Synthese (2016) 193:1833–1873
DOI 10.1007/s11229-015-0815-0

A normative framework for argument quality: argumentation schemes with a Bayesian foundation

Ulrike Hahn[1] · **Jos Hornikx**[2]

Received: 19 May 2014 / Accepted: 25 June 2015 / Published online: 22 July 2015

Abstract In this paper, it is argued that the most fruitful approach to developing normative models of argument quality is one that combines the argumentation scheme approach with Bayesian argumentation. Three sample argumentation schemes from the literature are discussed: the argument from sign, the argument from expert opinion, and the appeal to popular opinion. Limitations of the scheme-based treatment of these argument forms are identified and it is shown how a Bayesian perspective may help to overcome these. At the same time, the contributions of the standard scheme-based approach are highlighted, and it is argued that only a combination of the insights of different traditions will yield a complete normative theory of argument quality.

Keywords Argumentation · Rationality · Testimony · Evidence · Inference

1 Introduction

In order for humans to learn, make decisions, and interact with others, reasoning and argumentation are essential (Hahn and Oaksford 2012; Mercier and Sperber 2011). People use arguments in virtually all areas of their lives in order to convince others, and to convince themselves. It thus seems of fundamental importance to be able to distinguish good arguments from bad, so that we may end up convinced of the right things to do and believe. However, argument quality is not only of interest from the

✉ Ulrike Hahn
u.hahn@bbk.ac.uk

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

[2] Department of Communication and Information Sciences, Radboud University Nijmegen, Nijmegen, The Netherlands

perspective of rational standards that guide what we *should* do, but—comfortingly—there is considerable evidence to suggest that argument quality is a key factor in determining how convincing arguments are in actual practice (Hoeken et al. 2012, 2014; Hornikx and Hoeken 2007). While there may be occasions on which weak arguments command more influence than they deserve, by and large, so empirical research suggests, better arguments are more likely to meet with greater persuasive success. Identifying what arguments should count as weak and which as strong is thus a fundamental question with both theoretical and practical implications.

Despite its importance, and considerable research effort aimed at the issue, a comprehensive theory of argument quality has remained elusive. This is manifest in many theoretical contexts: in epistemological debate and in the study of practical reasoning, including the formal study of fallacies of reasoning and argumentation (e.g, Hamblin 1970); in the design of computational systems capable of engaging in argument (e.g., Rahwan and Simari 2009); in the context of empirical studies of persuasion, as conducted by psychologists (e.g., Eagly and Chaiken 1993; Petty and Cacioppo 1986) and communication researchers (O'Keefe 2002); and, finally, in the plethora of non-classical logics that have been developed in order to provide a broader coverage of reasoning and inference, in particular in light of uncertainty (see e.g., Prakken and Vreeswijk 2002, for a review).

Historically, two broad approaches to the issue of argument quality may be seen. The one proceeds from the abstract to the concrete, and has been dominant in the development of non-classical logics. Abstract features of what might constitute a good argument are identified, such as general desiderata of an inference relationship, and formal frameworks are developed on this basis. Along the way, there may be some consideration of intuitive examples, but there is typically limited contact with large numbers of real world examples. Starting from the other end, research traditions within argumentation studies or informal logic, by contrast, have collected and catalogued real world arguments and sought to develop appropriate classifications for them. Here, there have been intuitions about strength, but formal frameworks have been lacking.

Several more recent developments have sought to overcome these limitations, trying to combine rich, everyday examples, with more formal, computational treatments. Moving from specific to abstract, researchers have increasingly sought to combine so-called scheme-based approaches to argumentation (attempts to systematize and classify types of everyday argument which will be introduced in more detail below) with more formal, often computational treatments by incorporating schemes into computational tools for argument reconstruction (e.g., Reed and Rowe 2004; Rahwan and Simari 2009) and/or combining schemes with systems for plausible or default reasoning (e.g., see, e.g., Verheij 2004; Walton and Gordon 2005; Walton et al. 2008, Chaps. 11, 12). Moving from abstract to specific, the formal framework of Bayesian probability theory has seen ever increasing application to general questions about evidential strength within epistemology—such as whether diversity makes evidence stronger (e.g., Howson and Urbach 1993; Fitelson 1996; Earman 1992; Wayne 1995; Myrvold 1996), or the extent to which more coherent testimonies are more likely to be true or stem from sources that are reliable (e.g., Olsson 2002, 2005; Olsson and Schubert 2007; Bovens and Hartmann 2003). Most recently, the Bayesian framework has been applied to the catalogue of so-called fallacies of argumentation (e.g., Hahn

and Oaksford 2006a, 2007a), collections of everyday argument forms that may sometimes seem convincing, but should not do so. Finally, the Bayesian framework has been used as a basis for computational systems for argument generation (e.g., Korb et al. 1997; McConachy et al. 1998; McConachy and Zukerman 1999).

Both of these research strands, argumentation schemes and Bayesian argumentation, share a common point of departure in their recognition of the limitations of classical logic in the face of everyday reasoning, but have responded to the need for alternative norms for argument evaluation in different ways. In the remainder of this paper, we seek to establish how these two traditions may profitably be combined in order to provide a comprehensive treatment of argument quality that ranges from abstract considerations to specific, everyday examples, allowing one to both capture the richness of real-world argument, and to do so in a way that is formally and computationally explicit.

We begin by introducing both the scheme-based and the Bayesian approach to argumentation in more detail.

2 Argumentation schemes

In order to define what high-quality arguments are, standards or norms are needed against which argument quality can be judged. One type of standard that has been proposed to govern argumentation is procedural norms (see e.g., Alexy 1989; Eemeren and Grootendorst 2004; Walton 1998). As Hahn and Oaksford (2006b, 2012) have argued, these norms are certainly useful for determining to what extent a discussion can be considered to be rational (or fair, see e.g., Christmann et al. 2000a, b), but such norms cannot cover all aspects of argument quality. This is readily seen in the context of an overall dialogue with a series of individual arguments and counter-arguments. Here, procedural norms miss an evaluative component when it comes to the concluding stage of the discussion: in the final reckoning, which set of arguments is stronger? What should a rational agent's overall conclusion be? For this evaluation, *content* norms are needed.

The scheme-based approach to argumentation seeks to provide such norms for content evaluation. In this approach, which originates from Aristotle's notion of topics (*topoi*), different argument types are distinguished, such as the argument from example or the argument from analogy. Typologies within argumentation theory have identified a large number of different argumentation schemes (e.g., Garssen 1997; Hastings 1962; Kienpointner 1992; Perelman and Olbrechts-Tyteca 1969; Schellens 1985), as has research on communication and debate (e.g., Ehningner and Brockriede 1963; Rieke and Sillars 1984; Reinard 1991; Inch and Warnick 2009). Differences across typologies concern not only the number of schemes, but also the structure of these schemes. Table 1 provides a range of sample classification systems (for a detailed historical overview of the scheme literature see Walton et al. (2008), Chap. 8). Some of these classification systems are very simple, such as the pragma-dialectical account of Garssen (1997) which classifies all argument schemes into one of three main categories: sign, comparison, and cause. A relatively complex classification is provided by Perelman and Olbrechts-Tyteca (1969), whose classification has five levels. At the

Table 1 Some classification systems for argumentation schemes

System	Example
Hastings (1962)	
Verbal reasoning	Argument from example
Causal reasoning	Argument from sign
Verbal or causal reasoning	Argument from comparison
Perelman and Olbrechts-Tyteca (1969)	
Quasi-logical arguments	Arguments by comparison
Arguments based on the structure of reality	Pragmatic argument
Relations establishing the structure of reality	Example
Schellens (1985)	
Argumentation based on regularity	Causality
Argumentation based on rules	Appreciation rules
Pragmatic argumentation	Pragmatic argumentation
Unbound argumentation	Argument from example
Kienpointner (1992)	
Classification	Definition
Comparison	Similarity
Opposition	Contradictory
Causality	Cause
Schemes establishing rules	Inductive example
Schemes using or establishing rules	Authority
Garssen (1997)	
Sign	Sign
Comparison	Comparison
Causality	Cause

first level, there is the large class of arguments that 'associate' (all arguments given in Table 1) and a small class of arguments that 'dissociate' (which serve to dissociate elements that an audience sees as belonging together). Among the arguments that 'associate' then, are 'arguments based on the structure of reality' (see Table 1), which contain further sub-divisions, such as 'arguments based on sequential relations', and, then, below that level, individual schemes such as, for example, the 'argument from waste'. These two contrasting approaches—Garssen (1997) and Perelman and Olbrechts-Tyteca (1969)—illustrate that it is not straightforward to compare different classification systems. Though such systems may share basic schemes such as the argument from example or causal argument, the ways in which these schemes are related to each other diverge from one system to the other.

The scheme-based approach has a descriptive starting point: the schemes are "forms of argument (structures of inference) that represent structures of common types of arguments used in everyday discourse" (Walton et al. 2008, p. 1). However, this approach also has a normative component. Since Hastings (1962) in particular, schol-

ars have sought to formulate appropriate "critical questions" for each argumentation scheme. These critical questions serve as scheme-specific norms: an argument's quality depends on the responses an arguer gives to the questions asked. For the argument from expert opinion, for instance, Walton (1997) lists six critical questions, such as 'Is the expert biased?'.

The argumentation scheme approach has been a dominant framework in argumentation research, also influencing researchers in other fields such as artificial intelligence (AI) (e.g., Walton et al. 2008; Rahwan and Simari 2009), persuasion (e.g., Hoeken et al. 2012; Hornikx and Hoeken 2007) and educational psychology (e.g., Nussbaum 2011; Nussbaum and Edwards 2011). Even to its proponents, however, the approach is not free of problems. For instance, Walton et al. (2008) point out that there remain fundamental questions concerning the exact normative status of the critical questions. For the argument from expert opinion, they write: "Suppose the proponent answered all the six basic critical questions posed by the respondent in prior dialogue exchanges. Is the respondent obliged at that point to accept the appeal to expert opinion as reasonable?" (Walton et al. 2008, p. 31). What if there exists a relevant seventh question? Or what if the proponent answers the first five questions, and the sixth question is not addressed?

As we will argue in the paper, this completeness problem is not the only problem the scheme-based approach has encountered. This and other problems identified in the following mean, we think, that the scheme-based approach cannot stand on its own. As we will seek to demonstrate, however, argumentation schemes may be combined with a Bayesian perspective into an integrated, overall account of argument quality that has both a solid normative foundation and covers successfully actual, specific, argument forms used in everyday life. We next outline the Bayesian perspective to argument quality. We then analyse in detail three sample argumentation schemes from a Bayesian perspective.

3 Bayesian argumentation

The formal framework of Bayesian probability has swept a wide range of fields ranging from statistics (e.g., Bolstad 2004), computer science (e.g., Bishop 2006), cognitive science (e.g., Chater et al. 2006), to epistemology (Bovens and Hartmann 2003), and has found increasing application to argumentation (e.g., Hahn and Oaksford 2007a; Oaksford and Hahn 2004; Korb 2004; Zukerman 2009; Harris et al. 2012).

On a Bayesian perspective, probabilities represent degrees of belief. The probability $P(h)$ represents a degree of belief in the hypothesis h being true (see for an introduction e.g., Howson and Urbach 1993). Crucially, the framework offers a means for relating beliefs to other beliefs. Central here is the notion of conditional probability: $P(a|b)$ is the probability of event a conditional on b, that is, its probability given that b is true. The importance of the notion of conditional probability to evidential contexts rests on the fact that it allows one to calculate a degree of belief in a claim, such as hypothesis h, given evidence, e, for that hypothesis. That is, the notion of conditional probability provides a basic notion of the appropriate degree of belief in a hypothesis given particular evidence, $P(h|e)$.

At the heart of the Bayesian framework is Bayes' theorem as a normative standard for belief revision:

$$P(h|e) = \frac{P(h)P(e|h)}{P(h)P(e|h) + P(\neg h)P(e|\neg h)} \tag{1}$$

According to Bayes' theorem people's belief in conclusion h given the evidence e, $P(h|e)$, should normatively depend on their prior degree of belief in the conclusion, $P(h)$ (that is, how convinced they are independently of evidence e), and the likelihood ratio associated with the given evidence e: this likelihood ratio is the ratio of the so-called likelihood, $P(e|h)$, which in signal detection terms reflects test *sensitivity* (Green and Swets 1966), and the false positive rate $P(e|\neg h)$, that is the probability of obtaining evidence e even if h is false. Furthermore, $P(e|\neg h) = 1 - P(\neg e|\neg h)$, where $P(\neg e|\neg h)$ represents the so-called *specificity* of the evidence. Probabilities as degrees of belief reflect the degree of certainty one might have concerning a proposition, ranging from $P = 0$ for a claim that is necessarily false (such as a logical contradiction) to $P = 1$ for claims that are definitely true.[1]

The importance of the likelihood ratio can be seen from the fact that decreasing $P(e|\neg h)$ in the denominator will increase, all other things equal, the posterior $P(h|e)$. This is made particularly transparent by thinking in terms of odds (see also Pearl 1988):

$$\frac{P(h|e)}{P(\neg h|e)} = \frac{P(e|h)}{P(e|\neg h)} \times \frac{P(h)}{P(\neg h)} \tag{2}$$

in other words, the posterior odds, $P(h|e) : P(\neg h|e)$, are equal to the prior odds times the likelihood ratio. Consideration of odds, and the fact that odds can readily be converted into probabilities and vice versa[2] also makes clear that in evaluating the strength of evidence it is ratios that ultimately matter, not absolute values: what is important is that a piece of evidence is, say, twice as likely to be observed under the assumption that the hypothesis is true than that it is false. This is important, last but not least, because the assessment of ratios may be psychologically considerably more straightforward than estimating absolute values (see also Pearl 1988).

