COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2015) 1–38

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12276

The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach

Adam J. L. Harris,[a] Ulrike Hahn,[b] Jens K. Madsen,[b] Anne S. Hsu[c]

[a]*Department of Experimental Psychology, University College, London*
[b]*Department of Psychological Science, Birkbeck College, London*
[c]*School of Electronic Engineering and Computer Science, Queen Mary University, London*

Received 4 April 2014; received in revised form 1 April 2015; accepted 13 April 2015

Abstract

The appeal to expert opinion is an argument form that uses the verdict of an expert to support a position or hypothesis. A previous scheme-based treatment of the argument form is formalized within a Bayesian network that is able to capture the critical aspects of the argument form, including the central considerations of the expert's expertise and trustworthiness. We propose this as an appropriate normative framework for the argument form, enabling the development and testing of quantitative predictions as to how people evaluate this argument, suggesting that such an approach might be beneficial to argumentation research generally. We subsequently present two experiments as an example of the potential for future research in this vein, demonstrating that participants' quantitative ratings of the convincingness of a proposition that has been supported with an appeal to expert opinion were broadly consistent with the predictions of the Bayesian model.

Keywords: Argumentation; Appeal to authority; Appeal to expert opinion; Epistemic authority; Bayesian probability; Quantitative modeling

1. Introduction

Concepts of trust and expertise are important in human life. From an early age, we build our knowledge of the world and language from direct experience, but also from the testimonies of those in a better position to know than ourselves (e.g., our parents—see Harris & Corriveau, 2011, for a review of developmental research showing infants' and children's ability to differentiate between reliable and unreliable information sources). This reliance on testimony extends throughout adulthood—we let our doctor diagnose us; the weatherman forecasts for us; reviews direct us toward movies to see or books to buy;

Correspondence should be sent to Adam J. L. Harris, Department of Experimental Psychology, University College London, 26, Bedford Way, London WC1H 0AP, UK. E-mail: adam.harris@ucl.ac.uk

and more formally, witnesses (both expert and lay) influence decisions within the courtroom. We do not, however, blindly follow the advice or opinions of others. Central features of testimony that help us to determine its truth are the trustworthiness and expertise of the source, as well as how the testimony's content fits with our own opinion on the matter.

This fundamental importance of the concepts of trust and expertise has led to substantial coverage across a wide variety of psychological research areas. In the current paper, we will define "trust" and "expertise" as distinct concepts, but in past research the terms have often been used interchangeably. In addition to developmental psychology, other research areas that have been interested in these concepts include the following: judgment and decision making (e.g., Birnbaum & Mellers, 1983; Birnbaum & Stegner, 1979), reasoning research (Stevenson & Over, 2001; Wolf, Rieger, & Knauff, 2012), and social psychological research into persuasion and attitude change (e.g., Brinol & Petty, 2009; Chaiken, 1980; Hovland, Janis, & Kelley, 1953; McGuire, 1985; O'Hara, Netemeyer, & Burton, 1991; Petty & Cacioppo, 1981; Pornpitakpan, 2004—also incorporating direct applied research in advertising, for example, Braunsberger & Munch, 1998; Ohanian, 1991; Wiener & Mowen, 1986). Moreover, the concepts of trust and expertise are also of vital importance in the evaluation of legal testimony, and research has concerned both formalizations of how testimony *should* be viewed (e.g., Friedman, 1987; Hahn, Oaksford, & Harris, 2012; Lagnado, Fenton, & Neil, 2013; Schum, 1981, 1994; Walton, 2008a), and descriptive studies investigating the degree to which people are sensitive to different relevant aspects of a witness's testimony (e.g., Eaton & O'Callaghan, 2001; ForsterLee, Horowitz, Athaide-Victor, & Brown, 2000; Harris & Hahn, 2009; Krauss & Sales, 2001; see Wells & Olson, 2003, for a review). The importance of trust and expertise for humans, and hence its interest for researchers in psychology, predicts its importance for artificial intelligence systems, and hence its interest for computer scientists. For example, information search systems must delineate between trustworthy and non-trustworthy sources of information (e.g., Balakrishnan & Kambhampati, 2011; for a review of trust research in computer science, see Artz & Gil, 2007). Concepts pertaining to trust and expertise are also of central interest to a research area spanning psychology and computer science, as well as philosophy: namely, argumentation (e.g., Hahn, Harris, & Corner, 2009; Hahn et al., 2012; Walton, 1997, 2008a). Here, trust and expertise are of particular importance in the evaluation of the appeal to expert opinion, which has long been recognized as a distinct argument form (see e.g., Hastings, 1962; Schellens, 1985; Kienpointner, 1992; and, in particular, Walton, 1997, and references therein). This argument form is the focus of the theoretical and empirical work presented here, providing a bridging point for divergent disciplines who share a mutual interest in trust and expertise research.

1.1. The appeal to expert opinion

The "appeal to expert opinion" or "appeal to authority" "uses the opinion of a respected authority or expert on a subject as a positive personal argumentation to

support one's own side of an argument" (Walton, 2008b, p. 209). We use the term "appeal to expert opinion," since we follow the majority of textbook treatments (see Walton, 1997) in being specifically concerned with appeals to epistemic authority (what Walton, 1997, refers to as "cognitive authority"), as opposed to what Walton (1997) terms "administrative authority." Epistemic authority relates to the authority of those with superior knowledge in a specific field—experts. Administrative authority, by contrast, relates to those who have had authority bestowed upon them, and are thus in a position of power. For example, the police are in a position of administrative authority, according to the law of the land. Hence, if they inform you that you should drive more slowly, you should heed them because you are legally obliged to do so. By contrast, a medical doctor possesses epistemic authority pertaining to health matters. For that reason, if she were to advise me to take my medicine twice a day, I would be wise to comply, but not legally or morally obliged. I would likely comply because I perceive her to possess more knowledge than I do on the benefits of various medicine routines.[1]

Following a thorough review of the extant literature, Walton (1997) provides a scheme-based treatment of the appeal to expert opinion (on argumentation schemes more generally see, for example, Hastings, 1962; Kienpointner, 1992; Garssen, 2001; and within computer science, for example, Reed & Rowe, 2004; and Rahwan & Simari, 2009; for a comprehensive overview of the literature on argumentation schemes see Walton, Reed, & Macagno, 2008). From this scheme-based perspective, the appeal to expert opinion is viewed as a fallible, defeasible argument form, which is certainly not inherently fallacious, but can be considered stronger or weaker depending on the degree to which six key criteria are met (Table 1). Walton's focus is on the dialectical context of the argument—that is, the wider argumentative exchange within which the argument is put forward. In general, dialectical approaches to argumentation maintain that the evaluation of arguments must take into account the dialectical nature of argumentation as a structured exchange between proponents and opponents putting forward claims and counter-claims. In keeping with this focus, Walton presents his evaluation criteria as questions that an opponent in an argumentative exchange should raise to a proponent putting forward an appeal to expert opinion. The strength or weakness of a

Table 1
Six key criteria for the appeal to expert opinion (Walton, 1997, p. 223; Walton, 2008b, p. 218)

Expertise question:	How credible is the source as an expert source?
Field question:	Is the source an expert in the field that the issue concerns?
Opinion question:	What did the source assert that implies the conclusion?
Trustworthiness question:	Is the source a personally reliable source?
Consistency question:	Is the conclusion consistent with what other expert sources assert?
Backup evidence question:	Is the source's assertion based on evidence?

particular instance of this argument form can then be evaluated as a function of how satisfactorily these questions can be answered.

The questions outlined in Table 1 originate from Walton's (1997, p. 102) description of the argument form of the appeal to expert opinion being:

"E [for expert] is a genuine expert in *S* [the subject under discussion].

E asserts that *A*.

A is within *S*.

A is consistent with what other experts say.

A is consistent with available objective evidence (if any is known).

Therefore, *A* can be accepted as a plausible presumption."

Walton thus sees the appeal to expert opinion as a type of presumptive argument: So long as the premises in the first five lines hold, the conclusion holds—with the qualifier that *A* is plausible (rather than certain as in logical, deductive inference). The potential for each of these premises to be true to *some degree* rather than dichotomously true or false, combined with the recognition that non-fallacious instances of the appeal to expert opinion can range from very weak to very strong make a probabilistic treatment of this argument type appealing. Walton (1997, p. 121) himself states, "We leave open whether it could be analyzed as some species of subjective probability, of the kind studied in statistics" (though see additional critiques of the probabilistic approach in Walton, 2008a, pp. 33, 92–102). In this paper we aim to describe such a probabilistic analysis.

Hahn et al. (2012) presented a Bayesian formalization of the appeal to expert opinion, capturing its key characteristics (as identified by Walton, 1997, 2008a,b; see Table 1) within a Bayesian network, which we describe next. As we show in this paper, such a formalization allows one to move beyond the qualitative considerations of the scheme-based approach and to provide normative evaluation that highlights not just individual factors influencing argument strength, but also their interactions, and provides specific quantitative evaluation of the relative convincingness of individual exemplars of the appeal to expert opinion. In showing how the approach can be applied to specific examples, we show also how it can be used to guide detailed empirical investigation of people's actual evaluation of appeals to expert opinion. The two empirical studies we present show initial support for the contention that people's evaluations of this argument form are broadly consistent with the predictions derived from Hahn et al.'s (2012) Bayesian formalization. In the next section, we briefly introduce the Bayesian approach to argumentation, before describing how Walton's (1997) argumentation scheme for the appeal to expert opinion might be instantiated within a specific Bayesian model.

1.2. The Bayesian approach

Hahn and Oaksford (e.g., 2006, 2007a) put forward a Bayesian theory of informal argumentation. The Bayesian approach to argumentation, and to reasoning more generally, proposes that an individual's degree of belief in a particular proposition, or hypothesis, can be represented as a subjective probability between 0 and 1 (see also e.g., Evans & Over, 2004; Howson & Urbach, 1996; Oaksford & Chater, 1998, 2007). Upon receiving a new piece of evidence—for example, a testimony from an expert source—the normative way in which this evidence should be integrated with one's previous belief is given by Bayes' rule:

$$P(H|e) = \frac{P(H)P(e|H)}{P(H)P(e|H) + P(\neg H)P(e|\neg H)} \quad (1)$$

In Eq. 1, $P(H|e)$ represents one's posterior degree of belief that some hypothesis is true, given the evidence, e, which one has just received. From Eq. 1, it can be seen that this is a function of $P(H)$, one's initial, *prior*, belief in the hypothesis (before receiving evidence) and the relationship between $P(e|H)$, that is, the likelihood of receiving the evidence if the hypothesis is indeed true, and $P(e|\neg H)$, the likelihood of receiving the evidence if the hypothesis is in fact false. This relationship is captured by the likelihood ratio, $\frac{P(e|H)}{P(e|\neg H)}$. When relating to the reliability of a source, the more reliable a source is, the greater the likelihood ratio (on Bayesian approaches to source reliability, see Bovens & Hartmann, 2003; Corner, Harris, & Hahn, 2010; Hahn et al., 2009, 2012; Wang et al., 2011). In the basic case where e corresponds simply to an assertion of H by a source, a likelihood ratio below 1 indicates a source who is essentially a liar—they are more likely to provide positive evidence if the hypothesis is false than if it is true, whilst a likelihood ratio equal to 1 represents a maximally uninformative source—for they are equally likely to provide positive evidence if the hypothesis is false as if it is true. In the simplified appeal to expert opinion we will consider in our experiment, the expert is asserting H. We will make this clearer in our notation by replacing e with H_{rep}. Thus, the likelihood ratio concerns the relationship between the likelihood of the source reporting that H is true when indeed it is, $P(H_{rep}|H)$, and an erroneous report that H is true when, in fact, it is not, $P(H_{rep}|\neg H)$.

From the above, we have the core component of the Bayesian account of the appeal to expert opinion. One can see that the argument is likely to be more persuasive the greater the value of $\frac{P(H_{rep}|H)}{P(H_{rep}|\neg H)}$, and will have some influence on increasing one's belief in the truth of the hypothesis under consideration wherever this likelihood ratio is greater than one. Can we, however, more closely tie the Bayesian approach to those issues previously identified by philosophers (e.g., Table 1)? We can, and in so doing, we can make the account more amenable to empirical testing.

Hahn et al. (2012) developed a conceptualization of the appeal to expert opinion using a Bayesian network. The network proposed is shown in Fig. 1. The first aspect of this approach to note is that it is not a *different* approach to evaluating the likelihood ratios from that outlined above. Trustworthiness and expertise can be captured in the likelihood

ratio of $P(H_{rep}|evidence)$ (as spelled out in Hahn et al., 2009; see also, Schum, 1981), and $P(H_{rep}|H)$ can then be obtained through marginalizing out the conditional probabilities that depend on the *evidence* using the formula for marginalization and the chain rule for joint probabilities. Thus, in Fig. 1, evidence, trustworthiness, and expertise are intrinsic to the model. If only H_{rep} and H are explicit in the model; however, the other components can be considered to be captured extrinsically in the assignment of $P(H_{rep}|H)$ (see also Bovens & Hartmann, 2003).

