph as in Figure 2 where $H, Rep_1, ..., Rep_n$ are propositional variable, N_S ranges over a set $V_{N_S} \subseteq \{tt, ft, rand\}$, and B_S ranges over a finite set $V_{B_S} = \{b_1, ..., b_m\} \subseteq \mathbb{Q}$.*
- *$\Pr_{\mathrm{RM}}$ is a joint probability distribution over $H, Rep_1, ..., Rep_n, N_S, B_S$ such that:*
  - *$H \perp\!\!\!\perp N_S$ and for all $i \neq k$, $1 \leq i,k \leq n$ $Rep_i \perp\!\!\!\perp \{Rep_k, H, N_S\} | B_S$;*
  - *if $tt \in V_{N_S}$, then $\Pr(b_j | \mathrm{H} \cdot tt) = 0$ if $b_j \leq 0.5$, and $\Pr(b_j | \neg\mathrm{H} \cdot tt) = 0$ if $b_j \geq 0.5$;*
  - *if $ft \in V_{N_S}$, then $\Pr(b_j | \mathrm{H} \cdot ft) = 0$ if $b_j \geq 0.5$, and $\Pr(b_j | \neg\mathrm{H} \cdot ft) = 0$ if $b_j \leq 0.5$;*
  - *if $rand \in V_{N_S}$, $\Pr(b_j | \mathrm{H} \cdot rand) = \Pr(b_j | \neg\mathrm{H} \cdot rand)$;*
  - *for $1 \leq i \leq n$, $1 \leq j \leq m$, $\Pr(\mathrm{Rep}_i | b_j) = b_j$.*

The model above explicitly represents an agent's beliefs about a source's nature via the variable N_S which can assume one to three values depending on how one sets V_{N_S}. Then, for each pair of values for H and N_S a probability distribution over the source's possible behaviours B_S is specified. This conditional probability distributions represents an agent's beliefs in the likelihood of a source using different possible behaviours given its nature and the value of H. The following is also required: If the source is a truth-teller, the only behaviours considered possible are those such that the probability of a positive report is bigger (smaller) than 0.5 if H holds (does not hold). The idea is exactly the one outlined above that a truth-teller will give truth-conducive information more often than not, even though one may not know with certainty how good of a truth-teller one is. Similarly, if the source is a false-teller, they necessarily produce positive reports with a probability smaller (bigger) than 0.5 if H holds (does not hold). Moreover, if the source is a randomiser, then learning the behaviour of the source carries no information about the actual value of H since the same behaviour would be used no matter the value of H. Finally, a behaviour determines the likelihood of positive reports in line with the principal principle (Lewis, 1981; Lin, 2024).

Definition 2. *A reliability model RM is:* ***general*** *whenever* $V_{N_S} = \{tt, ft, rand\}$*;* ***truth-symmetric*** *iff for all* $b_j \in V_{B_S}$ *there is* $b_i \in V_{B_S}$ *such that* $b_j = 1 - b_i$ *and for all* $b_j, b_i \in V_{B_S}$ *with* $b_i = 1 - b_j$, $a^{\mathrm{H},tt}_{b_j} = a^{\neg\mathrm{H},tt}_{b_i}$ *and* $a^{\mathrm{H},ft}_{b_j} = a^{\neg\mathrm{H},ft}_{b_i}$*;* ***source-symmetric*** *iff* $a^{\mathrm{H},tt}_{b_j} = a^{\neg\mathrm{H},ft}_{b_j}$ *and* $a^{\neg\mathrm{H},tt}_{b_j} = a^{\mathrm{H},ft}_{b_j}$.

A reliability model is: general whenever a source could be a truth-teller, a false-teller or a randomiser; truth-symmetric whenever the likelihood of a positive report when H holds is the same as the likelihood of a negative report when H does not hold no matter whether the source is a truth or a false-teller; source symmetric in case the probability of any behaviour being employed by a truth-teller if H holds (does not hold) is the same as the probability of that behaviour being employed by a false-teller if H does not hold (holds).

