Argument Quality in Real World Argumentation

Ulrike Hahn

Dept. of Psychological Sciences,

Birkbeck College, University of London

***correspondence: u.hahn@bbk.ac.uk (Ulrike Hahn)**

Abstract

The idea of resolving dispute through the exchange of arguments and reasons has been central to society for millennia. We exchange arguments as a way of getting at the truth in contexts as diverse as science, the court room, and our everyday lives. In democracies, political decisions should be negotiated through argument, not deception, or even worse, brute force. If argument is to lead to the truth or to good decisions, then some arguments must be better than others and 'argument strength' must have some meaningful connection with truth. Can argument strength be measured in a way that tracks an objective relationship with truth, and not just mere persuasiveness? This article describes recent developments in providing such measures.

Keywords: argument quality, logic, probability, rationality

Argumentation: What it is and Why it Matters

Argumentation pervades all aspect of our lives: Arguing things through is at the center of our attempts to come to accurate beliefs about the world, decide on a best course of action and, in a society committed to resolving disagreement without force, convincing others of positions we would like to see accepted. In an argumentative exchange, one might employ many different means and tactics. Some, such as lies and distortions, may support short-term victories or promote narrow self-interest, but are unlikely to give rise to good long-term outcomes. Yet, it is not enough that an argument is endorsed with conviction, and in good faith, to make it compelling: it seems a poorer reason for voting for a candidate that she has six letters in her name than that she has previously successfully held office; it seems a poorer reason for believing a suspect committed the crime that she disliked the victim than that she was seen approaching the victim with a lethal weapon of the kind that killed him; and large-scale randomized controlled trials seem better for accurately judging the efficacy of a drug than anecdotal experience. In other words, some arguments or reasons intuitively seem stronger than others. Is this more than a personal preference? Is there a meaningful, objective standard against which some arguments can be judged as better or worse?

When faced with seemingly clear examples such as those just given, it seems obvious that the answer to this question should be 'yes'. Indeed, if all arguments were equally 'good', it seems impossible that argument could reliably guide us toward truth or good decisions: since arguments can be cooked up for any position, 'anything goes' if there is nothing to choose between them. So, a belief in meaningful standards for evaluating 'good argument' is arguably implied in the social practice of argument itself, and the privileged role which we accord it. Yet, two millennia of thought have struggled to articulate such standards. This article outlines the limitations of past attempts, describes progress made in providing such standards for increasingly large fragments of everyday informal argument, and identifies the key features for future accounts to cover the remainder.

Measuring Argument Quality

The Limits of Intuition

It is neither explanatory, nor inferentially justified, to say that arguments are strong simply because intuitively they seem strong. For one, intuitions about strength may diverge. Social psychologists studying persuasion, for example, have sought to establish 'argument quality' through pre-testing [1], but all this establishes is that a particular group of participants viewed some arguments as strong; it does not establish the key issue, namely that they were *right* to do so [2].

This is readily illustrated with so-called fallacies of argumentation [3, 4]. Fallacies are 'traps for unwary reasoners': arguments that might seem strong and convincing but, on closer scrutiny, are not. As will be discussed below, these fallacies provide an important testbed for explanations of argument strength. The very idea of a 'fallacy' indicates that people may not agree in their intuitions about strength. So whose intuitions should count and why [5]? The inference that what seems a good argument is a good argument, that is, arguments that happen to convince are arguments that ought to convince, is an inference from 'is' to 'ought'. Such inference is not compelling: One cannot conclude from the fact that there is litter on the street that one ought to litter (or vice versa) [6]. Hence intuition alone seems insufficient even if a majority were to share that intuition.

Argument evaluation requires an independently motivated, general, normative standard which indicates what arguments we should find convincing [7]. There must be reasons for why we should accept this general standard, and it is those reasons that will ultimately ground the quality evaluation of specific individual arguments. This normative standard then provides the yardstick against which argument quality can (objectively) be measured. Such a standard would not only be helpful to our daily lives, it is also crucial to a whole range of scientific goals. Cognitive psychologists wish to assess the quality of people's argumentation (e.g., [8,9,10]) as part of the long tradition of rationality-focussed research on reasoning, judgment and decision-making [11,12]. Educational psychologists who want to improve argument skills [13,14] must know what counts as good argument. Social psychologists studying persuasion are interested in the many factors that influence persuasive success *other* than the intrinsic quality of the argument [1]. This means they must be able to experimentally manipulate argument quality. For computer scientists, finally, argument quality matters, because artificial intelligence systems providing support in complex, high-level domains aim to provide high-quality advice [15,16], which requires the need to filter, weight and integrate information.