The reports of other experts (S2 and S3 in Fig. 1, where all experts' reports are conditionally independent given H) will result in different degrees of belief in the hypothesis, which can be captured in the prior probability, $P(H)$ in Eq. 1.[2] The Bayesian network is a principled way of representing the conditional dependencies between different concepts (Pearl, 2000). We next outline how Fig. 1 captures all the key criteria outlined in Table 1 (see also Hahn et al., 2012).

The first two questions, "expertise" and "field," are captured by the expertise node. In considering these two questions in a single node, we recognize that expertise is typically narrow in focus and therefore must be within the field in question so as to qualify as expertise (on the field specificity of expertise, see Ericsson & Lehman, 1996). The "opinion" question is captured by the probabilistic relationship between H and H_{rep}. In designating the expert's report as H_{rep}, we are modeling a situation in which the expert is

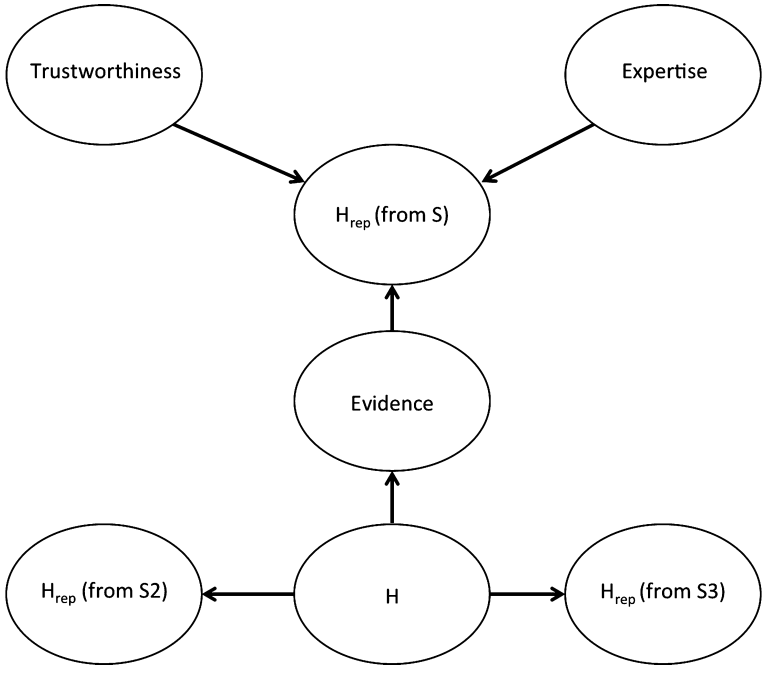

Fig. 1. The Bayesian network proposed as a formalization of the appeal to expert opinion by Hahn et al. (2012). The expert source is represented by the letter *S*. *S* is the expert whose report is being used in the argument, whereas S2 and S3 refer to other experts, so as to address the consistency question in Table 1.

directly asserting that *H* is true. In other situations, however, the expert might merely report a fact that is related to the truth of *H*. In Fig. 1, this would be akin to reporting *Evidence*. The degree to which this implies the truth of *H* would be captured in the probabilistic relationship between *H* and *Evidence*. The "trustworthiness" question is captured by the "*trustworthiness*" node. The "consistency" question is represented by the nodes "H_{rep} (*from S2*)" and "H_{rep} (*from S3*)." These nodes are "descendants" of node *H* demonstrating that they should be affected by the truth value of *H*. Finally, the "backup evidence question" is captured by the inclusion of the "*evidence*" node between *H* and H_{rep}. In reality, this evidence might take many forms, but, for the argument to be convincing, there must be a positive relationship between the "*evidence*" and *H*. Whereas the links (or "arcs") drawn in Fig. 1 demonstrate conditional dependencies between variables ("nodes"), it is worth drawing attention to those arcs that could have been included, but were not, as these are informative regarding the assumptions that we are making. These assumptions include the conditional independence of the different experts, represented by the lack of a link between the different H_{rep} nodes other than via *H*. Of course, in some situations, this will be an unrealistic assumption, as one expert might base his or her opinion primarily on another expert (from reading an academic paper, for example). Another arc that may have been drawn, but which we have decided to leave out would be from one (or both) of the nodes "H_{rep} (*from S2/3*)" to the "*expertise*" or "*trustworthiness*" node. A situation in which such an arc might be considered appropriate would be where other experts have explicitly stated that the expert cited in the appeal to expert opinion is (non-) trustworthy (for example). What is of import is that the arcs within a Bayesian Network represent assumed conditional dependencies, and these might differ from one situation to another. Fig. 1 does, however, seem a reasonable model to assume for a variety of situations, and, in particular, the assumptions seem reasonable for the situation we will be examining in our experiments.

That said, the model presented in Fig. 1 can be further simplified. Whereas Fig. 1 is explicit in capturing all the key criteria in Table 1, the core aspects of the argument form can be captured in the simplified model, Fig. 2. In Fig. 2, Walton's "Backup evidence question" is subsumed in the "*expertise*" node, as the expert would not be behaving as an expert on this occasion if they were to provide a report of evidence without evidence for that report.[3] Because "H_{rep} (*from S2/3*)" only have arcs to *H*, their effects can be represented solely by *H*, and thus the *consistency* question will simply be reflected in prior degrees of belief in *H*. The reason why consistency with other experts is a core criterion for the appeal to expert opinion is because the statements of other experts affect the likelihood of *H* (as captured by the links between these nodes). If *H* is known, no additional information is gained from "H_{rep} (*from S2/3*)." *H* is said to "screen off" the influence of these variables, and hence the "consistency" question can be captured by different prior degrees of belief in *H* (on screening off, see e.g., Pearl, 1988). The *Field* question is not explicit in either Fig. 1 or 2. We argue that the *Field* question is an integral part of the *Expertise* question and do not therefore represent it separately. Not only is expertise invariably narrow in its scope (see e.g., Ericsson & Lehman, 1996), but recipients have been shown to be sensitive to the particular scope of an expert's expertise in evaluating

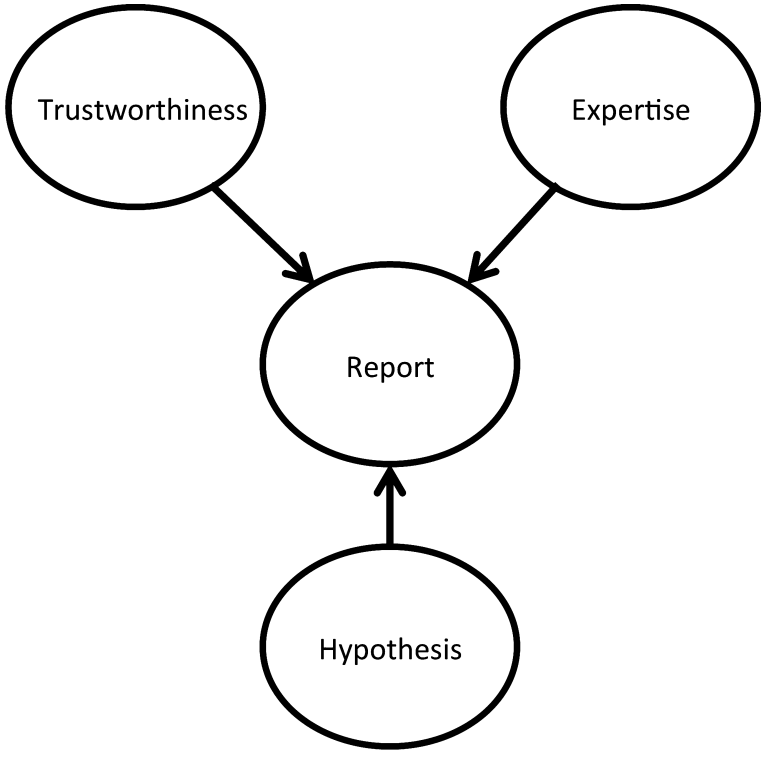

Fig. 2. A Bayesian Network including the two key factors for a probabilistic analysis of the appeal to expert opinion.

his or her claim (e.g., Maddux & Rogers, 1980; Pornpitakpan & Francis, 2001; but see, Hornikx, 2011; Hornikx & Hoeken, 2007; for cultural differences in this sensitivity). One component of the appeal to expert opinion not explicitly captured in either Fig. 1 or 2 is the reliability of the individual advancing the appeal to expert opinion. This is because we are directing the model to address the critical questions advanced by Walton (1997). Consequently, in the empirical work that follows, we assume that this argument proponent is perceived as a fully reliable source by participants—an assumption which we believe is tenable given the experimental setup. The simple models in Figs. 1 and 2 could, however, be straightforwardly extended to include consideration of the arguer's potential unreliability by including "H_{rep} (from arguer)" as a child of "H_{rep} (from S)," with corresponding trustworthiness and expertise parameters in Fig. 1.

The structure of our formalization of the appeal to expert opinion, in terms of the assumed conditional (in)dependencies can be straightforwardly read off from the figures. The critical values for inference via Bayesian Networks are the probabilities governing the different nodes. Because H_{rep} has three parents (H, Exp and T), the value of H_{rep} is conditional on the value of these nodes. We are proposing Figs. 1 and 2 as formalizations of the appeal to expert opinion. Consequently, we are modelling the situation in which an

expert has actually made a report concerning the truth or falsity of the hypothesis.[4] *P*(*Report*), *P*(*Rep*), in this formalization, concerns the value of this report—the expert either provides a confirmatory (*Rep*) or disconfirmatory (¬*Rep*) report. In the experiments reported, the probabilities for the conditional relationships between the variables in Fig. 2 were determined a priori in such a way that they matched the set-up in our experiments.

In assigning probabilities to capture expertise, for the experiments below we followed the approach of Bovens and Hartmann (2003). This approach (used also, for example, in Jarvstad & Hahn, 2011) takes the genuine expert to simply report the evidence accurately, whereas the maximally unreliable source (non-expert) reports in a way that bears no systematic relationship with the evidence. In other words, from the perspective of the argument's recipient, the non-expert functions like a randomizer: that is, the non-expert is as likely to report "yes" as he is to report "no," regardless of what the evidence actually says. In a novel development from Bovens and Hartmann's work on source reliability, we here distinguish expertise from trustworthiness (as implied in Walton, 1997, 2008a; see also, Fearnside & Holther, 1959; Schum, 1981).[5] Whereas expertise essentially functions to determine the likelihood of an honest mistake, trustworthiness captures the likelihood of systematic deceit. That is, we assume a trustworthy source will report the value of the hypothesis that they believe to be true, whilst a non-trustworthy source will report the opposite value. Consequently, a trustworthy expert will report the true value of the hypothesis, a non-trustworthy expert will report the opposite of this, and both trustworthy and non-trustworthy non-experts will report random values (the full conditional probability table is shown in Table 2). Note that, as with the choice of which dependencies to be included in the Bayesian Network, our stipulation of the conditional probabilities entails assumptions for the model. One could, of course, entertain different assumptions. Bovens and Hartmann, for example, include a parameter to represent inherent bias in a randomizer (non-expert) to provide a confirmatory report. A similar free parameter could be added to our model. Of course, any such parameter, like any more fine-grained modeling of a situation in general, is useful and appropriate only where the situation contains information that pertains to those distinctions. Hence, in the spirit of maintaining model simplicity (and therefore generality), we do not employ such a parameter here, because the probabilistic

Table 2
The conditional probability table assumed for Fig. 2

	Hypothesis = "True" (*H*)				Hypothesis = "False" (¬*H*)			
	Trustworthy (*T*)		Not Trustworthy (¬*T*)		Trustworthy (T)		Not Trustworthy (¬*T*)	
	Expert (*E*)	Not Expert (¬*E*)	Expert (*E*)	Not Expert (¬*E*)	Expert (*E*)	Not Expert (¬*E*)	Expert (*E*)	Not Expert (¬*E*)
Report = "Yes"	1	.5	0	.5	0	.5	1	.5
Report = "No"	0	.5	1	.5	1	.5	0	.5

Notes. The values represent the conditional *P*(*Rep*)—where *P*(*Rep*) corresponds to a report of "yes" and *P* (¬*Rep*) corresponds to a report of "no." Thus, for example, the value "1" in the top-left cell of the table shows that $P(Rep|H,T,Exp) = 1$.

relationships we assume between the variables seem representative of the experimental set-up that we will employ to investigate how people evaluate appeals to expert opinion.

The nature of our distinction between trust and expertise aligns well with Shafto, Eaves, Navarro, and Perfors's (2012) distinction between helpfulness (trustworthiness) and knowledgability (expertise). They found that children's use of adult testimony was better described by a model including both these components than one including knowledge alone. The conditional probabilities assumed in Table 2 give trustworthiness and expertise the same roles as helpfulness and knowledgability in Shafto et al. (2012).