Before considering *more-is-less* effects, we show how posterior beliefs are computed in general reliability models as both positive and negative reports are received from a source. We omit computations as they are simple applications of the rule of calculus on Bayesian networks (see Neapolitan (2003)). We need additional notation. For a set of n report variables $\{Rep_1, ..., Rep_n\}$, for all $0 \leq k \leq n$, $\Pr(\mathrm{Rep}^k \cdot \neg\mathrm{Rep}^{n-k})$ denotes the probability that all Rep_j with $j \leq k$ assume value Rep_j and all Rep_j with $j > k$ assume value $\neg\mathrm{Rep}_j$. Assuming that n reports are received the first k of which are positive (technically it does not matter that the positive ones are the first k), we use the notation $\Pr^{(k,n)}_{\mathrm{RM}}(x) := \Pr(x|\mathrm{Rep}^k \cdot \neg\mathrm{Rep}^{n-k})$ for x any value in the range of any of H, N_S, B_S. When it is clear that $k = n$ (assuming $n > 1$), we simply write $\Pr^{(n)}_{\mathrm{RM}}(x)$ to denote a posterior probability based on n only positive reports. For consistency, we assume $0^0 = 1$.

Given a general RM, for all $hyp \in \{\mathrm{H}, \neg\mathrm{H}\}$ and all $rel \in V_{N_S}$, the likelihood of receiving k positive reports out of n reports $\Pr(\mathrm{Rep}^k \cdot \neg\mathrm{Rep}^{n-k}|hyp, rel)$ is given by

$$\sum_{b_j \in V_{B_S}} a^{hyp,rel}_{b_j} \cdot b_j^k \cdot (1 - b_j)^{n-k} =: x^{hyp,rel}_{k,n}.$$

We now state the equations describing the dynamics of posterior beliefs in B_S and in H. The posterior belief in a given behaviour b_j is given by

$$\Pr^{(k,n)}_{\mathrm{RM}}(b_j) = \frac{b_j^k \cdot (1 - b_j)^{n-k} \cdot w_{b_j}}{\sum_{b_i \in V_{B_S}} b_i^k \cdot (1 - b_i)^{n-k} \cdot w_{b_i}},$$

where $w_{b_i} := \sum_{hyp \in \{\mathrm{H},\neg\mathrm{H}\}} \sum_{rel \in V_{N_S}} P(hyp) \cdot P(rel) \cdot a^{hyp,rel}_{b_i}$ for $b_i \in V_{B_S}$. The posterior belief in H is given by

$$\Pr^{(k,n)}_{\mathrm{RM}}(\mathrm{H}) = \frac{h}{h + \bar{h} \cdot E[k,n]}$$

where

$$E[k,n] := \frac{t \cdot x^{\neg\mathrm{H},tt}_{k,n} + f \cdot x^{\neg\mathrm{H},ft}_{k,n} + r \cdot x^{\neg\mathrm{H},rand}_{k,n}}{t \cdot x^{\mathrm{H},tt}_{k,n} + f \cdot x^{\mathrm{H},ft}_{k,n} + r \cdot x^{\mathrm{H},rand}_{k,n}}.$$

We now study how one's beliefs about the truth of a hypothesis change. The properties in Definition 2 play an important role in one's posterior beliefs. We use "$a \overset{<}{\underset{>}{=}} b$ iff $c \overset{>}{\underset{<}{=}} d$" to say that, at the same time, "$a < b$ iff $c > d$" and "$a = b$ iff $c = d$" and "$a > b$ iff $c < d$".