Applications to argumentation derive from the fact that e may readily be viewed as a reason for believing h: in evaluating the quality of an argument from e to h, we consider to what extent it rationally increases one's belief in conclusion h. From a Bayesian perspective, any kind of reason offered to support a conclusion can broadly be regarded as evidence for a hypothesis.

To illustrate Bayes' theorem with a simple example, let the claim or conclusion h be that 'London is a tourist destination', and the evidence e 'There are large numbers of people with digital cameras on the street'. According to Bayes' theorem, people's belief

[1] In first instance, probabilities—as degrees of belief—are subjective, and the probability calculus is about coherence, in the same way that classical logic is about the relationships between statements, not their truth or falsity per se. However, it is typically assumed that a rational agent should adopt as her subjective degree of belief objective probabilities (limit frequencies) where these are defined, see e.g., Lewis (1980).

[2] Posterior odds convert into posterior degrees of belief via the simple relationship $P(A) = Odds(A)/(1 + Odds(A))$.

in London as a tourist destination given the evidence e, should be based on their prior belief in touristic London, $P(h)$, on their belief in there being numerous people with digital cameras given that London is a tourist destination, $P(e|h)$, and their degree of belief in there being numerous digital camera carriers given that London is not a tourist destination, $P(e|\neg h)$. The more likely people think it is that there will be numerous people in the street armed with digital cameras given that London is a tourist destination, the more convinced they should be that London is, in fact, a tourist destination once they observe large numbers of people with digital cameras on the street. Likewise, the less likely they think it that there will be large numbers of camera carriers if London is *not* a tourist destination, the more they should think it is a tourist destination when these camera carriers are observed. Where $P(e|h) = P(e|\neg h)$ (that is, the likelihood ratio $= 1$), the evidence is simply non-diagnostic: if one thinks the evidence is as likely to be found regardless of whether the hypothesis is true or false, then observing it will not affect one's belief in its truth: for example, one may think it highly likely that there will be cars with four wheels in London, if it is a tourist destination; but if one thinks it is equally likely that there will be cars with four wheels if it is not, then the presence of four-wheeled cars will have no impact on one's belief about London's touristic status.

Bayesian probability has long been considered a normative framework for many different evidential contexts such as medical tests and diagnosis or judgments of risk (Hardman 2009).[3] It has also been applied within the philosophy of science to the analysis of scientific reasoning (e.g., Earman 1992; Howson and Urbach 1993). Its normative appeal derives, among other things, from formal results indicating that adherence to the probability calculus secures coherent degrees of belief that are not subject to Dutch books (translations into betting odds that will incur a sure loss, see e.g., Hajek 2008), its systematic link with calibration (Rosenkrantz 1992), and demonstrations that being Bayesian will minimize the inaccuracy of our beliefs, that is the extent to which we take things to be true when they are not and vice versa (Leitgeb and Pettigrew 2010). The probability calculus also seems to match many fundamental intuitions about evidence and evidential strength, as highlighted in the famous adage by Laplace et al. (1951) that it is but 'formalised common sense' (for thorough discussion of norms and their foundation with particular focus on argumentation, see Corner and Hahn 2013; and for more general discussion of the role of Bayesian models in the study of cognition, see Hahn 2014).

Bayes' theorem has been successfully applied to argumentation theory, in particular in the study of fallacies (e.g., Hahn and Oaksford 2006a, 2007a; Korb 2004; Oaksford and Hahn 2004). Central examples here include the so-called argument from ignorance ('ghosts exist, because no one has proven that they don't'), circular arguments, and slippery slope arguments.

A long-standing issue in fallacies research had been the absence of a systematic formal treatment (see Hamblin 1970 for discussion) that might explain *why* fallacious arguments are weak. Key to the application of the Bayesian framework in this context is the idea that many everyday arguments may be thought of as weak inductive arguments where a reason provides 'evidence', broadly speaking, for a claim.

[3] This is not to say, however, that alternative frameworks have not been put forward, see e.g., Schum (1994).

Fallacies such as the argument from ignorance are then not categorically different from other types of argument. Rather they are simply weak versions of inductive arguments that, with other content, may be perfectly acceptable. By demonstrating how the strength of a type of fallacy depends on the strength of the probabilistically relevant factors—sensitivity and specificity in particular—it is possible to provide an explanation for why arguments of that structure are typically weak. It is also possible to address a longstanding difficulty in fallacies research, namely the presence of 'exceptions': arguments with the same structure, which nevertheless seem acceptable. To illustrate with the case of the argument from ignorance, Hahn and Oaksford (2007a) show that for a broad range of possible values for sensitivity and specificity, arguments from direct evidence are stronger than corresponding arguments from indirect (negative) evidence. At the same time, they evaluate individual examples of arguments from ignorance that are sufficiently compelling that they figure prominently in everyday life ('this drug is safe, because no side effects have been observed in clinical trials').

The Bayesian framework thus not only provides abstract considerations about reason strength, but can be used to evaluate specific, concrete arguments. Such evaluation is sufficiently detailed to allow experimental study of the extent to which lay people's perceptions of argument strength match Bayesian prescriptions (see e.g., Oaksford and Hahn 2004; Hahn et al. 2005; Corner et al. 2011; Harris et al. 2013). Furthermore, a common finding here has been that they often match surprisingly well (see, e.g., Harris et al. 2012). These experimental studies provide detailed, concrete examples of how the Bayesian framework is operationalized in practice in order to provide both qualitative and quantitative predictions.

In general, the approach, by design, supports a graded notion of argument strength that deals naturally with the uncertainty inherent in most everyday debate. Moreover, it draws on an established inference mechanism—Bayesian inference as a general framework for inductive inference—to capture the inferential relationship between arguments and conclusions. As we will see in more detail, the scheme-based approach has sought to develop these capabilities from scratch, and one key question for consideration will be the extent to which it has met that challenge successfully.

The remainder of this paper is structured in four parts. First, we analyse merits and shortcomings of the scheme-based approach to argument quality by examining three examples: the argument from sign (Sect. 4), the argument from expert opinion (Sect. 5), and the appeal to popular opinion (Sect. 6). For each scheme, we also identify how a Bayesian perspective may help to address the limitations observed and provide a firmer normative footing. In the final Sect. 7, we summarise the general benefits that the Bayesian framework offers for the study of argument quality, highlighting how a scheme-based and a Bayesian perspective may profitably be combined.

4 The argument from sign

As outlined above, a variety of scheme-based frameworks have been developed over the years, and these differ widely in the number of distinct argument types they assume, ranging from 3 (Garssen 1997) to over 60 (Walton et al. 2008). The sample argumentation schemes we have selected for detailed discussion, however, are present in most

typologies. In our presentation of these schemes, we generally take Walton et al.'s (2008) treatment as a starting point given that their book provides the most comprehensive synthesis of both individual argumentation schemes, and typologies of these schemes.

The first of our examples, the argument from sign, is described by Walton et al. (2008, p. 329) as follows:

> A (a finding) is true in this situation.
> B is generally indicated as true when its sign, A, is true.
> B is true in this situation.

The first thing to note is that the intended inferential relationship between *A* and *B* in the argument from sign seems very similar to the general elements of evidence and hypothesis in Bayes' theorem above. One might consider camera-carrying individuals to be a sign of a tourist destination and the example could be rendered to match Walton et al.'s scheme description:

> In this situation, there are a large number of people with digital cameras on the street.
> A city, like London, is generally indicated to be a tourist destination when there are a large number of people with digital cameras on the street.
> London is a tourist destination.

It is clear that this scheme is about the co-occurrence of *A* and *B*. Hastings (1962, p. 56) underlines this in his discussion of the argument from sign when he writes that "there is a reciprocal correlation between the existence of the two situations, and either one may be taken as a sign of the other". The close relationship of this argumentation scheme with probability becomes even clearer when looking at the two critical questions (CQs) that Walton et al. (2008) associate with the argument from sign:

> CQ1: What is the strength of the correlation of the sign with the event signified?
> CQ2: Are there other events that would more reliably account for the sign?

These two critical questions are not unique to Walton et al. but match those formulated by Hastings (1962, p. 63), and are comparable to those of other accounts such as those of Ehninger (1974) and Rieke and Sillars (1984). In asking about correlation, CQ1 is asking about a statistical, probabilistic notion and it asks about the relative strength of that correlation: correlations may be stronger or weaker. Correspondingly, it seems intuitive that the stronger that correlation the more convincingly *A* raises our belief that *B* is true.

That some graded notion of strength is desirable to proponents of scheme-based approaches is clear from Hastings' (1962) comments on the argument from sign: "The argument is one of probability rather than certainty, and so it must always be implicitly stated as 'p implies q with x degree of probability'. The higher this degree of probability, the stronger the argument, and for sign reasoning the correlation must be of high degree: either certainty or high probability, else the sign is not reliable as an indication of the second event" (Hastings 1962, p. 60).

Yet, there seems to be no clear inferential link between CQ1 which asks about strength, and the scheme-based inference itself: how exactly can degrees of strength be related inferentially to the conclusion given that it is derived via subsumption under the generalization that "*B* is generally true when its sign *A* is true"? And how might degrees be attached to the conclusion "*B* is true in this situation"?

Many authors within the scheme-based tradition have been content with informal presentations of schemes; more recently, however, there has been an increasing use of schemes within computational contexts and, with that, an increasing interest in formalization. Walton et al. (2008) highlight the formalization of scheme-based argument as a major research agenda. To the extent that semi-formal or formal treatments have been attempted (see e.g., Walton et al. (2008), Chaps. 11, 12 for examples), authors have typically been adamant that the kind of inference drawn from a scheme such as the argument from sign represents a third form of reasoning that is distinct from "deductive forms of reasoning like *modus ponens* and inductive forms like arguing from a collected set of data to a statistical conclusion drawn from the data" (Walton et al. 2008, p. 10). Instead reasoning with schemes such as the argument from sign is 'tentative', 'presumptive', and 'defeasible' (see e.g., Walton 2001; Walton et al. 2008; Rescher 1976; Fox and Das 2000; Prakken 2005; Pollock 1995) and the probability calculus is typically explicitly disavowed.

One suggestion for formalization is to treat schemes such as the argument from sign as an example of a *defeasible modus ponens*, where the major premise is a statement that is only 'generally true' (e.g., Verheij 2004; Walton 2006; Walton and Reed 2002): in the case of the argument from sign the premise that "*B* is generally indicated as true, when its sign *A* is true". In many cases, these premises may be construed as (Aristotelian) *endoxa*, statements that are generally accepted as true by most or by the wise.

However, treating the inference as one of defeasible modus ponens, where the conclusion is only presumptively or plausibly true and may be overturned by further evidence (in particular from evidence that follows from consideration of the critical questions) does not solve the question of how to integrate the critical questions from an inferential perspective.

One might consider adding them as additional premises that need to be answered in the appropriate way for the inference to go through. Or one may seek to bring together premises and critical questions in a dialectical way (e.g., Walton et al. 2008) as we shall illustrate with the appeal to expert opinion below. Broadly, the scheme then licenses the basic inference, which is retained until challenged by an opponent posing a critical question. At this point the burden of proof shifts back to the proponent wishing to use the scheme, who must now provide further evidence to satisfactorily address the critical question in order for the inference to be re-instated.

This still leaves unclear how one might transmit '*degrees of probability*' to the claim that *B* is true, in the way that Hastings suggests. Assuming the parties come to agree that the correlation is moderate, how is that transmitted to the conclusion? CQ1 and CQ2 both naturally admit answers that are graded, not just binary, all or none. And how do responses to different critical questions interact? The correlation between *A* and *B* may be high (CQ1) but there may also be some other possible explanations (CQ2), and these may vary in how plausible they seem. In fact, it seems likely that

most of the time there will be potential alternative explanations for a given empirical phenomenon. So if the argument is to be of use in practice there must be some way of combining these outcomes: it will hardly ever be the case that the answer to CQ1 will simply be 'very high' and the answer to CQ2 be 'no'.

This question of how scheme and critical questions should interact is not limited to the argument from sign, and Walton et al. (2008) highlight the issue of how to formalize critical questions as one of the most important and most vexing in the context of devising formal, computational systems for argument.

We show next how these difficulties are handled by adopting a probabilistic, Bayesian perspective. The basic building block of such a treatment is the notion of conditional probability: what makes A relevant and informative to the truth of B is that the conditional probability $P(B|A)$ is greater than just $P(B)$. Hence, observing the sign A, raises our degree of belief in B. Where $P(B|A) = P(B)$, A and B are independent, and learning that A is the case will be uninformative for B. 'Correlation' is simply a more specific instance of this general relationship, and wherever A and B are independent, they will also be uncorrelated.[4] We next outline the formal relationship between correlation and the components of Bayes' theorem. This will make clear why it is odd to draw on a notion such as correlation (whether formally or informally) but reject a probabilistic, inductive, treatment of the argument.

For dichotomous variables (such as a claim which can be true or false, and a sign that is present or absent) correlation is captured by the so-called phi-coefficient, which is defined over a simple 2×2 contingency table as shown in Fig. 1 (see e.g., Falk and Well 1997). A contingency table is a way to display the relationship between two categorical variables.