Our theoretical treatment of the appeal to expert opinion is also consistent with applications of Bayesian Networks in the legal domain (e.g., Kadane & Schum, 1996), specifically with recent demonstrations that complex chains of inference in legal contexts might be broken down into simple components that recur repeatedly (Fenton, Neil, & Lagnado, 2013; Lagnado, 2011; Lagnado et al., 2013). A similarly generic model of the appeal to expert opinion is important here, because such appeals often play a hugely prominent role within the law (e.g., Godden & Walton, 2006; Walton, 1997, Chapter 6, 2008a).

These past applications concern the possibility of Bayesian formalization itself. Once such a formalization is in place, however, one may ask also to what extent people's intuitive judgments of appeals to expert opinion match that formal treatment. Because the Bayesian framework itself has a well-developed claim to normative status (founded on considerations of instrumental rationality, but also the minimization of inaccuracy of our beliefs see, for example, Lindley, 1994; Rosenkrantz, 1992; and, for a general discussion of the issue of norms of argumentation, Corner & Hahn, 2013), our formalization provides a candidate computational level theory of human behavior: that is, a characterization of what it is that people are seeking to do, not in terms of underlying psychological process but rather in terms of a characterization of the problem the behavior in question is seeking to solve. In other words, our formalization provides the basis for a potential rational analysis of people's judgments concerning appeals to expert opinion that seeks to understand such judgments as approximations to the rational norm (on rational analysis, see e.g., Anderson, 1990; Chater & Oaksford, 1999). Such an analysis does not assume that people consciously and deliberately conduct Bayesian calculations (indeed, many domains, such as vision, in which Bayesian analyses have been applied allow no introspective access to underlying processes at all, on perception as Bayesian inference, see, for example, Knill & Richards, 1996). Rather, it is entirely silent on the actual algorithms people might be employing. The goal is simply to provide a functional explanation for why those judgments are the way they are.

At the same time, possession of an appropriate computational-level theory will allow one to make successful predictions about human behavior. Finally, of course, from the perspective of those whose primary interest lies in normative considerations, that is, specification of what we should think of as "good arguments" (as is the case for the philosophical literature on argumentation), examining actual judgments of appeals to expert opinion and comparing them to Bayesian prescriptions provides insight into how good people are at informal argument evaluation.

The central aim of the experiments presented in this paper was therefore to demonstrate the amenability of the Bayesian approach to the development and testing of quantitative empirical predictions. The qualitative Bayesian predictions of the effects of the three key factors—source expertise, source trustworthiness, and the opinions of other experts (labeled "others' opinions" in the empirical sections)—are intuitive, in keeping with Laplace's verdict that the probability calculus is "nothing but formalized common sense" (Laplace, 1814/1951). Greater expertise or trustworthiness will increase the impact of the testimony on one's posterior degree of belief. Likewise, better fit with the opinions of other experts will have a positive effect (for a detailed exploration of the issue of coherence between testimonies, see Harris & Hahn, 2009). Given this intuitive nature of the qualitative predictions, the normative Bayesian predictions deviate from those of other putatively normative accounts (e.g., formal approaches to "plausible reasoning," such as Rescher, 1977; or, Pollock, 2001; see, for example Walton, 1997, 2008a and references therein) primarily where some of those alternative accounts make somewhat counter-intuitive predictions. This issue has been pursued elsewhere (Hahn et al., 2012). Here, we therefore focus on another feature of the Bayesian framework: its ability to make detailed quantitative predictions.

With its grounding in probability theory, the Bayesian framework has the capability to make quantitative predictions as to how convinced individuals should be by an argument. Because the Bayesian framework stipulates that subjective degrees of belief should be represented as probabilities, one must elicit these probabilities in some way. One approach is to engage in a model-fitting exercise to demonstrate that the quantitative pattern of results observed is obtainable from a Bayesian model—fitting parameters rather than attempting to elicit them empirically (as in Hahn & Oaksford, 2007a). A second approach is to present participants with an experimental scenario in which the quantitative parameters are defined in the problem (as in Harris & Hahn, 2009). Finally, one can ask participants for the necessary parameter values and use these to predict participants' posterior convincingness ratings—that is, $P(H|argument)$, in the current project, $P(H|Rep)$ (as in Harris, Hsu, & Madsen, 2012; see also, Fernbach & Erb, 2013). In Experiment 1, we adopt the latter approach as closely as possible, introducing only three free parameters (relating to just one variable). In Experiment 2, we elicit ratings from participants for that variable also and our model therefore includes no free parameters in that test.[6] The experiments demonstrate how implementing a scheme-based treatment within a Bayesian model enables quantitative testing of the suitability of that model as a computational-level description of people's argumentation skills. Furthermore, the results are encouraging in the sense of the closeness of the fit between behavior and prediction.

2. Experiment 1

Experiment 1 was designed to determine whether people are sensitive to the critical factors that should influence the convincingness of an appeal to expert opinion. More specifically, it was intended to test the degree to which participants' quantitative

probability ratings approximated the prescriptions of a Bayesian formalization of the appeal to expert opinion incorporating the notions of expertise and trustworthiness. The model being tested was the simplified model shown in Fig. 2. In this test, parameter estimates for the prior probabilities *P*(*Expertise*) and *P*(*Trustworthiness*) [hereafter, *P*(*Exp*) and *P*(*T*)] were elicited from participants, while *P*(*Hypothesis*) [hereafter, *P*(*H*)], was a free parameter. This prior degree of belief in the hypothesis also captures variance from additional sources of knowledge about the hypothesis—such as the manipulation in the present study of whether other experts agreed with the expert cited within the argumentation dialog.

2.1. Method

2.1.1. Participants

Twenty-nine males and 55 females, aged between 18 and 80 (median = 20), participated in the experiment without remuneration. The study was advertised on http://psych.hanover.edu/research/exponnet.html, a site for recruiting volunteers to participate in web-based experiments.

2.1.2. Design and materials

A 3 × 2 × 2 mixed design was employed, with others' opinions as a between-participant variable (3-levels) and expertise and trustworthiness as within-participant variables.

We here describe the materials in an order that makes the relevance of each measure clear, rather than chronologically. The chronology of the experiment is outlined in the Procedure and Fig. 4.

Participants were presented with appeals to expert opinion within a dialog (e.g., Fig. 3). Each dialog had the same structure, with a male proponent presenting an appeal to expert opinion (where the expert was always male) to a female recipient. Four different argument topics were used to incorporate the within-participant manipulations. The topics concerned fictional medical information: whether Proftanine would lower cholesterol; whether handling Kworgs causes skin blemishes; whether taking Antiprone causes insomnia; and whether exposure to Bongus causes painful swelling. The expertise manipulation was operationalized through having the expert be either a doctor (high expertise) or a musician (low expertise). Trustworthiness was manipulated through having the expert be a friend of the argument proponent (high trust) or an enemy of the argument proponent (low trust). Others' opinion was manipulated in the final sentence of the dialog. There was either no mention of other experts (as in Fig. 3), or other experts were said to agree [or disagree], through adding the sentence: "I also read in 'Science' magazine that a number of experiments have been completed across the world and it is now considered to be medical fact that eating Proftanine [does NOT] lower(s) cholesterol." The pairing of the drug-effect dialogs with a particular experimental condition (i.e., the expertise and trustworthiness of the source and whether others agreed) was implemented in a Latin square design such that, across participants, each experimental condition occurred in each drug-effect dialog. Each individual participant, however, only

saw one example of each dialog and each experimental condition, so as to minimize demand effects.

Participants provided posterior degree of belief ratings (hereafter, "convincingness ratings") by moving a slider in response to the question: "In light of the dialogue above, what do you think Anne's opinion should now be of Proftanine?" The slider was anchored at "Completely convinced it does NOT lower cholesterol" and "Completely convinced it DOES lower cholesterol," and recorded responses (participants saw no numbers) as a number between 0 and 100, thus providing a percentage value for $P(Hypothesis|Report)$ [$P(H|Rep)$].

In addition to rating the convincingness of an argument, participants provided prior belief ratings for the variables $P(Exp)$ and $P(T)$. We set up the experiment in such a way as to capture the conditional relationships that we assumed in Table 2. Thus, for the trustworthiness question, participants were asked:

> "Imagine the following scenario: Keith is a musician who is an old enemy of James and they are discussing whether taking Proftanine lowers cholesterol.
>
> How likely do you think it is that Keith would deliberately give James wrong information about whether taking Proftanine lowers cholesterol?"

While this may or may not concur with people's everyday understanding of the term trustworthiness, it captures the characteristics of the concept within the model under consideration (Fig. 2, Table 2). Participants answered by moving a slider on a scale anchored at "I'm completely convinced he would NOT deliberately give James wrong information." and "I'm completely convinced he WOULD deliberately give Jonathan wrong information." Responses to this question were subsequently subtracted from 100, to provide a percentage value for $P(T)$.

James: Do you think eating Proftanine lowers cholestrol?

Anne: I have no idea if eating Proftanine lowers cholestrol.

James: Well, I can tell you that eating Proftanine lowers cholestrol.

Anne: Why do you say that?

James: Keith told me that eating Proftanine lowers cholestrol

Anne: Who's Keith?

James: Keith is an old enemy of mine who is also a musician.

Fig. 3. An example argumentation dialog, from the "no mention of others," "low expertise," "low trustworthiness" condition.

For expertise, participants were asked: "How likely do you think it is that Keith is an expert on whether taking Proftanine lowers cholesterol?" and responses were made using a slider anchored at "I'm completely convinced he is NOT an expert" and "I'm completely convinced he IS an expert." We acknowledge that this question might not match as directly to the conditional relationships outlined in Table 2 as the Trustworthiness question. We contend, however, that providing a precise continuous estimate for this concept, as our participants must do, requires a precise definition of this concept—as precise probabilities of imprecise events are nonsensical (see e.g., Wallsten, 1990). We assume that the most appropriate precise definition is that provided in Table 2, in line with recent approaches within epistemology (e.g., Bovens & Hartmann, 2003). As such, our approach aims to describe the situation in such a way as to lead participants' subjective estimates of the relevant conditional probabilities in the direction of those outlined in Table 2. An alternative approach is to elicit these subjective probabilities from participants. We chose the current approach to reduce the complexity of the task for participants.

For completeness, participants also provided posterior ratings of the expertise and trustworthiness of the expert after having read the dialog containing the argument, thus providing estimates of $P(Exp|Rep)$ and $P(T|Rep)$. These were made using the same scales (in response to the exact same questions) as for their prior ratings of expertise and trustworthiness. These questions were presented beneath the dialog of the argument.

All non-central aspects of the experiment, such as the names used, and orders of presentation were randomized across participants.

2.1.3. Procedure

The experiment was run online using Adobe Flash. After consenting to participate in the study, participants first provided their prior ratings of trust and expertise for all four arguments they were to be presented with. Subsequently, they viewed each argument, rated its convincingness and, on the next page, provided posterior estimates of the expert's likely trustworthiness and expertise, before moving onto the next argument. Finally, participants provided their demographic details before receiving debrief information (see Fig. 4 for a schematic of the experimental procedure).

2.2. Results[7]

A 3 × 2 × 2 mixed ANOVA yielded significant main effects (in the predicted direction, see Table 3) of all three independent variables: Others' opinions, $F(2, 81) = 32.53$, $p < .001$, $\eta_p^2 = .45$; Expertise, $F(1, 81) = 34.38$, $p < .001$, $\eta_p^2 = .298$; Trustworthiness, $F(1, 81) = 11.14$, $p = .001$, $\eta_p^2 = .121$.[8] The only significant interaction was between Trustworthiness and Others' opinions, $F(2, 81) = 3.27$, $p = .043$, $\eta_p^2 = .075$. As Fig. 5 demonstrates, this interaction resulted from the trustworthiness manipulation having no effect when other experts disagreed with the position advanced in the dialog.

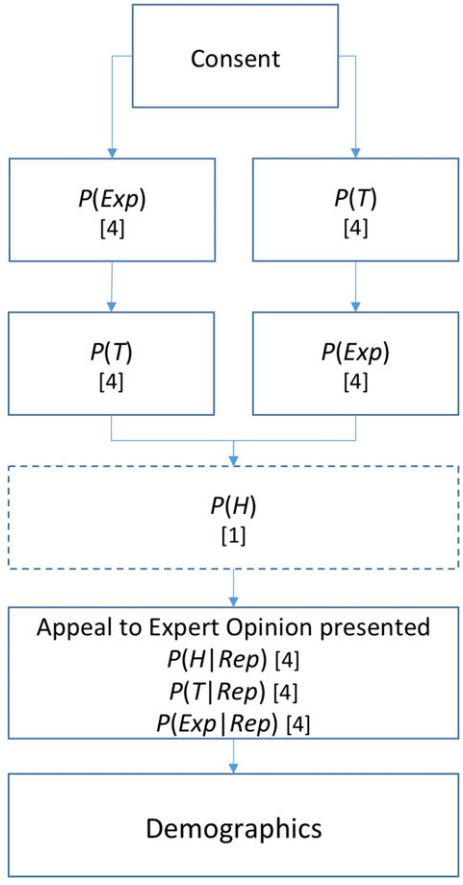

Fig. 4. A schematic showing participants' progression through the experiment. The bracketed numbers represent the number of questions answered at each point. Note that in the penultimate box, for each dialog, participants always answered *P*(*H*|*Rep*) first, before, on a subsequent page, answering the other two questions (order counterbalanced across participants). The dashed box represents a stage that was only present in Experiment 2.

