Synthese (2021) 198 (Suppl 23):S5773–S5801
https://doi.org/10.1007/s11229-020-02595-2

S.I. : RELIABILITY

Formal models of source reliability

Christoph Merdes[1] · **Momme von Sydow**[2] · **Ulrike Hahn**[3]

Received: 19 December 2018 / Accepted: 22 January 2020 / Published online: 2 March 2020

Abstract
The paper introduces, compares and contrasts formal models of source reliability proposed in the epistemology literature, in particular the prominent models of Bovens and Hartmann (Bayesian epistemology, Oxford University Press, Oxford, 2003) and Olsson (Episteme 8(02):127–143, 2011). All are Bayesian models seeking to provide normative guidance, yet they differ subtly in assumptions and resulting behavior. Models are evaluated both on conceptual grounds and through simulations, and the relationship between models is clarified. The simulations both show surprising similarities and highlight relevant differences between these models. Most importantly, however, our evaluations reveal that important normative concerns arguably remain unresolved. The philosophical implications of this for testimony are discussed.

Keywords Source reliability · Bayes · Testimony

1 Introduction

Imagine someone browsing their social media, and encountering a remote acquaintance endorsing the claim that vaccination of children causes autism. Let us assume that this reader, Anna, previously had a belief (or 'credence') in the opposite, namely that vaccination does not have this side effect. Given that belief, it seems reasonable that Anna lower her subjective reliability estimate (or epistemic 'trust'[1]) for this

[1] In the formal epistemology literature, 'trust' is often stipulated to mean subjective source reliability. We adopt this usage for convenience, but point out that there are alternative explications of epistemic trust in the philosophical literature (cf. Lackey 2011; Wilholt 2013).

✉ Christoph Merdes
christoph.merdes@fau.de

[1] ZiWiS, Friedrich-Alexander-Universität Erlangen Nürnberg, Bismarckstraße 8, 91054 Erlangen, Germany

[2] Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität München, Ludwigstr. 31, 80539 Munich, Germany

[3] Department of Psychological Sciences, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK

source. However, Anna's belief in the truth of the proposition at stake (the vaccination claim) should also be revised in light of the new testimonial evidence; in particular, it should be revised upwards by Anna, the focal agent, if she previously considered the source to be highly reliable.

The basic intuition, here, is that we are rationally required to update both our belief and our trust in a source based on a given report if the source's reliability has not been externally provided. This revision of trust in a source and credence in a proposition must be based on prior subjective estimates of both variables, as neither of them are known with certainty.[2] At least prima facie, given these intuitions, such an epistemic strategy seems rational. Moreover, recent experimental evidence suggests that lay people do, in fact, adopt something like this strategy in simple, scenario-based experimental tasks (Collins et al. 2018; Collins and Hahn 2019).

It should thus come as no surprise that multiple authors both within the formal epistemology literature (e.g., Olsson 2011; Bovens and Hartmann 2003) and in adjacent fields such as cognitive and developmental psychology (Shafto et al. 2012), have provided Bayesian formalizations of such a strategy. In all cases, these formalizations have served to elucidate fundamental issues (both normative and descriptive) with respect to testimony. Though these formalizations all conceptually implement the same basic strategy of joint estimation of hypothesis and source reliability, they differ in formal detail. This paper's purpose is to review, analyze and compare prominent formal models setting them against other models for dealing with less than fully reliable sources and the underlying problem of how to estimate source reliability.

Taking a step back, the problem is the following: determining how an agent, confronted with one or more reports from a source of information such as a witness, scientific expert or personal acquaintance, should revise

(1) Their credence in a hypothesis H the report is relevant for and
(2) Their estimation of the source's reliability, in particular, as it pertains to the trustworthiness of further reports from the same source.

In anticipation of formalization, a more stylized description of this problem suggests itself: imagine a task used by decades of psychological research on belief revision (Phillips and Edwards 1966). In front of you are two urns, one containing predominantly blue chips, the other predominantly red chips. One of these urns is selected, and your task is to work out which urn this is on the basis of successive draws from the chosen urn (with replacement after each draw). In the classic task, you as a participant are told the respective proportions of red and blue balls, and the experimenter is interested in how closely your belief revision matches the prescriptions of Bayes' rule given those likelihoods.

But now, imagine a further difficulty: the composition of the urn is not exactly known. In other words, you don't know the true diagnosticity of each red or blue ball that you receive. Furthermore, you also need to maintain the possibility that you could

[2] Further examples might be reading in an internet blog that particles move faster than the speed of light at CERN, or someone telling us that the US did not land on the moon; as these are contingent empirical facts, we cannot be entirely certain of their falsity. Hence, it seems reasonable that we update both our beliefs and, simultaneously, our epistemic trust in the report's source.

systematically mistake red for blue balls and vice versa. This is the belief revision problem for an agent faced with *unknown source reliability*.

This problem could arise with any real-world source. Objectively, measurement instruments may not be fully reliable, and the reliability of those measurement instrument may not be known. It may even be that the scale ends of a measurement instrument have been mixed up (resulting in full or partial anti-reliability), say, such that a reading of 'hot' is produced when it is cold, and vice versa. Even where well-intentioned, communicators sometimes get things wrong, and we may have little or nothing to go by in order to estimate the probability that what a source says is, in fact, true. It is, of course, reasonable to assume that, *on average*, what other people say is more likely to be true than not, because human communication would simply not be worthwhile if this were not the case. But it is often impossible to know *how* likely this is in a specific case, either because we do not know the source from previous interactions, or we do not know anything about their expertise with respect to the particular claim at issue. This is the fundamental fallibility of testimony that has dogged epistemological concern about this important influence on our everyday beliefs (see, e.g., Coady 1992).

From a philosophical point of view, source reliability is a key problem of social epistemology. Models of source reliability offer a more precise formulation of the problem of testimony (cf. Bovens and Hartmann 2003). Such models may be used to explore fundamental intuitions about testimonial evidence, such as the role of coherence (Olsson 2005; Bovens and Hartmann 2003). They can also be used to identify conditions under which being credulous is an effective or reasonable strategy, and when it is not (e.g., Zollman 2015; Hahn et al. 2018b). Such models may also offer more fundamental explanations about bootstrapping reliability and its limitations (cf. Elga 2007). Furthermore, formal model also apply to the discussion on how to deal with expert testimony as a lay person (cf. Goldman 2001; Hahn et al. 2012), and determine the conditions of successful lay evaluation of expert witnesses (e.g., Hahn et al. 2009; Harris et al. 2013, 2012, 2016; Fenton et al. 2013).

Formal models of source reliability have also been used for agent-based modelling within formal and social epistemology. Models of source reliability may be used to replace or supplement simpler agent models in representations of peer disagreement (Lehrer 1976; Douven 2010), and they have been at the heart of agent-based simulations investigating norms of assertion (Olsson and Vallinder 2013), the impact of over-confidence (Vallinder and Olsson 2014), group polarization (Olsson 2013; Hahn et al. 2018), the impact the structure of social networks on the accuracy of beliefs (Hahn et al. 2018a), or the impact of communication on the informedness of voters (Hahn et al. 2019). Agent-based models of belief or opinion dynamics are complex systems, and it is crucial to understand the individual-level behavior of the agent that the societal model employs.

In short, formal models of source reliability serve a number of research goals, and there is a host of issues that cannot be understood fully without a detailed understanding of these models. To this end, the paper examines the two most prominent Bayesian models of source reliability from the formal epistemology literature, the models of Olsson (2011) and Bovens and Hartmann (2003, Ch. 4) respectively. It describes the models and their basic properties, setting them against other Bayesian attempts to capture the fallibility of testimony. It then evaluates them both conceptually and

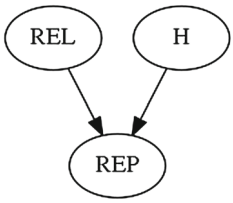

Fig. 1 Bayes net for a single witness report (REP) on a hypothesis (H) with reliability (REL)

through large-scale simulation. These results are used to assess the claim that these models provide adequate normative guidance on source reliability.

The paper proceeds as follows: First, we present the dimensions of comparison for models of source reliability. We then proceed to detailed model description and analysis. This involves both conceptual analysis across the key dimensions and the presentation of simulation results. These results reveal both differences and surprising similarities between the models. Finally, we discuss the implications of these results from both a technical and philosophical point of view. Specifically, we argue, based on the simulation results, that neither model solves the fundamental bootstrapping problem of testimonial knowledge. We close with a brief summary and an outlook on further developments and potential future research on formal models of source reliability.

2 Dimensions of comparison

2.1 General assumptions

Before describing both models in detail we start by highlighting general characteristics. First and foremost, the models under discussion assume that credences are plausibly modelled by probabilities. A further shared assumption of both models is standard Bayesian conditionalization, that is, they are committed to the claim that rationality requires updating credence by Bayes' rule:

$$p'(H) = p(H|E) = \frac{p(E|H) \cdot p(H)}{p(E)} \tag{1}$$

Finally, the two models share a common conception of the reliability problem itself, in as much as they both implement the intuitive strategy outlined with the initial vaccination scepticism example above. Both models use evidence reports to simultaneously revise beliefs both about the underlying hypothesis and the reliability of the source. Hence the basic structure of the source reliability problem as conceptualised by both the Olsson and the Bovens and Hartmann model may be represented by the Bayes net in Fig. 1. It requires assumptions about the nature of three relevant random variables of *REP, REL* and *H*.

