Bayesian Argumentation and the Value of Logical Validity

Benjamin Eva[*] Stephan Hartmann[†]

March 23, 2018

Forthcoming in Psychological Review

Abstract

According to the Bayesian paradigm in the psychology of reasoning, the norms by which everyday human cognition is best evaluated are probabilistic rather than logical in character. Recently, the Bayesian paradigm has been applied to the domain of argumentation, where the fundamental norms are traditionally assumed to be logical. Here, we present a major generalisation of extant Bayesian approaches to argumentation that (i) utilizes a new class of Bayesian learning methods that are better suited to modelling dynamic and conditional inferences than standard Bayesian conditionalization, (ii) is able to characterise the special value of logically valid argument schemes in uncertain reasoning contexts, (iii) greatly extends the range of inferences and argumentative phenomena that can be adequately described in a Bayesian framework, and (iv) undermines some influential theoretical motivations for dual function models of human cognition. We conclude that the probabilistic norms given by the Bayesian approach to rationality are not necessarily at odds with the norms given by classical logic. Rather, the Bayesian theory of argumentation can be seen as justifying and *enriching* the argumentative norms of classical logic.

[*] Department of Philosophy, University of Konstanz, 78464 Konstanz (Germany) – http://be0367.wixsite.com/benevaphilosophy – benjamin.eva@uni-konstanz.de.

[†] Munich Center for Mathematical Philosophy, LMU Munich, 80539 Munich (Germany) – http://www.stephanhartmann.org – s.hartmann@lmu.de.

Keywords: Argumentation, Reasoning, Bayesianism, Conditionals, Dual Source Models, Logical Validity.

1 Introduction

In science and in everyday life, arguments are used to convince others (as well as ourselves) of certain statements or propositions. A good argument supports a proposition (the conclusion of the argument) with reasons (the premises of the argument). There is a huge literature in cognitive science, philosophy, logic, and computer science that studies argumentation. We find, for example, classifications of argument schemes (Walton 2008, Walton 2013), theories of defeasible reasoning and argumentation (Pollock 1967, Pollock 1987), and studies of logical fallacies (Van Eemeren and Grootendorst 1995, Hamblin 1972, Walton 1995, Walton 2011). More recently, philosophers and psychologists have started to study argumentation from a Bayesian perspective (Eva and Hartmann 2018, Godden and Zenker 2016, Hahn and Hornikx 2016, Hahn and Oaksford 2007, Oaksford and Hahn 2006, Zenker 2013). This approach allows for the formulation of probabilistic measures of argument strength, and it demonstrates that many so-called 'fallacies' may nevertheless be very good arguments, in the sense that they considerably raise the probability of the conclusion. That is, deductively invalid argument schemes (such as affirming the consequent (AC) and denying the antecedent (DA)) can also provide considerable support for a conclusion. To what extent this is the case depends partially on the specific *context* (i.e. the prior probability distribution), and not only on the *logical structure* of the argument. Equally, we might support a conclusion with a logically valid argument scheme such as modus ponens (MP) or modus tollens (MT), even if the premises are uncertain. However, if the premises are not certain, then valid and invalid argument schemes will support the conclusion, if at all, only to a certain degree. This raises the questions 'why should we use valid argument schemes at all?' and 'what is the value of logical validity if we are in the business of making inferences with uncertain premises?'. This paper proposes a novel answer to these questions and presents a major extension to the Bayesian approach to argumentation, based on the idea that argumentation is learning (the premises of the argument) and that the learning in question proceeds in a conservative way via the minimization of some “distance” measure such as the Kullback-Leibler divergence.

The approach presented here ties in very naturally with recent advances in the psychology of reasoning. In particular, Oaksford and Chater (2007, 2009) have developed the ‘Bayesian approach to rationality’ (commonly referred to as the ‘new paradigm’), which contends that ‘cognition in general, and human everyday reasoning in particular, is best viewed as solving probabilistic, rather than logical, inference problems’ (Oaksford and Chater 2009: 69). The Bayesian approach overturns the previously dominant logical paradigm, in which the standards by which the rationality of human cognition are evaluated are articulated in purely logical terms. There is a wealth of experimental evidence (see Chapters 1–4 of Oaksford and Chater 2007) that appears to show that, by logical standards, humans are naturally prone to making a range of systematically irrational inferences. Furthermore, it is well known that (classical) logical inference is monotonic, in the sense that adding extra premises to a valid argument can never affect the argument’s validity, i.e. ‘adding premises can never overturn existing conclusions’ (Oaksford and Chater 2009: 72). But, as Oaksford and Chater note,

> [I]n reasoning about the everyday world... *non*-monotonicity is the norm: almost any conclusion can be overturned if additional information is acquired. Thus, consider the everyday inference from *It’s raining* and *I am about to go outside* to *I will get wet.* This inference is uncertain – indefinitely many additional premises (*the rain is about to stop; I will take an umbrella; there is a crooked walkway*) can overturn the conclusion, even if the premises are correct. The non-monotonicity of everyday inference is problematic for the application of logical methods to modelling thought. Non-monotonic inferences are not logically valid and hence fall outside the scope of standard logical methods. (Oaksford and Chater 2009: 72)

The Bayesian approach solves these problems by treating everyday inference as *probabilistic*, instead of logical. Thus, a rational inference, according to the Bayesian approach, is one where the premises significantly increase the probability of the conclusion. Clearly, this notion of rationality is not monotonic, and so is able to deal naturally with the non-monotonic inferences that are so characteristic of uncertain, everyday reasoning. Furthermore, the Bayesian approach allows us to interpret the behavior that is typically observed for certain reasoning tasks (such as the Wason selection task) as a rational form of probabilistic reasoning, whereas the standard logical approach requires us to interpret such behaviors as irrational (see Oaksford and Chater 2007). The Bayesian approach, in contrast to the logical approach, allows us to think of human beings as being exceptionally good at everyday reasoning.[1]

Here, we employ the probabilistic Bayesian paradigm to provide a novel analysis of the special role of logical validity in uncertain reasoning. In particular, we show that the argumentative norms of classical logic emerge naturally from the probabilistic norms given by the Bayesian approach. This result undermines the received wisdom that the logical and Bayesian perspectives on human reasoning give rise to distinct and incompatible norms. Furthermore, the fact that a general preference for logical validity emerges naturally from the Bayesian approach to argumentation undermines one of the key motivations for dual process theories of human reasoning, i.e. the fact that, ceteris-paribus, people tend to endorse logically valid inferences more readily than they do invalid ones (see e.g. Evans 1983, 1993, 2000, 2003, Singmann *et al.* 2016).

The remainder of the paper is organized as follows. Section 2 outlines the main idea of our proposed theory of Bayesian argumentation, according to which *argumentation is learning*. We show that this theory is consistent with classical logic (in the special case where we learn the premises with certainty) and outline the shortcomings of some existing approaches to argumentation in uncertain reasoning situations. We also motivate a general 'distance based' approach to learning conditional information, which generalises an extant account from the

[1] Of course, this is not the whole story, and there have been many attempts to capture the apparent non-monotonicity of everyday reasoning in a logical framework, for example by appealing to 'default assumptions' (see e.g. Reiter 1978, 1980, Stenning and Van Lambalgen 2008). It is not our aim here to provide a programatic justification of the probabilistic approach to reasoning in general, or to conduct a comprehensive analysis of rival frameworks. For canonical defences of the Bayesian approach to scientific and everyday reasoning, readers should consult e.g. Howson and Urbach (2005) and Oaksford and Chater (2007). For detailed explications of the advantages of a probabilistic as opposed to logical conception of argumentative norms, see e.g. Corner and Hahn (2013), Hahn, Harris and Oaksford (2013), Hahn and Hornikx (2016) and Hahn and Oaksford (2007).

literature and is also compatible with standard Bayesain updating methods, before going on to describe how the Bayesian approach to argumentation relates to experimental results regarding the acceptability of valid and invalid inferences. Section 3 studies what happens to the probability of the conclusion of an argument when we learn the minor premise from an unreliable source, comparing valid and invalid argument schemes, and Section 4 connects our approach to recent work on the non-monotonicity of dynamic conditional inferences. Section 5 considers the role of logical validity in contexts where the major premise is learned with non-extreme probability, and Section 6 then uses the results of previous sections to analyse the value of logical validity in argumentation involving uncertain premises, before showing how our approach improves upon existing results from the literature. In Section 7, we then go on to utilize these formal results to undermine some influential theoretical motivations for dual process models of human cognition. Finally, Section 8 summarizes the main results of this paper and outlines some open questions to be addressed in future research.

2 Argumentation as Learning

The fundamental starting point of our approach can be summarised by the slogan that '*argumentation is learning*'. According to this approach, *the strength of an argument is determined entirely by the extent to which evidence for the premises counts as evidence for the conclusion.* We aim to use this basic slogan to provide a general vindication of the special role played by valid arguments in human reasoning, even in uncertain contexts where logical validity doesn't appear to offer any inferential advantages. Consider, for example, MP

$$\begin{array}{c} \mathrm{A} \rightarrow \mathrm{B} \\ \mathrm{A} \\ \hline \mathrm{B} \end{array}$$

Here, the truth of the conclusion follows from the truth of the premises with necessity.

The argument is *logically valid.* However, we may not *trust the source* who provides us with the information that A, or we may consider a *disabling condition* that prevents the consequent (B) from obtaining when the antecedent (A) is instantiated. In these situations, it is not clear how we should interpret the fact that MP is almost universally accepted as a 'good argument'. Certainly, the mere fact that MP is logically valid is not enough, since logical validity only tells us that it is rational to infer the conclusion when we are certain of the truth of the premises.[2] More generally, it's unclear what these complications imply concerning the rationality of MP inferences in situations where we are uncertain about the truth of the premises.

Here is the main idea of our proposal: First, the agent has to *model* the reasoning situation, i.e. she (i) identifies the relevant variables, and (ii) specifies a prior probability distribution P over those variables that reflects her degrees of belief regarding the respective likelihoods of their truth/falsity. Next, the agent *learns* the premises of the argument from some (possibly only partially reliable) information source. The source may, for example, utter that A is the case or that B follows from A. This new information translates into a *constraint* on the posterior probability distribution P', e.g.

- If the agent learns that A is the case, then $P'(\mathrm{A}) = 1$ (if the source is perfectly reliable).
- If the agent learns that "If B, then C", then $P'(\mathrm{C}|\mathrm{B}) = 1$ (if no disabling conditions are considered in the agent's model and the source is perfectly reliable).

To clarify the motivation here, we take the aim of argumentation to be to *indirectly* support a given conclusion. This is done by providing evidence for the premises of an argument whose structure then forces those considering it to increase their credence in the conclusion. This is exactly what is captured by our slogan that 'argumentation is learning'. To argue is to provide evidence for the premises of arguments whose structure tells us something about the probability of the conclusion. More generally, we take this approach to constitute an accurate

[2] Another way of making the same point is to note that logical norms tell us only that we are allowed to infer the conclusions of *sound* arguments, i.e. valid arguments with true premises. But when we are uncertain about the truth of the premises, logic alone will never be enough to determine the rationality of an argument.

description of the *actual practice* of argumentation. If I want to convince you of a conclusion for which we lack direct evidence, the best way for me to do this is to present you with evidence for the premises of an argument whose structure allows you to transfer some of the weight of that evidence towards the desired conclusion. The success of such an argument will generally depend on a number of factors. In particular, it will depend on (i) the logical form of the argument, (ii) the strength of the evidence given in support of the premises, and (iii) your prior belief state (as encoded in your prior probability distribution).

After learning new information about the premises of the given argument, the agent will need to change their probabilistic beliefs to incorporate the learned information. Here, we argue that they should do this by adopting the posterior probability distribution P' that (i) is consistent with what the agent learned about the premises, and (ii) is obtained by minimizing a particular kind of "distance" measure (specifically, an f-divergence) such as the *Kullback-Leibler divergence* between the new probability distribution P' and the old probability distribution P (see Diaconis and Zabell (1982)). The idea here is that the agent changes her beliefs in a *conservative way*. She makes sure that the learned constraints are satisfied, but apart from this the changes should be as minimal as possible, i.e. the posterior probability distribution P' should be as close as possible to the prior distribution P. Other distance measures (including other f-divergences; see Csiszár (2008)) are also possible, but we consider it to be important that the chosen distance measure is consistent with conditionalization and Jeffrey conditionalization[3] (which holds for all f-divergences; see Csiszár (2008)). Since these normative constraints do not uniquely fix the proper distance measure, fixing the "right" measure is (at least to some extent) also an empirical question whose answer needs to be experimentally and philosophically justified. While this paper focuses on the Kullback-Leibler divergence

[3] Jeffrey conditionalization is the standard Bayesian approach to modelling learning that does not result in certainty, i.e. to modelling situations where an agent becomes more confident in some evidence E without becoming certain of E's truth. Formally, it is defined as follows. Suppose that, via some learning experience, I change my prior degree of belief $P(\mathrm{E})$ in E to a posterior value of $P'(\mathrm{E})$. Then, my posterior degree of belief in any proposition X will be given by $P'(\mathrm{X}) = P(\mathrm{X}|\mathrm{E})\, P'(\mathrm{E}) + P(\mathrm{X}|\neg\mathrm{E})\, P'(\neg\mathrm{E})$. Of course, when $P'(\mathrm{E}) = 1$, this is equivalent to standard Bayesian conditionalization. Jeffrey conditionalization is widely used in the Bayesian canon, and has been given a pragmatic justification by means of dynamic dutch book arguments (see Skyrms 1987).

(i.e. one particular f-divergence), we note that different f-divergences generally give rise to importantly different updating procedures in some of the learning scenarios considered in this paper. Exploring the differences in detail will be the focus of future work.

At this point, one might be tempted to ask what we hope to gain by moving to such a general picture. In particular, what does this approach achieve that is not already done by the standard Bayesian approaches of conditionalization/Jeffrey conditionalization? One major motivation is the following. It is well known that Bayesian conditionalization becomes very problematic when the evidence on which we are updating takes the form of a conditional proposition (see e.g. Popper and Miller 1983, Douven and Dietz 2011, Douven 2012, Eva, Hartmann and Rafiee Rad forthcoming).[4] Indeed, one currently prevalent view holds that the very idea that we should think of conditionals as propositions with determinate semantics is inherently flawed (see e.g. Adams 1975, Eddington 1995, Bennett 2003). Our approach allows one to update on conditional evidence in a very natural way[5] that doesn't require a propositional account of conditionals.[6] Since we are interested in providing a probabilistic characterisation of the value of deductive validity in situations where we have uncertain premises, and arguments typically involve conditional statements, this is crucial for our current purposes (the importance of the distance-based approach is made most apparent in Section 5, where we discuss argumentative contexts in which the major premise is learned with non-extreme

[4] To illustrate some of the problems here, suppose that there is gold in one and only one of four regions of a country: the South-West, the South-East, the North-East or the North-West. If we have no other information about where the gold is, we will assign all four regions equal probability, i.e. $P(\text{SW}) = P(\text{SE}) = P(\text{NW}) = P(\text{NE}) = 1/4$. Intuitively, the probability of the proposition 'if the gold is in the west, then it is in the south-west' is 1/2 in this situation, since SW and NW are equally probable. But now suppose somebody tells us that the probability of this conditional is 1/2. If we represent this learning experience by Jeffrey conditionalizing to set the probability of the material conditional $\text{W} \supset \text{SW}$ to 1/2, we get the disastrous result that our probability distribution will change significantly. But this is absurd, since we already believed that the conditional 'If the gold's in the west, then it's in the southwest' had probability 1/2. We learned something we already knew and so should not have changed our probability distribution at all.