2.2.1. *Quantitative comparisons of Bayesian predictions with observed ratings*

The main analyses of interest concerned the model fitting between the Bayesian predictions of how convincing the argument should be perceived to be, and participants' convincingness ratings (which were divided by 100 to provide values between 0 and 1 in all subsequent analyses). The model being tested is that shown in Fig. 2. As the report from the expert is provided, $P(Rep) = 1$, the remaining required parameters for the calculation of a Bayesian posterior, $P(H|Rep)$, are the priors $P(H)$, $P(T)$, and $P(Exp)$ and the conditional probability table for the network. As discussed above, the conditional probabilities were assumed a priori (see Table 2). $P(T)$ and $P(Exp)$ were provided by participants in the experiment as percentages (see Table 4). $P(H)$ was therefore the only free parameter in the model, and we estimated a best fitting value for it separately

Table 3
Mean convincingness of the dialog across the different experimental conditions in the two experiments.

			Mean		*SE*	
Others	Expertise	Trust	Experiment 1	Experiment 2	Experiment 1	Experiment 2
Disagree	Musician	Enemy	24.96	26.50	4.83	6.64
		Friend	22.57	30.83	4.20	5.92
	Doctor	Enemy	31.36	28.33	5.32	7.43
		Friend	32.68	35.78	4.15	7.08
No mention	Musician	Enemy	37.82	39.75	4.83	4.94
		Friend	45.25	50.05	4.20	4.52
	Doctor	Enemy	51.14	46.05	5.32	5.51
		Friend	72.11	67.30	4.15	6.17
Agree	Musician	Enemy	53.04	61.74	4.83	4.26
		Friend	65.50	70.37	4.20	3.63
	Doctor	Enemy	64.21	65.89	5.32	5.22
		Friend	74.32	80.89	4.15	3.64

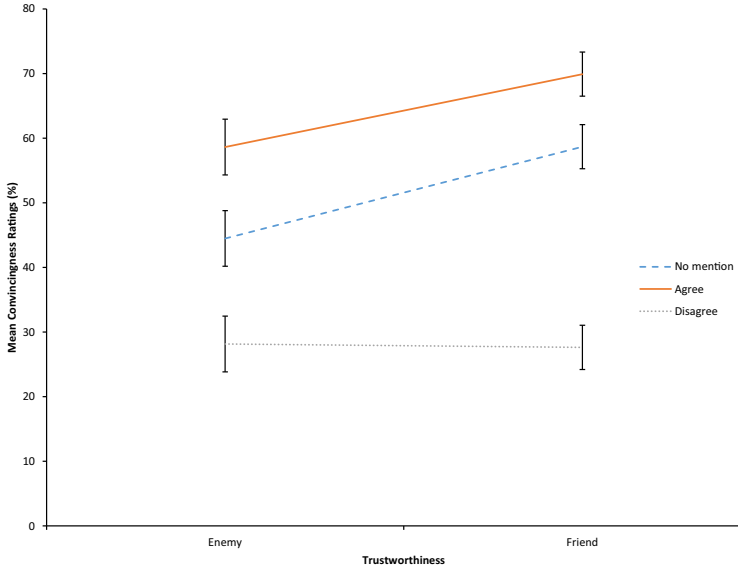

Fig. 5. The interaction between trustworthiness and others' opinions in Experiment 1. Error bars are ± 1 *SE*.

for each level of the Others' opinions variable (minimizing the mean squared error between participants' convincingness ratings and the Bayesian predictions). The choice of three free parameters was thus theoretically and empirically justified (as a significant effect of Others' opinions was observed in the ANOVA above). The best fitting values of $P(H)$ in the three conditions were .4125, .5565, and .1670 (no mention, others agree and others disagree, respectively). The Bayesian predictions can then be calculated using Eq. 1, with *Rep* substituted for *e*. In order to use Eq. 1, $P(Rep|H)$ and $P(Rep|\neg H)$ were

Table 4
Mean parameter values elicited from participants before exposure to the argumentation dialog (standard deviations in parentheses)

			Experiment 1		Experiment 2		
Others	Expertise	Trust	P(*Exp*)	P(*T*)	P(*Exp*)	P(*T*)	P(*H*)
Disagree	Musician	Enemy	19.96 (18.66)	38.14 (19.13)	33.11 (23.17)	37.78 (23.13)	19.56 (20.81)
		Friend	27.79 (22.15)	80.39 (20.87)	33.22 (24.39)	77.78 (27.64)	19.56 (20.81)
	Doctor	Enemy	68.43 (18.07)	59.11 (28.28)	68.72 (29.69)	66.50 (27.65)	19.56 (20.81)
		Friend	68.25 (25.37)	88.14 (15.96)	76.22 (19.94)	91.11 (11.70)	19.56 (20.81)
No mention	Musician	Enemy	22.86 (18.80)	55.29 (29.50)	34.25 (25.99)	39.15 (24.65)	48.65 (15.86)
		Friend	29.82 (21.61)	80.46 (20.94)	35.75 (24.66)	81.90 (21.54)	48.65 (15.86)
	Doctor	Enemy	62.82 (28.41)	61.82 (30.68)	78.30 (23.34)	59.85 (27.65)	48.65 (15.86)
		Friend	63.57 (29.14)	82.82 (25.46)	72.35 (23.33)	84.60 (18.34)	48.65 (15.86)
Agree	Musician	Enemy	19.86 (17.51)	46.50 (26.20)	30.53 (25.58)	60.26 (25.07)	77.79 (13.77)
		Friend	21.93 (23.87)	74.75 (27.50)	26.21 (24.29)	78.37 (20.61)	77.79 (13.77)
	Doctor	Enemy	67.86 (24.58)	69.00 (26.77)	62.89 (27.43)	71.16 (22.81)	77.79 (13.77)
		Friend	72.54 (20.85)	82.57 (20.09)	61.79 (28.97)	84.00 (16.80)	77.79 (13.77)

Note. $P(T)$ and $P(Exp)$ were asked before the manipulation of $P(H)$, so they should not be affected. They are split by $P(H)$ condition here for maximum transparency.

obtained through marginalising over *Exp* and *T*, using the values from Table 2. For example,

$$P(Rep|H) = P(Rep|H,Exp,T) \times P(Exp,T) + P(Rep|H,Exp,\neg T) \times P(Exp,\neg T) + P(Rep|H,\neg Exp,\neg T) \times P(\neg Exp,\neg T) + P(Rep|H,\neg Exp,T) \times P(\neg Exp,T) \quad (2)$$

Conditional independence of $P(Exp)$ and $P(T)$ was assumed (see Fig. 2) and so $P(Exp,T) = P(Exp) \times P(T)$.

A good fit was observed between the convincingness ratings provided by participants and the Bayesian predictions, accounting for 89% of variance across conditions ($p < .001$; Fig. 6). An inspection of Fig. 6 clearly shows that the Bayesian model performs least well when other experts are said to disagree with the report of the cited expert. Although the 95% confidence intervals overlap in all instances, the "predictions" line is clearly more affected by the evidence from the trustworthy doctor than is the "observed" line. Closer analysis of the raw data shows that there are a number of participants whose predicted convincingness is 1 (14 datapoints), because they have provided ratings indicating certainty for both the trustworthiness and expertise prior questions. When participants whose predicted ratings are 1 (there were none whose predicted ratings were zero) are excluded from the overall average across all conditions, the "peakiness" in the "predictions" line is reduced (see Fig. 7) and the model fit increases to explain 93% of the variance across conditions. This suggests that the clear outlier in the "predictions" for the "others disagree, friend, doctor" condition is predominantly driven by those partic-

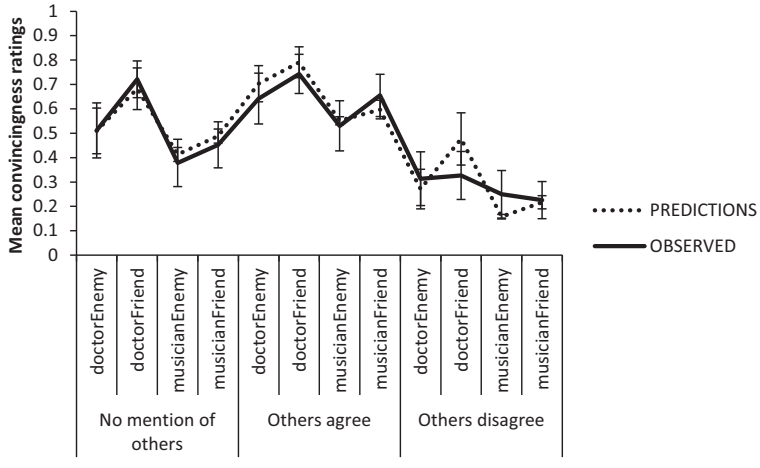

Fig. 6. The fit between the Bayesian predictions and the observed convincingness ratings in Experiment 1. Error bars show 95% confidence intervals.

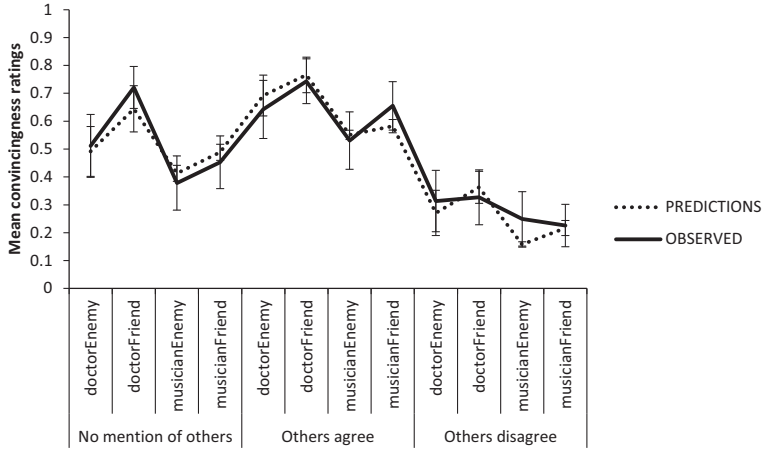

Fig. 7. The fit between the Bayesian predictions and the observed convincingness ratings in Experiment 1 after removing predicted values of 1. Error bars show 95% confidence intervals.

ipants who have provided what might be seen to be unrealistically high certainty ratings for their priors of expertise and trustworthiness—we suspect that, if pushed, these participants would provide ratings near to certainty, but acknowledge the *possibility* that the opposite state of affairs could hold true, which is not possible when estimates of certainty are provided.

The above analysis demonstrates a good fit between participants' convincingness ratings and the predictions of the Bayesian model. A further method for testing whether participants' ratings provide a good approximation to the Bayesian model is to examine *how* they differ. The predicted and observed ratings could differ in a random manner, or there

might be systematic effects that the observed ratings are sensitive to that are not predicted by the model (or vice versa). A straightforward way to test this is to enter "predicted/observed" as a variable in a 4-way ANOVA that also includes the three experimental variables. Interactions involving the "predicted/observed" variable demonstrate a situation in which participants' observed ratings are reacting systematically differently from the predictions of the model. The 4-way ANOVA (excluding predicted posteriors of 1, since they are likely to distort predictions disproportionately) suggested that the non-parallelism shown in the right most portion of Fig. 6 does correspond to a systematic difference between the Bayesian predictions and the observed ratings, as the 3-way interaction between "predicted/observed," "expertise," and "others' opinions" was significant, $F(2, 70) = 3.81$, $p = .027$. No other effects involving the "predicted/observed" variable reached significance (all ps $> .22$).[9]

There is then one significant point of systematic disagreement between model and data. Given, however, that this complex 3-way interaction was the only significant interaction out of seven possible interactions (three 2-way, three 3-way, and one 4-way), there is little systematic mismatch between participants and model, thus further emphasizing that, overall, participants' ratings are reasonable approximations to the Bayesian predictions.

The above analyses provide good support for the contention that people's responses to appeals to expert opinion are well approximated by a Bayesian theory. As a further test of this contention, participants' ratings of the trustworthiness of the expert and the expertise of the expert following the argumentation dialog, $P(T|Rep)$ and $P(Exp|Rep)$, were compared with the predictions of the Bayesian model (obtained via Eqs. 2 and 1, switching H with T and H with Exp, respectively). In this case, values of $P(H)$ were not refitted to the data, rather the *same parameters* as for the convincingness ratings were used.