Within the table, each cell entry represents a count of the number of times two states of each variable are observed together (for example, in Fig. 1, Panel (a), b represents the number of times H, a hypothesis, has been observed to be false while D, some data, are true). What matters for the presence or absence of a statistical correlation between two variables is the relationships between the cells, not their absolute values. So cell entries can simply be converted to relative proportions of the total number of observations. Converting the frequencies to probabilities in this way brings out the systematic connection between correlation and the components of Bayesian belief revision.

The phi-coefficient ranges from -1 to $+1$. A value of 0 indicates no relationship between the two variables, a value of 1 indicates they are perfectly correlated (i.e., A occurs if and only if B occurs), and values between 0 and 1 indicate a positive, but less strong correlation. A negative correlation between sign A and claim of interest B means that when A is present B is less likely to be the case than when A is absent (and vice versa).

[4] For continuous variables correlation is defined as $rAB = \frac{E(A,B)-E(A)E(B)}{\sqrt{E(A^2)-(E(A))^2}\sqrt{E(B^2)-(E(B))^2}}$ where $E(A)$ is the expected value of A, Specifically, independence implies zero correlation, but the converse is not necessarily true. Variables can be systematically related, and hence non-independent, in ways not captured by (linear) correlation (e.g., x and y in $y = sin(x)$). This also suggests that independence, as the more general notion, is preferable to correlation as the basis for the argument from sign.

Binary Correlation & Bayesian Belief Revision

(a)

	H	
D	True	False
True	a	b
False	c	d

(b)

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

When cell entries are normalized by total number of observations:

(c)

$$a = P(a) = P(D, H)$$
$$b = P(b) = P(D, \neg H)$$
$$c = P(c) = P(\neg D, H)$$
$$d = P(d) = P(\neg D, \neg H)$$

Fig. 1 *Panel a* shows the simple 2 × 2 contingency table over which the phi-coefficient (*Panel b*) is defined. The counts in each cell can be normalized by dividing each entry by the total number across all cells, converting each cell entry into a proportion. This allows systematic connections to probability to be drawn out (*Panel c*). Note that ',' represents logical 'and' and '¬' represents negation

If we replace A with D, and B with H, then the following relationships between cell entries and the posterior degree of belief in H, $P(H|D)$, hold by virtue of the definition of the phi-coefficient (Fig. 1 Panel b) and the definition of conditional probability:[5]

if $b = 0, c = 0$ and $ad \neq 0$ then $phi = 1$ and $P(H|D) = 1$.

In other words, a phi coefficient of 1, indicating perfect correlation, corresponds to a case where observing the data (sign) means H is true for certain, regardless of our prior degree of belief in H (with $P(D|H) = 1$ and $P(\neg D|\neg H) = 1$ and thus $P(D|\neg H) = 0$) making the data maximally diagnostic. By contrast, if $a = 0, d = 0$ and $bc \neq 0$, then the likelihood ratio is less than 1, so that observing D decreases our belief in H, $phi = -1$, and $P(H|D) = 0$, regardless of our prior degree of belief in H.

And, finally, when $ad = bc$, $phi = 0$, the likelihood ratio $LHR = 1$ and $P(H|D) = P(H)$, that is, when the variables are uncorrelated, observing D will not alter our degree of belief in H.

In between these cases, different degrees of correlation will map onto different degrees of belief in H given D, $P(H|D)$.

In short, there is a straightforward mapping between the statistical notion of correlation and the components of Bayes' theorem, in particular posterior degrees of belief, that is, our degree of belief in the claim in light of the argument provided. By virtue of that mapping, adopting a Bayesian perspective on the argument from sign also

[5] $P(A|B) = P(A, B)/P(B)$

supplies a direct inferential link between the strength of the correlation highlighted in CQ1 and our degree of belief in the claim.

Secondly, adopting a Bayesian perspective also makes clear that what really matters for an argument from sign is the *correlation*, *not* the *generalization* contained in the major premise, and that conceptualizing the relationship between claim and sign as one of defeasible inference is, in fact, awkward and unsatisfactory. As a reminder, the major premise in the defeasible modus ponens version of the argument is that "It is generally the case that B is true when A is true" (or: "H is true when D is true"). However, this is *not* the relationship that is critical and this can be seen from the fact that this premise may be satisfied trivially just because H is generally the case. If H is true most of the time, then it may well also generally be true when D is true. However, in this case D need carry no further information about H, so that H and D are entirely independent, that is $P(H|D) = P(H)$. From a Bayesian perspective, D then fails to provide a reason for believing H in this case, but the defeasible modus ponens inference nevertheless goes through.

One might counter that this case is harmless, because it will still be the case that claim H is likely to be true, which, as the conclusion of the inference, is the claim of interest. However, it does not make for a very intuitive notion of argument quality if inferences from *irrelevant* reasons are accepted as 'good arguments', and this is exactly what ensues:

(1) The sun will rise tomorrow, because cheese contains milk.

is a perfectly good argument from sign from the perspective of defeasible modus ponens, even though the ingredients of cheese are entirely informationally irrelevant to the behaviour of the sun. The problem is even more apparent when one considers that the inference is as good when the irrelevant premise is true as when it is false. In these cases where the major premise holds simply because the claim is already 'generally the case', nothing about the sign itself matters. As a consequence, Example (2) is as compelling a case for the sun's rising as is (1)!

(2) The sun will rise tomorrow, because the moon is made of cheese.

Intuitively, these are not good arguments because argumentation is about *belief change*. An argument should be viewed as good when it can further increase our degree of belief in a claim even when we are already quite convinced that the claim is true. From a Bayesian perspective, this will be reflected in the fact that the posterior, or conditional probability, that is our degree of belief in the claim given the argument, $P(H|D)$, will be higher than our belief without it. The extent to which the argument changes belief in H will be determined by the likelihood ratio. By contrast, arguments that *cannot* involve a change in rational belief are poor, however plausible the claim contained in the conclusion may be. This is true even where conclusions follow by logical necessity as is apparent in the case of circular arguments ('God exists, because God exists') which long troubled philosophers interested in argument quality. An analytic focus on probabilistic dependence as a basis for belief change allows one to distinguish circular arguments that are acceptable from ones that are poor (see, Hahn

and Oaksford 2007a; Atkinson and Peijnenburg 2010; Hahn 2011). The same is true here in the context of the argument from sign.[6]

Probabilistic dependence provides a notion of informational relevance (see also, Pearl 1988). The fact that premises are 'true' or 'generally true', in and of itself, does not. This limits the extent to which logical inference or defeasible inference modelled on logical inference can match fully basic intuitions about what constitutes 'a good argument'.

In summary, more is required for an argument from sign to provide a good reason for believing a claim than that the defeasible scheme set out in the argument from sign is fulfilled. That 'more'— informational relevance—is encapsulated in the critical question and is an integral part of the Bayesian framework. Adopting a Bayesian perspective thus provides both integration and refinement for this basic scheme.

Finally, what of CQ2 and the possibility of alternative explanations for the presence of the sign? This too can be captured naturally within a Bayesian framework. Moreover, it can be captured in a way that provides inferential integration with the other aspects of the scheme. Bayesian inference can involve multiple interrelated factors and naturally supports inference in both a backward and a forward direction: One can reason equally from hypothesis to data, $P(D|H)$ and from data to hypothesis, $P(H|D)$. Consequently, the phenomenon of "explaining away" is a natural consequence of Bayesian inference involving multiple variables (see, e.g., Pearl 1988): seeing that the ground is wet (=sign) raises our degree of belief that it has rained, but observing further that the sprinkler is on will lower our belief in rain.

The dependencies in this well-known example are represented in the simple Bayesian Belief Network of Fig. 2.[7] Assume, for illustration that, the prior probabilities are: $P(\text{rain}) = 0.1$, and $P(\text{sprinkler}) = 0.1$. Assume further the sample conditional probabilities shown in Table 2 for the relationships between the three nodes in the model of Fig. 2.

Figure 2 then shows the effects of different pieces of evidence, including a potential alternative explanation as might be received in response to the critical questions. As seen in panel (a), coming to know that the lawn is wet increases the probability of rain (hypothesis) from 0.1 (the prior) to 0.49. Panel (b) shows the situation on observing

[6] To put this more formally, one might think of a generalization such as "If A, then generally B" as saying that $P(B|A)$ is high. On observing A, the probability we should now assign to B will be $P(B|A)$, exactly as (defeasible) modus ponens suggests. However, whether A provides a *reason* for believing B, depends on whether $P(B|A)$ is greater than $P(B)$ in the first place, and that depends on the likelihood ratio being greater than 1, i.e., that $P(B|A) > P(B|\neg A)$. On modus ponens and other conditional inferences from a probabilistic perspective see e.g., Oaksford and Chater (1994), Evans and Over (2004) and, specifically in an argumentation context Hahn and Oaksford (2012).

[7] Bayesian Belief Networks simplify multi-variable computations by taking into account dependence and independence relations within a graphical representation (for an introduction see e.g., Pearl 1988 or Korb and Nicholson 2003). The nodes in a network such as that in Fig. 2 represent random variables. The directed arrows (links) between them signify (assumed) direct causal influences and the strengths of these influences are quantified by conditional probabilities. Each variable is assigned a link matrix that represents estimates of the conditional probabilities of the events associated with that variable given any value combination of the parent variables' states. These matrices together provide a joint distribution function: a complete and consistent global model, on the basis of which all probabilistic queries can be answered.

Table 2 Conditional probability table for the sign lawn and the two hypotheses sprinkler and rain

	No sprinkler		Sprinkler	
	No rain	Rain	No rain	Rain
Lawn wet	0.02	0.9	0.85	0.95
Lawn dry	0.98	0.1	0.15	0.05

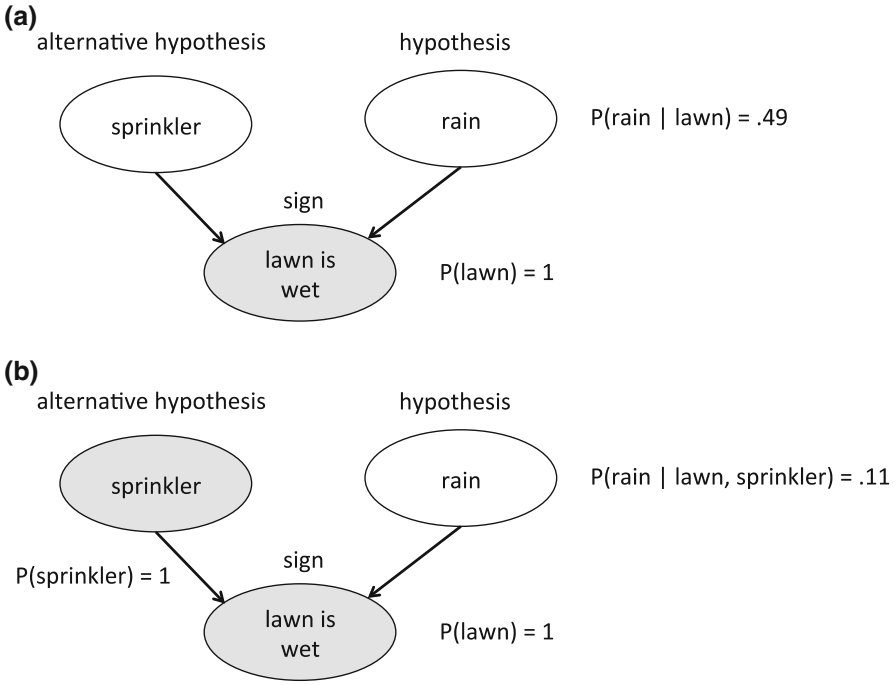

Fig. 2 An alternative explanation (sprinkler) in a sign argument. In light of an alternative hypothesis (sprinkler) for why the lawn is wet (sign) the probability of the hypothesis (rain) is lower than without the alternative explanation. *Panel a* shows the relevant probabilities without evidence of the alternative hypothesis (i.e., without knowing the present state of the sprinkler), *Panel b* shows the probabilities with it (i.e., on observing that the sprinkler is on).

also that the sprinkler is on: the posterior probability of the hypothesis 'rain' now decreases: P (rain|lawn, sprinkler) $= 0.11$.

Note also that the extent to which the posterior probability decreases depends on the probabilities involved: it is because there is a dependence between sprinklers being on and lawns being wet that the 'alternative hypothesis' gives rise to 'explaining away'. Where the probabilistic relationship between alternative hypothesis and sign is weaker, the effect will be weaker (or even non-existent).

Hence, the Bayesian framework captures naturally the fact that not any old alternative hypothesis will do, and this can be used to further illustrate the point about relevance just made above. Shown in Table 3 below is an alternative set of probabilities that could (hypothetically) govern the relationship between lawns, rain and sprinklers. In this possible world, the lawn is generally wet, so is also generally wet

Table 3 Alternative conditional probability table for the sign 'lawn' and the two hypotheses 'sprinkler' and 'rain'

	No sprinkler		Sprinkler	
	No rain	Rain	No rain	Rain
Lawn wet	0.86	0.96	0.85	0.95
Lawn dry	0.14	0.04	0.15	0.05

when the sprinkler is on. However, it is no more likely to be wet when the sprinkler is on than when it is off. As a result, learning that this alternative type of sprinkler is on leaves our beliefs virtually unchanged (P(rain|lawn, sprinkler) = 0.1105, whereas P(rain|lawn) = 0.1104).

Finally, the examples also serve as a reminder that Bayesian Belief Networks give the Bayesian framework the possibility not only of graphical representations, but of graphical representations that simplify computations in circumstances involving multiple elements, an issue we will return to below. By contrast, most of the graphical representations that have been applied to argumentation schemes within the scheme-based literature lack computational inference abilities, leaving this as a major issue for further research (see e.g., Walton et al. 2008, Chap. 12).