As we will see, the two models can be distinguished by whether they formalize the reliability of a source as point probabilities for a binary variable or a more complex probability distributions over a continuous variable. The Bayes net stipulates the

independence of *REL* and *H*. This assumption is shared by both models as a plausible condition on the relationship between the source and the claim at issue: violating this independence assumption would suggest the reliability of the source determining the state of the world at issue in the hypothesis, or, alternatively, this state of the world determining the reliability of the source. To return to one of our examples, non-independence would mean that vaccines causing autism could causally influence the characteristics of the reporting source, a fairly patent absurdity in the real world.

Two clarificatory notes are in order here: The independence assumption does not imply that a source could not be more reliable in *reporting* truthfully contingent on the truth of the hypothesis. Furthermore, there might be specific cases where non-independence is justified. One plausible example where the truth of a hypothesis may actually impact reliability concerns hypotheses about human psychology: imagine the focal agent is concerned with the hypothesis that human agents suffer from confirmation bias. If this hypothesis is true, the source of testimonial evidence is less likely to be reliable, and therefore, the hypothesis is not independent of the estimation of reliability. But for standard hypothesis—such as the vaccination scepticism example—the independence assumption seems well justified.

However, the models, at least in their application to date, differ in how they treat the independence/non-independence of *multiple reports*. For computational reasons, the Olsson model, which has been used primarily for agent-based modelling, considers multiple reports to be independent even when they come from the same source, in the sense that they are not modelled in a shared network. By contrast, applications of the Bovens and Hartmann model have consistently represented multiple reports within the same Bayes' net in order to capture appropriately relations between them. The consequences of this difference are discussed in greater detail below.

Finally, the models differ in their conception of what it means for a source to be 'unreliable': In general, a source could be reliable, if not perfectly so, it could be a pure randomizer (unmoored from the truth), or even a systematic liar. In principle, it would make sense for the focal agent to be able to learn all these possibilities, that is, reach all logically possible values for reliability. We call the ability of a model to learn all of these possibilities *reachability.* At the same time, however, an anti-reliable source, unlike a randomizer, still has a lawful relationship with the truth. Hence, one might want to exclude the systematic liar by assumption in order to avoid treating a source as anti-reliable on conceptual, technical or practical grounds. Technically, doing so can simplify the model. Practically, if we assume that there are few actual systematic liars in the world, this omission may actually increase the accuracy of our beliefs, as it avoids falsely classifying someone as a systematic liar. As outlined in greater detail below, the Olsson model specifically allows learning of anti-reliability while the basic Bovens and Hartmann model does not, instead conceptualising "unreliable" as 'randomizing'. At the same time, we will see below that an in principle ability to infer anti-reliability may not suffice for anti-reliability to be reachable in practice.

2.2 Behavior

With respect to the models' behavior, we will focus on three key aspects: First, we discuss the models' precise conception of reliability and how they construe the problem faced by the epistemic agent. Second, we analyse the accuracy of the resulting beliefs in a simulated world with a 'ground truth' such that the accuracy of beliefs can be measured. Third, we consider the effective reachability in both models.

By accuracy, we mean a measure of the agent's credence distance to the truth of the matter. The main concern is with the accuracy of belief in *H,* as this is the piece of world knowledge the agent is interested in learning. To measure accuracy, we utilize the Brier score (Brier 1950), which is effectively the squared error and constitutes a proper scoring rule.[3] Additionally, we will examine the models' ability to accurately reflect the true reliabilities of the source.

As just noted, while the ability to discern anti-reliability can be built into a model by assumption, it is not obvious that the model is effectively able to move on trajectories ranging through all parts of the parameter space (what we call 'reachability'). What will be of particular interest is whether a model is capable of *learning* that a source is anti-reliable given initial trust in that source.

3 Detailed model description

We next consider the Bovens and Hartmann (2003) and Olsson (2011) models in more detail.

3.1 Bovens–Hartmann-model

3.1.1 Informal description

The Bovens–Hartmann-model (BH from hereon) is built to represent an agent trying to distinguish between a random source of reports and a reliable one. Key to the model specification are two assumptions, described in Bovens and Hartmann (2003):

(1) A reliable source is *perfectly* reliable, that is, it reports that the hypothesis is true when it is true and says it is false when it is false. Hence, reliable sources make no mistakes. All the uncertainty is situated in the *focal agent* who is only partially informed about whether the source is reliable or not.
(2) If the source is unreliable, it is a randomizer, that is, whether or not it provides a positive report is determined at random (e.g., through a coin toss), and is unrelated to the true state of the world.

[3] While there is debate about choice of scoring rule, in particular whether squared or log error scores are more appropriate (e.g., Leitgeb and Pettigrew 2010), there is no need to employ multiple accuracy measures in our case, as those arguments largely relate to differences occurring when applying those measures to multiple propositions.

The probability that an unreliable source will randomly generate a positive report is determined by the so-called randomization parameter, a. This allows sources to be biased toward a particular type of report. Changes to this parameter can have significant effects on the overall behavior of the model (see e.g., Jarvstad and Hahn 2011). In this paper, we follow Bovens and Hartmann (2003) in simply assuming a value of 0.5 as a default.

Clearly, real world sources do not typically create answers at random. In this sense the BH model is not a plausible generative model of assertion (nor is it intended as such). Both the empirical literature on testimony (Pornpitakpan 2004) and other formal treatments (e.g., Schum 1994; Harris et al. 2016) have decomposed 'reliability' into more fine-grained concepts such as veracity, bias, and accuracy, that seek to more faithfully reflect how testimony comes about. However, the BH model is plausible as a minimal model that the *focal agent* can form about an unknown source. Such an agent will typically know little to nothing about the mental processes or factors by which the source will come to generate a report if unreliable. 'Randomization' represents a convenient way of capturing that lack of knowledge, formalising the idea that for a fully unreliable source the content of their report is uncorrelated with the truth. Moreover, if $P(\text{REL})$ is not known but inferred, the BH model can at least indirectly represent *degrees* of reliability as a mixture between a randomizing source and a reliable source. In other words, though the model assumptions are clearly idealizations, Bovens and Hartmann maintain that the model can still be utilized to represent partially reliable sources: from the point of view of the focal agent, the source is subjectively fallible such that for a given report, the agent cannot be certain whether it is correct.

3.1.2 Formal model[4]

The fully specified model must define the calculation of the posterior estimation of reliability, $P'(REL)$ and the posterior degree of belief in H, $P'(H)$. The focal agent holds the corresponding prior beliefs, $P(REL)$ and $P(H)$.[5] To calculate the posteriors, the four conditional probabilities implied by the Bayes net have to be determined. They follow from the two conditions stated informally above:

$$P(\text{REP} = 1 | H = 1, \text{REL} = 1) = 1$$
$$P(\text{REP} = 1 | H = 0, \text{REL} = 1) = 0$$
$$P(\text{REP} = 1 | H = 0, \text{REL} = 0) = a$$
$$P(\text{REP} = 1 | H = 1, \text{REL} = 0) = a$$

The corresponding probabilities for $REP = 0$ follow accordingly from the two assumptions. This leads to the following equations as shown by Bovens and Hartmann:

$$P'(H) = \frac{P(H) \cdot (P(REL) + (1 - P(REL) \cdot a))}{P(H) \cdot (P(REL) + (1 - P(REL) \cdot a)) + (1 - P(H)) \cdot (1 - P(REL) \cdot a)} \quad (2)$$

[4] Appendix A contains a standardized ODD description as an alternative representation of the model for reference and further clarification.

[5] We use $P(V)$ to denote $P(V = 1)$ as a shorthand.

and

$$P'(REL) = \frac{P(H) \cdot P(REL)}{P(H) \cdot P(REL) + (1 - P(REL)) \cdot a} \tag{3}$$

3.2 Olsson model

3.2.1 Informal description

The key difference between the Olsson model and the BH model is the representation of reliability. In the Olsson model, the focal agent maintains a distribution over possible reliabilities, ranging from perfect reliability to randomization (as in BH), through to anti-reliability, that is, a source that lies consistently.

This changes substantially what targets can plausibly be modelled, motivating the formally more complex interpretation of reliability: The BH model in its basic form presented above is inherently unable to classify a source as a systematic liar; hence, any application for which this possibility needs to be maintained deviates from the assumptions of the BH model.

But, as it is able to represent systematic anti-correlation with the truth, the Olsson agent can also wrongly take a source to be a liar. Even more problematic would seem to be that the focal agent can hold a *prior* on the hypothesis that implies the source's anti-reliability (at least for hypothesis incongruent evidence), or vice versa, as the agent utilizes their expectation to evaluate the evidential impact of a report. Since priors in a subjective Bayesian framework are unrestricted as long as they are chosen from the open interval (0, 1), this does not constitute a violation of the principles of Bayesian rationality; but the consequence that the choice of priors determines whether evidence is supporting or refuting evidence appears problematic. To better understand the exact structure and consequences, we have to turn to the formal model.