Values for $P(H)$ were specifically fit to the data so as to maximize the fit between participants' convincingness ratings and the Bayesian predictions, not their expertise or trustworthiness ratings. Consequently, we expect the fits between observed and predicted ratings of $P(Exp|Rep)$ and $P(T|Rep)$ to be less good than those for $P(H|Rep)$. The significant positive correlation between the predicted and observed values of expertise, $P(Exp|Rep)$, $r(10) = .90$, $p < .001$, indicating that 81% of variance (81.5% if predictions of certainty [either 0 or 1] are excluded—39 datapoints) in posterior expertise ratings across conditions was predicted by the model (Fig. 8), does, however, provide further support for the Bayesian theory as an approximation of how people evaluate argumentation, and update their beliefs. Once more, a 4-way ANOVA was conducted (excluding predicted posteriors of certainty, since they are likely to distort predictions disproportionately) to determine the systematicity of the differences between the Bayesian predictions and participants' observed posteriors. Out of seven possible interactions, two were significant (all other ps $> .16$), "predicted/observed" × "others' opinions," $F(2, 57) = 8.24$, $p = .001$, and "predicted/observed" × "expertise" × "others' opinions," $F(2, 57) = 4.25$, $p = .019$.[10]

For posterior ratings of trustworthiness, Fig. 9 shows excellent correspondence between participants' posterior ratings of the expert's *trustworthiness,* $P(T|Rep)$, and the Bayesian predictions (using the same parameters once more), $r(10) = .97$, $p < .001$, indicating that

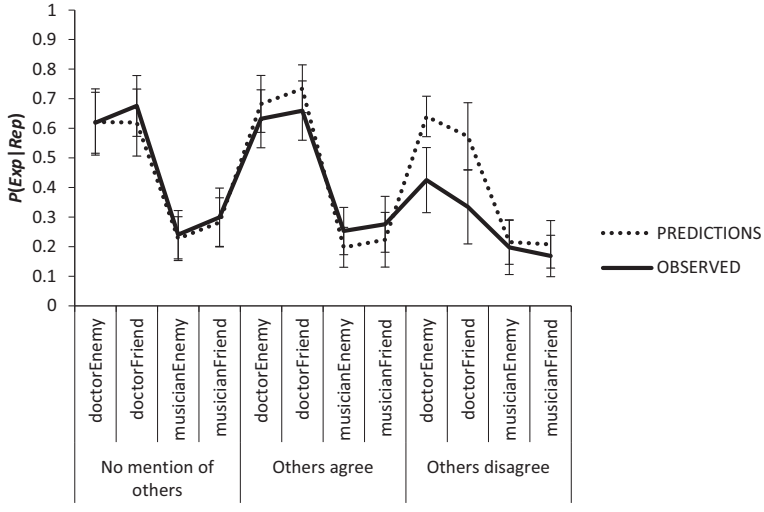

Fig. 8. The fit between the Bayesian predictions for posterior ratings of the expert's expertise and the observed ratings in Experiment 1. Error bars show 95% confidence intervals.

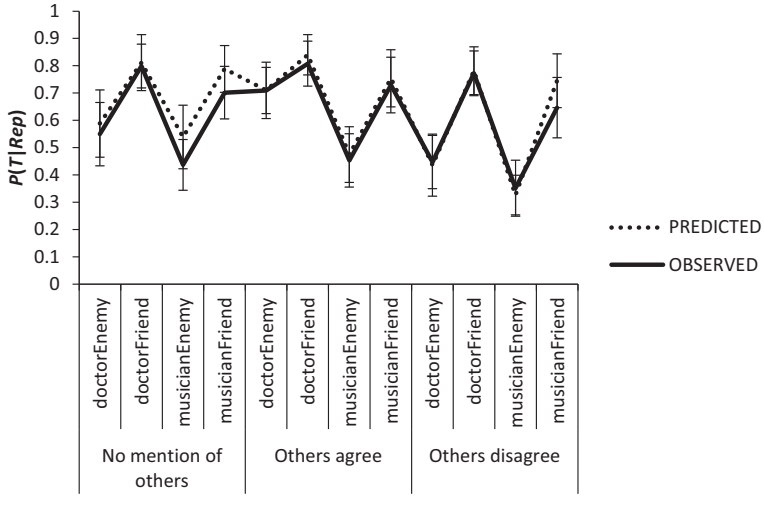

Fig. 9. The fit between the Bayesian predictions for posterior ratings of the expert's trustworthiness and the observed ratings in Experiment 1. Error bars show 95% confidence intervals.

94% of variance (88% if predictions of certainty [either 0 or 1] are excluded—65 datapoints) in trustworthiness posterior ratings across conditions can be explained by the Bayesian model. Furthermore, in a 4-way ANOVA (excluding predicted posteriors of certainty), there were no significant interactions involving the "predicted/observed" variable (all $ps > .28$).[11]

Finally, as the same parameter values were used for all fits, a correlation was computed across the three dependent variables simultaneously. It was found that the Bayesian model was able to account for 89% of the variance across all 36 datapoints, $r(34) = .94$, $p < .001$, with just the three free parameters. This value is the same if predictions of certainty (0 or 1) are excluded.

The analyses above assessed the fit of the model to the mean participant response in each experimental condition. One can also use an individual-level analysis, in which a correlation coefficient is computed for each participant and then the average of these is computed. Because "others' opinion" was a between-participants variable, this reduces the number of datapoints in the analysis. Were we to simply look at the posterior ratings of convincingness, this correlation would be based on only four datapoints per participant. We therefore only report the results of this analysis for the correlation incorporating *all* judgments made by a participant (i.e., $P(H|Rep)$, $P(T|Rep)$, and $P(Exp|Rep)$). The resulting analysis suggested that 38% of variance in participants' responses was explained by the Bayesian model, and for all but three participants the correlation with the model was positive, with this correlation coefficient being significant for 55 out of the total 84 participants. For the remaining 29 participants, some of them will have not engaged with the task, but the question of whether the ratings of a group of them might be well captured by an alternative model (e.g., another Bayesian model with alternative assumptions) is an example of the sort of further questions that this line of research can generate.

3. Experiment 2

Experiment 1 provided good support for the Bayesian model of the appeal to expert opinion as one that human reasoning approximates. Experiment 2 was designed as a more robust replication of Experiment 1. In Experiment 1, the model was fit to the data with three free parameters. This is common practice in model fitting. In Experiment 2, however, we provided a more rigorous test of the model by conducting a parameter free test. This was achieved by additionally asking participants to provide estimates for the free parameters in Experiment 1, $P(H)$.

3.1. Method

3.1.1. Participants

Twenty-three males and 34 females, aged between 16 and 54 (median = 23), who participated in the experiment without remuneration, were retained for analysis after excluding two participants under the age of 16 (in line with BPS ethical guidelines) and one who reported his/her age as 99(!). As a result of these exclusions, there was minor imbalance in the sample size for each condition with 18 participants in the "others disagree" condition, 19 in the "others agree" condition and 20 in the "no mention of others" condition. The study was advertised on http://psych.hanover.edu/research/exponnet.html, a site for recruiting volunteers to participate in web-based experiments.

3.1.2. Design, materials, and procedure

There were two changes from the method employed in Experiment 1. The first was to reword the question eliciting $P(Exp)$ so as to ensure it was a better match for the *expertise* variable as its characteristics are described in Table 2. Subsequently, participants were asked (for example), "How likely do you think it is that Keith *knows* whether taking Proftanine lowers cholesterol" and responded on a scale anchored at "I'm completely convinced he would NOT know" and "I'm completely convinced he WOULD know." This was to guard against participants' own idiosyncratic understandings of the term "expert."

The second change was more consequential. Participants provided their own estimates for the $P(H)$ parameter. After having provided their prior ratings for trust and expertise, participants were presented with a screen designed to elicit their ratings of the prior belief of the recipient of the argument. As discussed in the Introduction, the opinions of other experts can be captured in the prior probability, and the question we asked participants was designed to be faithful to that (see also note 2). Participants therefore read (for example—in the "others agree" condition):

> Anne has no idea whether taking Proftanine lowers cholesterol.
>
> She then reads in Science magazine that a number of experiments have been completed across the world and most experts are agreed that taking Proftanine lowers cholesterol.
>
> What do you think Anne's opinion should now be of Proftanine?[12]

Participants made their response using a slider anchored at "Completely convinced taking Proftanine does NOT lower cholesterol" and "Completely convinced taking Proftanine DOES lower cholesterol." Participants only provided one rating of the prior probability. The mention of other experts was specific to the experimental condition to which they were assigned (with the second sentence simply absent in the "no mention of others" condition), and the precise issue that they reported a prior probability for was randomized between participants. This design was chosen so as to reduce the potential for experimental pragmatics to play a role when participants might perceive that they are being asked the same question four times (i.e., if experts state... what should your degree of belief be?).

All other aspects of the experimental method were the same as Experiment 1 (see Fig. 4 for the complete experimental procedure).

3.2. Results

A 3 × 2 × 2 mixed ANOVA yielded significant main effects (in the predicted direction, see Table 3) of all three independent variables: Others' opinions, $F(2, 54) = 23.39$,

$p < .001$, $\eta_p^2 = .46$; Expertise, $F(1, 54) = 6.42$, $p = .014$, $\eta_p^2 = .106$; Trustworthiness, $F(1, 54) = 20.94$, $p < .001$, $\eta_p^2 = .279$. There were no significant interactions.

3.2.1. Quantitative comparisons of Bayesian predictions with observed ratings

The fit of the model was tested as in Experiment 1, with the conditional probabilities assumed a priori (see Table 2). The only difference was that all remaining parameters were estimated by participants (see Table 4), such that there were no free parameters in the model.

A good fit was again observed between the convincingness ratings provided by participants and the Bayesian predictions, accounting for 94% of variance across conditions ($p < .001$; Fig. 10). An inspection of Fig. 10 suggests, as in Experiment 1, although less strongly, that the Bayesian model performs least well when other experts are said to disagree with the report of the cited expert. Although the 95% confidence intervals overlap in all instances, the "predictions" line is clearly more affected by the evidence from the trustworthy doctor than is the "observed" line. In this instance, excluding datapoints that include predicted posteriors of certainty (0 or 1; 38 datapoints), either because participants have provided ratings indicating certainty for both the trustworthiness and expertise prior questions, or for the $P(H)$ question, had a negligible effect on the model fit.

As in Experiment 1, we carried out a 4-way ANOVA, incorporating "predicted/observed" as a variable to test whether there are systematic differences between the model's predictions and the observed ratings. The 4-way ANOVA (excluding predicted posteriors of 0 or 1, since they are likely to distort predictions disproportionately) suggested that there were no systematic differences between the Bayesian predictions and the observed ratings, as no interactions involving "predicted/observed" were significant, although the 3-way interaction between "predicted/observed," "trust" and "others' opinions" approached significance, $F(2, 40) = 2.56$, $p = .09$ (all other ps $> .395$).[13] Thus, there is no significant

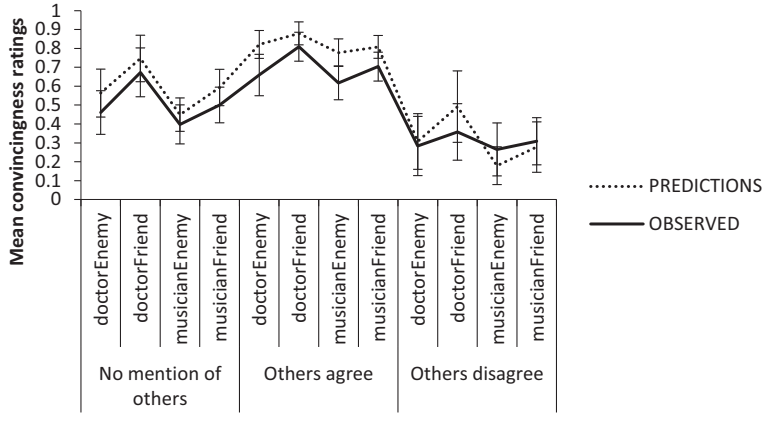

Fig. 10. The fit between the Bayesian predictions and the observed convincingness ratings in Experiment 2. Error bars show 95% confidence intervals.

systematic mismatch between participants and model, further suggesting that, overall, participants' ratings are reasonable approximations to the Bayesian predictions.

The above analyses provide good support for the contention that people's responses to appeals to expert opinion are well approximated by a Bayesian theory. As a further test of this contention, participants' ratings of the trustworthiness of the expert and the expertise of the expert following the argumentation dialog, $P(T|Rep)$ and $P(Exp|Rep)$, were compared with the predictions of the Bayesian model.

The significant positive correlation between the predicted and observed values of expertise, $P(Exp|Rep)$, $r(10) = .94$, $p < .001$, indicating that 88% of variance in posterior expertise ratings across conditions was predicted by the model (Fig. 11), provides further support for the Bayesian theory as an approximation of how people evaluate argumentation, and update their beliefs. Once more, a 4-way ANOVA was conducted to determine the systematicity of the differences between the Bayesian predictions and participants' observed posteriors. No interactions were significant, either when ratings of certainty were excluded (35 datapoints; all ps > .31), or when they were not (all ps > .11).