Proposition 3. *Let RM be a general reliability model. Then:*

(i) $\Pr^{(k,n)}_{\mathrm{RM}}(\mathrm{H}) \overset{<}{\underset{>}{=}} \Pr_{\mathrm{RM}}(\mathrm{H})$ *iff* $t \cdot (x^{\neg\mathrm{H},tt}_{k,n} - x^{\mathrm{H},tt}_{k,n}) \overset{>}{\underset{<}{=}} f \cdot (x^{\mathrm{H},ft}_{k,n} - x^{\neg\mathrm{H},ft}_{k,n})$*;*

(ii) if RM is source-symmetric, the following three hold:

$$\Pr^{(k,n)}_{\mathrm{RM}}(\mathrm{H}) = \Pr_{\mathrm{RM}}(\mathrm{H}) \quad \textit{iff } t = f,$$

$$\Pr^{(k,n)}_{\mathrm{RM}}(\mathrm{H}) > \Pr_{\mathrm{RM}}(\mathrm{H}) \quad \textit{iff } \big(t < f \textit{ and } x^{\neg H,tt}_{k,n} > x^{H,tt}_{k,n}\big)$$
$$\textit{or } \big(t > f \textit{ and } x^{\neg H,tt}_{k,n} < x^{H,tt}_{k,n}\big),$$

$$\Pr^{(k,n)}_{\mathrm{RM}}(\mathrm{H}) < \Pr_{\mathrm{RM}}(\mathrm{H}) \quad \textit{iff } \big(t < f \textit{ and } x^{\neg H,tt}_{k,n} < x^{H,tt}_{k,n}\big)$$
$$\textit{or } \big(t > f \textit{ and } x^{\neg H,tt}_{k,n} > x^{H,tt}_{k,n}\big).$$

(iii) if RM is truth-symmetric, $k = n/2$, $\Pr^{(k,n)}_{\mathrm{RM}}(\mathrm{H}) = \Pr_{\mathrm{RM}}(\mathrm{H})$.

Point (i) shows how one's posterior belief in H depends on one's prior belief that the source is a truth-teller or a false-teller, and on the likelihood that the set of incoming reports come from a truth-teller or a false-teller given different values for H. Point (ii) shows that source symmetry is sufficient for an agent to raise or lower their prior in H simply based on their prior belief about a source's nature, and one's expectations that the reports obtained would be observed if H holds or not. Finally, point (iii) highlights how truth-symmetry implies that one's prior distribution over H does not change as long as one receives as many positives as negative reports.

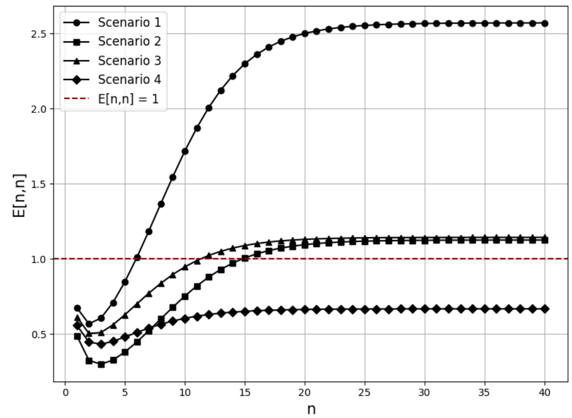
How More Evidence May Convince Less

Can our model account for *more-is-less* effects? We start with a general statement.

Proposition 4. *Assume* $k = n$*. For all* $\Pr_{RM}$ *and* $n > 1$*: (i)* $\Pr^{(1)}_{\mathrm{RM}}(\mathrm{H}) > \Pr^{(n)}_{\mathrm{RM}}(\mathrm{H})$ *iff* $E[1,1] < E[n,n]$*; (ii)* $\Pr^{(1)}_{\mathrm{RM}}(\mathrm{H}) > h$ *and* $\Pr^{(n)}_{\mathrm{RM}}(\mathrm{H}) < h$ *iff* $E[1,1] < 1$ *and* $E[n,n] > 1$.

To prove that *more-is-less* effects may arise in our model, we need to find two prior distributions: one satisfiying (i) in Proposition 4 where a stream of n positive reports brings less support for H than a single report; another satisfying (ii) in Proposition 4 whereby a single report supports H but n pieces eventually support its falseness.