3.2.2 Formal model[6]

Analogous to BH, the model is described by two equations. Credence in H is updated according to

$$P_{t+1}(H) = P_t(H|E) = \frac{M[\tau_t] \cdot P_t(H)}{M[\tau_t] \cdot P_t(H) + (1 - M[\tau_t]) \cdot (1 - P_t(H))} \tag{4}$$

where τ is the reliability distribution and M its mean. Subjective reliability is updated according to

$$\tau_{t+1}(x) = \tau(x|E) = \frac{x \cdot P_t(H) + (1 - x) \cdot (1 - P_t(H))}{M[\tau_t] \cdot P_t(H) + (1 - M[\tau_t]) \cdot (1 - P_t(H))} \tau_t(x) \tag{5}$$

The credence update (Eq. 4) is structurally very similar to basic Bayesian updating (Eq. 1), with the key difference that the static conditional probability (likelihood)

[6] For the ODD style model description, see again Appendix A.

of receiving a piece of evidence is replaced by the dynamic reliability model. The evolution of the reliability distribution (Eq. 5) is rather difficult to extract from the point-wise representation in Eq. 5; qualitatively, however, the distribution's mean increases with belief-congruent evidence and shrinks with belief incongruent evidence, while the variance of the distribution diminishes over time, modelling the increasing certainty about the degree of reliability of the source as the amount of evidence grows.

To fully specify the model, it is necessary to choose a particular probability distribution for τ. For the purpose of this paper, a beta distribution is used.[7]

4 Model comparison

4.1 Conceptual contrasts

Conceptually, both models can be applied to sources of any kind. In the social network simulations of Olsson (2011, 2013), for example, the model is used both to characterise the testimony agents receive from others, and, agents' own information gathering from the world. Likewise, Bovens and Hartmann (2003) use their model not just to capture testimonial reports, but also the reliability of scientific instruments. In that sense, the two models are entirely general models of source reliability.

However, it is worth considering further the models' application specifically to testimony. Conceptually, both models deal with the same aspect of testimony, leaving others aside. Specifically, Collins et al. (2018) distinguish two aspects of testimony and, with it, source reliability: the first is what they call the "testimonial aspect". Here, the testimony *is* the evidence, as when an expert tells us that something is the case. For example, a doctor may tell us that we suffer from a particular disease. The second is what Collins et al. call the "transmission aspect" of testimony. Here, the source *transmits* evidence and concerns about source reliability are focussed on the faithfulness of that transmission. For example, the doctor may tell us that our blood tests show elevated levels of a marker that itself provides evidence of that particular disease. We do not have access to that test ourselves (nor typically does the doctor, who merely receives a report on its outcome from the lab). In this latter case, the report is not on the status of a hypothesis itself, but rather on evidence for that hypothesis, such as a testable consequence. Both models described above deal only with the testimonial aspect. However, the transmission aspect can be modelled in the Bovens and Hartmann framework through a simple extension that includes a further node concerning a testable consequence in the network, *CON*, between hypothesis and report (see Fig. 2, and Chapter 4, Bovens and Hartmann 2003). It is less clear how this could be captured in the Olsson framework. So, although, the Olsson model and the basic BH model that are our main focus in this paper both exclude the transmission aspect, they differ in the extent to which they may readily be expanded to capture other aspects of testimony as well.

[7] For repeated dichotomous events (Bernoulli processes) and resulting Binomial distributions, the beta distribution is normally considered an appropriate prior distribution, since it is the conjugate distribution of the binomial distribution.

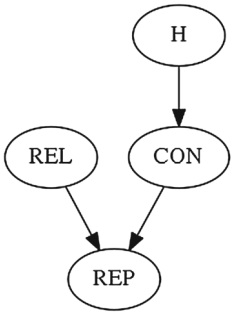

Fig. 2 Bayes net for a single witness report (REP) on a testable consequence (CON) of hypothesis (H) with reliability (REL) in the BH model

In their shared focus on 'pure testimony' (i.e., the 'testimonial' aspect), both the Olsson model and the BH model share also the fact that they seek to *unpack* the effective diagnosticity of a report as captured by the likelihoods $P(REP|H)$ and $P(REP|\neg H)$. These are the basic quantities in Bayes' rule defining the evidential value of the source's report. In the case of testimony, when we speak with a source with whom we have no prior familiarity or little familiarity, these likelihoods are unknown or not known exactly. They must consequently be *estimated*. As Hahn et al. (2018b) outline, there are two very different ways in which such an estimate may be formed. The first is what they call "outcome-based" estimation. Here, past predictions are squared with eventual outcomes to calculate predictive success. This is then used to estimate the relevant likelihoods. That estimate may be as simple as using the relative frequency of truthful reports as a direct stand-in for the likelihood, or a more sophisticated statistical procedure of Bayesian inference may be used where its assumptions are met (e.g., Kruschke 2010). Recast in terms of our earlier urn example, the agent uses repeated draws from the urn—without a partially reliable intermediary, hence 100% fidelity—to assess the likelihoods associated with the urn. Outcome-based update therefore requires definite knowledge on the actual outcomes, in this case, the color of the balls drawn. For a real-world example, one might consider the case of pregnancy tests, where pragmatically certain knowledge that a positive test result was eventually genuinely associated with pregnancy and vice versa for negative test results can be used to estimate the reliability of the test.

However, in many real-world circumstances, knowledge about past outcomes for the source in question may not be available, giving rise to a second way to form such estimates, illustrated in our opening vaccine example. Here, revising beliefs about the reliability of the source on the basis of the congruence of the report(s) with one's current uncertain beliefs provides a potential alternative. Because those beliefs are presently uncertain (in contrast to known outcomes) this second mechanism has been labelled "belief" or "expectation-based" updating (Collins et al. 2018; Hahn 2018b). As discussed above, both the Olsson and the Bovens and Hartmann model formalise such a strategy. To this end, both provide what Bovens and Hartmann (2003) call an *endogenous* model of reliability. The basic likelihoods $P(\text{REP}|H)$, $P(\text{REP}|\neg H)$ capture reliability exogenously, that is, leaving the inferential means by which they are determined outside of the model, drawing only on the estimates that result from those processes; by contrast, both BH and Olsson model include those inferential

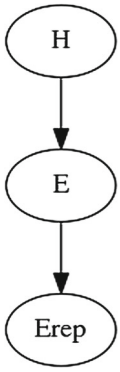

Fig. 3 A hierarchical model in which the reliability of the reporting source is captured explicitly. Three levels are distinguished: the underlying hypothesis H, the evidence E, and the source's actual report of that evidence E_{rep}

means *within* the model. Inference takes place simultaneously about both the truth or falsity of the hypothesis in question and the reliability of the source itself. At any given point in time, these more complex models determine an *effective likelihood* $P(\text{REP}|H)$, but that likelihood changes dynamically over time.

There are other Bayesian ways in the literature to unpack that effective likelihood.[8] For example, Schum (1981) or Pearl (1988) advocated the use of a simple hierarchical model (see Fig. 3) to capture the uncertainty about the veracity of a source's report. Alternatively, standard methods of Bayesian inference for estimating hidden parameters (e.g., Kruschke 2010), could, as mentioned above, be used to infer the effective likelihood from the data. However, the latter ultimately do very different things. Specifically, they generate estimates of the likelihood that change dynamically over time. However, the data are used only for those estimates; the data are not simultaneously used to learn about the truth or falsity of the hypothesis. By contrast, the simple hierarchical model of Schum and Pearl uses the report (i.e., 'data') only to make an inference about the hypothesis: the reliability of the source (i.e., the effective likelihood) remains fixed and unchanging.

Finally, the other popular procedure for dealing with the uncertainty of evidence in general and testimony in particular, Jeffrey conditionalization (Pearl 1998; Jeffrey 2004), also does not capture reliability endogenously. Whereas standard conditionalization assumes that one comes to know E with certainty ($P(E) = 1$ or $P(E) = 0$), Jeffrey conditionalization is less restrictive, and allows contingent observations themselves to have non-extreme probabilities (see Eq. 5).

$$P_{t+1}(H) = P_t(H|E) \cdot P_{t+1}(E) + P_t(H|E) \cdot P_{t+1}(E) \tag{6}$$

In other words, the uncertainty concerning the evidence is folded into a single quantity $P(E)$, which is combined with priors via total probability to yield a probability of the hypothesis in light of that evidence. As the determination of the uncertain

[8] A quite different perspective is taken by the Lehrer-Wagner model (cf. Lehrer 1976), which can be viewed as implementing a sequential process of reliability learning through an iterative weighted linear averaging procedure. However, this style of model differs in its target: It assumes that the agents processed all information from the world in advance, and only have to figure out the implications of their initial weight assessments and credences.

evidence $P(E)$ is entirely external to the inference, there is no model and hence no inferential procedure for deriving it or dynamically revising it over time. In short, whereas the hierarchical model of Schum (1981) unpacks testimony into a report (which is certain, thus allowing conditionalization) and the (uncertain) evidence being reported on, Jeffrey conditionalization treats the evidence (in this case testimony) as itself uncertain.

Both the BH and the Olsson model follow the Schum model in this regard but allow dynamic updating about reliability by implementing expectation-based updating. Set against these alternative suggestions for the treatment of uncertain evidence, the BH and the Olsson model may thus seem like minor formal variants of the same underlying intuition about expectation-based revision. Nevertheless, there are important, and consequential, differences between these two models as well.