[5] Namely, by minimising the distance to some constraint on the relevant conditional probability.

[6] It is worth pointing out that for the purposes of this paper, we remain entirely agnostic about the propositional status of conditionals. In particular, our approach does not commit us to the position (known as 'Adam's thesis') that the indicative conditional is a proposition whose probability is always given by the corresponding conditional probability (although it is mainly consistent with that position (but see section 5)). We require only that the rational response to learning an indicative conditional is to minimise an f-divergence to a constraint on the relevant conditional probability. This is an important distinction, since it means that our approach is not susceptible to Lewis's (1976) triviality results.

probability).

It should also be noted that the idea that one should always update their credence function as conservatively as possible is not new or controversial. Indeed, this idea is commonly used to motivate mainstream update rules such as conditionalization. The minimization of distance measures is a natural approach to allowing for the consideration of non-propositional evidence in a way that preserves this intuitive justification of update rules like conditionalization.[7] The fact that it is possible to recapture these rules in the special case where the evidence is propositional shows that *there is nothing to lose* by adopting a distance-based approach. But in so far as purely propositional update rules are unable to guide agents in learning experiences that involve non-propositional evidence, there is a great deal to gain. In the present case, we are able to provide a new analysis of the value of logical validity in a way that doesn't rely on the philosophically dubious practice of (Jeffrey) conditionalizing on conditional evidence.

2.1 Existing Approaches

Before illustrating the formal details of our extended Bayesian approach to argumentation, we will briefly consider a couple of existing approaches to probabilistic validity and argument strength.

The most natural Bayesian approach to measuring argument strength is simply to define the strength $S(\mathfrak{A})$ of any argument $\mathfrak{A}$ to be the degree to which $\mathfrak{A}$'s premises confirm its conclusion. But of course, this degree of confirmation will be a function of the conditional probability of the conclusion given the premises, and we've already argued that conditionalizing on conditionals is not a satisfactory approach.

According to Pfeifer (2013), $S(\mathfrak{A})$ should be thought of as a positive function of both (i) the precision of the argument, i.e. how tightly the probabilistic conditions given by the premises constrain the probability of the conclusion, and (ii) the probability of the conclusion, as given

[7] Of course, conditionalization has also been justified by pragmatic Dutch book arguments and arguments from accuracy (see e.g. Pettigrew (2016)). The possibility of extending these kinds of arguments to justify and discriminate between different f-divergences is an important one that we intend to pursue in future work.

by a suitably chosen average over the possible coherent probability values of the conclusion that respect the constraints imposed by the premises. Intuitively, this means that a good argument is one where the premises both give us a lot of information about the probability of the conclusion and give us good reason to assign the conclusion a high posterior probability. This approach bypasses the problem of needing to conditionalize on conditional premises[8] and allows us to quantify the strength of arguments with uncertain premises.

Crucially though, the coherence based approach to argument strength is purely *synchronic*. It measures only the extent to which assigning high probabilities to the premises forces one to assign high probabilities to the conclusion at a particular time. It tells us nothing about how *learning* the premises of an argument forces an agent to change their belief in the conclusion. In contrast, the Bayesian approach to argument strength described here is fundamentally *dynamic*. It concerns the way in which an agent is forced to change their belief in the conclusion when they gain new evidence for the premises.

In particular, we propose to define $S(\mathfrak{A})$ as a positive function of $P'(\mathrm{C}) - P(\mathrm{C})$, i.e. the strength of the argument is just a function of the degree to which learning the premises (with some increased probability, according to the procedure outlined in this paper) increases the probability of the conclusion C. This is essentially identical to the intuitive Bayesian approach of defining $S(\mathfrak{A})$ to be the degree to which the premises confirm the conclusion. The only difference is that the notion of confirmation has been generalised to allow for confirmation by conditional premises.[9] Of course, there are numerous ways to measure the degree to which the premises of an argument confirm the conclusion (see e.g. Fitelson (1999) for a detailed analysis of different Bayesian confirmation measures), the simplest being $S(\mathfrak{A}) = P'(\mathrm{C}) - P(\mathrm{C})$. The question of whether this choice of confirmation measure gives the best account of argument strength will, we contend, only be determined by experimental investigations into the way that people actually evaluate the strength of real arguments. Corner and Hahn (2009) have

[8] Indeed, this is one of the primary motivations cited by Pfeifer (2013).

[9] To be clear, $P'(\mathrm{C})$ here does not denote the conditional probability of C on the premises, but rather the posterior probability of C after we minimise the distance from the original probability distribution relative to some new constraints on the probabilities of the premises.

already used Bayesian methods to interpret the results of experiments studying the perception of argument strength in science and everyday reasoning, so this kind of question connects naturally with current work in the psychology of reasoning (for other experimental work examining adherence to Bayesian norms in argumentation see also: Hahn and Hornikx 2016, Oaksford and Hahn 2004, Hahn and Oaksford 2007, Harris *et al.* 2012 and Harris *et al.* 2016).[10]

Perhaps the most influential approach to probabilistic validity is forwarded by Adams (see e.g. Adams (1998)). Adams derives some useful characterisations of the confirmatory relationship between the premises and conclusions of valid arguments. However, these results are not sufficiently general to constitute a full vindication of the special role played by valid arguments in uncertain contexts. Specifically, Adams' results only apply to the special case in which the premises of the argument are learned with certainty. Furthermore, Adams' approach assumes that the premises of arguments are always propositional which, as we have already noted, is problematic. In Section 6 we show that the approach described here is able to provide a more general, psychologically salient and philosophically principled probabilistic elucidation of the value of logical validity in uncertain reasoning contexts.

2.2 The Simple Case

Let us now illustrate our approach to argumentation by looking at the four inference schemes Modus Ponens (MP), Modus Tollens (MT), Affirming the Consequent (AC), and Denying the Antecedent (DA). We introduce binary propositional variables A and B (in italic script) which have the values A and ¬A, and B and ¬B (in roman script), respectively. Prior to encountering an argument involving the propositions, the agent has beliefs about the propositions in question as well as about their dependencies. These beliefs are represented by a prior probability distribution P, with the parameters

[10] Note that Harris *et al.* (2012, 2016) test detailed quantitative Bayesian models of argument strength, as opposed to simple qualitative predictions.

$$P(\mathrm{A}) = a, \tag{1}$$

$$P(\mathrm{B}|\mathrm{A}) = p \quad , \quad P(\mathrm{B}|\neg\mathrm{A}) = q \tag{2}$$

With this, the prior distribution over the variables A and B is given by

$$\begin{aligned} P(\mathrm{A},\mathrm{B}) &= a\,p \qquad , \qquad P(\mathrm{A},\neg\mathrm{B}) = a\,\overline{p} \\ P(\neg\mathrm{A},\mathrm{B}) &= \overline{a}\,q \qquad , \qquad P(\neg\mathrm{A},\neg\mathrm{B}) = \overline{a}\,\overline{q}\,, \end{aligned} \tag{3}$$

where we have used the shorthand notation $P(\mathrm{A},\mathrm{B})$ for $P(\mathrm{A}\wedge\mathrm{B})$. We also use the shorthand $\overline{x}$ for $1-x$.

Next, the agent learns the premises of the argument, which prompts her to update P and obtain a posterior probability distribution P'.[11] Here we replace the variables a, p and q by the corresponding primed variables a', p' and q', respectively:

$$\begin{aligned} P'(\mathrm{A},\mathrm{B}) &= a'\,p' \qquad , \qquad P'(\mathrm{A},\neg\mathrm{B}) = a'\,\overline{p'} \\ P'(\neg\mathrm{A},\mathrm{B}) &= \overline{a'}\,q' \qquad , \qquad P'(\neg\mathrm{A},\neg\mathrm{B}) = \overline{a'}\,\overline{q'} \end{aligned} \tag{4}$$

P' is constrained by the probabilistic information implied by the premises. To illustrate the proposed analysis, we first consider MP, which has the following two premises:

Premise MP1: A → B. In probabilistic terms, this amounts to $P'(\mathrm{B}|\mathrm{A}) = p' = 1$.

[11] An important technical caveat needs to be made here. The agent's prior probability distribution P will tell us which probabilistic independencies obtain between the variables being considered. These independencies can be represented by a Bayesian network (see e.g. Bovens and Hartmann, 2003). We assume that this Bayesian network stays fixed across the given learning experience, i.e. the agent does not come to learn that some variables they thought to be correlated are in fact independent or vice-versa. This assumption plays no role here (where we consider only arguments involving two propositional variables), but it could play an important role in more complex settings where many propositions are considered simultaneously and the learning experiences are more complex. One might think of the fixed Bayesian network as encoding the agents' beliefs about the causal structure of the considered variables. For Bayesian approaches to argumentation and conditional reasoning that explicitly employ causal structure, see e.g. Ali *et al.* (2010, 2011), Hall *et al.* (2016), Fernbach and Erb (2013) and Oaksford and Chater (2017).

Premise MP2: A. In probabilistic terms, this amounts to $P'(\mathrm{A}) = a' = 1$.

Here we assume that the premises are learned with certainty, and that the agent then considers the effect of the learning process on the posterior probability of the conclusion. Thus, this case is closely related to reasoning tasks in which the participant is asked to evaluate a modus ponens argument from a purely deductive perspective, ignoring considerations about the actual believability of the premises and the conclusion. This is the kind of reasoning task that is typically associated with the 'deductive' or 'old' paradigm in the psychology of reasoning (see e.g. Evans 2002), where participants are often explicitly instructed only to draw conclusions that are logically entailed by the relevant premises. With this setup, the following result holds (all proofs in supplemental online materials):

Proposition 1 *An agent considers the propositions* A *and* B *and has a prior probability distribution* P *according to eqs. (3) defined over them. Learning the premises* **MP1** *and* **MP2** *then implies that the new probability of* B*, i.e.* $P'(\mathrm{B})$*, equals* 1.

This is in accordance with the fact that B follows deductively from **MP1** and **MP2**. The proposed procedure is consistent with classical logic.

Note that our premise **MP1** does not require that we represent the conditional as a proposition. We interpret the learning of the conditional A → B as a simple constraint on the conditional probability of B given A. We do not assume that any propositional representation of the conditional is possible. Next, we examine MT, i.e.

$$\begin{array}{c} \mathrm{A} \rightarrow \mathrm{B} \\ \neg \mathrm{B} \\ \hline \neg \mathrm{A} \end{array}$$

Here we encounter the following two premises:

Premise MT1: A → B. In probabilistic terms, this amounts to $P'(\mathrm{B}|\mathrm{A}) = p' = 1$.

Premise MT2: $\neg$B. In probabilistic terms, this amounts to $P'(\mathrm{B}) = 0$.

As expected, we can establish:

Proposition 2 *An agent considers the propositions* A *and* B *and has a a prior probability distribution P according to eqs. (3) defined over them. Learning the premises* **MT1** *and* **MT2** *then implies that the new probability of* $\neg$A*, i.e.* $P'(\neg\mathrm{A})$*, equals* 1.

Again, our procedure gives the right result, which encourages us to study arguments that are not deductively valid but that may nevertheless have some strength. To do so, let us first consider AC, i.e.

$$\begin{array}{c} \mathrm{A} \to \mathrm{B} \\ \mathrm{B} \\ \hline \mathrm{A} \end{array}$$

This argument has the following two premises:

Premise AC1: A → B. In probabilistic terms, this amounts to $P'(\mathrm{B}|\mathrm{A}) = p' = 1$.

Premise AC2: B. In probabilistic terms, this amounts to $P'(\mathrm{B}) = 1$.

Proposition 3 *An agent considers the propositions* A *and* B *and has a a prior probability distribution P according to eqs. (3) defined over them. Learning the premises* **AC1** *and* **AC2** *(and minimizing the Kullback-Leibler divergence between P' and P) then implies that the new probability of the antecedent* A*, i.e.* $P'(\mathrm{A})$*, equals* $P(\mathrm{A}|\mathrm{B})$.

As we want to infer the proposition A here, we conclude that AC is a good argument if the conditional probability $P(\mathrm{A}|\mathrm{B})$ is large.[12] This result is highly intuitive. It captures the

[12] An important clarification is needed here. When we say that an argument is successful or persuasive, we mean that the probability of the conclusion increases across the learning experience. The strength of the argument is given by the magnitude of this increase. Thus, an argument may have low strength even when the posterior probability of the conclusion is very high, i.e. if the prior probability was already high.

natural idea that AC is 'missing a premise', i.e. B → A, and the strength of the argument corresponds exactly to our confidence in this extra premise.

Finally, we study DA, i.e.

$$\begin{array}{c} \mathrm{A} \rightarrow \mathrm{B} \\ \neg\mathrm{A} \\ \hline \neg\mathrm{B} \end{array}$$

This argument has the following two premises:

Premise DA1: A → B. In probabilistic terms, this amounts to $P'(\mathrm{B}|\mathrm{A}) = p' = 1$.

Premise DA2: ¬A. In probabilistic terms, this amounts to $P'(\mathrm{A}) = a' = 0$.

Proposition 4 *An agent considers the propositions* A *and* B *and has a a prior probability distribution P according to eqs. (3) defined over them. Learning the premises* **DA1** *and* **DA2** *(and minimizing the Kullback-Leibler divergence between P' and P) then implies that the new probability of* ¬B, *i.e.* $P'(\neg\mathrm{B})$, *equals* $P(\neg\mathrm{B}|\neg\mathrm{A})$.