For posterior ratings of trustworthiness, the fit was not quite as good as in Experiment 1, but nevertheless, 83% of variance ($p < .001$) in participants' posterior ratings of the expert's trustworthiness was explained by the model (Fig. 12). In a 4-way ANOVA, only 1 out of 7 possible interactions was significant—with the "Others' opinion" variable, $F(2, 54) = 5.06$, $p = .01$ (all other ps > .06). When predictions of certainty were excluded (53 datapoints), this interaction became non-significant ($p > .13$), but the interaction with Trustworthiness reached significance, $F(1, 25) = 4.31$, $p = .048$ (all other ps > .13).

Finally, a correlation was computed across the three dependent variables simultaneously. It was found that the Bayesian model was able to account for 90% of the variance across all 36 datapoints, $r(34) = .95$, $p < .001$, with no free parameters.

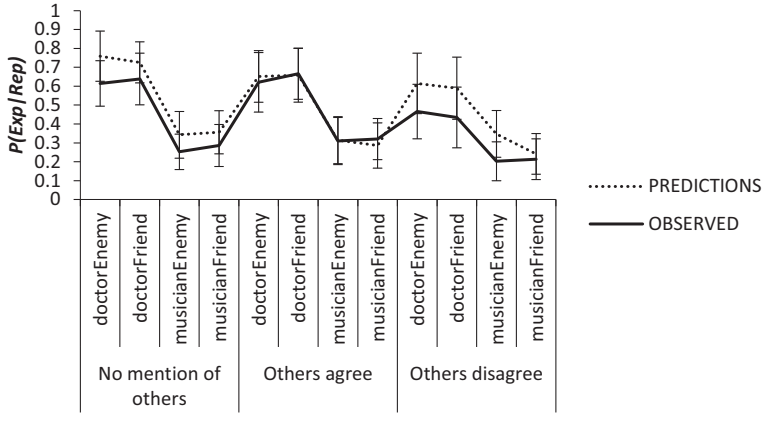

Fig. 11. The fit between the Bayesian predictions for posterior ratings of the expert's expertise and the observed ratings in Experiment 2. Error bars show 95% confidence intervals.

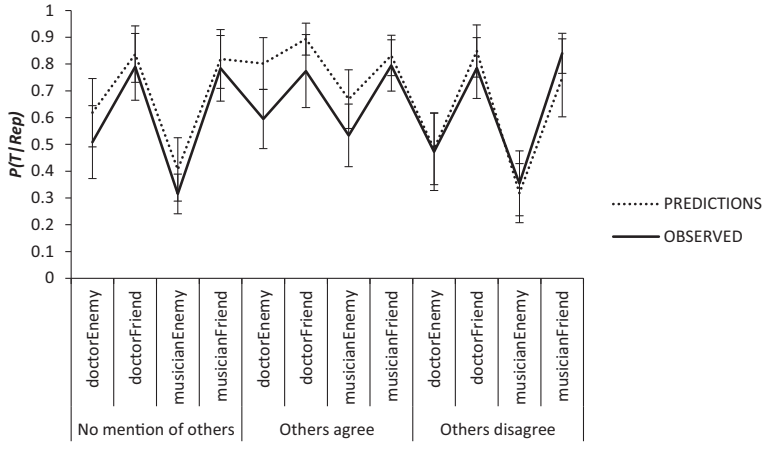

Fig. 12. The fit between the Bayesian predictions for posterior ratings of the expert's trustworthiness and the observed ratings in Experiment 2. Error bars show 95% confidence intervals.

In the same individual-level analysis as conducted for Experiment 1, across all judgments made by each participant ($P(H|Rep)$, $P(T|Rep)$, and $P(Exp|Rep)$), 38% of variance in participants' responses was found to be explainable by the Bayesian model. For all but three participants the correlation with the model was positive, with this correlation coefficient being significant for 40 out of the total 57 participants.

Finally, because no free parameters were fit for the data in Study 2, it provided ideal conditions for comparing the performance of the model in Fig. 2 with two models that are nested within Fig. 2 to determine the degree to which participants were sensitive to all the factors in the model. Specifically, we generated predictions from a model which ignored the information about perceived expertise, and one which ignored information about perceived trustworthiness (we specify the equations in Appendix S1 for maximum clarity). The model ignoring trustworthiness accounted for 70% of the data at the group level, while that ignoring expertise accounted for 84% of the data. Thus, the data were 286 times more likely under the full model than under the best performing nested model (Glover & Dixon, 2004). Although our design gave us only four datapoints per participant for convincingness ratings, we also sought to compare the individual-level correlations across the three different models. The good group-level fit observed could potentially result not from participants being sensitive to the three relevant factors, $P(H)$, $P(Exp)$, and $P(T)$, and the complex interactions between them, but from different subgroups of participants being sensitive to different individual parameters (in a less sophisticated way; see Page, 2007). In order to provide some preliminary evidence that this was not the case (although necessarily weak, given the limited number of datapoints per participant), we compared the number of participants whose four convincingness ratings were best explained by each of the three models. If the good group level fits result from a combination of distinct groups of participants each using simple strategies, there will be more participants whose ratings are best explained by the model excluding either trustworthiness or expertise than the full model. This was not the result we observed. For the 50 partici-

pants for whom such a correlation was possible, the results for 27 were best explained by the full model (15 by the model excluding trustworthiness and 8 by the model excluding expertise).

Overall, therefore, good support was observed for the contention that participants' judgments may be considered (somewhat) noisy estimates of the Bayesian model's predictions.

4. General discussion

In this paper, we have suggested an approach for a reconciliation between the scheme-based approach to argumentation (see Walton et al., 2008) and the Bayesian approach (see Hahn & Oaksford, 2007a) using the appeal to expert opinion as an example case. Furthermore, we have demonstrated how such an approach enables the formulation and testing of precise quantitative predictions. In two example experiments, our results demonstrated that participants' quantitative ratings of the convincingness of a conclusion following appeals to expert opinion were seemingly well predicted by a Bayesian network model. Furthermore, subsequent ratings of the expertise and trustworthiness of the expert sources were also well predicted by the model. This provides initial support for our contention that participants evaluate arguments in a way that is a close approximation to the model advanced here. Although the current experiments are intended primarily as example cases to demonstrate the amenability of a Bayesian quantitative representation of the scheme-based approach to testing, finding that participants' ratings did not diverge wildly from the predictions of a Bayesian model is in line with recent results in cognitive psychology (e.g., Gigerenzer, Hell, & Blank, 1988; Gigerenzer & Hoffrage, 1995; Griffiths & Tenenbaum, 2006; Harris & Hahn, 2009; Krynski & Tenenbaum, 2007). Of most relevance, Harris et al. (2012) provided empirical support for participants' Bayesian treatment of a simple variation of an *ad hominem* argument, using arguments set in a dialogical context as in the current study. These, and the current findings appear to be in line with Mercier and Sperber's (2011) argumentative theory of reasoning. Mercier and Sperber claim that the purpose of reasoning is to support argumentation. Hence, human reasoning performance should be at its best when set in an argumentation context. The results presented here, and in Harris et al. (2012), offer support for the contention that people are good Bayesian reasoners in an argumentation context.

The results presented here also provide further support for the Bayesian theory of argumentation more generally (Hahn & Oaksford, 2006, 2007a), which has previously offered empirically tested formalizations of the argument from ignorance (Hahn & Oaksford, 2007a; Hahn, Oaksford & Bayindir, 2005; Oaksford & Hahn, 2004), circular arguments (Hahn & Oaksford, 2007a), slippery slope arguments (Corner, Hahn, & Oaksford, 2011), and *ad hominem* arguments (Harris et al., 2012; Oaksford & Hahn, 2012). That participants' updating of their belief in the trustworthiness and expertise of the expert source was well predicted by the Bayesian model provides further support for the model, as do the totally parameter-free model fits observed in Experiment 2. While Experiment 2 had no free parameters that could be 'tweaked' to improve model fits, we did set the conditional

probabilities in the network a priori. These probabilities seemed appropriate for the experimental situation we set up and the questions we were asking. Future research might, however, profit from eliciting these conditional probabilities from participants, and testing the degree to which they are sensitive to changes in context. Alternatively, the quantitative parameters can be defined by the problem (as in Harris & Hahn, 2009; see also Harris et al., 2012, Experiment 3), suggesting another potential avenue for future research.

In addition to providing initial support for the present model as a computational-level description of people's ratings of a simple appeal to expert opinion, the present results are informative for maximizing the informativeness of future research endeavor within this area, by identifying aspects of the argument form that provide better or worse fits between model and participant data. Areas of worse fit are informative for model development. They can be suggestive of either erroneous assumptions on the part of the modeler, or else systematic human bias. In the present experiments, the model was seen to perform least well in a situation where the authority was a friendly expert who disagreed with other experts in the field (further highlighted by the significant result from the ANOVA analysis that demonstrates an interaction between "predicted/observed," "expertise," and "others' opinions"). Further research could focus on such situations, so as to determine the cause of this discrepancy (potential candidates might include conservative belief revision [e.g., Phillips & Edwards, 1966], on the side of systematic cognitive bias, or a missing dependency relation, on the side of an erroneous model assumption). Clearly, a potential outcome of such research is that evaluation of argumentation cannot be considered an approximation of normative Bayesian reasoning, but thus far we have obtained some support, and cited other support, for the contention that it can be.

Another important avenue for future research concerns the question of how to handle participants' ratings of 0 or 1 (certainty). In our analyses, we have also reported results where those participants are removed. While this removal generally increases model fits without systematically changing the relationship between model and data, in one instance it did. In Experiment 2, the pattern of ANOVA results for the trustworthiness posterior demonstrated that, while when no participants were excluded the effect of the "others' opinions" variable was systematically different in the model's predictions and the observed data, when participants' ratings of 0 or 1 were removed it was the effect of the "trustworthiness" variable that was different. Future research could potentially check participants' ratings of certainty by asking them a follow-up question delineating what this response actually means, to check that they are sure they want to provide that response rather than a response of (for example) .95. This is important, as identifying those variables whose effects are different in the model as opposed to the data can be highly beneficial for model development.

Finally, our experimental focus here has been on the relationship between the experimental conditions and the degree to which people's posterior degrees of belief in these different conditions are in line with the Bayesian predictions. Future research could focus on belief *change*. In such experiments, estimates of $P(H)$ should be elicited for all individual scenarios immediately before that argument is subsequently presented, so that the relationship between participants' estimates of $P(H)$ and $P(H|Rep)$ is perceptually evident to them on the response slider.

A standard concern with Bayesian approaches is the question of where prior degrees of belief should come from. In the "no mention of others" condition of the current experiments, Anne says that she has no idea whether taking Proftanine lowers cholesterol (for example). The mean $P(H)$ rating in Experiment 2 in this condition was .49. On the classical approach to probability this seems appropriate (dividing half one's degree of belief between H and $\neg H$). Another approach, however, is for her to use her knowledge of all possible substances. She would then recognize that only a very small proportion of these lower cholesterol and hence her prior should be very small indeed (e.g., $1 \times 10^{-\text{something}}$). One might question whether participants' estimates of $P(H)$ were therefore inflated because of the response scale we used (i.e., without the potential to provide infinitesimally small estimates). The distribution of responses for $P(H)$ (Appendix S2), however, does not suggest many participants saw such a response as appropriate. Recall that $P(H)$ was elicited after $P(Exp)$ and $P(T)$. In the latter questions, it is stated that two people are discussing H (e.g., whether Proftanine lowers cholesterol). From this, participants are likely to infer that Proftanine is not simply a random substance, but rather one that is currently under discussion as a potential cholesterol lowerer. Consequently, a value of .49 seems appropriate given this background, as do the values of .78 and .20 when experts agree and disagree, respectively.

The present paper, and the Bayesian approach in general, makes no claim as to the *process* by which people's ratings ended up being close approximations to the advanced model. The Bayesian formalization provides a statement of how people *should* update their belief, and the current data suggest that participants are appropriately sensitive to all the relevant parameters manipulated in this experiment. Consequently, these data can be seen to provide a potential challenge for the so-called dual-process theories of persuasion (the Heuristic Systematic Model [Chaiken, 1980]; the Elaboration Likelihood Model [e.g., Petty & Cacioppo, 1981]). Persuasion researchers have long been aware of the importance of both message content and source characteristics in effecting attitude change (Chaiken, 1980; Petty & Cacioppo, 1984, 1996). However, message and source characteristics have figured largely as tools in the development of process models of persuasion, with processing of message content associated with an "analytical route" to persuasion, and processing of message source with a "heuristic route." Hence, dual-process theories have typically viewed source and message factors in opposition.