As noted above, the first, and most obvious difference, lies in their respective definitions of what it means to be unreliable. For Bovens and Hartmann, the lowest reliability a source can attain is to be entirely non-diagnostic: reports are random with respect to the hypothesis in question. By contrast, the Olsson model considers this to be just one point along the spectrum from full reliability to full anti-reliability, where the source is perfectly anti-correlated with the truth. In other words, the effective likelihood ranges from 0 to 1 in the Olsson model, but only from 0.5 to 1 for BH. It is an interesting empirical question under which circumstances people may be willing to infer anti-reliability of sources in their everyday lives. The existence of so-called "backfire" or "boomerang" effects in the context of persuasion (Nyhan and Reifler 2010) suggest the possibility of anti-reliability in real world scenarios (though other inferences such as arguments from ignorance, see Harris et al. (2013), may also be in play here), and there is at least tentative evidence for anti-reliability effects in lab-based, experimental studies of argumentation (Collins et al. 2018).

The more fundamental difference between the two models is that while the Olsson model allows for a stochastic relationship between the report and the truth or falsity of the hypothesis, the BH model assumes *a fully deterministic relationship*, which the message recipient is merely uncertain about: if the source is reliable, then it accurately reports the true state of the hypothesis in question.

This is not a minor difference. To illustrate we return to the urn task introduced above. The task the Olsson model is addressing is, in effect, this: Imagine, once again, trying to learn whether the urn the experimenter has chosen is the one that contains predominantly red or predominantly blue balls. To inform your choice of hypothesis (predominantly red/predominantly blue) you receive draws from this urn and must revise your beliefs in light of those draws. However, the exact composition of the urn is also unknown; that is, you do not know the exact underlying proportion of red and blue balls, so you do not know how diagnostic a given draw is. Hence you are trying to simultaneously revise your beliefs both about whether the draws are coming from the predominantly red or the predominantly blue urn *and* about the likelihoods, that is, the underlying proportion of red balls. Furthermore, you are even willing to entertain the hypothesis that the colours are inverted (anti-reliability).

By contrast, the BH model assumes the urns contain *just* red balls in the one urn and just blue balls in the other. Once again, one of these urns is selected and you receive a draw. If your source is fully reliable, the ball comes from the urn itself, so it will

definitively tell you which urn you are faced with. If the source is unreliable, however, the ball isn't from the urn at all, and its colour is entirely random with respect to the chosen urn.

In effect, the Olsson model is conceptually trying to address, in entirety, the problem of coming to know the likelihoods, whereas the BH model is trying to determine whether or not a reporting source is fully reliable.

The full consequences of this distinction are best exemplified by considering a *set of reports* from the source, say {report 1 = 1, report 2 = 0, report 3 = 1}. For the Olsson model, that set is, in principle, compatible with a whole range of reliabilities; which posterior reliability is arrived at depends on the (current) priors for both the hypothesis and the source's reliability. By contrast, for the BH model, that data set is *logically inconsistent* with the source being reliable, so that the posterior degree of belief in the reliability of the source will necessarily go to 0, and the posterior for the hypothesis will equal the prior as the report has no diagnostic value—regardless of the priors for both hypothesis and reliability.

This contrast between the two models is related to a further fundamental difference which concerns the way the models have been applied when there are multiple reports from the same source. Theoretical applications of the BH model, whether in the context of conceptual analysis (Bovens and Hartmann 2003) or in application to human behavioral data (Harris et al. 2012; Harris et al. 2016; Jarvstad and Hahn 2011), have all included the multiple reports in the same Bayesian network. This is normatively correct, guarantees optimal inference, and means that there are no differences in outcome depending on whether the data are received all at once, or sequentially. By the same token, the order in which sequential data are received is entirely irrelevant.

In contrast, application of the Olsson model in past research has assumed a 'local', sequential process, where the model is applied iteratively to successive pieces of evidence, assuming simply the posterior of the last update as the new prior, and treating the evidence as otherwise unrelated. This leads to a mis-weighting of that evidence vis a vis the normative (global) model. As outlined above, the Olsson model assumes independence of hypothesis and reliability. But, on receipt of a report, these two variables become dependent (see the phenomenon of "explaining away" in common effect models, e.g., Pearl 1988). Normatively, this dependence also affects the probability of subsequent reports: receiving a first, positive, report affects the probabilities of receiving a second and third. A Bayes net that envisions the possibility of such reports from the outset appropriately factors in these dynamic changes. A purely 'local' sequential update does not, because it fails to factor in the emerging correlation between hypothesis and reliability, giving rise to either over- or underweighting of evidence, conditional on its consistency with expectation. Figure 4 illustrates the contrast:

Figure 4a implements a *global* viewpoint on the possible reports. Such a model includes *all possible* reports that could eventually be received from the source. Hence the updates for a string of sequential updates on evidence reports will lead to the same final result as if these reports had all been received simultaneously. Had they been received simultaneously and updating consequently took place in one step, there would, of course, be no intermediate update of $P(\mathrm{REL})$. As a result, the eventual outcome is the same regardless of whether the reports are received in batches or individually, and regardless of the order in which they arise. Consequently, this global

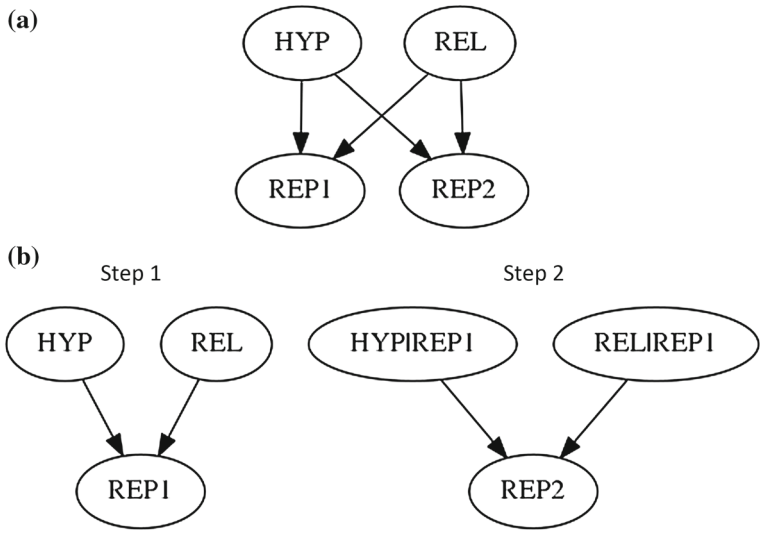

Fig. 4 In Fig. 4a, both reports are part of the initial Bayes net, that is, part of the algebra from the start. In Fig. 4b, the second report only emerges subsequently and is linked to the first report only because the new values of HYP and REL (i.e., HYP|REP1 and REL|REP1) have been copied across after the updates of step 1

perspective may be seen as the normatively correct approach on standard Bayesian analysis.

By the same token, however, the global perspective implies that the focal agent effectively uses the *reliability prior* for their belief update on *all* reports (as any sequential outcome is equivalent to updating on all the evidence at once). So even though the global model provides a posterior for *REL* after obtaining *all* reports, the fact that the same prior uniquely determines all reports conflicts with the idea that the focal agent is actually *dynamically adjusting* source reliability after each report in order to adequately evaluate future reports; the agent calculates a reliability posterior, but that reliability posterior does not have any direct impact on the subsequent belief updates. Any 'learning' of reliability in the global model is arguably epiphenomenal with respect to credence in the hypothesis itself.

If, in contrast, the posterior is used directly as the prior for the next update, one not only enters the world of truly sequential trust updating, but also the world of order effects. Figure 4b illustrates this *local* (inherently sequential) procedure. Here, both reliability and credence in H are updated after one report, and these posteriors are then used as priors to update on the second report, and so forth. Effectively, such a local model relaxes the independence of *REL* and H. As the focal agent estimates credences consistent with the report, these two variables become correlated for the following reports. This creates a dependence of the final state of the credences on the *order* in which reports arrive (Hahn et al. 2018b). It is hard to see this order dependence as rational (unlike seeming order dependence in the context of Jeffrey

conditionalization, where closer scrutiny reveals that other things have also changed, see Osherson 2002).[9]

Because it fails to fully respect dependencies between variables as a result of 'local' application (at least in its typical application), the Olsson model constitutes only a naïve Bayesian agent, rather than the Bayes optimal model.

These differences in use (global vs. local) between the two models make sense in the models' respective domains of application. The BH model has been used to explicate formally fundamental intuitions about evidence, seeking to probe their normative foundations. These analyses have typically involved stylised examples with only a few reports. By contrast, the Olsson model is the core component of agent-based simulations involving many agents in a social network. Computational reasons alone already mean that those agents must ignore the network structure: that is, they treat reports from other agents who might themselves be communicating (Olsson 2011) as independent from one another, even though communication creates dependencies. The local application with respect to multiple reports from a *single source* is in keeping with that limitation.

Furthermore, 'local' updating seems largely unavoidable for actual, real-world agents: it seems impossible to know in advance what kind of future evidence a source may eventually come to report on an issue. The only way an agent could deal with this lack of foresight is to remember all past reports and, after each report received, *retrospectively* form the appropriate global model and recompute using the initial priors for both reliability and hypothesis. This not only seems unrealistic in practice, it also, once again, renders the reliability revision process itself entirely moot.

At the same time, naïve Bayesian models often perform surprisingly well in practical machine learning contexts (e.g., Hand and Yu 2001). So it is of interest to determine how well they solve the source reliability problem in practice.