As we want to infer the proposition ¬B here, we conclude that DA is a good argument if the conditional probability $P(\neg\mathrm{B}|\neg\mathrm{A})$ is large. Again, this is a very natural result that captures our intuitions about when DA will be a good argument.[13]

Before moving on, it is worth offering a few final clarifying remarks about our approach. With regards to the slogan, '*argumentation is learning*', the central idea is that a reliable argument is one for which it is, in a certain sense, impossible to learn the premises without learning the conclusion. In the usual case where the premises are learned with certainty,

[13] Oaksford *et al.* (2000) conducted experiments where people are asked whether they endorse each of the four inference rules MP, MT, AC, and DA. They found that while nearly all respondents were entirely happy to endorse MP, only around 70% endorsed MT. Surprisingly, the invalid schemes DA and AC were endorsed by over 50% of respondents. This is an instance of a general phenomena, whereby naive reasoners tend to unanimously accept MP as a good inference. MT generally has much lower acceptance rates (although typically still above 50%), while AC and DA are quite often accepted as good inferences, with the acceptance rate of AC sometimes reaching that of MT (see e.g. Evans, 1993).

logically valid arguments are the only ones that satisfy this criterion. But in cases of 'ineffable learning' where the agent becomes more confident of the premises without learning anything for certain, the situation is rather more complicated. Our fundamental aim is to characterise different types of arguments by studying what happens to the probability of the conclusion when agents learn the premises in this more general way. If it is always the case that increasing the probability of the premises necessitates an increase in the probability of the conclusion, then the argument being considered is clearly reliable in the sense that any evidence for the premises automatically counts as evidence for the conclusion. Note that this is a property not of individual instances of arguments, but rather of abstract argument schemes. This distinction is an important one, especially when it comes to the notion of argument strength. The idea is that a particular instance of an argument is given by a particular agent with a prior probability distribution who then obtains new evidence for the premises of the argument and revises their probability estimates in line with that evidence to obtain a new probability for the conclusion. The argument strength is given by the change in the probability of the conclusion, and this change will be a function not just of the argument's logical structure, but also of the prior probability distribution and the new evidence. A 'reliable' argument scheme for which evidence for the premises always counts as evidence for the conclusion will always have positive argument strength, regardless of the prior distribution and the strength of the new evidence.

Thus, when we say that 'argumentation is learning', we mean that the strength of any particular instance of an argument form cannot be calculated without specifying both the prior distribution of the agent considering the argument and the nature of the new evidence obtained by that agent. Although the logical form can guarantee that the argument will never have a negative effect, regardless of the agent's prior distribution and the strength of the evidence, the actual success of the argument can only be precisely quantified relative to a particular learning scenario.

So far, we've only considered the special case where the premises are learned with certainty.

In the following sections, we consider the more general cases where the premises are learned with non-extreme probability (for example because the source from which the premises are learned is not perfectly reliable). There are a number of reasons to care about these more general cases. Firstly, as we stressed in the introduction, the vast majority of argumentation that occurs in science and everyday life is argumentation *under uncertainty*. Typically, people make arguments based on premises which are only probably true (at best), and a theory that applies only to arguments with certain premises will be of very limited interest in this regard. Secondly, it is unrealistic to suppose that people generally respond to learning conditionals (for example) by setting the relevant conditional probability to 1, regardless of their prior epistemic situation. By relaxing this kind of assumption, we are likely to be able to obtain a much more plausible description of actual reasoning practices. For example, Stevenson and Over (2001) conducted experiments in which participants were asked to evaluate prospective MP and MT arguments with the major premise 'If Bill has typhoid then he will make a quick recovery'. They found that the acceptability of the argument depended importantly on the reliability of the source that informed the participants of the major premise. The participants endorsed the inferences more readily if the source was a professor of medicine, compared to when the source was a first year medical student (related effects are discussed by e.g. Singmann and Klauer 2011). This clearly looks like a case where the extent to which the participants were convinced by the presented evidence for the major premise had a significant effect on the perceived strength of the argument. If we hope to plausibly account for effects like these, then we need to allow for the possibility that agents are only partially convinced by the presented evidence for the premises of the relevant argument (for an experimental examination of the role of source reliability in argumentation within a probabilistic framework, see also Hahn, Harris and Corner, 2009). We turn now to considering the role of logical validity in this more general setting.

3 Uncertain Minor Premises

In Section 2, we saw that the Bayesian approach gives intuitively correct results in the special case where we learn the premises with certainty, i.e. it allows us to infer the conclusion with certainty for valid schemes, but not for invalid ones. Let us now consider the more general case where the minor premise of the argument becomes more likely, but remains uncertain. That is, the new probability of the minor premise goes up, but it does not go up to 1. In that case, the probability of the conclusion of the argument can be smaller than 1 even if the underlying argument pattern is valid.

This raises the question 'what is the value of a logically valid argument in the light of uncertain premises?'. Is the use of a logically valid argument pattern in some sense better than using an argument pattern that is not logically valid? And if so, in which sense? This will be the guiding question of this section.

We consider again the prior probability distribution in eqs. (3). We then learn two premises of the given argument. As above, the posterior probability distribution P' is given by eqs. (4). The additional constraints either fix one of the parameters (a', p' or q') directly or some function of these parameters. In the latter case, we add a corresponding constraint to the respective Kullback-Leibler divergence, using the method of Lagrange multipliers. Without the constraints, the Kullback-Leibler divergence for the cases we consider is given by

$$D^0_{KL}(P'||P) = \Phi_a + a'\,\Phi_p + \overline{a'}\,\Phi_q. \tag{5}$$

with

$$\Phi_x := x'\,\log\frac{x'}{x} + \overline{x'}\,\log\frac{\overline{x'}}{\overline{x}}\,. \tag{6}$$

The Kullback-Leibler divergence of P' and P is also known as the 'relative entropy' between the two distributions. Intuitively, it measures how much information is lost when we try to approximate P' using P.[14] The smaller it is, the better the approximation.

Let us examine MP first. Here we learn the following two premises:

Premise MP1: A → B. In probabilistic terms, this amounts to $P'(\mathrm{B}|\mathrm{A}) = p' = 1$.

Premise MP2: A. In probabilistic terms, this amounts to $P'(\mathrm{A}) = a' > a$.

Then the following proposition holds:

Proposition 5 *An agent considers the propositions* A *and* B *and has a prior probability distribution P according to eqs. (3) defined over them. Learning the premises* **MP1** *and* **MP2** *(and minimizing the Kullback-Leibler divergence between P' and P) then implies that the new probability of* B*, i.e. $P'(\mathrm{B})$, is always greater than the prior probability $P(\mathrm{B})$.*

Hence, if we make a MP argument with uncertain premises, then the probability of the conclusion increases if the probability of the minor premise increases. Intuitively, this tells us that MP is fundamentally reliable, even in situations involving uncertain premises.

Let us now study DA. Here we learn

Premise DA1: A → B. In probabilistic terms, this amounts to $P'(\mathrm{B}|\mathrm{A}) = p' = 1$.

Premise DA2: ¬A. In probabilistic terms, this amounts to $P'(\neg\mathrm{A}) > P(\neg\mathrm{A})$ and hence $a' < a$.

Proposition 6 *An agent considers the propositions* A *and* B *and has a prior probability distribution P according to eqs. (3) defined over them. Learning the premises* **DA1** *and* **DA2** *(and minimizing the Kullback-Leibler divergence between P' and P) then implies that the new probability of* ¬B*, i.e. $P'(\neg\mathrm{B})$, is greater than the prior probability $P(\neg\mathrm{B})$ iff $a\,\overline{p} + (a' - a)\,\overline{q} < 0$.*

Note that $a\,\overline{p}$ is always positive while $(a' - a)\,\overline{q}$ is always negative as now $a' < a$. Hence the sum of both may be positive or negative and hence the probability of ¬B may go up or

[14] Note that despite being widely described as a 'probability distance measure' the KL divergence is actually not a metric, since it is not a commutative.

down as a result of a DA argument, depending on the context (i.e. on the prior probability distribution). Thus, DA, unlike MP, is not *reliable* in the sense that evidence for the premises of the argument need not count as evidence for the conclusion. Note, however, that our approach doesn't simply categorise DA as a 'fallacious' argument scheme. Rather, it provides us with precise conditions for when the argument can be used to provide legitimate support for the conclusion.

Next, we study MT. Here we learn

Premise MT1: $\mathrm{A} \rightarrow \mathrm{B}$. In probabilistic terms, this amounts to $P'(\mathrm{B}|\mathrm{A}) = p' = 1$.

Premise MT2: $\neg\mathrm{B}$. In probabilistic terms, this amounts to $P'(\neg\mathrm{B}) \geq P(\neg\mathrm{B})$.

Proposition 7 *An agent considers the propositions* A *and* B *and has a prior probability distribution P according to eqs. (3) defined over them. Learning the premises* **MT1** *and* **MT2** *(and minimizing the Kullback-Leibler divergence between P' and P) then implies that the new probability of* $\neg\mathrm{A}$*, i.e. $P'(\neg\mathrm{A})$, is always greater than the prior probability $P(\neg\mathrm{A})$.*

So MT, like MP, is characterized as being 'reliable' by our approach, in the sense that evidence for the premises always counts as evidence for the conclusion. Finally, we consider AC. Here we learn

Premise AC1: $\mathrm{A} \rightarrow \mathrm{B}$. In probabilistic terms, this amounts to $P'(\mathrm{B}|\mathrm{A}) = p' = 1$.

Premise AC2: B. In probabilistic terms, this amounts to $P'(\mathrm{B}) \geq P(\mathrm{B})$.

Proposition 8 *An agent considers the propositions* A *and* B *and has a prior probability distribution P according to eqs. (3) defined over them. Learning the premises* **AC1** *and* **AC2** *(and minimizing the Kullback-Leibler divergence between P' and P) then implies that the new probability of* A*, i.e. $P'(\mathrm{A})$, is greater than the prior probability $P(\mathrm{A})$ iff $a\,p - \delta\,P(\mathrm{A}|\mathrm{B}) < 0$ with $\delta := P'(\mathrm{B}) - P(\mathrm{B}) > 0$.*

Again, the invalid argument scheme has been characterized as unreliable. We have specific conditions for when the argument can be legitimately used to support the conclusion, but it is generally possible to obtain evidence for the premises that doesn't count as evidence for the conclusion.

In closing this section, we show that the results of Propositions 5 to 8 also hold if the material conditional is used instead of minimizing the Kullback-Leibler divergence.

Proposition 9 *The results of Propositions 5 to 8 also hold if one conditionalizes on the material conditional* $\neg\mathrm{A} \vee \mathrm{B}$ *and Jeffrey-conditionalizes on the corresponding minor premise.*

Hence, Propositions 1 to 4 also hold for the material conditional (cf. Suppes (1966)). So it seems that one obtains the intuitively correct verdicts for the reliability of valid and invalid arguments in argumentative contexts where the major premise is learned with certainty regardless of whether one represents the learning of conditional premises according to the distance-based approach or by simply conditioning on the material conditional. In Section 5 we turn to considering more general argumentative contexts where the way in which one represents the learning of conditional premises makes an important difference.

4 Dynamic Non-Monotonicity and Rigidity

At this stage, it is instructive to connect the Bayesian approach to argumentation described here to some existing work on the relationship between non-monotonic inferences and dynamic reasoning with conditionals. As we noted in section 2, the non-monotonicity of everyday inference constitutes one of the most powerful motivations for the Bayesian paradigm in the psychology of reasoning. It turns out that the dynamic nature of our Bayesian theory of argumentation is particularly well suited to analysing the role of non-monotonicity in human reasoning. Oaksford and Chater themselves note that 'a dynamical approach is required to deal with the central outstanding problems of providing a psychological theory of everyday conditional inference'. (Oaksford and Chater 2013: 348)

The following example (from Adams 1998) provides an illustrative starting point. Suppose that an agent has reason to believe with certainty that a given student either studies psychology (A) or quantity surveying (B), i.e. they are certain that A $\vee$ B is true. If they subsequently learn with certainty that $\neg$A is true (the student does not study psychology), they will be able to apply the classically valid inference of disjunctive syllogism to conclude that B is true, i.e. the student studies quantity surveying. The situation is of course more complicated when uncertainty is introduced into the picture. Specifically, suppose that the agent, instead of being certain of the truth of the disjunction A $\vee$ B, rather has a prior probability distribution given by $P(\mathrm{A} \wedge \mathrm{B}) = 0.5$, $P(\mathrm{A} \wedge \neg\mathrm{B}) = 0.4$, $P(\neg\mathrm{A} \wedge \mathrm{B}) = 0.01$, $P(\neg\mathrm{A} \wedge \neg\mathrm{B}) = 0.09$. Then their prior belief in the disjunction will be high, $P(\mathrm{A} \vee \mathrm{B}) = 0.91$. However, it is easy to see that $P(\mathrm{A} \vee \mathrm{B}|\neg\mathrm{A}) = 0.1 = P(\mathrm{B}|\neg\mathrm{A})$, and so the agent's posterior belief in B after learning $\neg$A will be low (much lower than the prior value $P(\mathrm{B}) = 0.51$). So when the agent learns $\neg$A, one of the premises of the disjunctive syllogism inference, she will subsequently become *less* confident of the conclusion of that inference. As Oaksford and Chater put it 'learning that the student does not study psychology makes it *less likely* that she studies quantity surveying. This is *non-monotonic* reasoning behaviour, i.e., learning new information has potentially lost a conclusion that was available before this information was lost' (Oaksford and Chater 2013: 349). One way of diagnosing the source of this non-monotonicity is to note that after learning $\neg$A, the agent becomes less confident of the disjunctive premise A $\vee$ B. Most of the formal results in this paper assume that the learning experience does not cause the agent to become less confident in any of the premises of the relevant argument. But of course, we need to say something about the more general situation where the learning experience leads to non-monotonic reasoning of the type illustrated by the previous example.