Figure 3 provides a number of scenarios satisfying the conditions above. Each scenario is characterised by a prior assignment of values to N_S and six conditional probability distributions for the values of $V_{B_S} = \{0, .25, .5, .75, 1\}$ given each



Scenario 1: $t = 0.7, f = 0.2, r = 0.1$

V_{BS}	*tt* H	*tt* ¬H	*ft* H	*ft* ¬H	*rand* H	*rand* ¬H
0	0	0.1	0.9	0	0	0
0.25	0	0.9	0.1	0	0	0
0.5	0	0	0	0	1	1
0.75	0.9	0	0	0.1	0	0
1	0.1	0	0	0.9	0	0

Scenario 2: $t = 0.8, f = 0.1, r = 0.1$

V_{BS}	*tt* H	*tt* ¬H	*ft* H	*ft* ¬H	*rand* H	*rand* ¬H
0	0	0.1	0.9	0	0	0
0.25	0	0.9	0.1	0	0	0
0.5	0	0	0	0	1	1
0.75	0.9	0	0	0.1	0	0
1	0.1	0	0	0.9	0	0

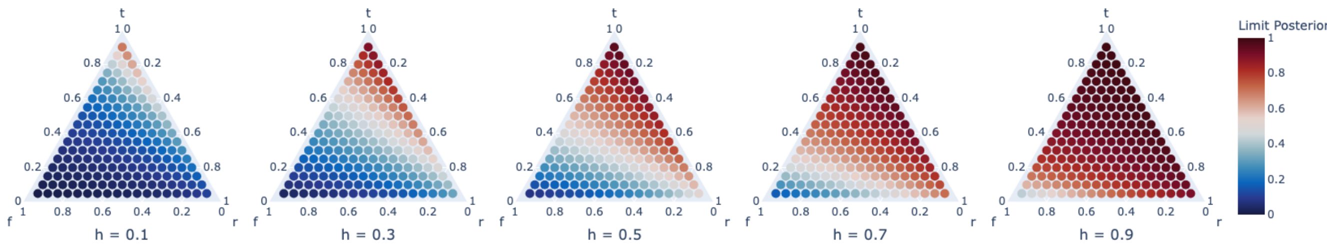
Scenario 3: $t = 0.7, f = 0.2, r = 0.1$

V_{BS}	*tt* H	*tt* ¬H	*ft* H	*ft* ¬H	*rand* H	*rand* ¬H
0	0	0.2	0.8	0	0	0
0.25	0	0.8	0.2	0	0	0
0.5	0	0	0	0	1	1
0.75	0.8	0	0	0.2	0	0
1	0.2	0	0	0.8	0	0

Scenario 4: $t = 0.7, f = 0.2, r = 0.1$

V_{BS}	*tt* H	*tt* ¬H	*ft* H	*ft* ¬H	*rand* H	*rand* ¬H
0	0	0.3	0.7	0	0	0
0.25	0	0.7	0.3	0	0	0
0.5	0	0	0	0	1	1
0.75	0.7	0	0	0.3	0	0
1	0.3	0	0	0.7	0	0

(a) More-is-less effect for different probability distributions.



(b) Example of limit posterior belief in H for different priors over N_S where $b_j > 0.5$ and $a^{\mathrm{H},tt}_{b_j} = 0.8, a^{\neg\mathrm{H},ft}_{b_j} = 0.6, a^{\mathrm{H},rand}_{b_j} = a^{\neg\mathrm{H},rand}_{b_j} = 0.1$.