Hence, we subsequently focus on local (sequential) updating. In order to examine this, we will take a machine learning perspective by defining a simple 'world'; we then simulate performance data from our artificial agents in that world and score their performance. We conduct such simulations both for the Olsson model and for a sequentialised, 'local' version of the BH model (implementing the procedure of Fig. 4b).

Crucially, once the BH model is applied locally, it no longer retains the original consistency constraint. The model in this regard now behaves just like the Olsson model despite its hard-coded logical constraints, because now only one response is considered at any given point in time. The two models can thus be compared more generally, both with respect to accuracy and updated trust, shedding light in particular on the impact of the possibility of viewing a source as 'anti-reliable' (present only in the Olsson model) and the consequences of this for reachability.

[9] One reason that order dependence is usually not something to worry about for medical tests is that their user does not expect to learn their reliability from their own experience in applying the test, but rather from extensive testing that was necessary to admit the test to the market.

4.2 Comparative performance in a simulated world

With these conceptual points in mind, we can now turn to both qualitative and quantitative comparison of model performance within a standardized simulated environment (for full details and additional explanations of the simulation experiments see Appendix B). For Olsson, examining the model in a simulated environment simply extracts a single agent from the social network context, in order to clarify its behavior. For the Bovens and Hartmann model, we developed the sequential variant from the basic model to enable a plausible comparison with the inherently sequential Olsson model.

Hahn et al. (2018b) already conducted such an investigation for the Olsson model. Specifically, they simulated a world in which the focal agent sequentially receives 10 pieces of evidence from the same source. Each piece of evidence was a testimonial report directly asserting the truth or falsity of the target hypothesis. After each piece of evidence, the agent revised its beliefs in both the hypothesis and the reliability of the source. These final beliefs were then compared to the 'true state of the world', the simulation ground truth, for evaluation. This process was then repeated over many such sequences of 10 pieces of evidence in order to provide stable estimates.

The evidence itself was generated according to the 'true' (objective) likelihood governing the source: that likelihood determined the probability that the simulation independently and stochastically generates a true as opposed to false report. In order to provide comprehensive insight into model behavior, the simulations probed behavior across a range of key parameters. Specifically, the simulations examined the range of possible objective likelihoods from 0 to 1 (in increments of 0.05). As there were multiple runs for each likelihood value, it was also possible to vary the base rate at which the hypothesis itself was true or false. This allowed Hahn et al. (2018b) to probe the effects of providing agents with an informative prior, namely true knowledge of the base rate, on both the accuracy of their beliefs in the hypothesis and on their subjective trust in the source. Specifically, the simulations combined the range of possible base rates from 0 to 1 (in increments of 0.05) factorially with all possible objective likelihoods. Finally, Hahn et al. (2018b) examined the impact of using different initial values for the subjective trust in the source. We focus here on the central results for an initial trust distribution of *beta*(2, 1). This distribution has an expected value of 2/3, representing a 'healthy dose of scepticism' concerning the source's reliability, while nevertheless considering the source to provide evidence that is more likely reliable than not.

Finally, to isolate more generally the effects of expectation-based updating, Hahn et al. (2018) also implemented a so-called 'fixed-trust' agent, which does not dynamically modify subjective trust. In order to otherwise match the agents, the fixed-trust agent started with a subjective trust of $p = 0.66$ and simply stuck with this throughout.

The simulation involved, for each parameter combination, 1000 simulated runs of length 10 for these agents, with the evidence at each step generated randomly according to the underlying objective likelihood. The resultant final beliefs were then compared to the true hypothesis on that run, with accuracy measured by the squared error (Brier score, Brier 1950).

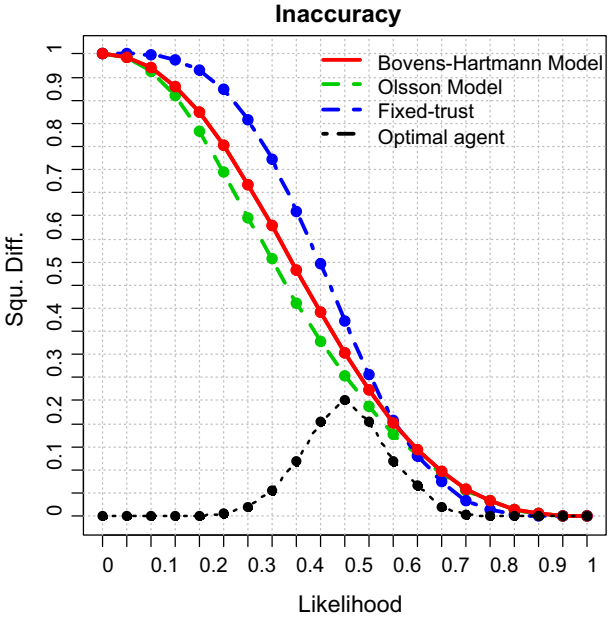

Fig. 5 A comparison of the mean (in)accuracy (squared error) values for the trust-updating agents in the Bovens–Hartmann (BH) model and the Olsson model as well as for fixed-trust agents and optimal agents. (The results for the optimal agent slightly differ from those by Hahn et al. (2018b). Here, we matched the optimal agent more closely to the other agents by taking the objective likelihoods to generate evidence stochastically (instead of using an exact frequency determined by the likelihood)). Shown are the results of simulating 1000 sweeps per data point at the underlying simulated, 21 true likelihoods (x-axis) and 21 base rates also used in the following surface plots. Results are shown aggregated across base-rates, which seemed irrelevant here. Hence each data point in the figure represents the mean across 21,000 runs of agents receiving 10 pieces of evidence. In all cases, the initial degree of belief in claim H itself is 0.5

Figure 5 summarises those results for the simulations in which the agents started with an uninformative prior regarding the hypothesis itself of $P(H) = 0.5$. Graphed are the results for the Olsson agent, the fixed-trust agent and an optimal agent that knows the true source reliability (i.e., the objective likelihood), as seen also in Fig. 4 of Hahn et al. (2018b). However, the figure additionally includes the results of the matching new simulations with the local BH agent.

On the x-axis is the range of objective likelihoods, on the y-axis is the mean accuracy of the final belief in the target hypothesis, averaged over the multiple runs. Figure 5 reveals that there is surprisingly little difference in accuracy between the fixed-trust and the two reliability-updating agents. Not only is the Olsson agent's performance remarkably similar to the fixed-trust agent's (as in Hahn et al. 2018b), the same is true for the fixed-trust and the BH agent, and BH and Olsson agents are almost identical. Examining specifically the range of objective reliabilities below 0.5, we see that *none* of the agents deal successfully with anti-reliability. Neither of the reliability updating agents manages to display adequate reachability, and neither does much better than

the fixed-trust agent.[10] Faced with a 'Cartesian demon' who is a systematic liar, none of the models can learn the anti-reliability of their source, even though the Olsson model is capable of representing it. This is because the models effectively have only the consistency of the data stream itself to revise their beliefs: once they have started out by (mildly) trusting the source, consistent anti-reliable information cannot be perceived as such. But such initial trust is unavoidable in practice, because an Olsson agent who is agnostic about the source's reliability (i.e., $p = 0.5$) cannot learn anything from that source. Unlike the Olsson agent, the BH agent could, in principle, start with a trust prior of 0.5 (because given the right values of the randomization parameter this need not translate into an effective likelihood ratio of 1); but this, of course, is of no help given that anti-reliability cannot be represented in the model.

Reachability becomes possible only when the models are provided with the true base rate of the underlying hypothesis as a prior. Figure 6 shows the extent to which base-rate knowledge allows the models to harness the full power of expectation-based updating for estimating trust. The top row shows trust (subj. likelihood) after 10 pieces of evidence for both the Olsson and the BH model for the models starting with the uninformative prior of $P(H) = 0.5$. The bottom row of Fig. 6 shows that same trust when models start with the base rate as their prior for $P(H)$. Performance is now displayed across different base rates, as the value of the base rate has a significant effect for the prior knowledge agents. For the Olsson model, the mean trust after 10 pieces of evidence approaches the objective likelihood at least for more extreme hypothesis base rates. This can be seen from the fact that the trust values come to lie almost on the diagonal at the front and the back face of the cube containing the corresponding landscape plot (Fig. 6, bottom right panel). Unsurprisingly, the BH model with its more restrictive representation scheme cannot match this performance.

However, the actual accuracy gains that the Olsson model achieves as a result are limited. Figure 7 shows further landscape plots across base rates and objective likelihoods to illustrate this. The top row plots show the mean posterior belief in the hypothesis after the 10 pieces of evidence for the Olsson and BH models, as well as for the fixed-trust agent. If the (mean) posteriors were inferred correctly, the mean posterior in this figures would correspond to the base rates (resulting in a diagonal plane with base-rate = posterior). The middle row of plots shows the corresponding mean accuracy of those beliefs. The bottom row, finally, shows the variance in the resultant accuracy. Whereas Fig. 5 showed the relationship between objective likelihood of the source (x-axis) and accuracy (y-axis) averaged across the range of base rates, the landscape plots of Fig. 7 shows this relationship at each possible value of the base rate (z-axis). Once the agents have access to base-rate knowledge to set their prior for the target hypothesis, the actual base rate has a significant impact on the final beliefs and, with that, the accuracies formed, because of the interaction between hypothesis prior and trust already seen in Fig. 6. As a result, the Olsson agent does noticeably better than both the BH agent and the fixed-trust agent with objective likelihoods below 0.5, that is, when the source is anti-reliable. Specifically, the differences between

[10] This also confirms the robustness claims in Hahn et al. (2018b) who argue that the results of their simulations reflect deep structural problems, and not merely details of the specific implementation of source reliability.