4.1 Variants of the argument from sign

It is a characteristic of the scheme-based approach that it presently contains a proliferation of schemes. However, the approach lacks clear guidelines for when and why a separate scheme should be assumed. In the case of the argument from sign, a number of closely related schemes illustrate these difficulties. These also bring out further the close relationship between the argument from sign and Bayesian inference.

A prime example is the 'argument from evidence to a hypothesis' which some have presented as an entirely separate argument scheme (e.g., Walton et al. 2008, p. 331; Walton 1996, pp. 67–70) comprising two subtypes, the argument from verification:

> *Major premise* If A (a hypothesis) is true, then B (a proposition reporting an event) will be observed to be true.
> *Minor premise* B has been observed to be true, in a given instance.
> *Conclusion* Therefore, A is true.

and the argument from falsification:

> *Major premise* If A (a hypothesis) is true, then B (a proposition reporting an event) will be observed to be true.
> *Minor premise* B has been observed to be false, in a given instance.
> *Conclusion* Therefore A is false.

The critical questions for both versions are given as:

> CQ1: Is it the case that if A is true, then B is true?
> CQ2: Has B been observed to be true?
> CQ3: Could there be some reason why B is true, other than its being because of A being true?

The argument from verification differs from the argument from sign only in fairly subtle ways. In particular, the term 'sign' is replaced by a generic event B, and the major premise is an unqualified conditional rather than a defeasible generalisation (i.e., the observable event B is true if the hypothesis is true, as opposed to merely 'generally' indicated).[8] Neither of these should be seen as compelling grounds for positing a separate scheme.

Concerning the presence or absence of the term 'sign' itself, the term 'sign' is nowhere defined in Walton et al. (2008). Moreover, the kinds of data used to test hypotheses (scientific or otherwise) would seem to typically be 'signs'—observable events that are indicative of the underlying phenomenon in question. This aspect thus would not seem to merit the distinction. Furthermore, much of our hypothesis testing revolves around evidence that does not follow from the hypothesis by necessity, and thus involves uncertainty. This is reflected in the extent to which the Bayesian, probabilistic framework has become central to epistemology and philosophy of science (e.g., Earman 1992; Howson and Urbach 1993; Bovens and Hartmann 2003). From that perspective, data that follow from a hypothesis by necessity simply constitute a limiting case where $P(D|H) = 1$, a case that is otherwise unremarkable.

Moreover, it is insufficient that the 'data' (event B) be implied by the hypothesis. Treated as a logical argument, the inference set out in the 'argument from verification' is a fallacy of classical logic: namely, the so-called affirmation of the consequent. The truth of a hypothesis (H) follows from the observation of data D only if D implies the hypothesis, not the other way around.

The problem is once again that of informational relevance, which may dissociate from logical validity as just discussed above.[9] D may always be true when H is true, but if it is also always true when H is false, then its presence is entirely non-diagnostic and will not affect our beliefs about H. In Bayesian terms, the posterior, $P(H|D)$, is affected not just by the likelihood (i.e. the sensitivity of the evidential test $P(D|H)$), but also by the false positive rate (i.e., $P(D|\neg H)$ see above), as captured by the likelihood ratio. Critical question CQ3 implicitly acknowledges this, but again fails to provide the overall integration into the inference that Bayes' theorem provides. Bayes' theorem links what is the focus of CQ3 with the overall inference from D to H in such a way that it allows degrees of uncertainty about CQ3 to be reflected in degrees of uncertainty about H. Both the argument from sign and 'the argument from evidence to a hypothesis' are more appropriately formalised as Bayesian inference, and within that context, they do not seem relevantly distinct.

Finally, it seems worth mentioning yet two further schemes that, once again, seem like close variants of the argument from sign, though Walton et al. (2008) list them completely separately (as schemes number 59 and 60, respectively, with the argument from sign as scheme number 30), namely: the argument from perception (Walton et al. 2008, p. 345), and the argument from memory (p. 346):

[8] Within the scheme-based tradition Hastings (1962, p. 143) also considers both schemes to be related to "causal relations which are used as generalizations to justify the conclusion on the basis of the premises".

[9] Further examples of the dissociation between logical validity and inductive strength to those given thus far are the so-called paradoxes of material implication, see Oaksford and Hahn (2007).

Person P has a φ image (an image of a perceptible property).
To have a φ image (an image of a perceptible property) is a prima facie reason to believe that the circumstance exemplify φ.
It is reasonable to believe that φ is the case.

Person P recalls ϕ.
Recalling ϕ is a prima facie reason to believe ϕ.
It is reasonable to believe ϕ.

These schemes were first proposed by Pollock (1995), who felt a set of novel schemes for common sense reasoning was needed in part because of his vociforous rejection of the probability calculus as an appropriate tool for capturing everyday reasoning.[10] It is ironic in light of this that one of the most influential and successful applications of Bayesian inference within psychology is to perception: specifically, it has become increasingly popular among perception researchers to characterize perception itself as Bayesian inference (see e.g., Knill and Richards 1996).

These further examples do not exhaust the range of schemes that seem closely related to the argument from sign. Whately (1846) and Hastings (1962, p. 126), for example, when discussing the argument from authority—the scheme we examine next—note that this argument scheme is "a form of sign reasoning: the fact that *X* asserts the conclusion is taken to be a sign that the conclusion is true". Inch and Warnick (2009) also consider the argument from authority to be a variety of, what they call, the coexistential argument. Furthermore, they even relate this argument to two other arguments: "As in analogy and generalization arguments, the inference moves from what is known (the sign) to what is unknown or less known (the condition or essence)" (Inch and Warnick 2009, p. 167). This overlap underlines the generality of argument from sign. Given, however, that 'signs', at the end of the day are just probabilistically relevant indicators, these examples serve at the same time to underline the extent to which Bayes' theorem provides a generic form of inductive inference of wide applicability to argumentation.

In summary, inspection of the argument from sign and other, closely related schemes illustrates that the scheme-based approach is conceptually less far from a Bayesian, probabilistic account than has been claimed (see Nussbaum 2011; Walton 2004; Walton et al. 2008, pp. 186–188; Pollock 1995), and that, in fact, a Bayesian perspective provides immediate solutions to some of the problems the scheme-based approach presently faces. Namely, it speaks to the question of how many schemes should be distinguished, how critical questions may be inferentially integrated with the basic inference of the scheme and with one another, and how the degree to which a critical question is met may be passed on to the conclusion. These issues are pursued further in our remaining two example schemes.

[10] These seem largely based on consideration of characteristics of probability in the context of logical inference, rather than, as advocated here, Bayesian conditionalization. For example, Pollock's arguments about how multiple, independent, premises lead rapidly to improbable conclusions assume that the relationship between premises and conclusions is conceived of as a logical inference from a conjunction, not as a conditional probability. In general, believing more things does not inherently imply greater risk of error, see e.g., Bovens and Olsson (2002).

5 The argument from expert opinion

The next argument scheme that we discuss here is one of the most common examples within the scheme-based approach to argumentation, namely the argument from expert opinion, or appeal to authority (Walton 1997; Walton et al. 2008, p. 14):

> Source E is an expert in subject domain S containing proposition A.
> E asserts that proposition A (in domain S) is true (false).
> A may plausibly be taken to be true (false).

Unlike the argument from sign just discussed, Walton et al.'s characterization of the argument from expert opinion explicitly, through its wording, draws attention to the plausible or defeasible nature of the conclusion. Experts, like other witnesses, are generally neither completely reliable nor completely unreliable. Moreover, the argument form is frequently presented as a prime example of the perceived need for a 'third way', that is, defeasible (or abductive, presumptive) arguments as a third category next to deductive and inductive arguments (Walton 1996; Walton et al. 2008). Again, why such a third way should be necessary remains unclear. Walton et al. describe a defeasible argument as "one in which the conclusion can be accepted tentatively in relation to the evidence known so far in a case, but may need to be retracted as new evidence comes in" (Walton et al. 2008, p. 2). But this is readily achieved via a probabilistic Bayesian reconstruction as an inductive argument. And, again, a Bayesian treatment provides a ready solution to the problem of how critical questions may be incorporated inferentially, as we will show below.

For the argument from expert opinion, Walton (1997, p. 223) lists six critical questions (similar questions can be found in Hastings 1962):

> How credible is E as an expert source?
> Is E an expert in the field that A is in?
> What did E assert that implies A?
> Is E personally reliable as a source?
> Is A consistent with what other experts assert?
> Is E's assertion based on evidence?

However, each of these questions has extensive further sub-questions that may need to be considered in evaluating the argument. Walton et al. (2008, pp. 92–93) discuss in detail how the appeal to expert opinion may be formalised and in that context provide a refined list, which is shown in Table 4 below:

These sub-questions are striking in their level of detail: clearly they will not be relevant in all cases of expert testimony. In other words, they are specified in anticipation of issues that may or may not arise. Yet, as such, they necessarily remain incomplete. It seems easy to extend the list of critical questions with questions that could be relevant such as 'Has E not unknowingly been given drugs before his assertion?', and the issue here is simply how far one wants to go in anticipating possible limitations to the expert's testimony (thus presumably drawing some line between a 'regular' and an 'extraordinary' course of affairs).

The underlying problem, here, is, of course, familiar: the fundamental non-monotonicity of everyday reasoning. Endowing computational systems with the

Table 4 Sub questions to the six basic CQ's for expert opinion

CQ1: Expertise question. How credible is E as an expert source?	
1.1.	What is E's name, job or official capacity, location and employer?
1.2.	What degrees, professional qualifications, or certification by licensing agencies does E hold?
1.3.	Can testimony of peer experts in the same field be given to support E's competence?
1.4.	What is E's record of experience, or other indications of practiced skill in S?
1.5.	What is E's record of peer-reviewed publications or contributions to knowledge in S?
CQ2: Field question. Is E an expert in the field that A is in?	
2.1.	Is the field of expertise cited in the appeal a genuine area of knowledge, or an area of technical skill that supports a claim to knowledge?
2.2.	If E is an expert in a field closely related to the field cited in the appeal, how close is the relationship between the expertise in the two fields?
2.3.	Is the issue one where expert knowledge in *any* field is directly relevant to deciding the issue?
2.4.	Is the field of expertise cited an area in which there are changes in techniques or rapid developments in new knowledge, and, if so, is the expert up to date in these developments?
CQ3: Opinion question. What did E assert that implies A?	
3.1.	Was E quoted as asserting A? Was a reference to the source of the quote given, and can it be verified that E actually said A?
3.2.	If E did not say A exactly, then what did E assert, and how was A inferred?
3.3.	If the inference to A was based on more than one premise, could one premise have come from E and the other from a different expert? If so, is there evidence of disagreement between the two experts (separately) asserted?
3.4.	Is what E asserted clear? If not, was the process of interpretation of what E said by the respondent who used E's opinion justified? Are other interpretations plausible? Could important qualifications have been left out?
CQ4: Trustworthiness question. Is E personally reliable as a source?	
4.1.	Is E biased?
4.2.	Is E honest?
4.3.	Is E conscientious?
CQ5: Consistency question. Is A consistent with what other experts assert?	
5.1.	Does A have general acceptance in S?
5.2	If not, can E explain why not, and give reasons why there is good evidence for A?
CQ6: Backup evidence question. Is E's assertion based on evidence?	
6.1.	What is the internal evidence the expert herself used to arrive at this opinion as her conclusion?
6.2.	If there is external evidence—for example, physical evidence reported independently of the expert—can the expert deal with this adequately?
6.3.	Can it be shown that the opinion given is not one that is scientifically unverifiable?

flexibility that human reasoners show in the face of potentially limitless exceptions is arguably the single biggest problem faced by computational systems (see e.g., Dennett 1984).

Research on argumentation within AI has taken different approaches to dealing with the non-monotonicity that characterises everyday argument (see e.g., Prakken and Vreeswijk 2002, for a review) and that is manifest in approaches to specific argument

forms such as the appeal to expert opinion. The dominant approach seeks to capture defeasible inference by defining appropriate relations over sets of statements, whereby a conclusion follows via some non-classical inference relationship if an argument may be constructed that survives all possible 'counterattacks' that could be constructed from that set. Another approach seeks to take into account the dialogical nature of argument by appealing to notions such as the 'burden of proof' (on both, including references to examples from the literature see Gordon et al. 2007). In either case, the third way tradition treats counter-arguments 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', see also Bex et al. 2003).

Gordon et al.'s (2007) exposition of their model of argumentation, named Carneades, specifically uses the appeal to expert opinion as an illustrative example. This work thus provides a ready illustration of how formalization of the scheme and the associated critical questions might be approached from a 'defeasible' reasoning, or 'third way', perspective. Specifically, the Carneades model ties argument evaluation to a graph representation of the dialectical structure of an argumentative exchange. In the context of such an exchange, statements are advanced and supported by arguments, which in turn may be attacked by counter-arguments and so on. Argument graphs keep track of the arguments that have been made and the relationships between them. Given a formal procedure for establishing such a graph, a formal method of argument evaluation may then be defined relative to that graph. Specifically, a statement is '*acceptable*' given the arguments for it, if the argument graph is a 'proof' of the statement (Gordon et al. 2007). It will be acceptable if a decision to accept the statement as true can be justified or explained given the arguments which have been put forward in the dialogue, where the definition of the acceptability of statements is recursive and depends on the relevant proof standard. Whether or not a statement's proof standard is satisfied depends on the defensibility of the arguments *pro* and *con* this statement. The defensibility of an argument, in turn, depends on whether or not its premises hold, which can depend on whether or not the premise's statement is acceptable.