So how can argument quality be measured? What normative standards might be devised? This is the focus of research on rational argument. The present article describes recent developments in this field, highlighting both their theoretical and practical impact.

Past Attempts

The fact that the assessment of argument quality poses both theoretical challenges and speaks to practical needs is nowhere more apparent than in the steady rise of courses (institutional and online), textbooks, and tests, of "critical thinking". Now an industry in its own right, critical

thinking education seeks to improve, among other things, argument evaluation: the ability to distinguish good arguments from bad [4]. Yet while it is easy to find examples of arguments that intuitively seem bad, it is hard to provide a meaningful account of exactly *why* they are bad. Typical guidance in critical thinking textbooks identifies arguments with a particular structure as problematic and cautions against their use, for example, the argument from ignorance ("Ghosts exist, because nobody has proven that they don't") or slippery slope arguments ("If we allow assisted dying, we will end up with involuntary euthanasia").

Historically, the lack of logical validity was identified as the source of weakness: fallacies were fallacies because they were mistaken for deductively valid arguments even though they were not [17]. In the ghosts example above, the premise does not (logically) imply the conclusion, so the premises may be true, yet the conclusions nevertheless false. This is indeed a 'limitation' of sorts, but it is one these arguments share with *most* of everyday argument [18, 19], including arguments we are entirely happy to base actions on: for example, we might consider medicines or GM foods to be safe, because (sufficiently) adverse effects have not been observed [19].

These examples, too, are arguments from ignorance, yet they play a fundamental role in guiding action. So, it cannot be their structure as 'arguments from ignorance' that makes them weak or strong. Rather, their *specific content matters*. This is true of non-deductive argument in general.

By contrast, logical validity rests only on very general, structural properties of arguments. An argument such as:

All men are mortal.	All X are Y.
Socrates is a man.	Z is X.
Therefore:	Therefore:
Socrates is mortal.	Z is Y.

is an example of a wider logical "scheme" for which the conclusion follows necessarily from the premises regardless of whether the argument is about men and mortality or the properties of Martians. As a result, argument strength does not vary between different examples of this argument scheme.

The same is not true for the vast majority of day-to-day arguments. These arguments are ampliative in the sense that the conclusion goes beyond the premises: hearing the engine run

(premise) provides reasonable grounds for believing someone in the car (conclusion), but it is nevertheless possible that nobody actually is. In other words, the truth of the premise cannot guarantee the truth of the conclusion. It simply makes it more likely or more plausible. And the extent to which it does, depends crucially on content features that go beyond those determining logical structure. *The structure of* an argument from ignorance about drug safety does not change depending on whether one clinical trial has been conducted or fifty, or if the trials were thorough or sloppy. Yet, intuitively, the support for the conclusion that the drug is safe does. Merely, pointing out that the argument is not logically valid does nothing to explain this crucial difference.

The limits of logic in everyday, real world argument were the focus of Toulmin's influential book *The Use of Argument* [20]. In its place, Toulmin proposed a framework that has been highly influential. Toulmin breaks arguments down into distinct basic components: the 'claim' (the conclusion to be established), data (the facts appealed to in order to support of the claim), 'warrants' (reasons that support the inferential link between data and claim), 'backing' (basic assumptions that justify particular warrants), 'rebuttals' (exceptions to the claim or the link between warrant and claim) and, finally, 'qualifiers' (indications of the degree of regularity with which claim may be stated, such as 'certain', 'highly probable', 'rare').

Figure 1 (panels A and B) show an example argument analysed by Hastings ([21], pg. 67) using Toulmin's framework. This framework seems useful for identifying the components of an overall argument, which in turn is a precondition for argument evaluation: one can't evaluate the argument before one knows what it is. So, it is no surprise that the Toulmin framework has been used for testing and improving argument skills (see e.g., [22]).