The first thing to note is that the monotonicity or non-monotonicity of an inference is not generally indicative of its strength. For example, suppose that an agent gains evidence for the major premise of an MP inference, A $\rightarrow$ B and is unsure how the learned evidence affects the probability of the minor premise A. This prompts them to raise the conditional probability $P(\mathrm{B}|\mathrm{A})$ while leaving the probability of A unconstrained. Typically, this will result in the probability of A decreasing and the probability of B increasing, i.e. the probability of the conclusion increases over the update while the probability of the minor premise decreases.[15] So the mere fact that the learning experience undermines one of the premises does not imply that the argument is unsuccessful. It can still have significant strength, although there is no guarantee that will always be the case.[16]

Another important issue concerns the relationship between non-monotonicity and the Bayesian 'rigidity' assumption. Consider a dynamic MP inference where the minor premise is learned by Bayesian conditionalization. Recall that this learning method satisfies the so called 'rigidity assumption' according to which the conditional probability of the conclusion given the minor premise remains unchanged over the update. Psychologically, this means that the degree of belief that agents assign to B after learning A is equal to their prior degree of belief in B *under the assumption* that A is true, i.e. that $P(\mathrm{B}|\mathrm{A}) = P'(\mathrm{B})$, where P' is the posterior distribution after learning A. However, Zhao *et al.* (2012) performed experiments in which the probabilities that participants assigned to B under the assumption that A is true were different to the probability they assigned to B when they actually *learned* that A is true. Supposing the truth of A was observed to generally have a lower impact on participant's estimation of the truth of B than actually learning the truth of A, i.e. $|P(\mathrm{B}|\mathrm{A}) - P(\mathrm{B})| < |P'(\mathrm{B}) - P(\mathrm{A})|$. In a sense this is unsurprising. For, as Zhao *et al.* write,

> Suppose that lions are discovered roaming your neighborhood; can you anticipate the probabilities you would attach to other events if such startling circumstances actually came to pass? (Zhao *et al.* 2012: 377)

This all suggests that models of conditional inference of the type proposed by e.g. Oaksford *et al.* (2000) may rest on a psychologically unrealistic assumption (rigidity). Oaksford and Chater (2007) took this limitation into account and considered a model where the new

[15] Oaksford and Chater (2013) consider a more radical illustration of this phenomenon. Specifically, they note that learning the conclusion of an MT inference (¬B) will typically lead an agent to completely reject the major premise (A → B) if they update by Bayesian conditionalization, since $\mathrm{P}(\mathrm{B}|\neg\mathrm{B},\mathrm{A}) = 0$. But if the prior value of $\mathrm{P}(\mathrm{B}|\mathrm{A})$ is high, the inference will still be strong in the sense that the probability of the conclusion (¬A) will increase.

[16] This is in contrast to the case we focus on in this paper, where it is assumed that the learning experience does not undermine the premises, and one can give guarantees about the success of valid arguments.

conditional probability value $P'(\mathrm{B}|\mathrm{A})$ obtained after updating on the minor premise is used in the update, instead of the original conditional probability $P(\mathrm{B}|\mathrm{A})$.[17] The new model outperformed its predecessor when it came to predicting the acceptance rates of MP, MT, DA and AC inferences (see Oaksford and Chater 2007, 2009), which suggests that a psychologically plausible model of conditional inference will need to take possible rigidity violations into account. Happily, the distance based Bayesian approach to argumentation described here is perfectly able to model these violations. For, one can always impose as a constraint that the posterior conditional probability $P'(\mathrm{B}|\mathrm{A})$ takes a psychologically plausible value with respect to the prior distribution and the learned information, so that the relevant update is no longer simple conditionalization. In more complicated argumentative scenarios, this will have a significant impact (which will depend importantly on the prior distribution and which f-divergence is used) on the strength of the relevant argument. Furthermore, this approach opens up an entirely new perspective on non-monotonic dynamic conditional inference. For, consider again a standard MP inference such as 'If the key is turned, the car starts. The key is turned. Therefore the car starts'. Given Zhao *et al.* (2012)'s results, it is perfectly plausible that people could reason in such a way that their conditional credence $P(\text{the car starts}|\text{the key is turned})$ is less than the credence that they would assign to the car starting after they actually learned that the key has been turned.[18] If we took this into account by imposing the constraint that the posterior probability of the car starting after learning that the key is turned is less than $P(\text{the car starts}|\text{the key is turned})$, we would be modelling a new form of non-monotonicity where learning the minor premise of the inference interferes with belief in the major premise. This form of non-monotonicity is entirely different to the well known 'strengthening of the antecedent' and can only be modelled dynamically using the kinds of distance minimization methods proposed here. Thus, our extended Bayesian theory of argumentation is able to

[17] Actually, the new model assumes that the update is rigid in the case of MP, but that rigidity is violated in this way for MT, AC and DA.

[18] Zhao *et al.* discuss mainly cases where the posterior degree of belief in the consequent after learning the antecedent is greater than the corresponding prior conditional probability, but of course this implies that there will also be propositions (e.g. the negation of the consequent) whose posterior probability is less than the original conditional probability.

provide novel models of non-monotonic conditional inference that do not rely on the psychologically implausible condition of universal rigidity.

5 Uncertain Major Premises

In Section 3, we saw that valid arguments seem to be distinguished by their reliability in situations where we learn the minor premise with non-extreme probability. It is natural to wonder whether the same is true in situations where the we learn the major premise with non-extreme probability. For example, it might be that the agent learns the conditional 'If A then B' from a partially reliable source or is aware of the possibility of disabling conditions (see e.g. Over and Stevenson 2001, Singmann and Klauer 2011). In cases like this, the agent will generally be inclined to increase the conditional probability $P(\mathrm{B}|\mathrm{A})$, but not all the way to 1. A number of new technical and conceptual subtleties arise in this kind of situation.

Firstly, note the conditional $\mathrm{A} \to \mathrm{B}$ is (according to most accounts) logically equivalent to the contrapositive form $\neg\mathrm{B} \to \neg\mathrm{A}$. Until now this equivalence has been perfectly preserved since $P(\mathrm{B}|\mathrm{A}) = 1$ if and only if $P(\neg\mathrm{A}|\neg\mathrm{B}) = 1$. It makes no difference whether one imposes the constraint $P(\mathrm{B}|\mathrm{A}) = 1$ or $P(\neg\mathrm{A}|\neg\mathrm{B}) = 1$ since they are equivalent. Thus all the results so far hold regardless of whether we represent the conditional in the major premise as $\mathrm{A} \to \mathrm{B}$ or $\neg\mathrm{B} \to \neg\mathrm{A}$. However, things change when we consider the case where the conditional is learned with non-extreme probability. For, the conditional probabilities $P(\mathrm{B}|\mathrm{A})$ and $P(\neg\mathrm{A}|\neg\mathrm{B})$ are generally very different. So, when an agent learns the conditional 'If A then B' from a partially reliable source, they have a choice about which conditional probability constraint to impose. They can either increase $P(\mathrm{B}|\mathrm{A})$ or $P(\neg\mathrm{A}|\neg\mathrm{B})$, and they will generally obtain different constraints on their posterior distribution depending on what they choose. This means that in the uncertain argumentative context where the agent learns the major premise with non-extreme probability, the argument forms we've been considering actually admit of two distinct interpretations: one where the major premise imposes a constraint on $P(\mathrm{B}|\mathrm{A})$

and one where it imposes a constraint on $P(\neg\mathrm{A}|\neg\mathrm{B})$. We will shortly see that the choice of conditional probability constraint has significant implications for the reliability of the relevant argument.

The first natural setting to consider is where the agent learns the minor premise with certainty but learns the major premise with non-extreme probability. In this case, modus ponens has the following premises:

Premise MP1: A → B. In probabilistic terms, this amounts to $P'(\mathrm{B}|\mathrm{A}) > P(\mathrm{B}|\mathrm{A})$.

Premise MP2: A. In probabilistic terms, this amounts to $P'(\mathrm{A}) = 1$.

By analogy with the results of Section 3, one would expect that the constraints implied by **MP1** and **MP2** would guarantee that $P'(\mathrm{B}) \geq P(\mathrm{B})$, which would ensure the reliability of MP in this kind of argumentative context. Surprisingly, it turns out that the probability of B can go up, down or stay the same, depending on the details of the prior distribution.[19] So logical validity *does not* ensure reliability in argumentative contexts where the agent learns the minor premise with certainty but only learns the major premise with non-extreme probability.

The fact that logical validity is not sufficient for reliability in this particular context does not mean that logical validity is generally useless in argumentative contexts where we learn the major premise with non-extreme probability. We turn now to considering contexts where the agent learns the major premise with some increased non-maximal probability but *learns nothing about the minor premise.* First, consider again MP. In these argumentative contexts, modus ponens takes the following form:

Premise MP1: A → B. In probabilistic terms, this amounts to either (i) $P'(\mathrm{B}|\mathrm{A}) \geq P(\mathrm{B}|\mathrm{A})$, or (ii) $P'(\neg\mathrm{A}|\neg\mathrm{B}) \geq P(\neg\mathrm{A}|\neg\mathrm{B})$.

[19] This is easy to see: **MP1** and **MP2** imply that $P'(\mathrm{B}) = p'$. Hence, using eqs. (3), we find that $P'(\mathrm{B}) - P(\mathrm{B}) = a\,(p' - p) + \overline{a}\,(p' - q)$, which can be negative if $p < p' < q$. Specifically, if (i) A has a low prior probability, (ii) there is a strong negative correlation between A and B, and (iii) the increase in the conditional probability $P(\mathrm{B}|\mathrm{A})$ is small, it is possible that increasing **MP1** and **MP2** can lead to a decrease in $P(\mathrm{B})$. This makes intuitive sense. For, if A and B are strongly anti-correlated and we increase the probability of A, that will of course lead to a decrease in the probability of B. If the corresponding increase in $P(\mathrm{B}|\mathrm{A})$ is sufficiently small, it won't be sufficient to overcome the effect created by the negative correlation in the prior distribution.

Premise MP2: A. In probabilistic terms, this amounts to $P'(\mathrm{A}) = P(\mathrm{A})$.

The constraint implied by **MP2** represents the fact that nothing is learned about the minor premise. Note that **MP1** now has two possible interpretations, one corresponding to the standard conditional form, one corresponding to the contrapositive.

Proposition 10 *An agent considers the propositions* A *and* B *and has a prior probability distribution P according to eqs. (3) defined over them. Learning the premises* **MP1**(*i*) *and* **MP2** *(and minimizing the Kullback-Leibler divergence between P' and P) then implies that the new probability of* B*, i.e. $P'(\mathrm{B})$, is always greater than the prior probability $P(\mathrm{B})$. However, if the agent learns* **MP1**(*ii*) *rather than* **MP1**(*i*)*, $P'(\mathrm{B}) < P(\mathrm{B})$ can hold.*

Thus, modus ponens is reliable in the argumentative context where the major premise is learned with non-maximal probability (and nothing is learned about the minor premise), *when the conditional is given a non-contrapositive reading.* If the conditional is interpreted in a contrapositive form, MP will not be reliable in such a setting. So the success and reliability of argument schemes in contexts where the major premise is uncertain depends crucially on whether or not one adopts a contrapositive reading of the conditional. These considerations simply don't arise in the special case where the major premise is learned with certainty.

Let us now turn to MT. In this case, we learn:

Premise MT1: A → B. In probabilistic terms, this amounts to either (i) $P'(\mathrm{B}|\mathrm{A}) \geq P(\mathrm{B}|\mathrm{A})$, or (ii) $P'(\neg\mathrm{A}|\neg\mathrm{B}) \geq P(\neg\mathrm{A}|\neg\mathrm{B})$.

Premise MT2: ¬B. In probabilistic terms, this amounts to $P'(\neg\mathrm{B}) = P(\neg\mathrm{B})$.

Proposition 11 *An agent considers the propositions* A *and* B *and has a prior probability distribution P according to eqs. (3) defined over them. Learning the premises* **MT1**(*ii*) *and* **MT2** *(and minimizing the Kullback-Leibler divergence between P' and P) then implies that the new probability of* ¬A*, i.e. $P'(\neg\mathrm{A})$, is always greater than the prior probability $P(\neg\mathrm{A})$. However, if the agent learns* **MT1**(*i*) *rather than* **MT1**(*ii*)*, $P'(\neg\mathrm{A}) < P(\neg\mathrm{A})$ can hold.*

So MT is exactly analogous to MP in these kinds of argumentative contexts, i.e. it is a reliable argument scheme *under one interpretation of the conditional probability constraint given by the major premise.* Specifically, MT is reliable when one adopts the contrapositive interpretation of the major premise. Next, consider the invalid scheme AC. Here we learn:

Premise AC1: A → B. In probabilistic terms, this amounts to either (i) $P'(\mathrm{B}|\mathrm{A}) \geq P(\mathrm{B}|\mathrm{A})$, or (ii) $P'(\neg\mathrm{A}|\neg\mathrm{B}) \geq P(\neg\mathrm{A}|\neg\mathrm{B})$.

Premise AC2: B. In probabilistic terms, this amounts to $P'(\mathrm{B}) = P(\mathrm{B})$.

Proposition 12 *An agent considers the propositions* A *and* B *and has a prior probability distribution P according to eqs. (3) defined over them. Learning the premises* **AC1**(*i*) *(or* **AC1**(*ii*)*) and* **AC2** *(and minimizing the Kullback-Leibler divergence between P' and P) then allows for the possibility that the new probability of* A*, i.e. $P'(\mathrm{A})$, can be less than the prior probability $P(\mathrm{A})$.*

So AC is unreliable in these argumentative contexts, *under either interpretation of the major premise.* This suggests an important a-symmetry between valid and invalid arguments in contexts where the major premise is learned with non-extreme probability (and nothing is learned about the minor premise). Specifically, we conjecture that, for any valid argument, there is at least one legitimate construal under which the argument is reliable in these contexts. The same is not true for invalid arguments, which will be unreliable under any construal of the major premise.

Finally, consider DA. Here, we learn

Premise DA1: A → B. In probabilistic terms, this amounts to either (i) $P'(\mathrm{B}|\mathrm{A}) \geq P(\mathrm{B}|\mathrm{A})$, or (ii) $P'(\neg\mathrm{A}|\neg\mathrm{B}) \geq P(\neg\mathrm{A}|\neg\mathrm{B})$.

Premise DA2: ¬A. In probabilistic terms, this amounts to $P'(\neg\mathrm{A}) = P(\neg\mathrm{A})$.

Proposition 13 *An agent considers the propositions* A *and* B *and has a prior probability distribution P according to eqs. (3) defined over them. Learning the premises* **DA1**(*i*) *(or*

DA1(*ii*)*) and* **DA2** *(and minimizing the Kullback-Leibler divergence between* P' *and* P*) then allows for the possibility that the new probability of* ¬B*, i.e.* $P'(\neg\mathrm{B})$*, can be less than the prior probability* $P(\neg\mathrm{B})$*.*

Again, this is in line with the conjecture. The invalid scheme DA is unreliable under either interpretation of the major premise. So it seems that valid arguments can be distinguished from their invalid counterparts in contexts where the major premise is uncertain.

It is also worth noting that this analysis would be impossible if we represented the change in the probability of the major premise by Jeffrey conditionalizing on the material conditional. For, Jeffrey conditionalizing on the material conditional will generally affect the posterior probability of both the antecedent and the consequent. In the argumentative contexts that we are considering, the agent learns *nothing* about the plausibility of the minor premise, and this is represented by the constraint that the probability of the minor premise stays fixed across the update. Such a constraint is generally incompatible with increasing the probability of the material conditional via Jeffrey conditionalization. Furthermore, the material conditional does not allow for any distinction between the contrapositive and non-contrapositive forms. Thus a proper analysis of logical validity in situations where the major premise is uncertain clearly requires a distance-based approach to probabilistic updating of the type advocated here.