Figure 3: Example of *more-is-less* effect (a) and of limit posterior in H (b).

pair of values for H and N_S. For simplicity, the scenarios are truth-symmetric but similar examples are easily found without assuming truth-symmetry. To illustrate, consider Scenario 1. There, an agent considers much more likely that the source is a truth-teller than a false-teller or a randomiser ($t > f > r$). At the same time, they consider more likely that if the source is a truth-teller, they are very likely (with a probability of .9) to provide truth-conducive reports not always but 75% of the time. If the source is a false-teller, the agent believes the source would be a really good false-teller, much more likely (with a probability of .9) to never provide truth-conducive information. The line with round markers in the picture describes how $E[n,n]$ varies with the number of positive reports n: While one or two positive reports increase one's posterior belief in the truth of H ($1 > E[1,1] > E[2,2]$), already three positive reports bring less confirmation than 2, and four reports bring less confirmation than a single report. Point (i) in Proposition 4 is satisfied. For a large enough n, $E[n,n]$ becomes even bigger than one, satisfying (ii) in Proposition 4. The other scenarios show how prior beliefs about a source's nature or behaviours can change the amount of evidence one needs for either (i) or (ii) to be satisfied. Scenario 4 shows that (i) may be satisfied without (ii) also being satisfied.

Why do *more-is-less* effects arise? Consider again the general case where one receives n reports k of which are positive.

Proposition 5. *If* $\lim_{n\to\infty} \frac{k}{n} = b_i$, $b_i \in V_{B_S}$ *and* $b_i \in (0,1)$, *then*

$$\lim_{n\to\infty} \Pr^{(k,n)}_{\mathrm{RM}}(b_i) = 1,$$

and

$$\lim_{n\to\infty} \Pr^{(k,n)}_{\mathrm{RM}}(\mathrm{H}) = \begin{cases} \dfrac{h}{h + \bar{h}\dfrac{a^{\neg\mathrm{H},ft}_{b_i}\cdot f + a^{\neg\mathrm{H},rand}_{b_i}\cdot r}{a^{\mathrm{H},tt}_{b_i}\cdot t + a^{\mathrm{H},rand}_{b_i}\cdot r}} & \text{if } b_i > 0.5, \\[2ex] \dfrac{h}{h + \bar{h}\dfrac{a^{\neg\mathrm{H},tt}_{b_i}\cdot t + a^{\neg\mathrm{H},rand}_{b_i}\cdot r}{a^{\mathrm{H},ft}_{b_i}\cdot f + a^{\mathrm{H},rand}_{b_i}\cdot r}} & \text{if } b_i < 0.5, \\[2ex] h & \text{otherwise.} \end{cases}$$

The proposition above shows that as more evidence is provided, the reasoner becomes progressively certain of which behaviour the source is employing, and in turn their posterior belief in H approaches a limit value (it is easy to check that if $n = k$, the limit of $\Pr^{(n)}_{\mathrm{RM}}(\mathrm{H})$ and $\Pr^{(n)}_{\mathrm{RM}}(b_j)$ for $n \to \infty$ are also described by the expressions above provided $b_j = 1$). The limit posterior in H depends on: the behaviour one has come to learn the source is employing, in particular, on its likelihood under different values for H and N_S; the agent's prior beliefs about N_S. Thus, the likelihood that one assigns to the behaviour one eventually learns may eventually take over one's initial prior about a source's nature. If the behaviour is believed with sufficient strength not to be that of a truth-teller, then one is led to eventually lower their confidence in H even if fewer pieces of evidence may have initially increased it.

Discussion

We have shown that *more-is-less* effects may result from rational Bayesian updating. If source reliability is modeled through the two layers of source nature and source behavior, *more-is-less* effects arise from prior assumptions about how behavior reflects nature. For instance, in the scenarios in Fig. 3, *more-is-less* effects reflect the assumption that truth-tellers are less likely than false-tellers to produce consistent sets of reports – where reports are either all positive or all negative. Is such a prior assumption necessary for *more-is-less* effects to arise? And is such an assumption plausible? While we conjecture that the answer to the first question is negative, we leave a rigorous formal answer to future work. However, even if such prior assumption turned out to be necessary, we still think it is plausible: Not only is it possible that false-tellers may sometimes have a better access to the truth than truth-tellers, but false-tellers may also be more consistent because unconstrained by noisy and heterogeneous evidence – which may instead determine more mixed reports from truth-oriented sources.