Persuasion research has not formulated clear, general predictions about what should happen in circumstances where people might analytically evaluate *both* source and message (though special cases such as the processing of ambiguous messages have been considered, Chaiken & Maheswaran, 1994). A fundamental point to be taken from the present study is thus that the relationship between source and message characteristics is a subtle one. Our participants updated their perceptions of the source's characteristics in relation to the message provided, and they did so under circumstances that process-models of persuasion might consider to trigger "heuristic processing": namely, the evaluation of fictitious issues in which they had no personal stake or involvement. Nevertheless, they appeared to process the characteristics of the source in light of the content of the message (the likely truth of the message in this experiment being manipulated through the "others'

opinions" variable), suggesting that the analytic and heuristic route to persuasion are not completely separable (see also Hahn et al., 2009; Harris et al., 2013; Jarvstad & Hahn, 2011, Experiment 2; Reimer, Mata, & Stoecklin, 2004). Persuasion researchers have entertained the possibility that source considerations and message content may influence each other in evaluation (e.g., Chaiken & Maheswaran, 1994; Petty & Wegener, 1999). The fact that they seem to do so suggests that how they do so must be properly understood. The Bayesian framework can complement social psychological process models through its ability to make clear predictions about complex, non-additive relationships between source characteristics and message content, providing a framework for evaluating perceptions of trust and expertise.

4.1. Analyzing the appeal to expert opinion

With its inherently defeasible form, the central characteristics of the appeal to expert opinion appear to be well suited to a Bayesian Network formalization. The network presented in Fig. 2 and examined in the empirical work here is a very simple network for the appeal to expert opinion. The network in Fig. 1 is a slightly richer representation, but there are still plenty of ways in which the complexity of the network could be enhanced to provide a fuller representation of the argument. As an example, Walton (1997) outlines not only the six critical questions shown in Table 1, but questions 1, 2, and 3 have additional subquestions (five, four, and four further questions, respectively). These subquestions provide a means for assessing the truth status of the higher level questions. For example, the second subquestion for question 1 from Table 1 (the expertise question) asks, "What degrees, professional qualifications, or certification by licensing agencies does *E* [the expert] hold?" (Walton, 1997, p. 223). Such a question is easily captured in a Bayesian network through adding an additional node for qualifications, whose parent node is expertise (see Fig. 13)—for expertise enables one to gain qualifications, and qualifications are thus evidence of expertise. Expertise makes the likelihood of qualifications more likely, and therefore the existence of qualifications can be seen as evidence of expertise. In the present work, we have simplified the situation by only considering the probabilities of the end results (the major critical questions), but the Bayesian network can readily be fleshed out so as to explicitly represent these additional sources of evidence that enable the higher level inferences of expertise and trustworthiness. Fenton et al. (2013) have proposed a series of idioms to represent different aspects of the reasoning process in a legal fact-finding context. The separate groups of subquestions proposed by Walton could be considered as distinct idioms that could be included in a fully explicit Bayesian network formalization of the appeal to expert opinion. In the current experiments, we directly elicited degrees of belief for trustworthiness and expertise of a particular individual. To fully extend the network, it might be necessary to model these characteristics as distributions. Shafto et al. (2012), for example, fit parameters of "balance" (how helpful/knowledgable people are perceived to be in general) and "uniformity" (degree to which people are assumed to differ in their helpfulness/knowledge) to define a beta distribution for these characteristics.

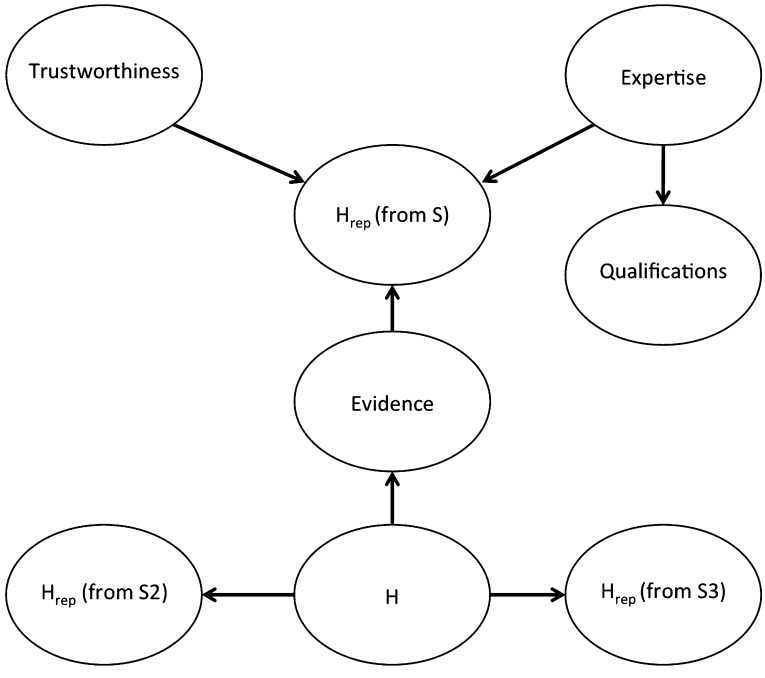

Fig. 13. Including Qualifications as evidence of expertise in a Bayesian network of the appeal to expert opinion.

Walton's (1997, 2008b) account of the appeal to expert opinion is explicitly dialectical in nature. Consequently, the critical questions represented in Table 1 are intended to be exactly that, *questions*. The Bayesian approach provides a formalization that demonstrates how the answers to these questions are important. The questioning will enable recipients of the appeal to expert opinion to gather the knowledge that is necessary for them to evaluate the convincingness of the argument—an evaluation process whose key components appear to be conceptualisable within a Bayesian account. This is important, because the issue of how critical questions could be formalized in a way that makes them amenable to treatment in a computational system has been of long-standing interest (see, e.g., Verheij, 2003a,b; Walton & Gordon, 2005; Walton et al., 2008, chapters 11 and 12). Typically, approaches to this problem have sought to develop defeasible generalizations of classical logic. Questions and counter-questions are then conceptualized as "attacks" that mean a particular premise can no longer be used (either by shifting the "burden of proof," or via some notion of "defeat"). Two fundamental problems for such approaches are the fact that argument quality itself would seem to matter in determining whether or not a claim is defeated or a burden of proof shifted (on the burden of proof in this context more generally, see Hahn & Oaksford, 2007b): for example, an entirely irrelevant proposition, advanced as a counter-claim, should neither be sufficient to defeat a claim nor to bring about a shift in the burden of proof. This means some independent means of

evaluating the content of the proposition and its strength as an argument is still required (see also Hahn & Oaksford, 2007a). Secondly, the point has been made that critical questions, in particular further subquestions, might, in principle be able to proceed indefinitely (see, e.g., Walton et al., 2008, chapter 11 on this point).

The Bayesian framework deals naturally with both of these concerns. It provides a generalization of classical logic in the sense that propositional inference is a limit case. However, as an intensional calculus (see e.g., Pearl, 1988), it deals naturally with argument *content*. It is the specific content of premises and claims that determines the probabilistic relations, and hence inferential relationships, between them. Thus, the framework captures naturally not only relevance relationships but also the summary consequences on degree of belief of amalgamating multiple pieces of evidence of varying strength, whether these be conflicting or mutually supporting (see also Hahn et al., 2012). Concerning the question of the possibility of ever further questions, finally, the Bayesian framework is helpful because, although it does not tackle fully the problem of never-ending exceptions (such exceptions will never be fully enumerated and hence explicitly modeled), it allows one to nevertheless reason in the face of such exceptions because probabilities *summarize* uncertainty, and thus can also summarize expectations of the possibility of relevant exceptions (see Pearl, 1988, chapter 1, for discussion of this point). Using the Bayesian framework to formalize schemes and critical questions thus allows one to address key challenges for formalization that the literature on schemes has previously identified.

At the same time, however, this is not to reject dialectical considerations in the context of the appeal to expert opinion, or informal argument more generally. For one, as the pragma-dialectic approach to argumentation has sought to identify, there are rules for a dialogical argumentative exchange that are most conducive to the resolution of a difference of opinion in a critical discussion (e.g., Van Eemeren & Grootendorst, 2004). A central tenet of this approach is that both parties in an argumentation exchange should be able to advance arguments, and the greater authority of one over another should not stand in the way of the truth-finding objective. Walton (1995; see also Van Eemeren & Grootendorst, 2004) defines a fallacy as being an argument that interferes with the proper goal of the dialog being engaged in—in this case the rational truth-finding goal of the critical discussion. Walton (1997, 2008b) therefore proposes not to label all weak instances of the appeal to expert opinion (as could be identified via the present Bayesian Network approach) as argument *fallacies*. Rather, the term "fallacy" (and thus the identification of the textbook fallacy of *ad verecundiam*) should be reserved for situations in which an appeal to expert opinion is used in a way that prevents the argument opponent from pursuing a line of critical questioning. Walton argues that appealing to an expert's opinion typically constitutes the *ad verecundiam* fallacy when the critical questioning of the argumentation opponent is suppressed by the proponent of the appeal. In these instances, the appeal to expert opinion is a "decisive blocking or shutting-down type of move in argumentation that blocks off the respondent's ability to raise any further questions or meaningfully or effectively take part in attempting to support his side of the issue any further in the dialogue" (Walton, 2008b, p. 61). The effect of the argument when used in this

way is perhaps understood in terms of its proponent following it up with the statement, "Well, who are you to have an opinion? You're no expert are you?" (see also, Walton, 2008b), and thus the silencing of the opponent is complete.

It is difficult to see what the Bayesian Network formalization laid out in Figs. 1 and 2 can have to say in this regard. We prefer to see the probabilistic Bayesian approach and the pragmatic, dialectical approach as addressing two different important questions pertaining to the evaluation of the appeal to expert opinion and, indeed, argumentation more generally (see also Hahn & Oaksford, 2006, 2007a,b). Dialectical rules are important for ensuring that the maximum relevant information possible is present within a critical discussion. Without such rules, the usefulness of argumentation in arriving at good conclusions is greatly compromised. Argument *content*, however, seems best evaluated through the formalization of the relationships between propositions as provided by the well-established norms of probability theory and hence the Bayesian approach to argumentation.

Notes

1. We are working with the assumption that the general practitioner has my best interests at heart.
2. If the opinions of other sources are sought rather than known a priori, then they are, technically speaking, not part of the *prior* belief. According to our model (Fig. 1), the effect on the evaluation of the argument will, however, be the same.
3. Šorm (2010) also maintains that the grounding of the expert's testimony in evidence should be considered a subcriterion of the expertise and trustworthiness criteria.
4. Of course, in general, an expert may also simply not speak to an issue. There is then no expert opinion to appeal to. This case, too, may inferentially be of interest as an argument from ignorance (see, e.g., Hahn & Oaksford, 2007a; Harris, Corner, & Hahn, 2013). Inclusion of this third option (a "non-response") demands three possibilities for the evidence: positive evidence, negative evidence, and "non response." Appropriate formalizations of this are presented, for example, in Hahn and Oaksford (2007a). Functionally, the inclusion of a third alternative, "non-response" leaves unaffected inferences that are drawn from positive evidence (as can be verified by consulting the equations in Hahn & Oaksford, 2008, p. 131). The strength of these inferences is determined only by the relative probability of obtaining the positive evidence if the hypothesis were true, as opposed to if it were false (and likewise for explicit negative evidence where the expert says "no"). Moreover, it is only the ratio between these probabilities that matters, not their absolute values. Whether the possibility of a "non-response" is included or not thus makes no difference to any of the cases examined in our paper.
5. Note that there is also empirical evidence that people do treat expertise and trustworthiness independently (e.g., O'Hara et al., 1991; Wiener & Mowen, 1986).

6. The fact that the links in a Bayesian Network (e.g., Fig. 2) have a clear, understandable interpretation as conditional probability relationships makes the parameterization of a Bayesian model more natural than the parameterization of, for example, a Markov Random Field Model, such as proposed in Butterfield, Jenkins, Sobel, and Schwertfeger (2009) to model children's trust in testimony. This makes the relation between the model and the behavior to be modeled more transparent. It also means that the probabilities necessary for parameterization can be elicited in a number of ways. Note also that Butterfield et al. (2009) did not allow for a variation in trustworthiness, stating that "we are concerned with collaboration between agents acting cooperatively, there is no reason for agents to deceive each other. . ." (p. 43).
7. The collected data for both experiments are included in the online supplementary materials.
8. Although not all studies have reported a significant effect of manipulating the expertise of the authority (e.g., Hoeken, Timmers, & Schellens, 2012), in Hoeken et al.'s study, the expertise manipulation was much more subtle than the expertise manipulation in the current experiments, with a freshman studying nutrition still likely to be perceived as a source of some expertise on the subject of nutrition—in a way that a musician would not be perceived as an expert on a medical matter in the present experiment.
9. If certainty ratings are not excluded, the same interaction remains significant, $F(2, 81) = 3.87$, $p = .025$, and the 3-way interaction between "predicted/observed," "trustworthiness," and "others' opinions" was also significant, $F(2, 81) = 4.48$, $p = .014$.
10. If certainty ratings are not excluded, the same interactions remain significant: "predicted/observed" × "others' opinions," $F(2, 81) = 8.04$, $p = .001$; "predicted/observed" × "expertise" × "others' opinions," $F(2, 81) = 4.48$, $p = .014$. In addition, there is also a significant "observed/predicted" × "expertise" interaction, $F(1, 81) = 12.26$, $p = .001$.
11. This result was the same when certainty ratings were not excluded (all ps > .17).
12. An alternative approach would have been to elicit the prior solely in the absence of any other information (as in the "no mention of others" condition) and capture the effect of other experts by eliciting participants' subjective conditional probabilities of the relationship between H and H_{rep} (*from S1*) and H_{rep} (*from S2*). . . (see Fig. 1). We chose the current approach to reduce the load on participants, and in line with the reasoning of note 2.
13. The pattern of results is unchanged if certainty ratings are not excluded. The 3-way interaction between "predicted/observed," "trust," and "others' opinions" again approached significance ($p < .06$; all other ps > .135).