The idea, commonly referred to as 'Adams' thesis' that we should interpret a conditional A → B as imposing a constraint on the conditional probability $P(\mathrm{B}|\mathrm{A})$ has been hugely influential in logic, philosophy, and the psychology of reasoning. However, the results presented here suggest that there exist argumentative contexts in which the major premise A → B is more naturally interpreted as imposing a constraint on the contrapositive conditional probability $P(\neg\mathrm{A}|\neg\mathrm{B})$. Thus, if we want to maintain both Adam's thesis and the position that MT is reliable in the same argumentative contexts as MP, we are forced to conclude that agents interpret the indicative conditional contrapositively when making MT style inferences.[20] We

[20] We conjecture that this insight may be useful in responding to some of the recent criticisms of Adam's

leave the detailed philosophical analysis of this suggestive idea for another day.[21]

6 The Value of Logical Validity

Let us summarize what we have shown so far. First, in Section 3 we showed the following.

Theorem 1 (Summary) *An agent considers the premises* A *and* B *and entertains a prior probability distribution P according to eqs. (3) defined over them. If the probability of the minor premise increases and the probability of the major premise goes to* 1*, then the new probability of the conclusion will always be greater than the prior probability of the conclusion for logically valid argument schemes MP and MT. For the logically invalid argument schemes AC and DA, the probability of the conclusion can be smaller or larger than the corresponding prior probability.*

This suggests the following conjecture.

Conjecture 1: *For any valid argument scheme, increasing the probability of at least one minor premise, ensuring that the probability of no other minor premise goes down and imposing the conditional probability constraints implied by the major premises guarantees that the probability of the conclusion will increase. For logically invalid argument schemes, the conclusion can either increase or decrease in probability, depending on the prior distribution.*

However, it is not hard to see that Conjecture 1 is false. Consider the simple valid argument forms of disjunction introduction and conjunction elimination. These both satisfy Conjecture

thesis in the literature (see e.g. Douven 2017, Douven and Verbrugge 2010, 2012, 2013, Skovgaard Olsen *et al.* 2016). For instance, one powerful criticism of Adams' thesis is that it implies that the conditional has a high probability/acceptability when $P(\mathrm{B}|\mathrm{A})$ is high but sill less than $P(\mathrm{B}|\mathrm{A})$, i.e. when B has a high prior probability that is slightly reduced by conditioning on A. In this case A is negatively relevant for B, i.e. it makes it more likely to be false, but Adams' thesis still requires that the conditional 'If A then B' has high probability, which seems strange. One might be able to solve this problem by arguing that the probability of the conditional goes not by $P(\mathrm{B}|\mathrm{A})$ but rather by $P(\neg\mathrm{A}|\neg\mathrm{B})$ (which will typically be lower in these cases).

[21] Of course, a staunch advocate of Adam's thesis could simply reject the idea that the probability of the conditional ever goes by the contrapositive conditional probability. If one takes this stance, then the results in this section can be seen as highlighting a new formal a-symmetry between MP and MT. As we noted in section 2, Oaksford *et al.* (2000) have already highlighted a significant a-symmetry between the acceptance rates of these argument schemes, so this view may also possess some merit.

1 unproblematically (proof omitted). However, their invalid counterparts, disjunction elimination (DE) and conjunction introduction (CI), both violate it. In particular, consider DE, which has the form A ∨ B, therefore A. We have the following result:

Proposition 14 *Increasing the probability of* A ∨ B *and minimizing the Kullback-Leibler divergence to the prior guarantees that the probability of* A *also increases. So DE violates Conjecture* 1.

The situation is just the same with CI, so it looks like our attempts to distinguish between valid and invalid arguments with uncertain minor premises is bound to fail. Valid arguments might have the desirable property that evidence for their minor premises always counts as evidence for their conclusion, but this condition is only necessary, it is not sufficient. There are also invalid arguments that are reliable in this important sense.

However, there is something special about CI and DE that appears to be crucial to their violation of the conjecture. Specifically, they are arguments whose conclusion entails all of the premises. They attempt to convince us of a hypothesis (conclusion) by providing evidence for its logical consequences. Thus, these arguments can be seen as instantiations of the hypothetico-deductive form of reasoning that Hempel took to be characteristic of the scientific method (see e.g. Hempel 1943, Sprenger 2011). Perhaps then it should not be surprising that these kinds of argument, despite being logically invalid, share with logically valid arguments an important form of reliability. This reflects the crucial role that such arguments are often thought to play in science. So we now have two fundamentally reliable types of argument: logically valid arguments and (possibly invalid) hypothetico-deductive arguments whose conclusions imply their premises. The following result suggests a natural generalisation of the conjecture:

Theorem 2 *Let* Γ *be the set of premises of a valid or hypothetico-deductive argument, with conclusion* ϕ*. Then, if we learn the premises of the argument with probability* $P'(\Gamma) \geq P(\Gamma)$ *and update by Jeffrey conditionalization, this learning will necessitate a corresponding increase* *in the probability of the conclusion, i..e $P'(\phi) \geq P(\phi)$. However, if the argument is invalid and not hypothetico-deductive, it will always be possible to find cases where $P'(\phi) < P(\phi)$, as long as the conclusion of the argument does not imply all of the premises.*

Now, this result is not adequate for our purposes, since it assumes that the premises of the arguments are always propositional, which is not the case for arguments with conditional major premises. However it does suggest a natural generalisation to our approach. In particular, we forward the following conjecture.

Conjecture 2: *For any valid argument or hypothetico-deductive argument schemes, increasing the probability of at least one minor premise, ensuring that the probability of no other minor premise goes down and setting the conditional probability constraints implied by the major premises to* 1 *guarantees that the probability of the conclusion will increase. For logically invalid non-hypothetico-deductive argument schemes, the conclusion can either increase or decrease in probability, depending on the prior distribution.*

This conjecture may not come as a surprise and one may wonder whether Ernest Adams didn't prove all this already in his book *A Primer of Probability Logic* (1998). In this book, Adam defines the *uncertainty of a proposition* ϕ as $u(\phi) := 1 - P(\phi)$ and proves the following two theorems:

Theorem 3 (Adams' Static Uncertainty Sum Rule) *If $\phi_1, \phi_2, \ldots, \phi_n$ are 'prior premises' and ϕ is a logical consequence of $\phi_1, \phi_2, \ldots, \phi_n$, then $u(\phi) \leq u(\phi_1) + \cdots + u(\phi_n)$.*

This theorem applies to our probability functions P and P'. Note, however, that we do not assign a probability to the conditional A → B. All we assume is that learning the conditional implies a certain probabilistic constraint on the new probability distribution P', i.e. $P'(\mathrm{B}|\mathrm{A}) = 1$. Hence, the theorem is not relevant for our purposes.

Theorem 4 (Adams' Dynamic Uncertainty Sum Rule) *If $\phi_1, \phi_2, \ldots, \phi_n$ are 'prior premises', ι is a new premise, and ϕ is a logical consequence of $\phi_1, \phi_2, \ldots, \phi_n$ and ι, then*

$u(\phi) \leq u(\phi_1|\iota) + \cdots + u(\phi_n|\iota)$.

This theorem assumes that the updating, i.e. the move from P to P', proceeds via conditionalization. However, conditionalization does not provide a satisfactory account of the learning of conditionals, which is why this theorem is also inapplicable to our project. So our conjecture can be seen as a generalisation of Adams' dynamic uncertainty rule to allow for the representation of conditional probabilistic constraints, where the conditional is not represented as a truth functional proposition. It should also be noted that Adams' dynamic rule assumes that one learns the evidence with certainty. Thus, it is not able to capture evidential uncertainty, where we learn the premises of the argument with non-maximal probability, which is exactly the kind of case we are interested in. The characterisation of logically valid arguments given by Conjecture 2 is extremely natural. It says that valid and hypothetico-deductive arguments are characterized by the property that evidence for the minor premises always counts as evidence for the conclusion. This provides a highly intuitive vindication of the special role of valid arguments in uncertain reasoning. Surprisingly, we've seen that the Bayesian approach to argumentation also identifies hypothetico-deductive arguments as a special class of reliable arguments. This suggests an interesting connection to the analysis of scientific reasoning and argumentation, where hypothetico-deductive arguments are thought to play a particularly important role (for detailed discussions of the confirmatory power of hypothetico-deductive reasoning, see e.g. Gemes (1998), Schurz (1991) and Sprenger (2011)). Finally, recall that in Section 5 we proved the following:

Theorem 5 (Summary) *An agent considers the premises* A *and* B *and entertains a prior probability distribution P according to eqs. (3) defined over them. If the probability of the minor premise stays the same and the probability of the major premise goes up, then the new probability of the conclusion will always be greater than the prior probability of the conclusion for logically valid argument schemes MP and MT, under one of the two possible interpretations of the major premise. For the logically invalid argument schemes AC and DA, the probability* *of the conclusion can be smaller or larger than the corresponding prior probability, for both of the two possible interpretations of the major premise.*

This suggests the following conjecture:

Conjecture 3: *For any valid argument scheme, increasing the probability of at least one major premise and holding fixed the probability of all the minor premises guarantees that the probability of the conclusion will increase, under at least one construal of the relevant major premises. For logically invalid non-hypothetico deductive argument schemes, the conclusion can either increase or decrease in probability (depending on the prior distribution), under either construal of the relevant major premises.*

7 Implications for Dual-Process Accounts of Reasoning

The results described above have significant implications for the currently popular dual-process accounts of reasoning. According to dual-process accounts, human reasoning is grounded in two distinct cognitive systems. System 1 processes[22] are typically assumed to be fast and automatic and only their final results are consciously registered. It is commonly conjectured that System 1 is shared with humanity's evolutionary ancestors and primarily utilizes long term memory. In contrast, System 2 processes are typically construed as slow, deliberative and distinctively human. System 2 is thought to be constrained by working memory capacity, and seems to be correlated with measures of general intelligence (see e.g. Evans (2000), Stankovich and West (1997)).

[22] In recent years, some authors have adopted the terminology of 'type 1' and 'type 2' processes (see e.g. Evans and Stanovich 2013), but here we stick with the traditional language of 'system 1' and 'system 2' processes.

7.1 Bayesian Argumentation and the Single Function View of Human Cognition

It has been argued that 'single-level probabilistic treatments' of human reasoning such as that given by the Bayesian paradigm are fundamentally incompatible with dual-process accounts, since they construe human reasoning as being governed by a single cognitive system (Evans 2007). This view is rejected by Oaksford and Chater (2012), who argue that the Bayesian paradigm is consistent with a particular interpretation of dual-process models. Specifically, they contend that a single function probabilistic model of human cognition is perfectly compatible with a dual process model of the *implementation* of reasoning. Although the fundamental objective of human reasoning may be to solve probabilistic inference tasks, it is still possible that the implementation of this central function is best explained by a dual process model which posits two distinct systems of the type described above. The dual-process model is only incompatible with the Bayesian approach to reasoning if one takes System 1 and System 2 to serve *distinct cognitive functions.* It is with this version of dual-process accounts of reasoning that we are presently concerned.

A central motivation for positing a dual-process account of reasoning comes from experiments in which participants untrained in formal logic are asked to assess the validity of particular instances of logical argument schemes. The crucial result is the observation of a 'belief-bias effect' in the participants' assessment of the arguments. Evans *et al.* (1983) introduced a methodology whereby participants are presented with syllogisms of the following four types.

Type 1: Valid Argument and Plausible Conclusion

- Example: No police dogs are viscious, some highly trained dogs are viscious, therefore some highly trained dogs are not police dogs.

Type 2: Valid Argument and Implausible Conclusion

- Example: No nutritional things are inexpensive, some vitamin tablets are inexpensive, therefore some vitamin tablets are not nutritional.

Type 3: Invalid Argument and Plausible Conclusion

- Example: No addictive things are inexpensive, some cigarettes are inexpensive, therefore some addictive things are not cigarettes.

Type 4: Invalid Argument and Implausible Conclusion

- Example: No millionaires are hard workers, some rich people are hard workers, therefore some millionaires are not rich people.

In the experiments, participants are explicitly instructed to treat the problem as a logical reasoning task, i.e. they should only accept the conclusions that follow with necessity from the relevant premises. Despite this instruction, participants (undergraduate students) consistently demonstrate significant 'belief bias', i.e. their assessments of the acceptability of the conclusion are strongly influenced by the independent plausibility of the conclusion, not only by the logical form of the argument. Evans *et al.* (1983) observed that acceptance rates for the conclusions of type 1 syllogisms were typically significantly higher than they were for type 2 syllogisms. Similarly, the acceptance rates for type 3 syllogisms were typically far higher than for type 4 syllogisms, and commonly higher than for type 2 syllogisms. Thus, we see that the independent plausibility of the conclusion tends to make a major difference to participants' assessments of the arguments. For both valid and invalid arguments, participants are significantly more likely to endorse the conclusion if it is independently plausible. Furthermore, participants are sometimes more willing to endorse invalid arguments with plausible conclusions than they are valid arguments with implausible conclusions. As Evans puts it, 'It is clear that participants are substantially influenced by both the logic of the argument and the believability of its conclusion, with more belief-bias on invalid arguments' (Evans, 2003). It is subsequently argued that this belief-bias is best explained by a dual-process account: 'It appears that both logical and belief-based processes are influencing the task and may be in competition with one another. In the dual-process account, these are attributed to Systems 2 and 1, respectively' (Evans, 2003).

This line of argument seems to explicitly support a dual-function view according to which System 1 and System 2 are engaged in fundamentally different cognitive functions. While System 1 attempts to solve probabilistic 'belief-based' inference tasks, System 2 is concerned with solving deductive logical inference tasks. Since logical form and independent plausibility both seem to play a crucial role in participants' assessments of argument instances, it takes both systems to adequately account for human reasoning behavior, or so the argument goes. But this line of reasoning can be resisted in light of the results presented in this paper. Belief-bias effects demonstrate that (1) *ceteris-paribus*, people endorse valid arguments over invalid ones, but this preference can be overridden if the valid argument has an implausible conclusion and the invalid argument has a plausible conclusion, (2) *ceteris-paribus*, people endorse arguments with plausible conclusions over arguments with implausible ones, but this preference can be overridden if the first argument is valid and the second is invalid. While it may seem that in order to account for both (1) and (2), it is necessary to postulate two separate cognitive systems, one governed by inductive probabilistic norms and the other governed by deductive logical norms, there is another alternative. Specifically, if it is the case that a *ceteris-paribus* preference for logically valid arguments emerges naturally from a preference for arguments with plausible conclusions, then both (1) and (2) can be easily explained by a single-function model of cognition of the type offered by the Bayesian paradigm. For, in this case, the observed preference for logically valid arguments, far from being incompatible with the Bayesian paradigm, is actually *explained* by participants treating the problem as a probabilistic inference task. In this paper, we have demonstrated that the Bayesian approach to human reasoning is perfectly capable of characterising and explaining the special role of logically valid arguments. Thus, the preference for logical validity is a natural and expected side effect of a preference for arguments with believable conclusions. We do not need to posit two distinct sets of norms governing the human reasoning in order to explain the belief-bias effects. The preference for logical validity is a natural side effect of the probabilistic norms given by the Bayesian paradigm. Thus, the Bayesian approach to argumentation described here undermines one of the fundamental motivations for the dual-process approach to human reasoning (under the interpretation whereby Systems 1 and 2 perform distinct cognitive functions).