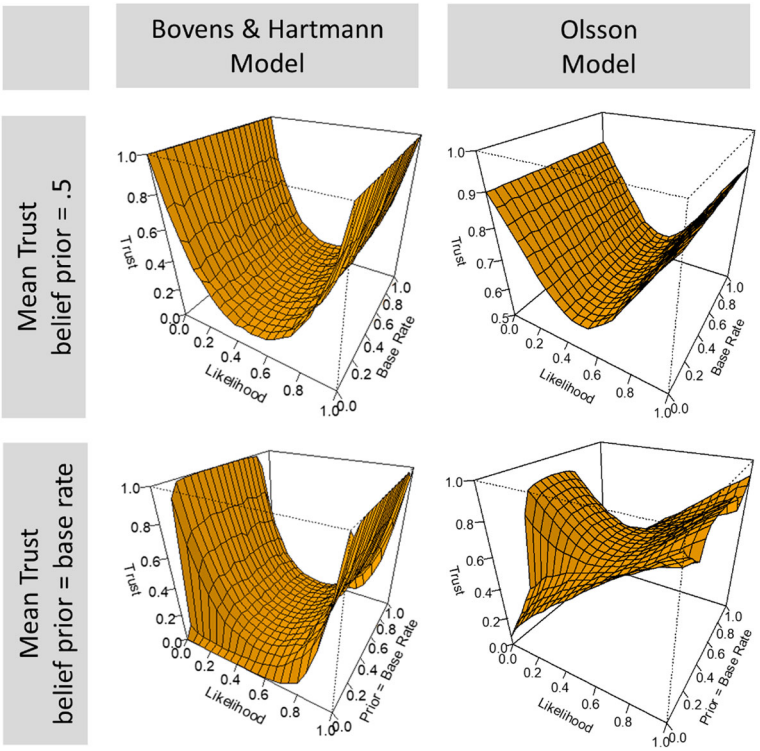

Fig. 6 Comparison of simulation results for trust (subjective reliability representation) for the Bovens and Hartmann Model and the Olsson Model across likelihoods and base rates. The upper panels show mean trust for a belief prior of 0.5. The lower panels show results for a situation with previous knowledge: It shows the same agents when they know the true base rate and use this as the prior

BH and Olsson agents indicate the value of being able to represent 'anti-reliability'. Nevertheless, there remain high levels of error when faced with anti-reliable sources across considerable regions of the space. This is one of the reasons to distinguish reachability and accuracy; although it appears intuitive that an ability to represent anti-reliability should, ceteris paribus, increase accuracy, this is not observed throughout all relevant parameter values in the actual models. Moreover, at high levels of source reliability (i.e., obj. likelihood) and extreme base rates, the Olsson model with base-rate knowledge starts to do *worse*, not only than the BH and fixed-trust models, but also than the Olsson model without base-rate knowledge (cf. Fig. 5).

Figure 8 helps explain why. To put model performance into context, Fig. 8 brackets the space of possible results by showing the accuracy of a simulated agent who simply disregards the evidence altogether and responds with the base rate only, on the one hand, and an optimal agent who starts with knowledge of the hypothesis base rate as a prior and has knowledge of the true, objective likelihood on the other hand.

These two extremes essentially span the range of possible performance and different approaches to trust will fall somewhere in between. In particular, consideration of the base-rate only agent is illuminating because comparing its behavior to that of the

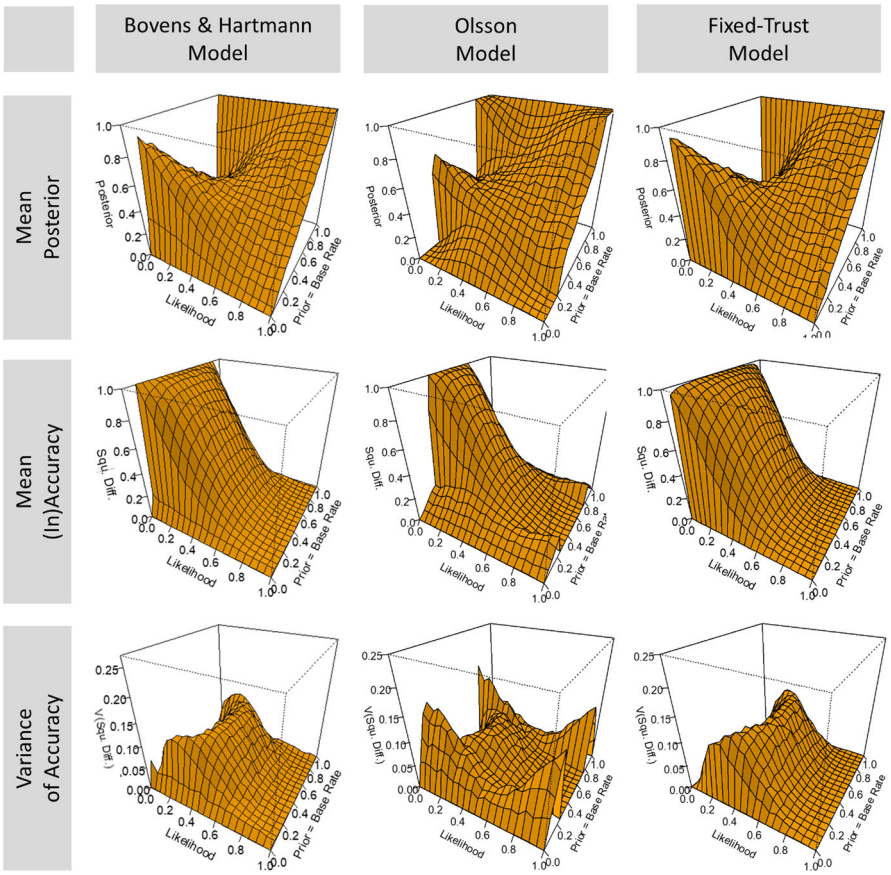

Fig. 7 Comparison of simulation results for the sequential (local) trust updating Bovens and Hartmann model, the Olsson model, and the fixed-trust agent with respect to the target hypothesis. The simulations show results across the range of possible likelihoods and base rates. Each data point in each graph represents the mean (variance) of 1000 runs of agents receiving 10 pieces of evidence at the underlying objective generative likelihood (x-axis) and base rates. The agents' prior belief in the hypothesis is set to the true base rate. The simulations all use an initial trust of 0.66 (or a beta distribution B with $\alpha = 2$, $\beta = 1$ and an expected value of 0.66). Row 1 shows posterior degrees of belief in H, row 2 accuracy as measured by squared error between posterior belief and true value of H, row 3 the variance thereof. (The data underlying the second and third column of graphs are from Hahn et al. (2018b), Fig. 6)

Olsson model (but also the BH model) at extreme base rates indicates that for those base rates, the two expectation-based updating models are so heavily influenced by those expectations that they virtually ignore the data. This underscores further the limitations of a purely expectation-based updating strategy when it comes to estimating source reliability.

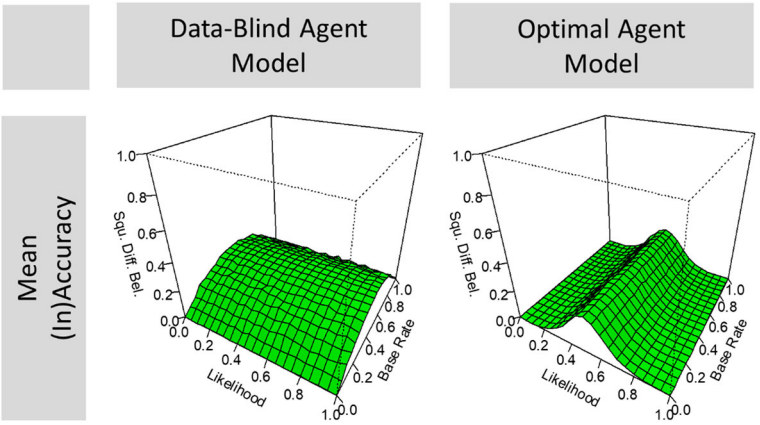

Fig. 8 Performance of the optimal agent (left panel) and a data-blind agent (right panel) who responds merely with the true underlying base rate

5 Conclusions

In this paper, we have critically evaluated the two main models of source reliability in the formal epistemology literature. Though the models are similar in spirit in that both implement so-called expectation-based (or belief-based) updating, there are formal differences that are consequential for model performance. The BH model, in its standard version, only represents trust versus randomisation, whereas the Olsson model is meant to learn the true likelihood embodied in the data generating process, including the representation of anti-reliability. This implies that, in its global formulation, the BH model dismisses all reports once an inconsistency in the data stream is identified. This feature is eliminated in the sequential version.

When both models are applied as local, sequential, naive Bayesian agents in our simulations without informing the agents with prior knowledge, the resultant accuracy of both models is arguably more similar than it is distinct. Moreover, relative to both, the credulous fixed-trust agent remains surprisingly competitive (for a similar result in a very different context, see also Zollman 2015).