However, beyond the rather trivial fact that an argument is 'better' if a reason (warrant) is provided than if it is not, the framework says virtually nothing about evaluation. In particular, it says nothing about whether the warrant itself is more or less compelling. The example in Figure 1C illustrates this; it has all the features highlighted by the Toulmin model, yet reasons, warrants, and backing provide no actual support to the claim.

This is a necessary consequence of the dialectical approach Toulmin takes. In order to move beyond classical logic, Toulmin imports dialectical notions from the courtroom. Argumentation is an activity that happens between two (or more) parties, and dispute is to be resolved through originally legal concepts such as the burden of proof [20,24]. It is the parties who determine what the individual arguments are, because they are the ones who put them forward. What makes

something a putative reason, is that it is offered as one. However, that is not enough to determine whether or not that reason is good.

Like classical logic, Toulmin's model fails to engage meaningfully with the actual *content* of claim, data, warrant and backing [23]. Consequently, it has nothing to say about the central question of the degree to which a reason provides genuine support, or, for that matter, whether a rebuttal provides a sufficient counter-argument to overturn a warrant (reason).

This same deficiency characterises other approaches to argument quality that have been framed solely around dialectical considerations. So-called pragma-dialectical theories of argumentation [25,26] try to identify procedural rules that govern argumentative discourse (such as "freedom rules" that allow participants to put forward arguments). Trying to understand these conventions is an important project in its own right. What does not work, however, is leveraging these conventions to explain why an argument such as "ghosts exist, because nobody has proven that they don't" is fallacious. On this, it is argued that the argument is weak because it is an illegitimate attempt to "shift a burden of proof". But this fails to explain why the argument seems weak even when read in isolation (without a wider dialectical exchange). It also fails to explain why the argument from ignorance "this drug is safe, because no side effects have been found in clinical trials" does not suffer the same deficiency. In fact, appeals to the burden of proof mistake cause and effect: a poor argument might constitute an illegitimate attempt to shift a burden of proof, but that presupposes that it is a poor argument, it does not explain or define why it is poor in the first place.

In short, argument evaluation must be able to do more than track structural relationships between arguments, or how they relate to dialectical, procedural 'moves', it must also be able to somehow *weigh* them. Only then can one derive any overall conclusion. An obvious path (taken by various computational approaches to argument within artificial intelligence, [27]) is to assign numerical weights – numbers that reflect the strength of an argument. But how should such weights rationally be assigned and how should weights for multiple arguments be combined into an overall evaluation? According to what rules?

Bayesian Argumentation

Bayesian argumentation seeks to provide an answer to these problems (see Box 1). Bayes' rule provides an update rule for revising beliefs in light of evidence. This allows one to think of argument evaluation as learning: how much one's beliefs should change on hearing this argument

or piece of evidence [28,29]. In other words, one can treat argument evaluation as the (hypothetical or actual) process of determining how great a belief change an argument brings about, with stronger arguments giving rise to greater change, all other things being equal: specifically, the prior degree of belief being equal. This will be the case when comparing arguments for the same claim. In order to compare strength across claims, measures that factor out the prior are also available (see, e.g., what [29] call "argument force", or suitable measures of confirmation [30]).

What determines belief change is the diagnosticity of this piece of information. To illustrate with an example: when considering the claim that a drug is safe the argument that 'no side effects have been found in clinical trials' is of no value if *no* clinical trials have been conducted. And it is of no value precisely because in this case the observation "no side effects" is as likely to observed if the drug is safe than if it is not. By the same token, if 10 studies have been conducted it is much more likely that they will be observed if the drug is not safe, than if only one such study has been conducted. The example may seem so obvious that it is easy to miss how powerful a notion diagnosticity is. What it forces one to do, is not only consider a generative model of how the evidence (argument) would come about if the underlying hypothesis (claim) were true, but also how it could come about if it were *not*, thus forcing consideration of alternatives. This distinguishes the Bayesian approach from attempts to capture informal argument through defeasible generalization, the latter of which are vulnerable to irrelevant generalizations due to this omission, see for discussion [31]. Moreover, the concept of diagnosticity applies whether there is a direct, causal link between evidence and claim or the evidence is merely correlated with other, causally efficacious factors. Shoe size correlates with reading ability (via an underlying link with age), hence shoe size supports inferences about reading, even though there are no general laws that would connect reading directly with shoes. In other words, diagnosticity provides a general-purpose measure of informational relevance.