7.2 Response to Singmann *et al.* (2016)

Singmann *et al.* (2016) developed a formal dual-source model (DSM) of reasoning with conditionals that employs a weighted sum whose components correspond to belief-based and logic-based (or 'content-based' and 'form-based') information. Thus, we read

> [T]here are content-independent effects of different argument forms that are not adequately captured by Bayesian models. Hence we propose that reasoning is influenced by two different and independent cognitive processes – a probabilistic process in line with extant Bayesian models, which we call *knowledge-based*, and a content-independent process driven by the form of the argument, which we call *form-based*. In this view, reasoners' evaluations actually reflect a mixture of form-based and knowledge-based information. (Singmann *et al.* 2016: 62).

Singmann *et al.* go on to apply the DSM to provide a reanalysis of a study by Markovits *et al.* (2016). In the study, participants were first asked to either assess an AC inference (Experiment 1) or a DA inference (Experiment 2). The contents involved in the inferences were entirely fictitious (involving an alien planet), which ensured that the participants had no strong prior beliefs about the truth of the premises/conclusion. After assessing the assigned inferences in Experiment 1/2, the participants were then given relative frequency information about the occurrence of either $A \wedge B$ and $\neg A \wedge B$ (in Experiment 1) or $\neg A \wedge B$ and $\neg A \wedge \neg B$ (in Experiment 2). They were then asked to assess the same inference they had previously assessed for a second time. For both of the experiments, some of the participants were given a high probability condition prompting a high value for the relevant conditional probability ($P(\mathrm{A}|\mathrm{B})$ for Experiment 1 and $P(\neg\mathrm{B}|\neg\mathrm{A})$ for Experiment 2) and some of them were given a low probability condition prompting a low value for the relevant conditional probability. As one would expect, it was observed that the acceptance rate for the initial inferences decreases for both high and low probability conditions when the participants are instructed to evaluate them from a deductive perspective, since even the high probability condition suggests the possibility of counterexamples. In contrast, when participants were instructed to evaluate the inferences from a probabilistic perspective, the acceptance rates were observed to decrease only when the low probability condition was given, which is also a natural result. Singmann *et al.* (2016) used a DSM model to describe the data and achieved a goodness of fit index of $R^2 = 0.86$. They then performed a meta-analysis in which they compared the performance of the chosen DSM model to three competing Bayesian models, each one corresponding to a different updating rule for learning the new conditional probability value. One of the models they considered corresponded to the learning rule whereby the agent minimised the Kullback-Leibler divergence using the learned conditional probability value as a constraint. They showed that the DSM model outperformed all of the considered Bayesian models, which had goodness of fit indices of around $0.79 - 0.8$. This suggests that the DSM may provide a better account of human reasoning than its Bayesian counterparts, including the minimization of the Kullback-Leibler divergence. It is not our aim here to provide a detailed analysis of this argument, but we would like to note that the approach developed in this paper may be able to contribute to a Bayesian response. Specifically, as we noted in section 2, the Kullback Leibler divergence is just one of a large family of probabilistic distance measures (the 'f-divergences'), all of which are compatible with Bayesian conditionalization, and all of which will give different results when updating on conditional information. The Kullback-Leibler divergence has no special status here, and we believe that it will be partly an empirical matter to determine which f-divergence gives the best account of human reasoning. And since there are many alternative f-divergences (e.g. the inverse KL-divergence, the Hellinger distance, the total variation distance, α-divergence, etc) it is reasonable to expect that some of them may be able to outperform the Kullback-Leibler divergence in matching people's evaluation of inferences. Furthermore, it should be noted that although the formal results described here all concern the Kullback-Leibler divergence, there are strong indications that the same or closely analogous results will be obtainable for other f-divergences.[23]

To summarise, the Bayesian approach to argumentation described here allows for a Bayesian model that explains the *ceteris-paribus* preference for logically valid arguments, thereby undermining one of the main motivations for the dual function view of human cognition. Furthermore, by utilising alternative f-divergences, the Bayesian can hope to improve on the empirical adequacy of their existing models of dynamic conditional inference (as described by e.g. Singmann and Klauer (2016)).

8 Conclusion

In this paper, we have presented a major extension to the Bayesian approach to argumentation that (amongst other things):

1. Utilizes a new class of learning rules that are better suited to modelling conditional inferences than standard Bayesian methods.

2. Demonstrates how a preference for logically valid arguments arises naturally from probabilistic reasoning norms, thereby undermining one of the most important theoretical motivations for dual-function models of human cognition.

3. Allows for the definition of precise and well motivated measures of argument strength whilst taking seriously philosophical concerns about (Jeffrey) conditionalizing on indicative conditionals.

[23] This is a topic of ongoing research.

4. Explicates the special role of logical validity in uncertain reasoning contexts, whilst also justifying the privileged status of hypothetico-deductive arguments in scientific and everyday reasoning, and accounting for the fact that invalid arguments can often have significant strength.

5. Opens the possibility of significantly extending the Bayesian approach to the psychology of reasoning by comparing the ability of different f-divergences to adequately model reasoning patterns.

6. Allows the Bayesian to naturally model failures of rigidity and thereby provide a more psychologically plausible model of argumentation and dynamic non-monotonicity.

More generally, the approach outlined here supports a dialogical view of argumentation, according to which argumentation is a dynamic process in which interlocutors attempt to convince each other of certain conclusions (see Hahn and Oaksford 2007, Van Eemeren and Grootendorst 2004).[24] The success of these arguments will depend not only on their logical form, but also on the epistemic context of the interlocutors (as encoded in their prior beliefs), the argumentative context and the strength of the evidence given in support of the premises. By utilising the distance based approach to learning, we have also greatly extended the range and variety of argumentative phenomena that fall under the ambit of the Bayesian approach to argumentation. Crucially, we now have a general framework for studying a far wider range of argumentative contexts than those that have previously been considered.

Finally as well as raising many important theoretical questions, the newly extended Bayesian theory of argumentation opens a number of novel avenues for empirical research, including for example:

1. Studying which generalised Bayesian learning rules provide the best account of the way that people update their beliefs upon learning conditional information, particularly in dynamic argumentation contexts such as that described in Markovits *et al.* (2016).

[24] In this sense, our approach shares a certain affinity with the influential view of argumentation forwarded by Mercier and Sperber (2011).

2. Studying whether people evaluate the reliability of general argument schemes and the strength of individual arguments in line with the norms developed here, i.e. whether subjects exhibit a general preference for hypothetico-deductive as opposed to other kinds of invalid arguments and whether presenting the major premises of arguments in a contrapositive form makes an important difference to the assessment of argument strength.

Acknowledgements

Several friends and colleagues helped us greatly during the writing of this article. We are especially thankful to Ulrike Hahn, Karl Christoph Klauer, Jan Sprenger, Mike Oaksford, Henrik Singmann, David Over, Reuben Stern and an anonymous referee. Our research was supported by the Researcher in Residence program of LMU Munich's Center for Advanced Studies in the academic year 2016/17 (project: "Reasoning and Argumentation in Science") and by the Alexander von Humboldt Foundation, for which we are grateful. Stephan Hartmann would also like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme "Probability and Statistics in Forensic Science" where work on this paper was undertaken. This work was also supported by EPSRC grant no EP/K032208/1.

Appendix

Proofs

We begin with a definition. Let $S_1, \ldots, S_n$ be the possible values of a random variable S over which probability distributions P' and P are defined. Then the Kullback-Leibler divergence between P' and P is given by

$$D_{KL}(P'||P) := \sum_{i=1}^{n} P'(S_i) \log \frac{P'(S_i)}{P(S_i)} \,. \tag{7}$$

In the remainder, we use the abbreviation "KL" for the Kullback-Leibler divergence. If a constraint is added via a Lagrange multiplier, we use the letter "L" for the function which we minimize.

In several of the following proofs we use the abbreviation

$$\Phi_x := x' \log \frac{x'}{x} + \overline{x'} \log \frac{\overline{x'}}{\overline{x}} \tag{8}$$

Note that

$$\frac{\partial \Phi_x}{\partial x'} = \log \left(\frac{x' \, \overline{x}}{\overline{x'} \, x} \right) \tag{9}$$

and that $\partial \Phi_x / \partial x' = 0$ implies $x' = x$.

Proposition 1

We conclude from **MP1** that $p' = 1$, and from **MP2** that $a' = 1$. Hence, the posterior distribution P' is given by

$$\begin{aligned} P'(\mathrm{A}, \mathrm{B}) = 1 \quad &, \quad P'(\mathrm{A}, \neg\mathrm{B}) = 0 \\ P'(\neg\mathrm{A}, \mathrm{B}) = 0 \quad &, \quad P'(\neg\mathrm{A}, \neg\mathrm{B}) = 0 \,, \end{aligned} \tag{10}$$

from which we see immediately that $P'(\mathrm{B}) = 1$. Note that, in this case, no distance measure needed to be minimized to obtain the result. ▮

Proposition 2

To determine the posterior probability distribution P', we assume again that the Bayesian Network does not change as a result of learning the two premises and replace the variables a, p and q by the corresponding primed variables a', p' and q'. With this, we conclude from **MT1** that $p' = 1$, and from **MT2** that $P'(\mathrm{B}) = a'\, p' + \overline{a'}\, q' = a' + \overline{a'}\, q' = 0$. Hence, $a' = q' = 0$, which leads to the posterior distribution

$$\begin{aligned} P'(\mathrm{A}, \mathrm{B}) = 0 \qquad &, \qquad P'(\mathrm{A}, \neg\mathrm{B}) = 0 \\ P'(\neg\mathrm{A}, \mathrm{B}) = 0 \qquad &, \qquad P'(\neg\mathrm{A}, \neg\mathrm{B}) = 1\,. \end{aligned} \tag{11}$$

From this we see immediately that $P'(\mathrm{A}) = 0$ and hence $P'(\neg\mathrm{A}) = 1$ (again without minimizing a distance measure), in accordance with the fact that ¬A follows deductively from MT1 and MT2. ▮

Proposition 3

AC1 entails that $p' = 1$ and **AC2** entails that

$$P'(\mathrm{B}) = a' + \overline{a'}\, q' = 1. \tag{12}$$

Hence, we have to minimize

$$L = \Phi_a + a' \, \log \frac{1}{p} + \overline{a'}\, \Phi_q + \lambda(a' + \overline{a'}\, q' - 1)\,. \tag{13}$$

We differentiate L with respect to q' and obtain

$$\frac{\partial L}{\partial q'} = \overline{a'} \left(\log \left(\frac{q'\, \overline{q}}{\overline{q'}\, q} \right) + \lambda \right). \tag{14}$$

Setting this expression equal to zero yields

$$q' = \frac{q}{q + \overline{q}\, x}, \tag{15}$$

with $x := e^{\lambda}$. We insert eq. (15) into eq. (13) and obtain:

$$L = \Phi_a + a' \log \frac{1}{p} - \overline{a'} \log(q + \overline{q}\, x) \tag{16}$$

Next, we differentiate this expression with respect to a', set the result equal to zero and obtain

$$a' = \frac{a\, p}{a\, p + \overline{a}\,(q + \overline{q}\, x)}. \tag{17}$$

Next, we insert eqs. (15) and (17) into eq. (12) and conclude that $x = 0$. Hence,

$$\begin{aligned} a' &= \frac{a\, p}{a\, p + \overline{a}\, q} \\ &= \frac{P(\mathrm{A}, \mathrm{B})}{P(\mathrm{B})} \\ &= P(\mathrm{A}|\mathrm{B})\,. \end{aligned} \tag{18}$$

The posterior probability distribution is then given by

$$\begin{aligned} P'(\mathrm{A}, \mathrm{B}) = P(\mathrm{A}|\mathrm{B}) \qquad &, \qquad P'(\mathrm{A}, \neg\mathrm{B}) = 0 \\ P'(\neg\mathrm{A}, \mathrm{B}) = P(\neg\mathrm{A}|\mathrm{B}) \qquad &, \qquad P'(\neg\mathrm{A}, \neg\mathrm{B}) = 0\,. \quad \blacksquare \end{aligned} \tag{19}$$

Proposition 4

DA1 entails that $p' = 1$ and **DA2** entails that $a' = 0$. Hence, the posterior probability distribution is given by

$$\begin{aligned} P'(\mathrm{A},\mathrm{B}) = 0 \quad &, \quad P'(\mathrm{A},\neg\mathrm{B}) = 0 \\ P'(\neg\mathrm{A},\mathrm{B}) = q' \quad &, \quad P'(\neg\mathrm{A},\neg\mathrm{B}) = \overline{q'}\,. \end{aligned} \tag{20}$$

To find q', we minimize the Kullback-Leibler divergence between P' and P, which is given by

$$\begin{aligned} KL &= q' \log \frac{q'}{\overline{a}\,q} + \overline{q'} \log \frac{\overline{q'}}{\overline{a}\,\overline{q}} \\ &= \Phi_q - \log \overline{a}\,. \end{aligned} \tag{21}$$

Hence,

$$\frac{\partial KL}{\partial q'} = \log \left(\frac{q'}{\overline{q'}} \, \frac{\overline{q}}{q} \right) \,. \tag{22}$$

Setting the expression on the RHS equal to zero and solving for q' yields $q' = q = P(\mathrm{B}|\neg\mathrm{A})$. Hence, the posterior probability distribution is given by

$$\begin{aligned} P'(\mathrm{A},\mathrm{B}) = 0 \quad &, \quad P'(\mathrm{A},\neg\mathrm{B}) = 0 \\ P'(\neg\mathrm{A},\mathrm{B}) = P(\mathrm{B}|\neg\mathrm{A}) \quad &, \quad P'(\neg\mathrm{A},\neg\mathrm{B}) = P(\neg\mathrm{B}|\neg\mathrm{A})\,. \end{aligned} \tag{23}$$

Finally, we compute $P'(\neg\mathrm{B}) = P(\neg\mathrm{B}|\neg\mathrm{A})$. ■

Proposition 5

As a' and p' are now fixed, we only have to minimize $KL = D^0_{KL}(P'||P)$ with respect to q', i.e.