In this context, Gordon et al. (2007, p. 885) distinguish three different types of proof standards (which they do not presume to be exhaustive):

1. *Scintilla of evidence (SE)* A statement meets this standard iff it is supported by at least one defensible pro argument.
2. *Best argument (BA)* A statement meets this standard iff it is supported by some defensible pro argument with priority over all defensible con arguments.
3. *Dialectical validity (DV)* A statement meets this standard iff it is supported by at least one defensible pro argument and none of its con arguments are defensible.

These different proof standards allow the model to formalise critical questions in a procedurally flexible way. Specifically, different kinds of premises can be distinguished: premises that must always be supported with further grounds ('ordinary premises'); premises that can be assumed until they have been questioned ('assumptions'); and 'exceptions', which are premises that do not simply hold because there is an absence of specific evidence to the contrary. Exceptions (semantically a particular type of undercutter with a single premise) need to be proven (i.e., a decision maker has to be convinced that the exception is true) or, at least, made plausible (i.e., a decision maker

would have reason to think it is true). Whether or not a critical question is treated as an assumption or an exception is thus highly consequential to the outcome of the inference (see also, Verheij 2003b).

When a side in the dialogue provides an argument instantiating an argumentation scheme, the burden is on that side to prove its ordinary premises and (once challenged) its assumptions, after which the burden rests with the opponent to defeat the argument by rebutting it or pointing out exceptional circumstances. Thus, merely *asking* a critical question does not immediately shift the burden of proof back to the proponent, so an argument may nevertheless generate the default if no further evidence is provided (on this issue see also, Walton and Gordon 2005).

In other words, each critical question must be classified as an assumption, or an exception. Specifically, Gordon et al. (2007) offer the following classification for the six basic critical questions associated with the appeal to expert opinion:

Premise E is an expert in the subject domain S containing the proposition A.
Premise E asserts A.
Assumption E is a credible expert.
Exception E is not reliable.
Exception A is not consistent with the testimony of other experts.
Assumption A is based on evidence.
Conclusion A.

Thus for the expertise and backup questions the burden of proof requires the proponent's side to provide an answer, once the question is raised. However, the proponent bears a burden of proof for the trustworthiness and consistency questions only if backup evidence is provided in the challenge, on the rationale that it is easier to produce evidence of instances of unreliable or inconsistent behavior than it is to prove that no instances of such behavior have occurred. These questions are thus modelled as exceptions on the basis of "the general principle of allocating the burden of production to the party with better access to the evidence" (Gordon et al. 2007, p. 887).

In principle, this then allows tracking through a specific dialectical exchange involving an appeal to expert opinion and coming to a final valuation of whether the conclusion *A* is to be accepted or not. Given further, that this is arguably the most detailed formal, inferential treatment of a scheme within the defeasible reasoning, 'third way' tradition, it is worth bringing out clearly its present limitations.

For one, the conclusion is a binary one of 'acceptable' or not. Though all critical questions clearly admit of degrees, these are not passed through to the final valuation. Moreover, any valuation within the model is only possible on specification of a *priority ordering* over arguments (see the definition of 'Best Argument' above) which decides on the relative strength in terms of conflict. This is, at present, entirely separate and not derived from other information in the model. Furthermore, in principle, each subquestion too would need to be classified as an ordinary premise, assumption or exception to the scheme (see Walton et al. 2008, p. 382), and the subquestions themselves raise the possibility that a party might "go on asking critical questions indefinitely" which in turn raises the question of whether it is ever the case that an argument matching a scheme can be "finally evaluated so that the process of questioning ends, and we can say definitively whether the argument is strong or weak or meets whatever cri-

teria of success we have adopted in a particular case?" Walton et al. (2008, p. 380). The Carneades model deals with this issue to the extent that it makes proof in the system contingent on questions *actually raised*. But while this may seem reasonable in a dialectical exchange, it is unclear what this should mean where a scheme such as the appeal to expert opinion is used in the reasoning of a single reasoner. Here, the distinction between assumption and exception collapses, and the reasoner would seem to bear the burden of proof either way. However, one may ask why, in general epistemic contexts (as opposed to contexts such as legal fact finding which are subject to pressures and goals other than simply 'finding truth'), there should be any difference between the single reasoner case and the dialogical exchange. If the approach is to be a model of the appeal to expert opinion *in general*, then shouldn't those questions that are truth relevant always matter? In other words, if the critical questions genuinely capture strength related criteria, should they not always be asked if the inference is to be normatively justified? These issues highlight some of the concerns surrounding whether or not the legal notion of 'burden of proof' is really the appropriate tool for epistemic contexts concerning truth (for wider discussion of this point and a more general critique of the notion of burden of proof in argumentation see Hahn and Oaksford 2007b).[11]

We can contrast Gordon et al.'s formalization with a formal Bayesian treatment of the scheme. Hahn et al. (2013) demonstrate how the appeal to expert opinion may be captured in a Bayesian Belief Network that incorporates all six critical questions (see Fig. 3).

Central in the model is the relationship between the hypothesis H, the *Evidence*, and the source's report about the hypothesis, *Hrep*. Hahn et al. (2013) argue that CQ1 (How credible is E as an expert source?) and CQ2 (Is E an expert in the field that A is in?) can typically be combined as one factor *Expertise*, since the source can only be credible as an expert if she is indeed an expert in that field. In Fig. 3, the source directly asserts H in *Hrep* (this is CQ3: What did E assert that implies H?). The source's reliability (CQ4: Is E personally reliable as a source?) is formalized as Trustworthiness in the model. The critical question about what other experts have said (CQ5: Is A consistent with what other experts assert?) is modelled as further reports (*Hrep*) from other sources (here $S2$ and $S3$). Finally, CQ6 (Is $E's$ assertion based on evidence?) is captured by the *Evidence* node.

This basic network can be expanded further to model the sub-questions (see Harris et al. 2015). For example, one may link to the *Expertise* node an additional node for 'qualifications'. Expertise enables one to gain qualifications, and qualifications are thus evidence of expertise. Expertise makes possession of qualifications more likely, and therefore the existence of qualifications can be seen as evidence of expertise. This captures the subquestion of "what degrees, professional qualifications, or certification by licensing agencies does E [the expert] hold?".

[11] Hahn and Oaksford (2007b) argue, among other things, that the notion of burden of proof is inherently tied to action, stemming in law from the need to make a decision. Where a decision is required, the utilities associated with various courses of action provide 'burdens of proof'. Where a decision is not immediately required, the notion is forced, and there are no normatively compelling reasons for determining either levels of proof required, or who should carry them.

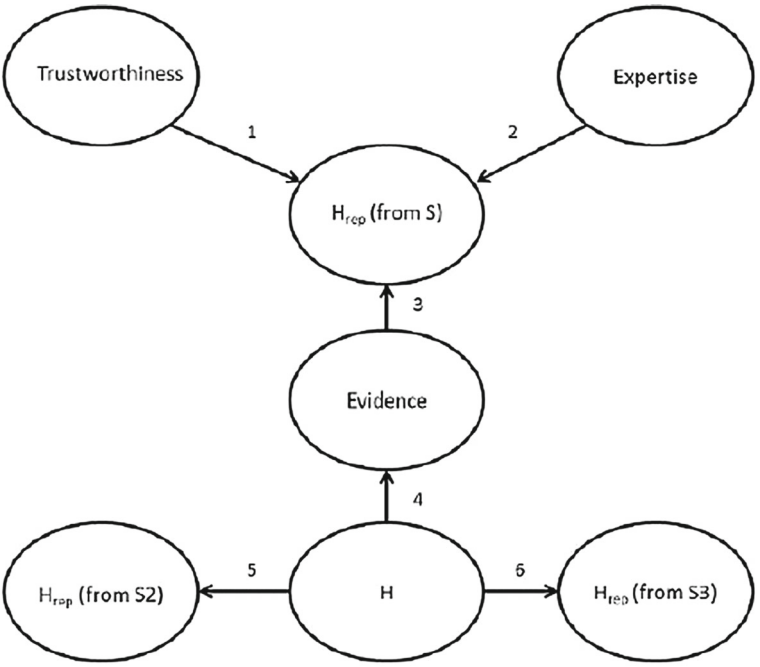

Fig. 3 A Bayesian belief network for the argument from expert opinion (Hahn et al. 2013)

In general, this makes clear that this (and other) sub-questions serve to establish the status of the parent node ('Expertise'). They thus allow one to revise one's beliefs about expertise, but they are not essential for the inference: if expertise is established, learning that qualifications exist is of no further relevance (i.e., the issue is 'screened off').

As in the examples given earlier (Fig. 2 above) the network is instantiated by specifying the conditional probabilities deemed appropriate to the particular case. Once these are given, inference takes the form of conditionalising on whatever evidence (argument) is received (i.e., the relevant variable takes the value P (variable) $= 1$, and the probabilities of all other nodes in the network are revised via Bayes' rule). There may or may not, for example, be an evidence report from a further expert ($S2$ or $S3$). If there is not, these nodes remain inferentially inert; if there is, however, then those reports are factored into the overall evaluation, which is quantitative and yields a degree of belief in H that reflects all information that has been accumulated, with each piece of information reckoned with and combined according to its relative strength. Whereas on Carneades' burden of proof approach arguments and subarguments become blocked off in entirety (on cases where this becomes problematic see also Gordon et al. 2007), the Bayesian approach takes factors *pro* and *con*, wherever they emerge in the overall argument, and weighs them quantitatively in the final, overall evaluation.

This is not only more flexible, but it seems normatively appropriate: many pieces of weak evidence together may outweigh a stronger consideration that stands against

them.[12] Moreover, multiple arguments may interact to produce sub-additive or super-additive effects. Hahn et al. (2009), for instance, show—both from a Bayesian normative perspective and with empirical, experimental results from psychological studies—that source reliability and strength of evidence interact. Reliable sources benefit from stronger as opposed to weaker evidence, but this effect is very much smaller for less reliable sources.

No 'third way' approach to date allows these kinds of interactions, whether the impact of critical questions is modelled by shifting the burden of proof, or via some notion of 'defeat'. The fundamental problem for such approaches to the non-monotonicity embodied by critical questions is the fact that argument quality itself matters in determining whether or not a claim is defeated or a burden of proof shifted: for example, an entirely irrelevant proposition, advanced as a counter-claim, should be sufficient neither to defeat a claim nor to shift 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 and Oaksford 2007a). The idea of a 'priority ordering' across arguments only partially solves the problem, precisely because it does not allow one to capture interactions between arguments other than in a very crude way. Moreover, quantitative schemes for evaluation that have been proposed which seek to take into account interactions between arguments (such as Theophrastus' rule, Rescher 1976; Pollock's 'weakest link principle' 1995; or Walton's (2008) MAXMIN rule) can be shown to yield counterintuitive or even paradoxical results (see Hahn et al. 2013 for examples). It is not enough simply to assign numeric weights to arguments (as proposed for Carneades for example in Walton and Gordon 2014), because not all assignments of weights across a set of arguments can be considered rational, and it is exactly constraints on what should qualify as rational assignments that the probability calculus provides. It is by virtue of this that the probability calculus allows one not only to capture the evidential impact of individual pieces of evidence but also their interactions.

These limitations of 'third way' approaches stem in part from the fact that many have been dubious about assigning numbers to arguments in the first place, querying where such 'numbers might come from'. The fact that the Bayesian framework has, by now, seen considerable application in real world contexts would seem to address that concern. To name just some particularly pertinent examples, Kadane and Schum (1996) provide a (heroic) reconstruction of the entire body of evidence in the famous Sacco and Vanzetti trial in Bayesian terms. Moreover, the Bayesian Belief Networks they construct allow them also to evaluate what would change under alternative assignments of probabilities, and thus to determine how sensitive conclusions are to the exact numbers assumed ('sensitivity analysis'). Second, the fact that Bayesian models may provide good fits of people's judgments of argument strength (e.g., Harris et al. 2012) suggests that people are sensitive to the relevant magnitudes whether they typically have introspective access to explicit numbers or not. We thus do not consider difficulties

[12] Carneades can handle such accrual of evidence for cumulative arguments if an argument for every member of the powerset of the pieces of evidence is included in the argument graph, see also Gordon and Walton (2009).

in obtaining probabilities sufficient to challenge the use of the Bayesian framework for normative concerns.

The Bayesian framework 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,[13] 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. Moreover, it provides a natural framework for dealing with the question of where 'priority orderings' come from in the form of the likelihood ratio: if a state of affairs is more likely to arise if the claim in question is true than if it is false, then that state of affairs supports the claim, and the greater that ratio the more it does so, both inherently and relative to other factors that may be brought into play. In other words, the Bayesian framework tells us what to consider in evaluating reasons and hence where to look. This is true not just of the relative strength between arguments (i.e., establishing the 'priority ordering') but, by the same token, true for determining what is relevant and hence a potential critical question in the first place. This latter issue of 'where the critical questions come from' and how we know, normatively, that we have the right ones, is pursued further in our final example scheme, the appeal to popular opinion.

Before moving on to that, however, we note that the Bayesian framework is helpful also with respect to the possibility of ever further questions, because, although it does not tackle fully the problem of never-ending exceptions in that such exceptions will never be fully enumerated (and hence explicitly modelled), it allows one to nevertheless reason in the face of such exceptions because probabilities *summarise* uncertainty, and thus can also summarise expectations of the possibility of relevant exceptions (see Pearl 1988, Chap. 1, for discussion of this point). Using the Bayesian framework to formalise schemes and critical questions thus allows one to address key challenges for formalization that the literature on schemes has previously identified.