References

Anderson, J. R. (1990). *The adaptive character of human thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Artz, D., & Gil, Y. (2007). A survey of trust in computer science and the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, *5*, 58–71.

Balakrishnan, R., & Kambhampati, S. (2011). SourceRank: Relevance and trust assessment for deep web sources based on inter-source agreement. In *Proceedings of the international conference on World Wide Web (WWW)* (pp. 227–236). New York: ACM.

Birnbaum, M. H., & Mellers, B. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology*, *37*, 48–74.

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise and the source's point of view. *Journal of Personality and Social Psychology*, *37*, 48–74.

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

Braunsberger, K., & Munch, J. M. (1998). Source expertise versus experience effects in hospital advertising. *Journal of Services Marketing*, *12*, 23–38.

Brinol, P., & Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology*, *20*, 49–96.

Butterfield, J., Jenkins, O. C., Sobel, D. M., & Schwertfeger, J. (2009). Modeling aspects of theory of mind with Markov random fields. *International Journal of Social Robotics*, *1*, 41–51.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*, 752–766.

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgement. *Journal of Personality and Social Psychology*, *66*, 460–473.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65.

Corner, A., & Hahn, U. (2013). Normative theories of argumentation: Are some norms better than others? *Synthese*, *190*, 3579–3610.

Corner, A., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory & Language*, *64*, 133–152.

Corner, A., Harris, A. J. L., & Hahn, U. (2010). Conservatism in belief revision and participant skepticism. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1625–1630). Austin, TX: Cognitive Science Society.

Eaton, T. E., & O'Callaghan, M. G. (2001). Child-witness and defendant credibility: Child evidence presentation mode and judicial instructions. *Journal of Applied Social Psychology*, *31*, 1845–1858.

Ericsson, K. A., & Lehman, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, *47*, 273–305.

Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford, UK: Oxford University Press.

Fearnside, W. W., & Holther, W. B. (1959). *Fallacy: The counterfeit of argument*. Englewood Cliffs, NJ: Prentice-Hall.

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

Fernbach, P. M., & Erb, C. D. (2013). A quantitative causal model theory of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1327–1343.

ForsterLee, L., Horowitz, I. A., Athaide-Victor, E., & Brown, N. (2000). The bottom line: The effect of written expert witness statements on juror verdicts and information processing. *Law and Human Behavior*, *24*, 259–270.

Friedman, R. D. (1987). Route analysis of credibility and hearsay. *The Yale Law Journal*, *96*, 667–742.

Garssen, B. (2001). Argument schemes. In F. van Eemeren (Ed.), *Crucial concepts in argumentation theory* (pp. 81–99). Amsterdam: Amsterdam University Press.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 513–525.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, *11*, 791–806.

Godden, D. M., & Walton, D. (2006). Argument from expert opinion as legal evidence: Critical questions and admissibility criteria of expert testimony in the American legal system. *Ratio Juris*, *19*, 261–286.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767–773.

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

Hahn, U., & Oaksford, M. (2006). A Bayesian approach to informal reasoning fallacies. *Synthese*, *152*, 207–223.

Hahn, U., & Oaksford, M. (2007a). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*, 704–732.

Hahn, U., & Oaksford, M. (2007b). The burden of proof and its role in argumentation. *Argumentation*, *21*, 39–61.

Hahn, U., & Oaksford, M. (2008). Inference from absence in language and thought. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind* (pp. 121–142). Oxford, UK: Oxford University Press.

Hahn, U., Oaksford, M., & Bayindir, H. (2005). How convinced should we be by negative evidence? In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 887–892). Stresa, Italy: Cognitive Science Society.

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, the Netherlands: Springer.

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

Harris, P. L., & Corriveau, K. H. (2011). Young children's selective trust in informants. *Philosophical Transactions of the Royal Society B*, *366*, 1179–1187.

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*, 1366–1372.

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

Hastings, A. C. (1962). *A reformulation of the modes of reasoning in argumentation*. Unpublished dissertation. Evanston, IL: Northwestern University.

Hoeken, H., Timmers, R., & Schellens, P. J. (2012). Arguing about desirable consequences: What constitutes a convincing argument? *Thinking and Reasoning*, *18*, 394–416.

Hornikx, J. (2011). Epistemic authority of professors and researchers: Differential perceptions by students from two cultural-educational systems. *Social Psychology of Education*, *14*, 169–183.

Hornikx, J., & Hoeken, H. (2007). Cutural differences in the persuasiveness of evidence types and evidence quality. *Communication Monographs*, *74*, 443–463.

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion: Psychological studies of opinion change*. New Haven, CT: Yale University Press.

Howson, C., & Urbach, P. (1996). *Scientific reasoning: The Bayesian approach* (2nd ed.). Chicago, IL: Open Court.

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

Kadane, J. B., & Schum, D. A. (1996). *A probabilistic analysis of the Sacco and Vanzetti evidence*. New York: John Wiley & Sons.

Kienpointner, M. (1992). *Alltagslogik: Struktur und Funktion von Argumentationsmustern*. Stuttgart-Bad Cannstatt: Friedrich Frommann.

Knill, D. C., & Richards, W. (Eds.) (1996). *Perception as Bayesian inference*. Cambridge, UK: Cambridge University Press.

Krauss, D. A., & Sales, B. D. (2001). The effects of clinical and scientific expert testimony on juror decision making in capital sentencing. *Psychology, Public Policy, and Law*, *7*, 267–310.

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*, 430–450.

Lagnado, D. A. (2011). Thinking about evidence. In P. Dawid, W. Twining, & M. Vasaliki (Eds.), *Evidence, inference and enquiry*. Oxford, UK: Oxford University Press/British Academy.

Lagnado, D. A., Fenton, N., & Neil, M. (2013). Legal idioms: A framework for evidential reasoning. *Argument and Computation*, *4*, 46–63.

Laplace, S. (1951). A philosophical essay on probabilities. In F. W. Truscott & F. L. Emory, Trans (Eds.). (pp. 148). New York: Dover. (Original work published 1814).

Lindley, D. (1994). Foundations. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 3–15). Chichester, UK: John Wiley & Sons.

Maddux, J. E., & Rogers, R. W. (1980). Effects of source expertness, physical attractiveness, and supporting arguments on persuasion: A case of brains over beauty. *Journal of Personality and Social Psychology*, *39*, 235–244.

McGuire, W. J. (1985). Attitudes and attitude change. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (3rd ed., Vol. 2, pp. 233–346). San Diego, CA: Academic Press.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*, 57–111.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning*. Hove, UK: Psychology Press.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, UK: Oxford University Press.

Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, *58*, 121–131.

Oaksford, M., & Hahn, U. (2012). Why are we convinced by the ad hominem argument? Bayesian source reliability and pragma-dialectical discussion rules. In F. Zenker (Ed.), *Bayesian argumentation* (pp. 39–60). Dordrecht, the Netherlands: Springer.

Ohanian, R. (1991). The impact of celebrity spokespersons' perceived image on consumers' intention to purchase. *Journal of Advertising Research*, *31*, 46–54.

O'Hara, B. S., Netemeyer, R. G., & Burton, S. (1991). An examination of the relative effects of source expertise, trustworthiness, and likeability. *Social Behavior and Personality: An International Journal*, *19*, 305–314.

Page, S. E. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufman.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.

Petty, R. E., & Cacioppo, J. T. (1981). *Attitudes and persuasion: Classic and contemporary approaches*. Boulder, CO: Westview Press.

Petty, R. E., & Cacioppo, J. T. (1984). Source factors and the elaboration likelihood model of persuasion. *Advances in Consumer Research*, *11*, 668–672.

Petty, R. E., & Cacioppo, J. T. (1996). *Attitudes and persuasion: Classic and contemporary approaches*. Boulder, CO: Westview Press.

Petty, R. E., & Wegener, D. T. (1999). The elaboration likelihood model: Current status and controversies. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 41–72). New York: Guilford Press.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simply probability inference task. *Journal of Experimental Psychology*, *72*, 346–354.

Pollock, J. L. (2001). Defeasible reasoning with variable degrees of justification. *Artificial Intelligence*, *133*, 233–282.

Pornpitakpan, C. (2004). The persuasiveness of source credibility : A critical review of five decades' evidence. *Journal of Applied Social Psychology*, *34*, 243–281.

Pornpitakpan, C., & Francis, J. N. P. (2001). The effect of cultural differences, source expertise, and argument strength on persuasion: An experiment with Canadians and Thais. *Journal of International Consumer Marketing*, *13*, 77–101.

Rahwan, I., & Simari, G. R. (2009). *Argumentation in artificial intelligence*. Dordrecht, the Netherlands: Springer.

Reed, C. A., & Rowe, G. W. A. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, *13*, 961–979.

Reimer, T., Mata, R., & Stoecklin, M. (2004). The use of heuristics in persuasion: Deriving cues on source expertise from argument quality. *Current Research in Social Psychology*, *10*, 69–83.

Rescher, N. (1977). *Dialectics: A controversy-oriented approach to the theory of knowledge*. Albany, NY: State University of New York Press.

Rosenkrantz, R. D. (1992). The justification of induction. *Philosophy of Science*, *59*, 527–539.

Schellens, P. J. (1985). *Redelijke argumenten: Een onderzoek naar normen voor kritische lezers*. Dordrecht, the Netherlands: Foris.

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*. New York: Wiley.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, *15*, 436–447.

Šorm, E. (2010). *The good, the bad and the persuasive: Normative quality and actual persuasiveness of arguments from authority, arguments from cause to effect and arguments from example*. Nijmegen, NL: LOT.

Stevenson, R. J., & Over, D. E. (2001). Reasoning from uncertain premises: Effects of expertise and conversational context. *Thinking and Reasoning*, *7*, 367–390.

Van Eemeren, F. H., & Grootendorst, R. (2004). *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge, UK: Cambridge University Press.

Verheij, B. (2003a). Dialectical argumentation with argumentation schemes: Towards a methodology for the investigation of argumentation schemes. In F. H. van Eemeren, A. Blair, C. Willard, & F. Snoeck Henkemans (Eds.), *Proceedings of the 5th Conference of the International Society for the Study of Argumentation (ISSA 2002)* (pp. 1033–1037). Amsterdam: Sic Sat.

Verheij, B. (2003b). Deflog: On the logical interpretation of prima facie justified assumptions. *Journal of Logic and Computation*, *13*, 319–346.

Wallsten, T. S. (1990). The costs and benefits of vague information. In R. M. Hogarth (Ed.), *Insights in decision making* (pp. 28–43). Chicago, IL: University of Chicago Press.

Walton, D. (1995). *A pragmatic theory of fallacy*. Tuscaloosa, AL: University of Alabama Press.

Walton, D. (1997). *Appeal to expert opinion: Arguments from authority*. University Park, PA: Pennsylvania State University Press.

Walton, D. (2008a). *Witness testimony evidence: Argumentation, artificial intelligence, and law*. Cambridge, UK: Cambridge University Press.

Walton, D. (2008b). *Informal logic: A pragmatic approach* (2nd ed.). New York: Cambridge University Press.

Walton, D., & Gordon, T. F. (2005). Critical questions in computational models of legal argument. In P. E. Dunne & T. Bench-Capon (Eds.), *International workshop argumentation in artificial intelligence and law* (pp. 103–111). Nijmegen, the Netherlands: Wolf Legal.

Walton, D. N., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge, UK: Cambridge University Press.

Wang, D., Abdelzaher, T., Ahmadi, H., Pasternack, J., Roth, D., Gupta, M., Han, J., Fatemieh, O., Le, H. & Aggarwal, C. (2011). On bayesian interpretation of fact-finding in information networks. In *14th International Conference on Information Fusion (Fusion 2011)*. Chicago, IL

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology*, *54*, 277–295.

Wiener, J. L., & Mowen, J. C. (1986). Source credibility: On the independent effects of trust and expertise. *Advances in Consumer Research*, *13*, 306–310. Available at: http://www.acrwebsite.org/volumes/display.asp?id=6509. Accessed July 17, 2012.

Wolf, A. G., Rieger, S., & Knauff, M. (2012). The effects of source trustworthiness and inference type on human belief revision. *Thinking and Reasoning*, *18*, 417–440.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. The equations behind the nested models used as a comparison in the results section of Experiment 2.

Appendix S2. Distributions of estimates of P(H) in Experiment 2 across the three "others' opinions" conditions.

Data S1. All ratings (raw data) provided by participants in both experiments.