$$\begin{aligned} \frac{\partial KL}{\partial q'} &= \overline{a'}\,\frac{\partial \Phi_q}{\partial q'} \\ &= \overline{a'}\,\log\left(\frac{q'}{\overline{q'}}\,\frac{\overline{q}}{q}\right). \end{aligned} \tag{24}$$

Setting this expression equal to zero yields $q' = q$. Hence

$$P'(\mathrm{B}) = a' + \overline{a'}\,q\,. \tag{25}$$

Noting that $P(\mathrm{B}) = a\,p + \overline{a}\,q$ we conclude that

$$\begin{aligned} \Delta_B^{MP} &:= P'(\mathrm{B}) - P(\mathrm{B}) \\ &= a\,\overline{p} + (a' - a)\,\overline{q}\,. \end{aligned} \tag{26}$$

As both terms in this sum are greater than zero (note that $a' > a$), we obtain that $P'(\mathrm{B}) > P(\mathrm{B})$. ∎

Proposition 6

The proof proceeds as the previous proof with $\Delta_B^{DA} := P'(\neg\mathrm{B}) - P(\neg\mathrm{B}) = -\Delta_B^{MP}$. However, as $a' < a$, we see that Δ_B^{DA} may be smaller or larger than zero, depending on the parameters that characterize the prior probability distribution. ∎

Proposition 7

The second premise implies $P'(\mathrm{B}) \leq P(\mathrm{B})$ with

$$P(\mathrm{B}) =: b = a\,p + \overline{a}\,q \quad , \quad P'(\mathrm{B}) =: b' = a'\,p' + \overline{a'}\,q'\,.$$

Hence the new probability distribution P' has to satisfy the constraint

$$a' + \overline{a'}\,q' - a\,p - \overline{a}q + \delta = 0, \tag{27}$$

with some positive number $\delta := b - b'$. We therefore have to maximize

$$L = \Phi_a + a'\,\Phi_p + \overline{a'}\,\Phi_q + \lambda\,(a' + \overline{a'}\,q' - a\,p - \overline{a}q + \delta), \tag{28}$$

with a Lagrange parameter λ. We differentiate L with respect to q' and set the resulting expression equal to zero. From this we get

$$q' = \frac{q}{q + \overline{q}\,x}, \tag{29}$$

with $x := e^{\lambda}$. Inserting this into eq. (28), we get

$$L = \Phi_a + a'\,\log\frac{1}{p} + \overline{a'}\,\log\frac{1}{q + \overline{q}\,x} + \lambda\,(1 - a\,p - \overline{a}q + \delta). \tag{30}$$

We differentiate L from eq. (30) with respect to a' and obtain:

$$\frac{\partial L}{\partial a'} = \log\left(\frac{a'}{\overline{a'}} \cdot \frac{\overline{a}\,(q + \overline{q}\,x)}{a\,p}\right)$$

Setting this expression equal to zero, we obtain:

$$a' = \frac{a\,p}{a\,p + \overline{a}\,(q + \overline{q}\,x)} \tag{31}$$

We insert eqs. (29) and (31) into eq. (27) and obtain:

$$a\,p + \overline{a}\,(q + \overline{q}\,x) = \frac{a\,p + \overline{a}\,q}{a\,p + \overline{a}\,q - \delta} \tag{32}$$

Hence,

$$\begin{aligned} a' &= \frac{a\,p\,(a\,p + \overline{a}\,q - \delta)}{a\,p + \overline{a}\,q} \\ &= P(\mathrm{A}|\mathrm{B})\,P'(\mathrm{B})\,. \end{aligned} \tag{33}$$

With a bit of algebra we see that

$$\begin{aligned} \Delta^{MT}_{\neg A} &:= P'(\neg\mathrm{A}) - P(\neg\mathrm{A}) \\ &= a - a' \\ &= a\,\overline{p} + \delta \cdot P(\mathrm{A}|\mathrm{B}) \\ &= a\,\overline{p} + (\overline{b'} - \overline{b}) \cdot P(\mathrm{A}|\mathrm{B})\,. \end{aligned} \tag{34}$$

As $\delta = \overline{b'} - \overline{b} > 0$, we conclude that $\Delta^{MT}_{\neg A} > 0$. ▮

Proposition 8

The proof proceeds as the previous proof. The only difference is that the sign of δ flipped, which makes it now possible that the LHS of the inequality mentioned in the theorem is greater than zero or smaller than zero, depending on the context (i.e. on the prior probability distribution). ▮

Proposition 9

Consider MP first. Using Jeffrey Conditionalization, we calculate

$$P^*(\mathrm{B}) = \pi_1 \cdot P'(\mathrm{A}) + \pi_2 \cdot P'(\neg\mathrm{A}),$$

with

$$\begin{aligned}
\pi_1 &:= P(\mathrm{B}|\neg\mathrm{A} \vee \mathrm{B}, \mathrm{A}) = \frac{P(\mathrm{A}, \mathrm{B}, \neg\mathrm{A} \vee \mathrm{B})}{P(\mathrm{A}, \neg\mathrm{A} \vee \mathrm{B})} = \frac{P(\mathrm{A}, \mathrm{B})}{P(\mathrm{A}, \mathrm{B})} = 1 \\
\pi_2 &:= P(\mathrm{B}|\neg\mathrm{A} \vee \mathrm{B}, \neg\mathrm{A}) = \frac{P(\neg\mathrm{A}, \mathrm{B}, \neg\mathrm{A} \vee \mathrm{B})}{P(\neg\mathrm{A}, \neg\mathrm{A} \vee \mathrm{B})} = \frac{P(\neg\mathrm{A}, \mathrm{B})}{P(\neg\mathrm{A})} = P(\mathrm{B}|\neg\mathrm{A}).
\end{aligned}$$

Hence,

$$P^*(\mathrm{B}) = a' + \overline{a'} \cdot q,$$

which is identical with $P'(\mathrm{B})$ from eq. (25). Hence, the results of Theorems 5 and 6 follow. Next, we consider MT and calculate

$$P^*(\mathrm{A}) = \pi_3 \cdot P'(\neg\mathrm{B}) + \pi_4 \cdot P'(\mathrm{B}),$$

with

$$\begin{aligned}
\pi_3 &:= P(\mathrm{A}|\neg\mathrm{A} \vee \mathrm{B}, \neg\mathrm{B}) = \frac{P(\mathrm{A}, \neg\mathrm{B}, \neg\mathrm{A} \vee \mathrm{B})}{P(\neg\mathrm{A} \vee \mathrm{B}, \mathrm{B})} = 0 \\
\pi_4 &:= P(\mathrm{A}|\neg\mathrm{A} \vee \mathrm{B}, \mathrm{B}) = \frac{P(\mathrm{A}, \mathrm{B}, \neg\mathrm{A} \vee \mathrm{B})}{P(\neg\mathrm{A} \vee \mathrm{B}, \mathrm{B})} = \frac{P(\mathrm{A}, \mathrm{B})}{P(\mathrm{B})} = P(\mathrm{A}|\mathrm{B}).
\end{aligned}$$

Hence,

$$P^*(\mathrm{A}) = P(\mathrm{A}|\mathrm{B}) \cdot P'(\mathrm{B}),$$

which is identical with $P'(\mathrm{A})$ from eq. (33). Hence, the results of Theorems 7 and 8 follow.

■

Proposition 10

Consider the case **MP1**(i) first. We then have to find q' such that $KL = a\,\Phi_p + \overline{a}\,\Phi_q$ is minimized. We obtain $q' = q$. Hence, as $p' > p$, it follows that $P'(\mathrm{B}) = a\,p' + \overline{a}\,q > a\,p + \overline{a}\,q = P(\mathrm{B})$.

Next, consider the case **MP1**(ii). The constraint $P'(\neg\mathrm{A}|\neg\mathrm{B}) > P(\neg\mathrm{A}|\neg\mathrm{B})$ implies that

$$\overline{p}\,\overline{q'} - \overline{q}\,\overline{p'} - \delta = 0, \tag{35}$$

with $\delta > 0$. We therefore have to find p' and q' such that

$$L = a\,\Phi_p + \overline{a}\,\Phi_q + \lambda\,(\overline{p}\,\overline{q'} - \overline{q}\,\overline{p'} - \delta)$$

is minimized. Differentiating L with respect to p' and q' and setting the resulting expressions equal to zero yields

$$\frac{p'\,\overline{p}}{\overline{p'}\,p} = \left(\frac{1}{x}\right)^{\overline{q}/a} \quad \text{and} \quad \frac{q'\,\overline{q}}{\overline{q'}\,q} = x^{\overline{p}/\overline{a}},$$

with $x := \mathrm{e}^{\lambda}$. Hence,

$$\left(\frac{p'\,\overline{p}}{\overline{p'}\,p}\right)^{\overline{p}/\overline{a}} \cdot \left(\frac{q'\,\overline{q}}{\overline{q'}\,q}\right)^{\overline{q}/a} = 1. \tag{36}$$

We now set $p = a$ and $q = \overline{a}$ and obtain from eq. (36) that $q' = \overline{p'}$. Using eq. (35) we obtain $p' = a + \delta$ and therefore $q' = \overline{a} - \delta$. Hence, in this case, $P(\mathrm{B}) = a\,p + \overline{a}\,q = a^2 + \overline{a}^2$ and $P'(\mathrm{B}) = a\,p' + \overline{a}\,q' = a^2 + \overline{a}^2 + \delta \cdot (a - \overline{a}) = P(\mathrm{B}) + \delta \cdot (a - \overline{a})$. We conclude that, for **MP1**(ii), $P'(\mathrm{B})$ can be less than $P(\mathrm{B})$, e.g. if $a < 1/2$, $p = a$ and $q = \overline{a}$. ▮

Proposition 11

Here it is more convenient to use the following parameterization of the prior distribution P (and accordingly for P'):

$$P(\mathrm{B}) = b \quad , \quad P(\mathrm{A}|\mathrm{B}) = p \quad , \quad P(\mathrm{A}|\neg\mathrm{B}) = q. \tag{37}$$

Consider the case **MT1**(ii) first. The constraint $P'(\neg\mathrm{A}|\neg\mathrm{B}) > P(\neg\mathrm{A}|\neg\mathrm{B})$ implies that $q' < q$. We then have to find p' such that $KL = b\,\Phi_p + \overline{b}\,\Phi_q$ is minimized. We obtain $p' = p$. Hence, $P'(\mathrm{A}) = b\,p + \overline{b}\,q' < b\,p + \overline{b}\,q = P(\mathrm{A})$.

Next, consider the case **MT1**(i). The constraint $P'(\mathrm{B}|\mathrm{A}) > P(\mathrm{B}|\mathrm{A})$ implies that

$$p'\,q - p\,q' - \delta = 0, \tag{38}$$

with $\delta > 0$. We therefore have to find p' and q' such that

$$L = b\,\Phi_p + \overline{b}\,\Phi_q + \lambda\,(p'\,q - p\,q' - \delta)$$

is minimized. Differentiating L with respect to p' and q' and setting the resulting expressions equal to zero yields

$$\frac{p'\,\overline{p}}{\overline{p'}\,p} = \left(\frac{1}{x}\right)^{q/b} \quad \text{and} \quad \frac{q'\,\overline{q}}{\overline{q'}\,q} = x^{p/\overline{b}},$$

with $x := \mathrm{e}^{\lambda}$. Hence,

$$\left(\frac{p'\,\overline{p}}{\overline{p'}\,p}\right)^{p/\overline{b}} \cdot \left(\frac{q'\,\overline{q}}{\overline{q'}\,q}\right)^{q/b} = 1. \tag{39}$$

We now set $p = \overline{b}$ and $q = b$ and obtain from eq. (39) that $q' = \overline{p'}$. Using eq. (38) we obtain $p' = \overline{b} + \delta$ and therefore $q' = b - \delta$. Hence, in this case, $P(\mathrm{A}) = b\,p + \overline{b}\,q = 2\,b\,\overline{b}$ and $P'(\mathrm{A}) = b\,p' + \overline{b}\,q' = 2\,b\,\overline{b} + \delta \cdot (b - \overline{b}) = P(\mathrm{A}) + \delta \cdot (b - \overline{b})$. We conclude that, for **MT1**(i), $P'(\neg\mathrm{A})$ can be less than $P(\neg\mathrm{A})$, e.g. if $b > 1/2$, $p = \overline{b}$ and $q = b$. ∎

Proposition 12

The proof follows from the proof of Theorem 10. ▮

Proposition 13

The proof follows from the proof of Theorem 11. ▮

Proposition 14

We note that $P(\mathrm{A} \vee \mathrm{B}) = P(\mathrm{A}) + P(\neg\mathrm{A}, \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}|\neg\mathrm{A})\, P(\neg\mathrm{A})$ and similarly for $P'(\mathrm{A} \vee \mathrm{B})$. Hence the following constraint applies:

$$a' + \overline{a'}\, q' - a - \overline{a}\, q - \delta = 0, \tag{40}$$

with $\delta > 0$. We therefore have to minimize

$$L = \Phi_a + a'\, \Phi_p + \overline{a'}\, \Phi_q + \lambda(a' + \overline{a'}\, q' - a - \overline{a}\, q - \delta). \tag{41}$$

We differentiate L by p', set the resulting expression equal to zero and obtain

$$p' = p. \tag{42}$$

Similarly, we obtain

$$q' = \frac{q}{q + \overline{q}\, x}, \tag{43}$$

with $x := e^{\lambda}$. Inserting eqs. (42) and (43) into eq. (41), we obtain

$$L = \Phi_a - \overline{a'}\, \log(q + \overline{q}\, x) + \lambda\, (\overline{a}\, \overline{q} - \delta). \tag{44}$$

We differentiate L with respect to a', set the resulting expression equal to zero and obtain

$$a' = \frac{a}{a + \overline{a}\,(q + \overline{q}\,x)}. \tag{45}$$

Inserting eqs. (43) and (45) into eq. (40) finally yields

$$a' = \left(1 + \frac{\delta}{a + \overline{a}\,q}\right)\, a\,. \tag{46}$$

Hence $a' > a$ iff $\delta > 0$. Not surprisingly, the same result obtains if we Jeffrey-conditionalize on ¬A ∨ B (proof omitted). ▮

Theorem 2

Let $\Gamma \vdash \phi$ be a valid argument scheme, and suppose we learn Γ with probability $P'(\Gamma) \geq P(\Gamma)$, and let $\alpha = P'(\Gamma) - P(\Gamma)$. Then,

$$\begin{aligned}
P'(\phi) &= P(\phi|\Gamma)\,P'(\Gamma) + P(\phi|\neg\Gamma)\,P'(\neg\Gamma) \\
&= \frac{P(\phi,\Gamma)}{P(\Gamma)}\,P'(\Gamma) + \frac{P(\phi,\neg\Gamma)}{P(\neg\Gamma)}\,P'(\neg\Gamma) \\
&= P'(\Gamma) + \frac{P(\phi,\neg\Gamma)}{P(\neg\Gamma)}\,P'(\neg\Gamma) \\
&= P(\Gamma) + \alpha + \left(\frac{P(\phi,\neg\Gamma)}{P(\neg\Gamma)}\right)(P(\neg\Gamma) - \alpha) \\
&= P(\Gamma) + \alpha + P(\phi,\neg\Gamma) - \frac{\alpha P(\phi,\neg\Gamma)}{P(\neg\Gamma)}\,.
\end{aligned}$$