As should now be clear, Bayesian belief networks are more than schematic formalizations of relationships between evidence (argument) and hypotheses (conclusions), as are widely used in the analysis of (complex) argumentation (e.g., Snoeck Henkemans 2000). Rather, like the argument graphs of the Carneades model, they are inferential tools. They capture, via the conditional probabilities, whether evidence/arguments decrease or increase support for a claim, and they arguably do so in ways that capture subtle interactions between different aspects of an overall argument. 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 also, e.g., Verheij 2004, 2003a; Walton and Gordon 2005; Walton et al. 2008, Chaps. 11, 12). The Bayesian framework would seem to be able to provide progress on this goal.

[13] By contrast, Walton and Gordon (2014) explicitly highlight 'relevance' as a key issue that still needs to be formally modelled within Carneades.

6 The appeal to popular opinion

Our final sample scheme is the so-called 'appeal to popular opinion'. This argument scheme is drawn from the traditional catalogue of fallacies of argumentation (see e.g., Woods et al. 2004), where it is also known as the ad populum argument or the bandwagon fallacy. Though the argument has traditionally been conceived of as a fallacy, and continues to figure as such in many textbooks on critical thinking (e.g., Inch and Warnick 2009), authors have more recently stressed that the argument may, in many cases, be plausible. Specifically, Walton (1999) argued from a broadly dialectical perspective that the argument may sometimes be acceptable. His treatment is further developed from a scheme-based perspective in Walton et al. (2008). At the same time, Hahn and Oaksford (2006a) highlight the ad populum argument as a further fallacy to which their Bayesian, formal treatment of the fallacies might in future be extended. In the following we present the detailed scheme-based treatment of the appeal to popular opinion provided by Walton et al. (2008) and develop a more in-depth Bayesian perspective on this fallacy, following on from Hahn and Oaksford's initial (2006a) considerations.

6.1 The ad populum scheme(s)

The basic scheme for the appeal to popular opinion given by Walton et al. (2008, p. 123; see also Walton 1999, p. 224; Freeman 1995, p. 70) runs as follows:

> S1: Everybody (in a particular reference group) accepts that A. Therefore, A is true (or you should accept A).
> S2: Everybody (in a particular reference group) rejects A. Therefore, A is false (or you should reject A).

As always, the scheme is accompanied by a number of critical questions (see also, Walton 1989, p. 89) designed to probe whether the argument is presumptively valid in a given case:

> CQ1: Does a large majority of the cited reference group accept A as true?
> CQ2: Is there other relevant evidence available that would support the assumption that A is not true?
> CQ3: What reason is there for thinking that the view of this large majority is likely to be right?

In marked contrast to the critical questions (and sub-questions) associated with the appeal to expert opinion these critical questions are striking in their lack of detail. CQ2 and CQ3 seem to involve only very general considerations, which do not seem specific to this particular scheme: Asking whether there is counter-evidence to *A* (CQ2) is a relevant consideration with respect to *any* claim, and the word 'group' does not even appear in the question. CQ3 does mention groups, but it effectively poses as a critical question the very thing the scheme is supposed to provide.

Schemes are intended to be arguments that are presumptively true, and thus inferences for which there is a normative basis. In posing the question of what reason there

is for thinking that the majority is likely to be right in this case, it is precisely that normative justification that is being sought. Consequently, it remains entirely unanswered by this scheme-based treatment why the argument should be presumptively acceptable in the first place.

A second feature of these critical questions that is worth noting is that they can, once again, be linked back to the core Bayesian quantities of sensitivity and specificity. CQ2, in asking for counter-evidence against A, is effectively asking about the chance that the group acceptance of A (which constitutes the relevant evidence in this argument) may be present even though A is false. In probabilistic terms this is $P(acceptance|\neg A)$, which can readily be seen to be an instance of the false positive rate $P(e|\neg h)$ defined above. CQ3, in probing the relationship between group acceptance and truth of claim A, $P(acceptance|A)$, effectively seeks to ascertain the sensitivity $P(e|h)$. Couching the questions in these terms immediately further illustrates three important advantages that we have already seen in our earlier examples. First, it once again makes clear that they can be fulfilled to degrees. The connection between majority acceptance and the truth of the claim, for example, may be more or less strong. Second, it once again makes more specific the relationship between the two considerations posed in CQ2 and CQ3. As noted above, the ratio between sensitivity and the false positive rate, the so-called likelihood ratio, determines the diagnosticity of the evidence, and how much our beliefs should change on receipt of that evidence. As long as that ratio is greater than 1, the evidence—majority acceptance—supports the claim. Again, the extent to which it does, however, is graded, ranging from 'very weak support' to 'strong support'. Third, the link provides a firm inferential foundation via Bayes' theorem—an inferential foundation that, unlike scheme-based plausible reasoning, is already well-established and worked out in detail.

From a scheme-based perspective, there is, however, more to the appeal to popular opinion than just the basic type. Walton et al. (2008) go on to list a further nine subtypes of the argument, a further illustration of the proliferation of argument schemes that may be seen as a common feature of the scheme-based approach more generally (e.g., Walton et al. 2008, identify five different types of slippery slope argument).

First is the sub-type of "Position-to-know ad populum argument" (ibid. p. 126):

> Everybody in this group G accepts A.
> This group is in a special position to know that A is true.
> Therefore, A is (plausibly) true.

The second sub-type introduced is the "Expert opinion ad populum argument":

> Everybody in this group G accepts A.
> G is a group of experts in a domain of knowledge.
> Therefore, A is true.

If argumentation schemes are to provide the basic theoretical framework for our understanding of argumentation, then there should be clear criteria for the declaration of types and subtypes, and theoretical and practically relevant consequences that are associated with a classification of types and subtypes. Walton et al. (2008, p. 125) state that "the subtypes can be considered as resting on a typology of groups and rhetorical

tactics". This, however, provides little guidance for these specific cases, as it leaves unanswered when or why a particular type of group might be viewed to constitute a distinct 'type'.

In fact, one may wonder how being in 'position-to-know' is distinct from expertise, and why one is not merely an instance of the other. Walton et al. provide a specific example of the position-to-know ad populum argument. It involves the population of Cedar Rapids who think the lake is a good place to swim in the summer. It seems natural enough to consider the population of Cedar Rapids as relative experts in their own lake, and, from this perspective it seems equally plausible to assimilate this example under the 'Expert-opinion ad populum' scheme. Moreover, nothing seems to follow from classifying the example into one sub-form as opposed to the other, as there are no specific critical questions associated with either, and it would thus seem that exactly the same considerations and inferential relationships obtain.[14]

Yet a further subtype is the 'Deliberation ad Populum Argument":

> Group G has deliberated intelligently and extensively on whether to accept proposition A.
> Everybody in G accepts A.
> Therefore, A is (plausibly) true.

This type is presented as closely related to the preceding two knowledge-based versions, and no new critical questions are proposed, but both the extra depth of deliberation and the uniformity of opinion are taken to provide an additional sign that the opinion of the majority is preferable (Walton et al. 2008, p. 127).

Adopting a probabilistic, Bayesian perspective on these putative schemes allows progress both in terms of scheme typology, and in terms of the critical questions associated with the schemes. This progress, as we will see, stems from a fundamental advantage of the Bayesian perspective, namely that it connects to a large body of insight on probability and statistics that is of direct inductive relevance.

6.2 When groups know best

The extent to which groups may form accurate judgments, and in many cases form judgments that are more accurate than not only their average, but even their best members, has been a topic of longstanding research within political science, psychology, forecasting, computer science, and statistics (see e.g., Clemen 1989). Of particular relevance to the question of whether or not to adopt a particular claim as true is Condorcet's (1785) jury theorem.

Condorcet's theorem shows that given two alternatives, of which only one is correct such as the truth or falsity of a claim, a group verdict based on simple majority will under specified circumstances outperform the individual judges in terms of accuracy. In fact, as group size increases, it will converge on the truth.

[14] This is not to deny that there may be contexts, such as the law, in which distinguishing between being 'in a position to know' and 'being expert' might be meaningful (see e.g., Godden and Walton 2006). However, in order to justify different argument schemes there must minimally be some consequential difference to either the basic inference or the critical questions.

Condorcet's basic result assumes n voters whose choices are independent of one another, and a probability p that each voter will pick the correct alternative, which is assumed to be the same for all voters. If that probability is greater than 0.5 (assuming prior odds for the alternatives that are even), then the probability that the group choice, P_N, will be correct, will not only be higher than p (i.e. the group verdict will be more accurate than the individual voters), but it will increase rapidly with group size N, and will approach infallibility in the limit.

This is true regardless of *how much* the voters know, as long as they know something (if their accuracy is at chance, i.e. $p = 0.5$, then the group verdict too will remain equal to tossing a coin; and, of course, if they are systematically biased against the right option, i.e., $p < 0.5$, then the reverse holds: P_N will be even lower).

So it is not required that the individual voters are very knowledgable or accurate. Their degree of competence merely affects how high the probability of a correct choice is for a group of given size, and how rapidly the asymptote of perfect knowledge is approached as group size increases.

In light of this, it seems more natural to view the three schemes 'basic ad populum', 'position to know ad populum' and 'expert opinion ad populum' not as qualitatively distinct subtypes, but merely as different instances of the same basic argument form that may vary continuously in argument strength, from 'not at all convincing' to 'establishing a fact with (near) certainty': individual members' accuracy (p) is higher in each case so the same group size will lead to a greater probability that the group verdict is correct, all other things being equal.

Vis a vis the primary critical question 'CQ1: Does a large majority of the cited reference group accept A as true?' it is worth stressing again that Condorcet's result only requires a simple majority, not 'a large majority'. However, within the basic Condorcet framework just outlined, the larger the majority, the more likely the chosen alternative will be correct (as long as $p > 0.5$, see Grofman et al. 1983, Theorem III), though in the more realistic case, where voters are of varying competence, *all* voters must have individual competences above 0.5 for greater majorities to guarantee greater likelihood of the group choice being correct.

The assumption of equal competence for all voters (i.e., a common value of accuracy p) is only one way in which the assumptions of Condorcet's theorem seem unrealistic. The other is the assumption of voter independence. In the real world, people's judgments may be correlated because they share common information, because they communicate with each other, and because they follow supposed 'experts' or opinion leaders, or belong to certain schools of thought. In fact, herein lies the most important source of limitation on the accuracy of group judgment, and one that is not at all reflected in the critical questions presently associated with the ad populum schemes. In the limit, non-independence can mean that a group verdict is not really a group judgment at all, but simply reflects a single opinion. Imagine a case where group members base their verdicts exclusively on the judgment of a single expert: in this case the group's judgment will be no more (or less) accurate than that single expert.

Even if group members base their judgments only partly on that expert, however, it will decrease (other things being equal) the group's accuracy relative to the levels it would attain if votes were independent. Assume, for the moment, equal levels of individual competence, p, for all group members (including its opinion leader) when

they cast their votes independently. However, assume also that group members have a certain probability d, of deferring to an opinion leader and simply adopting that leader's judgment. Now the group's competence will again collapse to that of the leader as soon as d is sufficiently large relative to p. To provide an example by Grofman et al. (1983), if $d = 0.2$ and $p = 0.6$, then the expected value of P_N is 0.6, *regardless of group size* N. In other words, the group majority is only exactly as competent as the leader, since the leader's voting bloc will be likely to determine the outcome of the vote.

Of course, in the more realistic case of unequal competence, benefits to overall accuracy may ensue if the opinion leader is more accurate than some (or even all) of the other group members (though there will still be costs of non-independence).

Ladha (1992) provides a more general, and more realistic version of Condorcet's theorem, which incorporates both differences in individual competence and non-independence of voters. Amazingly, whether the ampliative effect of group judgement holds or not, is determined only by the *average* probability of correct responding, $\bar{p}$, within the group, and by the *average* level of inter-dependence between raters, $\bar{r}$. Specifically, Ladha shows that it is possible to calculate a threshold, $T(n, \bar{p})$, for a given group size n and mean competence, $\bar{p}$, such that levels of average inter-dependence below that threshold will guarantee that the overall group accuracy (P_N) will be higher than the average accuracy of the individual group members ($\bar{p}$):

$$\text{if } \bar{p} > 0.5, \text{ and } \bar{r} < T(n, \bar{p}) \text{ then } P_N > \bar{p}$$

where,

$$T(n, \bar{p}) = \bar{p} - \frac{n}{n-1} \quad \frac{(\bar{p} - 0.25)(1 - \bar{p})}{\bar{p}} \tag{3}$$

and r_{ij} is the probability that voters i and j will vote simultaneously for the correct alternative ($r_{ij} = P(Vote\_i = correct), P(Vote\_j = correct)$, and the average across all such pairs is $\bar{r}$), which will be equal to the product of the individuals' probability of responding correctly, $p_i p_j$ if the votes are independent, greater than it if they are positively correlated, and smaller if they are negatively correlated.

Furthermore, the group verdict P_N will approach the correct answer with certainty as group size approaches infinity and mean degree of independence $\bar{r}$ approaches $\bar{p}^2$. In other words, despite some degree of non-independence, group 'infallibility', is still possible, as long as the inter-dependence is not too high.

As a consequence, individuals may contribute to overall group accuracy both by raising mean accuracy *and* by lowering mean interdependence. Furthermore, because it is only the average individual competence and the averagege inter-voter independence that matter, and not a minimum level of competence or independence to be met by each member, individuals may *improve* group accuracy even if their competence is considerably *below* the mean competence in the group if only they sufficiently benefit mean independence.