Why do *more-is-less* effects fail to emerge for the models in Figure 1? Interestingly, both models are special cases of our framework. To emulate the model in Figure 1(a), it is sufficient to fix $V_{N_S} = \{tt\}$ and $V_{B_S} = \{p, q\}$, and then assume that $a_p^{\mathrm{H},tt} = 1$ and $a_q^{\neg\mathrm{H},tt} = 1$. In this situation, where one is certain that the source is a truth-teller who produces positive reports with likelihood p (q) if H ($\neg$H), *more-is-less* effects are impossible since both a source's nature and behaviour are known with certainty. To recover the BH model, fix $V_{N_S} = \{tt, rand\}$ and $V_{B_S} = \{0, a, 1\}$ and assume that $a_1^{\mathrm{H},tt} = 1$, $a_0^{\neg\mathrm{H},tt} = 1$, and $a_a^{\mathrm{H},rand} = a_a^{\neg\mathrm{H},rand} = 1$. *More-is-less* effects cannot arise here since a stream of only positive report is possible only if the source is a truth-teller, thus the evidence always confirms H.

Our framework bears similarity to the models in Olsson (2011) and in Harris et al. (2016). In particular, Olsson (2011) represents reliability with a probability distribution over the likelihoods of receiving correct reports, i.e., Pr(Rep|H) and Pr($\neg$Rep|$\neg$H) assumed to be equal. These likelihoods then regulate the probability of reports via the principal principle. Beyond these similarities, there are a number of relevant differences: In Olsson (2011), multiple reports from a source are processed sequentially (Merdes et al., 2021) and not via a single Bayesian network; Source nature is not explicitly modeled via a corresponding random variable; Principles alike truth-symmetry and source-symmetry in Definition 2 are always assumed. While we suspect *more-is-less* effects may arise in the Olsson model, it is not obvious that our justification would also apply there. The second model we mentioned by Harris et al. (2016) develops a Bayesian model of appeal to expert opinion: There, reports are evaluated depending on the trust an agent attaches to the reporting source, and on the source's credibility as an expert, both modeled via binary random variables. It remains open whether the model by Harris et al. (2016) (and the model by Olsson (2011)) can be considered as a special case of our framework, and in which circumstances appeal to experts may lead to resisting persuasion by evidence.

The finding outlined in the previous section that a larger amount of reports provides *higher-order* evidence of a source's nature plays a role in failures of *varied evidence* to provide more support for a hypothesis than less varied evidence (Osimani & Landes, 2023). There, one generally considers how dependencies between distinct reports affect the support a hypothesis receives (Bovens & Hartmann, 2003; Claveau, 2013; Landes, 2020). In our case, we have shown that not only a larger body of evidence may sometime bring less support for a hypothesis, but that it may even end up disconfirming it. Thus, one's background beliefs about the interplay between a source's nature and behaviour may crucially determine whether any support *at all* can be had by access to more evidence.

While *more-is-less* effects may have a Bayesian justification, the issue remains whether they hinder the formation of true beliefs. Even if one's inferences are sound, assumptions about how source behaviour reflects source nature may not match reality. This may help to grasp the disruptive power of misinformation campaigns such as those aiming to raise doubts whether a scientific consensus on pressing scientific issues exists (Oreskes & Conway, 2010; Reutlinger, 2026). Our analysis suggests that the effectiveness of such disinformation strategies may reside not just in how they affect one's beliefs about what evidence would likely be observed if a scientific claim were true, but also in how they reconfigure one's beliefs about what kind of evidence the scientific community would provide if they were a truth-telling source. This latter consequence would render additional evidence useless, while interventions to change one's background assumptions about science behaviour as a source may be more effective.

Conclusion

We presented a novel Bayesian model of source reliability which justifies why a larger body of supporting evidence may persuade one to a lesser degree than a smaller body of evidence. The reason is this: As more evidence comes in, a reasoner progressively learns the reporting behaviour of a source; the behaviour itself constitutes then evidence of the nature of the source, which may sometimes be unlikely that of a truth-teller. Thus, further evidence is distrusted.