So,

$$
\begin{aligned}
P'(\phi) - P(\phi) &= P'(\phi) - P(\Gamma) - P(\phi, \neg\Gamma) \\
&= P(\Gamma) + \alpha + P(\phi, \neg\Gamma) - \frac{\alpha\, P(\phi, \neg\Gamma)}{P(\neg\Gamma)} - P(\Gamma) - P(\phi, \neg\Gamma) \\
&= \alpha - \frac{\alpha\, P(\phi, \neg\Gamma)}{P(\neg\Gamma)} \\
&\geq \alpha - \frac{\alpha\, P(\neg\Gamma)}{P(\neg\Gamma)} \\
&= 0.
\end{aligned}
$$

Next, suppose that the argument is hypothetico-deductive, i.e ϕ entails Γ. Then,

$$
\begin{aligned}
P'(\phi) &= P(\phi|\Gamma)P'(\Gamma) + P(\phi|\neg\Gamma)\, P'(\neg\Gamma) \\
&= \frac{P(\phi, \Gamma)}{P(\Gamma)}\, P'(\Gamma) + \frac{P(\phi, \neg\Gamma)}{P(\neg\Gamma)}\, P'(\neg\Gamma) \\
&= \frac{P(\phi, \Gamma)}{P(\Gamma)}\, P'(\Gamma) \\
&= \frac{P(\phi, \Gamma)}{P(\Gamma)}\, (P(\Gamma) + \alpha) \\
&= P(\phi, \Gamma) + \alpha\, \frac{P(\phi, \Gamma)}{P(\Gamma)} \\
&= P(\phi) + \alpha\, \frac{P(\phi)}{P(\Gamma)} \\
&\geq P(\phi).
\end{aligned}
$$

Finally, suppose that the argument is neither valid nor hypothetico-deductive. In this case, it is easy to see that we can have a prior distribution for which $P(\Gamma) = P(\Gamma, \neg\phi) = 0.9$ and $P(\phi) = P(\phi, \neg\Gamma) = 0.1$. As usual, we calculate $P'(\phi) = P(\phi|\Gamma)\, P'(\Gamma) + P(\phi|\neg\Gamma)\, P'(\neg\Gamma)$. Now, $P(\Gamma) = 0.9$. So, if we set $\alpha = 0.1$, then $P'(\neg\Gamma) = 0$ and $P'(\phi) = 0 < P(\phi)$. So for any invalid argument, $P'(\Gamma) \geq P(\Gamma)$ doesn't guarantee $P'(\phi) \geq P(\phi)$, as desired. This completes the proof. ▮

References

Adams, E. (1975). *The Logic of Conditionals. Synthese Library*, Vol. 86. Boston: D. Reidel.

Adams, E. (1998). *A Primer of Probability Logic.* Chicago: The University of Chicago Press.

Ali, N., N. Chater and M. Oaksford (2011). The Mental Representation of Causal Conditional Reasoning: Mental Models or Causal Models. *Cognition* 119(3): 403–418.

Ali, N., A. Schlottmann, C. Shaw, N. Chater and M. Oaksford (2010). Conditionals and Causal Discounting in Children. In M. Oaksford and N. Chater (eds.): *Cognition and Conditionals: Probability and Logic in Human Thinking.* Oxford: Oxford University Press.

Bennett, J. (2003). *A Philosophical Guide to Conditionals.* Oxford: Oxford University Press.

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

Brandom, R. B. (2008). *Between Saying and Doing. Towards an Analytic Pragmatism.* Oxford: Oxford University Press.

Byrne, R. M. (1989). Suppressing valid inferences with conditionals. *Cognition* 31: 61–83.

Corner A. and U. Hahn (2009). Evaluating Science Arguments: Evidence, Uncertainty, and Argument Strength. *Journal of Experimental Psychology: Applied* 15(3):199–212.

Corner A. and U. Hahn (2013). Normative Theories of Argumentation: Are Some Norms Better than Others? *Synthese* 190(16): 3579–3610.

Csiszár (2008). Axiomatic Characterizations of Information Measures. *Entropy* 10: 261–273.

Diaconis, P. and S. Zabell. (1982). Updating Subjective Probability. *Journal of the American Statistical Association* 77: 822–830.

Douven, I. (2012). Learning Conditional Information. *Mind and Language* 27(3): 239–263.

Douven, I. (2017). How to Account for the Oddness of Missing-Link Conditionals. *Synthese* 194(5): 1541–1554.

Douven, I. and R. Dietz (2011). A Puzzle about Stalnaker's Hypothesis. *Topoi* 30: 31–37.

Douven, I. and S. Verbrugge (2010). The Adams Family. *Cognition* 117(3): 302–318.

Douven, I. and S. Verbrugge (2012). Indicatives, Concessives, and Evidential Support. *Thinking and Reasoning* 18(4): 480–499.

Douven, I. and S. Verbrugge (2013). The Probabilities of Conditionals Revisited. *Cognitive Science* 37(4): 711–730.

Eddington, D. (1995). On Conditionals. *Mind* 104: 235–329.

Eva, B. and S. Hartmann (2018). When No Reason For is a Reason Against. *Analysis.*

Eva, B., S. Hartmann and S. Rafiee Rad (forthcoming). Learning from Conditionals. MCMP Working Paper.

Evans, J.S.B.T. (1983). On the Conflict Between Logic and Belief in Syllogistic Reasoning. *Memory and Cognition* 11: 295–306.

Evans, J.S.B.T. (1993). The Mental Model Theory of Conditional Reasoning: Critical Appraisal and Revision. *Cognition* 48: 1–20.

Evans, J.S.B.T. (2000). In: J.A. Garcia-Madruga *et al.* (eds.): *Mental Models In Reasoning*, pp. 41–56. UNED: Madrid.

Evans, J. S. B. T. (2002). Logic and Human Reasoning: An Assessment of the Deduction Paradigm. *Psychological Bulletin* 128(6): 978–996.

Evans, J.S.B.T. (2003). In Two Minds: Dual-Process Accounts of Reasoning. *Trends in Cognitive Science* 7(1): 454–459.

Evans, J.S.B.T. (2007). *Hypothetical Thinking.* Psychology Press: Hove.

Evans, J.S.B.T. and K. Stanovich (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science* 8(3): 223–241.

Fernbach, P. and C. Erb (2013). A Quantitative Causal Model Theory of Conditional Reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39(5): 1327–1343.

Fitelson, B. (1999). The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity. *Philosophy of Science* 66: 362–378.

Gemes, K. (1998). Hypothetico-Deductivism: The Current State of Play. *Erkenntnis* 49: 1–20.

Godden, D. and F. Zenker (2016). A Probabilistic Analysis of Argument Cogency. *Synthese*: 1–26. https://doi.org/10.1007/s11229-016-1299-2.

Hahn, U., A. J. Harris, and A. Corner (2009). Argument Content and Argument Source: An Exploration. *Informal Logic* 29(4): 337-367.

Hahn, U., A. J. Harris, and M. Oaksford (2013). Rational Argument, Rational Inference. *Argument and Computation* 4(1): 21-35.

Hahn, U. and J. Hornikx (2016). A Normative Framework for Argument Quality: Argumentation Schemes with a Bayesian Foundation. *Synthese* 193 (6): 1833–1873.

Hahn, U. and M. Oaksford (2007). The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. *Psychological Review* 114 (3): 704–732.

Hahn, U. and M. Oaksford (2007). A Bayesian Approach to Informal Argument Fallacies. *Synthese* 152(2): 207–236.

Hall, S., N. Ali, N. Chater and M. Oaksford (2016). Discounting and Augmentation in Causal Conditional Reasoning: Causal Models or Shallow Encoding. *PLOS ONE* 11(12): e0167741. doi: 10.1371/journal.pone.0167741

Hamblin, C. (1972). *Fallacies.* London: Methuen.

Harris, A. J., U. Hahn, J. K. Madsen and A. S. Hsu (2016). The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach. *Cognitive Science* 40(6): 1496–1533.

Harris, A. J., A. S. Hsu, and J. K. Madsen (2012). Because Hitler did it! Quantitative Tests of Bayesian Argumentation Using ad Hominem. *Thinking and Reasoning* 18(3): 311–343.

Hempel, C. G. (1943). A Purely Syntactical Definition of Confirmation. *Journal of Symbolic Logic* 8: 122–143.

J. Hornikx and U. Hahn (2012). Reasoning and Argumentation: Towards an Integrated Psychology of Argumentation. *Thinking and Reasoning* 18 (3): 225–243.

Howson, C. and P. Urbach (2005). *Scientific Reasoning: The Bayesian Approach.* Open Court.

Klauer, K. C., S. Beller and M. Hütter (2010). Conditional Reasoning in Context: A Dual-Source Model of Probabilistic Inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36(2): 298–323.

Lewis, D. (1976). Probabilities of Conditionals and Conditional Probabilities. *Philosophical Review* 85 (3): 297–315.

Markovits, H., J. Brisson and P. L. De Chantal (2015). Deductive Updating is not Bayesian. *Journal of Experimental Psychology: Learning, Memory and Cognition* 29: 949–956.

Mercier, H. and D. Sperber (2011). Why do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences* 34 (2): 57–74.

Oaksford, M. and N. Chater (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Oaksford, M. and N. Chater (2009). Precis of *Bayesian Rationality: The Probabilistic Approach to Human Reasoning. Behavioral and Brain Sciences* 32: 69–120.

Oaksford, M. and N. Chater (2013). Dynamic Inference and Everyday Conditional Reasoning in the New Paradigm. *Thinking and Reasoning* 19(3-4): 346–379.

Oaksford, M. and N. Chater (2017). Causal Models and Conditional Reasoning. In M. Waldmann (ed.): *The Oxford Handbook of Causal Reasoning*. Oxford: Oxford University Press.

Oaksford, M., N. Chater and J. Larkin. (2000). Probabilities and Polarity Biases in Conditional Inference. *Journal of Experimental Psychology:Learning, Memory and Cognition* 23: 441–458.

Oaksford, M. and U. Hahn (2004). A Bayesian Approach to the Argument from Ignorance. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie experimentale*, 58(2): 75–85.

Oaksford, M. and U. Hahn (2006). A Bayesian Approach to Informal Argument Fallacies. *Synthese* 152 (2): 207–236.

Over, D. and J. Evans (2003). The Probability of Conditionals: The Psychological Evidence. *Mind and Language* 18 (4): 340–358.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence.* Oxford: Oxford University Press.

Pfeifer, N. (2013). On Argument Strength. In: F. Zenker (ed.): *Bayesian Argumentation: The Practical Side of Probability.* Dordrecht: Springer, pp. 185–193.

Pollock, J. (1967). Criteria and our Knowledge of the Material World. *Philosophical Review* 76: 28–62.

Pollock, J. (1987). Defeasible Reasoning. *Cognitive Science* 11: 481–518.

Popper, K. and D. Miller (1983). A Proof of the Impossibility of Inductive Probability. *Nature* 302: 687–688.

Reiter, R. (1978). On Reasoning by Default. In: *Association for Computational Linguistics, Proceedings of the 1978 Workshop on Theoretical Issues in Natural Language Processing*, pp. 210–218.

Reiter, R. (1980). A Logic for Default Reasoning. *Artificial Intelligence* 13: 81–132.

Rottman, B., and R. Hastie (2014). Reasoning about Causal Relationships: Inferences on Causal Networks. *Psychological Bulletin* 140: 109–139.

Schurz, G. (1991). Relevant Deduction. *Erkenntnis* 35: 391–437.

Singmann, H. and K. C. Klauer (2011). Deductive and Inductive Conditional Inferences: Two Modes of Reasoning. *Thinking and Reasoning* 17(3): 247–281.

Singmann, H., K. C. Klauer and S. Beller (2016). Probabilistic Conditional Reasoning: Disentangling Form and Content with the Dual-Source Model. *Cognitive Psychology*, 88: 61–87.

Skovgaard Olsen, N., H. Singmann and K.C. Klauer (2016). The Relevance Effect and Conditionals. *Cognition* 150: 26–36.

Skyrms, B., (1987). Dynamic Coherence and Probability Kinematics. *Philosophy of Science* 54: 1–20.

Sprenger, J. (2011). Hypothetico-Deductive Confirmation. *Philosophy Compass.* DOI: 10.1111/j.1747-9991.2011.00409.x

Stankovich, K. E. and R. F. West (1997). Reasoning Indeendently of Prior Belief and Individual Differences in Actively Open-Minded Thinking. *Journal of Educational Psychology* 89: 342–357.

Stenning, K. and M. van Lambalgen (2008). *Human Reasoning and Cognitive Science.* Cambridge: MIT Press.

Stevenson, R. J. and D. E. Over (2001). Reasoning from Uncertain Premises: Effects of Expertise and Conversational Context. *Thinking and Reasoning* 7: 367–390.

Suppes, P. (1966). Probabilistic Inference and the Concept of Total Evidence. In: J. Hintikka and P. Suppes: *Aspects of Inductive Logic.* Amsterdam: North Holland, pp. 49–65.

Van Eemeren, F. and R. Grootendorst (1995). The Pragma-Dialectical Approach to Fallacies. In: H V. Hansen and R. C. Pinto (eds.): *Fallacies: Classical and Contemporary Readings.* Penn State University Press.

Van Eemeren, F. and R. Grootendorst (2004). *A Systematic Theory of Argumentation. The Pragma-Dialectical Approach.* Cambridge: Cambridge University Press.

Van Fraassen, B., R. I. G. Hughes and G. Harman (1986). A Problem for Relative Information Minimizers Continued. *British Journal for the Philosophy of Science* 37: 453–463.

Walton, D. (1995). *A Pragmatic Theory of Fallacies.* Tuscaloosa: University of Alabama Press.

Walton, D. (2008). *Argumentation Schemes.* Cambridge: Cambridge University Press.

Walton, D. (2011). Defeasible Reasoning and Informal Fallacies. *Synthese* 179: 377–407.

Walton, D. (2013). *Methods of Argumentation.* Cambridge: Cambridge University Press.

Zhao, J., V. Crupi, K. Tentori, B. Fitelson and D. Osherson (2012). Updating: Learning Versus Supporting. *Cognition* 124: 373–378.

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