Conversely, adding further expertise to the group will only be beneficial if it does not also increase too much the average interdependence. Further information that improves accuracy will be most effective where it is non-shared information between group members, and if its effects are too homogenizing it may actually decrease

accuracy. Adding in the influence of an expert's opinion may thus be detrimental even if that expert is far more competent than the other group members if the expert's influence is too strong.

In conclusion it should be stressed that although Condorcet jury theorems provide a particularly relevant body of results for considering appeals to popular opinion, they constitute but a fraction of a sprawling literature on the accuracy of group verdicts (see e.g., Clemen 1989). It is worth mentioning also the sizeable literature on the accuracy of group estimates of numerical quantities. Here, a classic empirical finding was that the group mean would provide an estimate of quantities such as the weight of a pig or the number of beans in a jar (see e.g., Galton 1907; Gordon 1924) that was more accurate than most, or even all of the individual estimates from which it was derived (see Lorge et al. 1958, for a review). It took experimental psychologists several decades to realise that these results were a statistical consequence of averaging, deriving from the fact that the standard error of the mean of several judgments is smaller than the standard deviation of the judgments themselves, so that statistical combinations of judgments will cancel out unsystematic judgment error (see e.g., Stroop 1932). In this context too, the negative effect of inter-rater correlation can be analytically tracked (see e.g., Hogarth 1978; see also Treynor 1987 for a nice numerical example of how the influence of financial experts will increase error in market price even though the expert's individual judgment may be superior). Research in this area continues to this day, both under the heading of 'wisdom of the crowds' (Surowiecki 2004; see also Page 2005), and in the context of the empirical study of online prediction markets (see e.g., Forsythe et al. 1992). Last but not least, this work provides insight into the extent to which requirements for group accuracy that emerge from the formal results (in particular with respect to key factors such as degree of independence) are typically met in different real world judgment contexts.

All of this has three main implications for the argument scheme 'appeal to popular opinion'. First, it adds a genuine critical question of fundamental importance to the evaluation of the scheme, namely, to what extent are the verdicts in the group likely to be independent? Do the group members possess private/unique information, and are they diverse in backgrounds, experience and beliefs in ways that reduce positive correlations among the members (remembering that it is not just the addition of un-correlated votes that may be beneficial but also negatively correlated ones).

Second, consideration of the relationship between probability of being correct, group size, and correlation of opinion adds further weight to the claim that the 'basic ad populum', the 'position to know ad populum', the 'expert opinion ad populum', and the 'deliberation ad populum' should not be viewed as distinct types. The same factors of group size, member competence, and member independence govern directly all three. The relative accuracy of the group judgments will depend on more than the level of knowledge of the group's members, namely its degree of independence. A poorly informed group may outperform relative experts if their judgements are based on knowledge that is fairly independent, whereas the experts' judgments are not. The power of group judgments lies in the information aggregating properties of majority voting, and relative expertise matters only all other things being equal. By the same token, it should be clear that 'extensive deliberation' is not, per se, a reason to believe

that the group's verdict is likely to be true. By increasing inter-dependence, group deliberation may make the group verdict less likely to be accurate, in keeping with the frequent empirical finding within experimental psychology that groups that deliberate or interact as a group tend to perform less well than 'statisticized groups' that are merely aggregate judgments drawn from individuals making judgments in isolation (see e.g., the results of Gigone and Hastie 1997, survey of such studies). Of course, deliberation may sometimes help, but for this to occur, the benefits to individual accuracy must outweigh the cost of increased inter-dependence (see e.g., the experiments in Joensson et al. 2015).

The third, and final, implication is that a probabilistic perspective brings together not only what, on a traditional scheme-based view, are different variants of the same argument type ('basic ad populum', 'position to know ad populum', 'expert ad populum'), but also different schemes and their associated considerations themselves. Questions of 'expertise' pervade different argument forms, and because they are formalized in a consistent, unifying manner, they bring together different argument schemes conceptually and in practice. One can ask about the importance of expertise within a group of decision-makers in the context of an ad populum argument, and one can compare the putative strength of an ad populum argument to one based on a single expert's expertise, as in the appeal to expert opinion. In all of these cases, 'expertise' is ultimately reducible to considerations of sensitivity and specificity, as described in the context of the 'argument from sign'. In this sense, probability theory provides a unifying perspective.

7 The benefits of merger

In discussing three argumentation schemes, we have raised a number of issues related to the scheme-based approach to argument quality. In particular, we have sought to highlight what are presently unresolved problems for the scheme-based approach. At the same time, we have aimed to indicate how a Bayesian perspective may help to solve these problems. In this section, we summarize the most important benefits the Bayesian approach brings to the scheme-based tradition.

7.1 A well-founded normative status

As discussed, the Bayesian approach has a normative status that derives from formal results, which relate Bayesian inference to the calibration and accuracy of our beliefs. Moreover, the Bayesian approach connects systematically to a large body of knowledge within statistics, machine learning, and formal epistemology, which exploits those normative foundations and brings them to bear in the analysis of specific questions and problems. By contrast, the scheme-based approach to argument quality is based on intuition, which is often misleading. We saw this particularly in the context of the appeal to popular opinion where factors that have been taken to make an inference stronger (such as the presence of group deliberation) can formally and empirically be shown to often have an adverse effect.

7.2 Probabilities are useful (and quite natural!) ways to capture relationships in the world and our beliefs about them

The probability calculus provides a straightforward representation of degrees of belief and with it, for an argumentation context, degrees of convincingness. It also captures readily relationships between variables as is essential for argumentation. Furthermore, as an intensional calculus it connects naturally with the content of statements, not just their formal syntactic relationships. This is at the heart of what gives the probability calculus inferential capabilities that are difficult to mimic by non-classical, logical systems. Furthermore, probabilities conceived of as degrees of belief (not just objective 'chances') are not limited to overtly numerical or statistical contexts as our examples show. Finally, the probability calculus and, with it, well-defined aggregation rules are already there. There is no need to invent such systems from scratch.

7.3 Meaningful relationships between premises and conclusion

The premises and the conclusion are meaningfully related to each other in a Bayesian account of argument quality: most importantly, the Bayesian framework captures the fact that premises must have informational relevance for the conclusion if an argument is to be strong. Informational relevance is captured naturally by probabilistic relationships. Defeasible modus ponens fails to provide a fully adequate approach; the extent to which a premise is generally true, in and of itself, is insufficient for providing evidential or argumentative support as we showed in the discussion of the example from sign (Sect. 4). Furthermore, the Bayesian framework allows one to assign to a conclusion degrees of belief that arise naturally from uncertainty associated with the reasons given.

7.4 A principled basis for critical questions

The Bayesian framework is useful both for identifying what the relevant critical questions are (see Sect. 5 on expert opinion, and Sect. 6 on appeal to popular opinion) and the inferential impact of critical questions. This includes both their individual impact and combinations across critical questions, or more specifically, the variables or factors these questions target. In the scheme-based approach to argument quality, the critical questions can typically never be met, strictly speaking. If the question is 'Is E personally reliable as a source?' for the argument from expert opinion, it is hard to see at what moment in time a person can be fully convinced to answer either 'yes' or 'no'. The belief in a person's reliability is a personal, but quantifiable probability. Bayesian formalization captures this uncertainty and allows evidence supplied in response to different critical questions to interact, both with each other, and the overall conclusion of the scheme. Concerning what critical questions should be asked, we provided examples where factors that genuinely make a difference to argument strength (e.g., the role of independence in the context of the appeal to popular opinion) were missed, providing evidence of the kinds of refinements that adopting a Bayesian perspective will allow.

7.5 A systematic basis for distinguishing different schemes

As arguments relate to different types of content in the world, models have to reflect such content variability. One major question is that of how many such schemes there should be. This question is not new; it has been raised ever since schemes were proposed (e.g., Katzav and Reed 2004). We agree with Walton et al. (2008, p. 13) who write that

> The existing formulations of the argumentation schemes are not very precise or systematic, perhaps because they have arisen out of practical concerns in dealing with real cases. New work is needed to refine, classify, and formalize these schemes.

We provided examples, both in the context of the argument from sign (Sect. 4) and the appeal to popular opinion, of how and why a Bayesian perspective informs decisions about types of schemes. Specifically, we provided examples of where scheme variants were based on considerations that were mis-judged in terms of their inferential impact (e.g., deliberation in the case of popular opinion), so that there were no normative grounds for that variant, and we also provided examples (in the context of the argument from sign), where different variants of the scheme seemed to be insufficiently distinct from an inferential perspective to merit separate schemes.

7.6 Computationally explicit

Finally, the Bayesian framework is computationally explicit. For each of our example schemes, we gave indications how actual posterior degrees of belief in a conclusion may be computed. This is not just a desirable 'add-on' that furthers a longer term project of making use of schemes in computational systems. Rather, it is an essential component of the normative project itself. Normative guidance is incomplete as long as it consists only of a number of considerations or 'topoi' without specification of what exactly is to be done with these. Without a formal, explicit account of how one is to get from a scheme or set of schemes and critical questions to an actual belief in a conclusion, the prescriptive force of schemes is extremely limited. In fact the actual reasoning remains completely underconstrained: there is little point being told to consider an issue, without being told what to do with the results of one's consideration. In other words, it is the 'bottom line' or 'solution' that needs to be constrained. Here, our examples suggest that there is no need for a computational 'third way', nor do present 'third way' approaches (whether these involve defeasible modus ponens or procedural dialectical treatment) seem to provide comprehensive solutions to the problem. In the examples we considered, the alternative, 'third way' tools (defeasible modus ponens—in the context of the argument from sign, shifting the burden of proof—in the context of the appeal to expert opinion), potentially yielded counter-intuitive evaluations of argument strength and/or formally captured less of the problem and thus seemed less suitable to the task of formalising argument quality than the probability calculus.

8 Conclusions

Walton et al. (2008, p. 39) propose five desiderata for a comprehensive theory of argumentation schemes. According to these, a comprehensive theory should be

(1) "*rich* and sufficiently exhaustive to cover a large proportion of naturally occurring argument"
(2) "*simple*, so that it can be taught in the classroom and applied by students"
(3) "*fine-grained*, so that it can be usefully employed as both a normative and an evaluative system"
(4) "*rigorous*, and fully specified, so that it might be represented in a computational language"
(5) "*clear*, so that it can be integrated with traditional diagramming techniques"

We believe that the combined approach of analysing the schemes from a Bayesian perspective allows a theoretical treatment that meets all five of these. First, it is *rich* (1) in the sense that it will cover most schemes albeit in a smaller number than the 60 that are presented in Walton et al. (2008). Second it is *simple* because there is a simple, unifying core in probabilities as degrees of belief, the notion of conditional probability, and Bayes' theorem. This simplicity can also, we believe, be capitalised on in educational contexts. At root this machinery requires no more than grasping the idea that events or states of affairs might be represented by numbers coupled with an ability to multiply and divide those numbers. However, the basic considerations do not require even that because they can be communicated on a qualitative level via examples because they reflect patterns of inference we already naturally draw in many contexts (as we saw in the case of the sprinkler, lawn and rain example; see on qualitative uses of Bayesian analysis also Corner and Hahn 2009). Third, the combined approach is *fine-grained* enough to provide both normative and evaluative guidance as seen both in the examples discussed in this paper, and in related work on fallacies of argumentation (e.g., Hahn and Oaksford 2007a). This is possible because the Bayesian framework not only has a well-developed normative basis, but because it is sufficiently computationally explicit that probabilistic reconstruction captures not just the quantitative impact of individual variables, but also their interrelationships. This *rigour* (4) means not only that implementation in formalized (computer) languages is possible, but such formalizations, and working computational systems based on them, already exist, not necessarily in the context of argumentation, but in the wider context of the ever-increasing influence of Bayesian methods within AI more generally (see, e.g., Korb and Nicholson 2003). Finally, though substantive differences between many types of argument diagramming or visualization exist, and not all of these lend themselves equally to combination with probabilistic inference, Bayesian Belief Networks themselves provide visualisation (5) (see, e.g., Pearl 1988) and the parallels between this and other approaches to displaying argument or evidential connections have long been known (see e.g., Schum 1994, for explicit discussion of both parallels and differences between different types of frameworks for displaying inferential links between arguments and/or evidence).

Hence, a Bayesian perspective on the catalogue of argumentation schemes, once systematically applied across the catalogue, will deliver, we think, a comprehensive

theory of informal argument. In bringing together the scheme-based tradition and Bayesian argumentation, we think, the rich body of research on actual arguments found in everyday argumentative practice that the scheme-based tradition has amassed will be given the theoretical analysis and rigour it deserves. While it is Bayesianism that carries the normative weight, the examples discussed illustrate, we think, how different schemes raise interesting and important questions, both in terms of formalization and in terms of normative considerations. It is the richness that comes only when considering specific instances of real world argumentation that a complete theory of argumentation ultimately needs.

Acknowledgments We would like to thank Frank Zenker for helpful comments on a draft of this manuscript, and Tom Gordon for helpful discussion.The first author was partially supported by the Swedish Research Council's Hesselgren professorship, and the second author was partially supported by the Centre for Language Studies (Nijmegen).

Compliance with Ethical Standards

Conflict of interest There are no potential conflicts of interest associated with this research.