What the simulations of the sequential models show is that, first and foremost, neither model is able to fundamentally solve the problem of source reliability in circumstances where there is no reference class of relevant past predictive success that enables outcome-based estimates of the relevant likelihoods. Neither model offers an effective solution to bootstrapping an accurate reliability estimate in order to evaluate the truth of a given hypothesis in light of evidential reports.

This constitutes a significant normative gap. As Hawthorne (1994) points out, Bayesian confirmation theory relies on the scientific community agreeing on the likelihoods; where there is no solid basis for that the entire endeavour remains up in the air. Likewise, present day concerns about the integrity of our everyday information environments, whether these concern politics, vaccines or climate change, seem unlikely to be fully solved if normative solutions remain elusive. And, finally, the limitations

of the models examined here have potential consequences for the normative standing of the wider explanatory projects within epistemology that these models have served.

Two distinct normative limitations emerged from our analyses, one through simulation, the other through conceptual analysis. First, our simulations illustrate that an ability to represent anti-reliability formally is not enough to guarantee reachability in practice. Needless to say, a model that cannot represent anti-reliability necessarily fails in cases where a source actually is anti-reliable. The BH model could easily be revised to make reliability a ternary variable (reliable, randomizer, liar),[11] bringing it more in line with the Olsson model in this regard. But as the simulations with the Olsson model show, the representational capacity itself is not enough. More fundamentally, the focal agent's decision about which kind of situation they are in is unduly determined by the initial trust—which itself must be constrained for pragmatic reasons. Providing the agents with relevant prior knowledge about (the base rate of H) provided gains for, at least, some regions of the parameter space. But even that did not provide reachability in general: there remain significant regions of the space where the Olsson model cannot learn that a source is anti-reliable.

Second, comparing the BH and the Olsson model in their original formulation made salient the global/local dimension. Only global versions of the expectation-based updating models could qualify as optimal Bayesian models, so only they could be deemed fully rational in the standard Bayesian sense. In order to avoid over/underweighting of the evidence, the optimal model must represent the possibility of all future reports within the initial algebra or recalculate beliefs entirely anew after every report. This guards against order effects as experienced by the naïve Bayesian versions, and only on the global version are the results of sequential and batch updating equivalent. But the very fact that considering all of the evidence at once yields the same result for the global model indicates that the potential revisions to reliability are not actually *inferentially relevant*. This results in an undue influence of the reliability prior. Only the initial *prior* for reliability influences the posterior belief in the hypothesis. This means that, unlike priors for the hypothesis, the prior for reliability doesn't 'wash out' as more evidence is received. Even more worryingly, it also means that the global models do not actually provide a procedure for reliability updating. If final results are ultimately determined only by the initial reliability, the global model cannot be said to be a model of adaptively learning the reliability of its sources. In short, deeper consideration of the global/local issue reveals that neither local nor global versions of these models provide fully adequate normative solutions, albeit for very different reasons: the local model because it is subject to systematic mis-weighting of evidence and order effects; the global model because it arguably fails to adequately address the problem it is trying to solve.

As a result, current formal models of source reliability still fall short of a compelling normative treatment of testimony. This is not to deny that both the Bovens and Hartmann and the Olsson model have already proved invaluable in clarifying and probing the issues. The limitations emerging, in our view, reflect deep and challenging philosophical issues, not simple model defects. And these issues could not have been formulated with this clarity without the models in question. Hence they have signifi-

[11] Such an extension is discussed by Olsson (2005, ch. 4.3) with regard to the problem of coherence.

cantly furthered understanding of the problems posed by testimony. We thus consider these models, and future models like them, essential to the philosophical project both of understanding testimony and deriving satisfactory normative approaches to it.

However, it may be that these particular models prove more enduring in the context of more descriptive-explanatory applications. Part of the fundamental appeal of the models lies in the fact that human beings seem to engage in something like expectation-based updating, whether one ultimately comes to view this as normative or not. Simulations with naïve Bayesian agents thus offer the possibility of deep insight into phenomena such as group polarization (Olsson 2013) or the impact of changes to the topology of our everyday information networks through the rise of social media (e.g., Hahn et al. 2018a).

At the same time, such simulations may help further develop our normative intuitions by supplementing *ex ante* assumptions about rationality with *ex post* evaluation based on the observed consequences of a strategy. This style of argument is already standard in the form of thought experiments. Simulations provide further evidence of unforeseen consequences for strategies that seem relevant to normative evaluation: the fact that expectation-based updating offers so little benefit over simple, fixed-trust strategies, for example, would have been difficult to foresee. Likewise, the fact that the benefit of considerable advance knowledge on the truth or falsity of the target hypothesis helps calibrate trust but still does comparatively little for accuracy would have been practically impossible to see. These results are disappointing for the expectation-based updating strategy. If simulations then additionally highlight further consequences such as the rise of polarization (Olsson 2013; Hahn et al. 2018), this may shed further doubt on the desirability and hence rationality of expectation-based estimates of source reliability.

In short, what intuitively seems rational in the abstract, may seem less so, once the application of a strategy is made progressively more concrete. Formal models allow the systematic exploration of descriptive consequences, both for individual agents and assemblies of such agents in social networks; from their application, a revised set of benchmarks for an adequate normative solution may eventually take shape.

Acknowledgements Open Access funding provided by Projekt DEAL. This project was funded by the Humboldt Foundation's "Anneliese Meier Research Award" to Ulrike Hahn.

Funding Funding was provided by Alexander von Humboldt-Stiftung.

Appendix A: details of the models

A.1 Overview

A.1.1 Purpose

This model represents an agent receiving evidence from the world and simultaneously trying to learn the truth about a certain hypothesis H and the reliability of the evidence on H it receives.

A.1.2 State variables and scales

The agents are defined by their subjective degree of belief in H, their estimation of the reliability of their data source (represented depending on the agent type by either a point probability or distribution) and their belief revision rule (see Submodels for details).

The world external to the agent is defined by the base rate of H being true across instantiations of the model and the objective likelihood of generating a piece of evidence representative of the status of H in a given instantiation.

Model time is discrete and the mapping between model and target time is determined by the frequency of inquiry (i.e. if collecting one piece of evidence takes 1 month in a given research domain, one timestep of the model corresponds to 1 month). The model operates only on a single time scale.

Overview[12]:

- $P_{obj}(H = 1)$: The probability of H being true across the relevant class of worlds (base rate).
- $P_{obj}(E = 1|H = 1)$: The objective reliability of the agent's data source, which it is trying to learn (objective likelihood).
- $P_{subj}(H = 1)$, short: $b_t(H)$: The agent's subjective credence in H at t.
- $P_{subj}(E = 1|H = 1)$, short R_t: The agent's subjective estimation of the reliability of its source (trust) at time t.

A.1.3 Process overview and scheduling

The model contains the following processes:

1. *Data generation* A piece of evidence is randomly generated by the world and passed on to the agent (all evidence is binary).
2. *Updating* This process consists of two subprocesses.
   (a) *Belief revision* The credence in H is updated according to the agent's current state and the data.
   (b) *Trust revision* The agent's estimation of its source's reliability is updated according to its current state and the data.

[12] We adapted variable names to highlight the structural similarities between the models at the expense of similarity to the original publications.

Belief and trust revision are executed synchronously.

A.2 Design concepts

- *Sensing* Agents only access the state of the world (represented by the value of H) via the evidence they receive. In addition, they are only capable to process binary evidence.
- *Stochasticity* All randomness is encapsulated in the data generation process, determined by the base rate and the assumed objective reliability. It represents the variation in data source reliability and fruitfulness; it represents any kind of random error in observation.
- *Observation* For the purpose of our experiments, only the final belief and trust are recorded. It is, however, both possible and meaningful to observe intermediate states of the model.

A.3 Details

A.3.1 Initialization

Technically, both degree of belief and trust could be initialized with any mathematically admissible value (though the models become trivial if any probabilities are instantiated as 1 or 0). If not declared otherwise, the initial values for the experiments are:

- Prior belief: uniform ($b_0(H = 1) = b_0(H = 0) = 0.5$).
- Prior trust: moderately trusting ($r_0 = 0.66$ for BH, $r_0 = \text{beta}(2, 1)$ with expected value or mean $M[r_0] = 2/3$ for O).

A.3.2 Submodels

Data is generated via a Bernoulli process. At the begin of a model run, the value of $H \in \{0, 1\}$ is drawn according to $P_{obj}(H)$. Then, whenever the agent receives data, a piece of evidence is drawn according to $P_{obj}(E = 1|H = 1)$.

It is assumed that $P(E{=}1|H{=}1){=}P(E{=}0|H{=}0)$ (both for the agent's subjective estimate and the objective probabilities of the data generating process).

On receiving a piece of data, the agent updates belief and trust; there are two types of agents, we call them Bovens–Hartmann-agents (BH) and Olsson-agents (O).

A.3.2.1 Bovens–Hartmann The key assumption for this model is that every source is either perfectly reliable or a strict randomizer:

- Every reliable source is perfectly reliable, i.e. reports in accordance with the true value of H with probability 1.
- Every source that is not perfectly reliable is a randomizer, i.e. will report 1 with probability a and 0 with probability $1–a$, regardless of the value of H.