In future work, we plan to test our model in psychological experiments: Building on existing experimental designs (Harris et al., 2016; Oaksford & Chater, 2020), we wish to test whether our two-level account of source reliability adequately describes the way humans reason about reliability. We will also pursue further theoretical analyses to characterise when resistance to evidence is a necessary consequence of one's background beliefs. Finally, we will embed our model in social simulations to study how one's background beliefs about source's nature and behaviour affect macro-level social phenomena such as polarisation or consensus formation.

Acknowledgments

This research was supported by the German Research Foundation (DFG) under grants HA3000/29-1 and HA3000/30-1. We thank Leon Assaad, Lilian von Bressensdorf, Rafael Fuchs, Ulrike Hahn, Ludi van Leeuwen, and Klee Schöppl for helpful comments, as well as the anonymous reviewers for their valuable feedback.

References

Baccini, E., Christoff, Z., Hartmann, S., & Verbrugge, R. (2023). The wisdom of the small crowd: Myside bias and group discussion. *Journal of Artificial Societies and Social Simulation*, *26*(4), 7.

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

Claveau, F. (2013). The independence condition in the variety-of-evidence thesis. *Philosophy of Science*, *80*(1), 94–118.

Harris, A. J. L., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35 5*, 1366–73.

Harris, A. J., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The appeal to expert opinion: Quantitative support for a bayesian network approach. *Cognitive Science*, *40*(6), 1496–1533.

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

Henderson, L., & Gebharter, A. (2021). The role of source reliability in belief polarisation. *Synthese*, *199*, 10253–10276.

Howson, C., & Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*. Open Court Publishing.

Jarvstad, A., & Hahn, U. (2011). Source reliability and the conjunction fallacy. *Cognitive Science*, *35*(4), 682–711.

Jeffrey, R. C. (1990). *The logic of decision*. University of Chicago press.

Landes, J. (2020). Variety of evidence. *Erkenntnis*, *85*(1), 183–223.

Lewis, D. (1981). A subjectivist's guide to objective chance. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs: Conditionals, belief, decision, chance and time* (pp. 267–297). Springer Netherlands.

Lin, H. (2024). Bayesian Epistemology. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Summer 2024). Metaphysics Research Lab, Stanford University.

Merdes, C., Von Sydow, M., & Hahn, U. (2021). Formal models of source reliability. *Synthese*, *198*, 5773–5801.

Neapolitan, R. E. (2003). *Learning bayesian networks*. Prentice Hall.

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, *17*(2), 141–161.

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

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*, 303–330.

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, *71*(1), 305–330.

O'Connor, C., & Weatherall, J. O. (2019). *The misinformation age: How false beliefs spread*. Yale University Press.

Olsson, E. J. (2011). A simulation approach to veritistic social epistemology. *Episteme*, *8*(2), 127–143.

Oreskes, N., & Conway, E. (2010). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Publishing.

Osimani, B., & Landes, J. (2023). Varieties of error and varieties of evidence in scientific inference. *The British Journal for the Philosophy of Science*, *74*(1), 117–170.

Pearl, J. (1990). Jeffrey's rule, passage of experience, and neo-bayesianism. In H. E. Kyburg, R. P. Loui, & G. N. Carlson (Eds.), *Knowledge representation and defeasible reasoning* (pp. 245–265). Springer Netherlands.

Reutlinger, A. (2026). Unreasonable doubt. how strategic science skeptics exploit the argument from disagreement. *Philosophy of Science*, 1–47.

Sprenger, J., & Hartmann, S. (2019). *Bayesian philosophy of science*. Oxford University Press.

Stanovich, K. E. (2021). *The bias that divides us: The science and politics of myside thinking*. MIT Press.

Sunstein, C. R. (2018). *#Republic: Divided democracy in the age of social media* (NED - New edition). Princeton University Press.