References

Alexy, R. (1989). *A theory of legal argumentation*. Oxford: Clarendon Press.
Atkinson, D., & Peijnenburg, J. (2010). Justification by infinite loops. *Notre Dame Journal of Formal Logic*, *51*, 407–416.
Bex, F., Prakken, H., Reed, C., & Walton, D. (2003). Towards a formal account of reasoning about evidence: Argumentation schemes and generalisations. *Artificial Intelligence and Law*, *11*, 125–165.
Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.
Bolstad, W. M. (2004). *Introduction to Bayesian statistics*. Hoboken, NJ: Wiley.
Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford: Oxford University Press.
Bovens, L., & Olsson, E. J. (2002). Believing more, risking less: On coherence, truth and non-trivial extensions. *Erkenntnis*, *57*, 137–150.
Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*, 287–291.
Christmann, U., Mischo, C., & Flender, J. (2000a). Argumentational integrity: A training program for dealing with unfair argumentational contributions. *Argumentation*, *14*, 339–360.
Christmann, U., Mischo, C., & Groeben, N. (2000b). Components of the evaluation of integrity violations in argumentative discussions: Relevant factors and their relationships. *Journal of Language and Social Psychology*, *19*, 315–341.
Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*, 559–583.
Corner, A., & Hahn, U. (2009). Evaluating science arguments: Evidence, uncertainty, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied*, *15*, 199–212.
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 and Language*, *64*, 153–170.
de Condorcet, N. C. (1785). *Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix*. Paris: Imprimerie Royale.
Dennett, D. C. (1984). Cognitive wheels: The frame problem of AI. In C. Hookaway (Ed.), *Minds, machines and evolution* (pp. 129–151). Cambridge: Cambridge University Press.
Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Belmont, CA: Thompson/ Wadsworth.
Earman, J. (1992). *Bayes or bust?*. Cambridge, MA: MIT Press.

Ehninger, D. (1974). *Influence, belief, and argument: an introduction to responsible persuasion*. Glenview, IL: Scott Foresman.
Ehningner, D., & Brockriede, W. (1963). *Decision by debate*. New York: Dodd, Mead.
Evans, J. St B T., & Over, D. E. (2004). *If*. Oxford: Oxford University Press.
Falk, R., & Well, A. D. (1997). Many faces of the correlation coefficient. *Journal of Statistics Education*, *5*(3), 1–18.
Fitelson, B. (1996). Wayne, Horwich and evidential diversity. *Philosophy of Science*, *63*, 652–660.
Forsythe, R., Nelson, F., Neumann, G. R., & Wright, J. (1992). Anatomy of an experimental political stock market. *The American Economic Review*, *82*, 1142–1161.
Fox, J., & Das, S. (2000). *Safe and sound*. Menlo Park: AAAI Press.
Freeman, J. B. (1995). The appeal to popularity and presumption by common knowledge. In H. V. Hansen & R. C. Pinto (Eds.), *Fallacies: Classical and contemporary readings* (pp. 263–273). University Park: University of Pennsylvania Press.
Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.
Garssen, B. J. (1997). *Argumentatieschema's in pragma-dialectisch perspectief: Een theoretisch en empirisch onderzoek*. Amsterdam: IFOTT.
Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin*, *121*, 149–167.
Godden, D. M., & Walton, D. N. (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.
Gordon, K. (1924). Group judgments in the field of lifted weights. *Journal of Experimental Psychology*, *7*, 398–400.
Gordon, T. F., & Walton, D. (2009). Proof burdens and standards. In I. Rahwan & G. Simari (Eds.), *Argumentation in artificial intelligence* (pp. 239–260). Berlin: Springer.
Gordon, T., Prakken, H., & Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence*, *171*, 875–896.
Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psycho-physics*. New York: Wiley.
Grofman, B., Owen, G., & Feld, S. L. (1983). Thirteen theorems in search of the truth. *Theory and Decision*, *15*, 261–278.
Hahn, U. (2011). The problem of circularity in evidence, argument and explanation. *Perspectives on Psychological Science*, *6*, 172–182.
Hahn, U. (2014). The Bayesian boom: Good thing or bad? *Frontiers in Cognitive Science*, *5*, Article 765.
Hahn, U., Harris, A. J. L., & Corner, A. J. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*, 337–367.
Hahn, U., & Oaksford, M. (2006a). A Bayesian approach to informal argument fallacies. *Synthese*, *152*, 207–236.
Hahn, U., & Oaksford, M. (2006b). Why a normative theory of argument strength and why might one want it to be Bayesian? *Informal Logic*, *26*, 1–24.
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. (2012). Rational argument. In Holyoak & Morrison (Eds.), *The Oxford handbook of thinking and reasoning*. Oxford: Oxford university Press.
Hahn, U., Oaksford, M. & Bayindir, H. (2005). How convinced should we be by negative evidence? In *Proceedings of the 27th annual meeting of the cognitive science society*.
Hahn, U., Oaksford, M., & Harris, A. J. (2013). Testimony and argument: A Bayesian perspective. In F. Zenker (Ed.), *Bayesian argumentation* (pp. 15–28). Dordrecht: Springer.
Hajek, A. (2008). Dutch book arguments. In P. Anand, P. Pattanaik, & C. Puppe (Eds.), *The handbook of rational and social choice* (pp. 173–196). Oxford: Oxford University Press.
Hamblin, C. L. (1970). *Fallacies*. London: Methuen.
Hardman, D. (2009). *Judgment and decision making: Psychological perspectives*. Chichester: BPS Blackwell.
Harris, A. J. L., Corner, A., & Hahn, U. (2013). James is polite and punctual (and useless): A Bayesian formalisation of faint praise. *Thinking & Reasoning*, *19*, 414–429.
Harris, A. J. L., Hahn, U., Hsu, A. S., & Madsen, J. K. (2015). The appeal to expert opinion: Quantitative support for a Bayesian network approach. *Cognitive Science*.

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, Northwestern University, Evanston, IL.

Hoeken, H., Sorm, E., & Schellens, P. J. (2014). Arguing about the likelihood of consequences: Laypeople's criteria to distinguish strong arguments from weak ones. *Thinking and Reasoning*, *20*, 77–98.

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

Hogarth, R. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, *21*, 40–46.

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

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach*. La Salle, IL: Open Court.

Inch, E. S., & Warnick, B. H. (2009). *Critical thinking and communication: The use of reason in argument* (6th ed.). Boston: Pearson.

Joensson, M., Hahn, U. & Olsson, E. (2015). The kind of group you want to belong to: Effects of group structure on group accuracy. *Cognition*. *Online first*.

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

Katzav, J., & Reed, C. A. (2004). On argumentation schemes and the natural classification of arguments. *Argumentation*, *18*, 239–259.

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: Cambridge University Press.

Korb, K. (2004). Bayesian informal logic and fallacy. *Informal Logic*, *23*, 41–70.

Korb, K. B., & Nicholson, A. E. (2003). *Bayesian artificial intelligence*. Boca Raton: CRC Press.

Korb, K. B., McConachy, R. & Zukerman, I. (1997). A cognitive model of argumentation. In: *Proceedings of the 19th annual conference of the cognitive science society* (pp. 400–405).

Ladha, K. K. (1992). The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, *36*, 617–634.

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

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

Lewis, D. (1980). A subjectivist's guide to objective chance. In Richard C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. II, pp. 263–293). Berkeley: University of California Press.

Lorge, I., Fox, D., Davitz, J., & Brenner, M. (1958). A survey of studies contrasting the quality of group performance and individual performance, 1920–1957. *Psychological Bulletin*, *55*, 337–372.

McConachy, R., & Zukerman, I. (1999). Towards a dialogue capability in a Bayesian argumentation system. *ETAI 3—Electronic Transactions of Artificial Intelligence (Section D)*, *3*, 89–124.

McConachy, R., Korb, K. B., & Zukerman, I. (1998). Deciding what not to say: An attentional-probabilistic approach to argument presentation. In *Proceedings of the 20th annual conference of the cognitive science society* (pp. 669–674), Madison, Wisconsin.

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

Myrvold, W. C. (1996). Bayesianism and diverse evidence: A reply to Andrew Wayne. *Philosophy of Science*, *63*, 661–665.

Nussbaum, E. M. (2011). Argumentation, dialogue theory, and probability modeling: Alternative frameworks for argumentation research in education. *Educational Psychologist*, *46*, 84–106.

Nussbaum, E. M., & Edwards, O. V. (2011). Critical questions and argument stratagems: A framework for enhancing and analyzing students' reasoning practices. *Journal of Learning Sciences*, *20*, 443–488.

O'Keefe, D. J. (2002). *Persuasion: Theory and research* (2nd ed.). Thousand Oaks, CA: Sage.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.

Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, *58*, 75–85.

Oaksford, M., & Hahn, U. (2007). Induction, deduction and argument strength in human reasoning and argumentation. In A. Feeney, & E. Heit (Eds.), Inductive reasoning (pp. 269–301). Cambridge University Press.
Olsson, E. J. (2002). What is the problem of coherence and truth? *Journal of Philosophy*, *94*, 246–272.
Olsson, E. J., & Schubert, S. (2007). Reliability conducive measures of coherence. *Synthese*, *157*, 297–308.
Olsson, E. J. (2005). *Against coherence: Truth, probability, and justification*. Oxford: Oxford University Press.
Page, S. E. (2005). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton: Princeton University Press.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufman.
Perelman, C., & Olbrechts-Tyteca, L. (1969). *The new rhetoric: A treatise on argumentation*. Notre Dame, IN: University of Notre Dame Press.
Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer.
Pollock, J. L. (1995). *Cognitive carpentry: A blueprint for how to build a person*. Cambridge: MIT Press.
Prakken, H. (2005). AI & law, logic and argument schemes. *Argumentation*, *19*, 303–320.
Prakken, H., & Vreeswijk, G. A. W. (2002). Logics for defeasible argumentation. In D. M. Gabbay & F. Guenthner (Eds.), *Handbook of philosophical logic* (2nd ed., Vol. 4, pp. 219–318). Dordrecht/Boston/London: Kluwer Academic Publishers.
Rahwan, I., & Simari, G. R. (Eds.). (2009). *Argumentation in artificial intelligence*. Dordrecht: Springer.
Reed, C., & Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal of Artificial Intelligence Tools*, *13*, 961–980.
Reinard, J. C. (1991). *Foundations of argument: Effective communication for critical thinking*. Dubuque, IA: William C. Brown.
Rescher, N. (1976). *Plausible reasoning*. Assen: Van Gorcum.
Rieke, R. D., & Sillars, M. O. (1984). *Argumentation and the decision making process*. New York: Harper Collins.
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: Foris.
Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Evanston, IL: Northwestern University Press.
Snoeck Henkemans, A. F. (2000). State-of-the-art: The structure of argumentation. *Argumentation*, *14*, 447–473.
Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology*, *15*, 550–562.
Surowiecki, J. (2004). *The wisdom of crowds*. New York, NY: W.W. Norton & Company Inc.
Treynor, J. L. (1987). Market efficiency and the bean jar experiment. *Financial Analysts Journal*, *43*, 50–53.
van Eemeren, F. H., & Grootendorst, R. (2004). *A systematic theory of argumentation. The pragma-dialectical approach*. Cambridge: 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* (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.
Verheij, B. (2004). Dialectical argumentation with argumentation schemes: An approach to legal logic. *Artificial intelligence and Law*, *11*, 167–195.
Walton, D. N. (1989). *Informal logic*. Cambridge: Cambridge University Press.
Walton, D. N. (1996). *Argumentation schemes for presumptive reasoning*. Mahwah, N.J.: Erlbaum.
Walton, D. N. (1997). *Appeal to expert opinion: Arguments from authority*. University Park, PA: Penn State Press.
Walton, D. N. (1998). *The new dialectic: Conversational contexts of argument*. Toronto: University of Toronto Press.
Walton, D. N. (1999). *Appeal to popular opinion*. University Park, PA: Penn State Press.
Walton, D. M. (2001). Abductive, presumptive, and plausible arguments. *Informal Logic*, *21*, 141–169.
Walton, D. N. (2004). *Relevance in argumentation*. Mahwah, NJ: Erlbaum.

Walton, D. N. (2006). *Fundamentals of critical argumentation*. Cambridge: Cambridge University Press.
Walton, D. N. (2008). *Witness testimony evidence: Argumentation, artificial intelligence, and law*. Cambridge: Cambridge University Press.
Walton, D., & Gordon, T. F. (2005). Critical questions in computational models of legal argument. In P. E. Dunne, & T. Bench-Capon (Ed.), *International workshop on argumentation in artificial intelligence and law* (pp. 103–111). Nijmegen: Wolf Legal Publishers.
Walton, D., & Gordon, T. F. (2014). How to formalize informal logic. In Mohammed, D., & Lewiński, M. (Eds.), Virtues of argumentation. In: *Proceedings of the 10th international conference of the Ontario society for the study of argumentation (OSSA), 22–26 May 2013*(pp. 1–13). Windsor, ON: OSSA.
Walton, D., & Reed, C. (2002). Argumentation schemes and defeasible inferences. In *Workshop on computational models of natural argument, 15th European conference on artificial intelligence* (pp. 11–20).
Walton, D. N., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge: Cambridge University Press.
Wayne, A. (1995). Bayesianism and diverse evidence. *Philosophy of Science*, *62*, 111–121.
Whately, R. (1846). *Elements of rhetoric: Comprising an analysis of the laws of moral evidence and of persuasion, with rules for argumentative composition and elocution/c by Richard Whately*. B. Fellowes.
Woods, J., Irvine, A., & Walton, D. N. (2004). *Argument: Critical thinking, logic and the fallacies*, Revised Edition. Toronto: Prentice Hall.
Zukerman, I. (2009). Towards probabilistic argumentation. In I. Rahwan & G. R. Simari (Eds.), *Argumentation in artificial intelligence* (pp. 443–462). Dordrecht: Springer.