According to these assumptions, reliability is represented by the binary variable *REL*.

The resulting model is described by the following equations to calculate posterior belief in H, b_{t+1} and posterior trust in the source, r_{t+1}, conditioning on a report that $H = 1$ is true:

$$b_{t+1}(H) = b_t(H|E) = \frac{b_t(H)(r_t + (1 - r_t)a)}{b_t(H)(r_t + (1 - r_t)a) + (1 - b_t(H))(1 - r_t)a}$$
$$r_{t+1} = r_t(REL|E) = \frac{b_t(H)r_t}{b_t(H)(r_t + (1 - r_t)a) + (1 - b_t(H))(1 - r_t)a}$$

A.3.2.2 Olsson The Olsson variant of updating uses a point probability to represent the degree of belief in H as well, but a probability distribution $r(x)$ for the estimation of source reliability. Otherwise, the probabilities have been adapted to mirror the notation we used for the Bovens–Hartmann variant as closely as possible. $M(r)$ denotes the mean (or expectation) of the distribution, here of the trust values r.

$$b_{t+1}(H) = b_t(H|E) = \frac{M[r_t]b_t(H)}{M[r_t]b_t(H) + (1 - M[r_t])(1 - b_t(H))}$$
$$r_{t+1}(x) = r_t(x|E) = \frac{xb_t(H) + (1 - x)(1 - b_t(H))}{M[r_t]b_t(H) + (1 - M[r_t])(1 - b_t(H))} r_t(x).$$

A.3.2.3 Simulation algorithm

1. Initialize priors (b_0 and r_0).
2. Draw H according to $P_{obj}(H)$.
3. REPEAT n times:
   (a) Generate evidence E according to $P_{obj}(E|H)$.
   (b) Set b to $b(H|E)$
   (c) IF AGENT-TYPE is BH:
   (d) Set r to $r(REL|E)$
   (e) ELSE IF AGENT-TYPE = O:
   (f) FORALL x set $r(x)$ to $r(x|E)$
4. END.

Appendix B: details of the similation experiments

In this paper, we mainly present results from two simulation experiments that were run in parallel for the BH and O models.

Both experiments systematically vary central parameters (for additional settings, cf. A.3.1 and A.3.2) and involve variants, investigating BH, O, and fixed-trust agents. All agents sequentially obtained 10 pieces of evidence. For every parameter combination, this updating procedure was repeated 1000 times.

B.1 Experiment 1—without prior knowledge

In Experiment 1 we varied the generative probability of H, $P_{obj}(H = 1)$, (i.e., the base rate) using [0, 0.05,…, 0.95, 1] and the true reliability, $P_{obj}(E|H)$, using [0, 0.05, …, 0.95, 1]. For each of the resulting 441 parameter combinations, 1000 model runs were conducted.

B.2 Experiment 2—with some prior knowledge

Experiment 2 varied the parameters as in Experiment 1, but investigated what happens if the agents have knowledge of base rate of H (without knowing the actual status of H). Thus the prior $b_0(H)$ was set to the varied base rate, $P_{obj}(H)$ with [0, 0.05, …, 0.95, 1]. Experiment 2 likewise involved 441,000 model runs.

B.3 Dependent variables

The recorded variables, associated with the parameter setting for each run, store information on whether the hypothesis was actually true or false, the resulting subjective belief and the resulting subjective reliability. After the simulation we aggregated these values over agents with the same parameter combinations. From the dependent variables, we calculate the Brier score as (in)accuracy measure; that is the squared difference between an agent's subjective belief and the true status of H (either 0 or 1), $(b_t(H) - H)^2$.

B.4 Figures

Figure 5 shows for Experiment 1 the mean squared differences (inadequacy) the two models and the fixed-trust agent (for an optimal agent who knows the true likelihood). The results are aggregated over the reliabilities that turned out to be irrelevant here.

The surface plots in Fig. 6 compare mean trust ($M[r_t]$), after updating. The upper panels show results for Experiment 1, the lower ones for Experiment 2, the left ones for BH, and the right ones for O. Each panel shows results covering base rates, $P_{obj}(H)$, and objective reliabilities, $P_{obj}(E|H)$.

The surface plots in Fig. 7 present for Experiment 2 mean beliefs, mean squared differences, and variance of squared differences for BH, O and fixed-trust agents.

Finally, Fig. 8 shows the results of two additional simulations of the mean inaccuracy (squared differences), exploring the same basic parameter space as in Experiment 1. However, now results are shown either for data-blind agents (i.e., agents who do not use data but base rate only) or optimal agents (who know the actual likelihood). See main text for the interpretation of the figures.

References

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

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3.

Coady, C. A. J. (1992). *Testimony: A philosophical study*. Oxford: Oxford University Press.

Collins, P. J., & Hahn, U. (2019). We might be wrong, but we think that hedging doesn't protect your reputation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000786.

Collins, P. J., Hahn, U., von Gerber, Y., & Olsson, E. J. (2018). The bi-directional relationship between source characteristics and message content. *Frontiers in Psychology, 9,* 18.

Douven, I. (2010). Simulating peer disagreements. *Studies in history and philosophy of science part A, 41*(2), 148–157.

Elga, A. (2007). Reflection and disagreement. *Noûs, 41*(3), 478–502.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive Science, 37*(1), 61–102.

Goldman, A. I. (2001). Experts: Which ones should you trust? *Philosophy and Phenomenological Research, 63*(1), 85–110.

Hahn, U., Hansen, J. U., & Olsson, E. J. (2018a). Truth tracking performance of social networks: How connectivity and clustering can make groups less competent. *Synthese*. https://doi.org/10.1007/s11229-018-01936-6.

Hahn, U., & Harris, A. J. (2014). What does it mean to be biased: Motivated reasoning and rationality. *Psychology of Learning and Motivation, 61,* 41–102.

Hahn, U., Harris, A. J. L., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic, 29,* 337–367.

Hahn, U., Merdes, C., & von Sydow, M. (2018b). How good is your evidence and how would you know? *Topics in Cognitive Science, 10*(4), 660–678. https://doi.org/10.1111/tops.12374.

Hahn, U., Oaksford, M., & Harris, A. J. L. (2012). Testimony and argument: A Bayesian perspective. In F. Zenker (Ed.), *Bayesian Argumentation* (pp. 15–38). Dordrecht: Springer.

Hahn, U., von Sydow, M., & Merdes, C. (2019). How communication can make voters choose less well. *Topics in Cognitive Science., 11,* 194–206.

Hand, D. J., & Yu, K. (2001). Idiot's Bayes—Not so stupid after all? *International Statistical Review, 69*(3), 385–399. https://doi.org/10.2307/1403452.

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.

Harris, A. J. L., Corner, A. J., & Hahn, U. (2013). James is polite and punctual (and useless): A Bayesian formalization of faint praise. *Thinking & Reasoning, 19,* 414–429.

Harris, A. J. L., Hsu, A. S., & Madsen, J. K. (2012). Because Hitler did it! Quantitative tests of Bayesian argumentation using "ad hominem". *Thinking & Reasoning, 18,* 311–343.

Hawthorne, J. (1994, January). On the nature of Bayesian convergence. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (Vol. 1994, No. 1, pp. 241–249). Philosophy of Science Association.

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

Jeffrey, R. (2004). *Subjective Probability: The Real Thing*. Cambridge: Cambridge University Press.

Kruschke, J. K. (2010). *Doing Bayesian data analysis*. Burlington: Academic Press.

Lackey, J. (2011). Testimony: Acquiring knowledge from others. In A. Goldman & D. Whitcom (Eds.), *Social epistemology: Essential readings* (pp. 71–91). Oxford: Oxford University Press.

Lehrer, K. (1976). When rational disagreement is impossible. *Noûs, 10*(3), 327–332.

Leitgeb, H., & Pettigrew, R. (2010). An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science, 77*(2), 236–272.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303–330.

Olsson, E. J. (2005). *Against coherence: Truth, probability, and justification*. Oxford: Oxford University Press.

Olsson, E. J. (2011). A simulation approach to veritistic social epistemology. *Episteme, 8*(02), 127–143.

Olsson, E. J. (2013). A Bayesian simulation model of group deliberation and polarization. In F. Zenker (Ed.), *Bayesian argumentation, the practical side of probability* (pp. 113–133). Dordrecht: Springer.
Olsson, E. J., & Vallinder, A. (2013). Norms of assertion and communication in social networks. *Synthese, 190*(13), 2557–2571.
Osherson, D. (2002). Order dependence and Jeffrey conditionalization. http://www.princeton.edu/~osherson/papers/jeff3.pdf. Retrieved January 10, 2017.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo: Morgan Kaufman.
Phillips, L., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72,* 346–354.
Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34,* 243–281.
Schum, D. A. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance, 27,* 153–196.
Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Evanston: Northwestern University Press.
Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science, 15*(3), 436–447.
Vallinder, A., & Olsson, E. J. (2014). Trust and the value of overconfidence: A Bayesian perspective on social network communication. *Synthese, 191*(9), 1991–2007.
Wilholt, T. (2013). Epistemic trust in science. *The British Journal for the Philosophy of Science, 64*(2), 233–253.
Zollman, K. J. (2015). Modeling the social consequences of testimonial norms. *Philosophical Studies, 172*(9), 2371–2383.