This is what ultimately allows the probabilistic framework to provide a formal, explanatory account of the traditional catalogue of fallacies of argumentation (see Table 1). It has long been observed that the thread running through the catalogue is (failures of) relevance (see also [26,32,33,34]). Appeals to authority are often fallacious, because mere authority (in particular in the sense of power) need have no systematic connection with truth. In the language of diagnosticity: that an authority claims something to be the case may be little or even no more likely if the claim is true, than if it is false. In this case, authority has little to no relevance. However, where the authority rests on relevant expertise, the opinion of that authority is likely to be highly diagnostic. A probabilistic perspective centered around diagnosticity captures these distinctions.

Because diagnosticity captures relevance, the probabilistic framework arguably offers an account of the majority of fallacies in the catalogue (see Table 1). Crucially, capturing relevance through diagnosticity is not merely a re-labelling exercise: it tells us where to look in order to determine relevance and it provides a numerical measure that supports quantitative inference.

Diagnosticity distinguishes between instantiations of these putative fallacies that do indeed seem weak, from those that seem acceptable, thus revealing that 'fallacies' are not fallacies simply because of their structure, but because of their specific content. At the same time, however, the influence of structure is clarified. The probabilistic framework allows one to compare structurally different types of argument, specified over the same set of facts in the world, that is, over the same probability distribution. The probabilistic treatment not only forces clarification of structurally different types of argument from ignorance (see Box 2), it shows also that they will vary in strength when applied in the same context. Different people may vary in the exact underlying probabilities they assign, but those probabilities determine the strength not just of a particular argument, but also other, systematically related arguments.

Furthermore, the probability calculus allows correction in cases where intuition goes wrong. For example, a probabilistic treatment distinguishes clearly between cases of circular argument and circular justifications that genuinely support belief change and ones that do not [36], even where intuition may be hopelessly at sea. The probability calculus may, at heart, be just "formalised common sense" as LaPlace [56] famously put it, but it provides a guide also where intuition breaks down.

By providing a treatment of fallacies, the probability calculus demonstrably expands normative standards for everyday argument significantly beyond what classical logic or Toulmin like accounts can do. To be clear, though it has been outlined how the treatment extends to the bulk of the fallacies in the catalogue [34], only a handful have received detailed probabilistic analysis. In each case, however, this has not just revealed important new distinctions (e.g., Box 2), it has also fostered new empirical work aimed at probing lay intuitions: there is now a growing body of experimental work on arguments from ignorance [59, 60], ad hominem arguments [42,51,61], slippery slope arguments [9,62,63], circular arguments [18,29,64], or appeals to popular opinion [52]. That each of these schemes has been found to contain theoretical and empirical richness of its own should encourage future work.

However, the probabilistic framework is not limited to the fallacies and recent work has started to expand the Bayesian framework to everyday, informal argument schemes that are putatively good,

not bad. The literature on informal argument has catalogued 60+ schemes [65]. Though not all of these may end up being theoretically distinct, there is also reason to believe that the catalogue is incomplete and that there are, for example, distinct types of causal argument that are not included in this typology [24]. Hahn and Hornikx [31] demonstrate for several sample schemes, including the practically important appeal to expert opinion [10,66], how a probabilistic treatment may help achieve the long-held goal within the scheme-based tradition of a comprehensive treatment of informal argument that (1) is computationally explicit, (2) has a solid normative foundation, and (3) can be used for education. This is further supported by recent demonstration of how the general notion of argument 'cogency' can be recast in probabilistic terms [67].

Justifying the Normative Standard

As demonstrated, the Bayesian framework significantly expands the portion of everyday, real-world argument for which there is an applicable normative standard. The Bayesian framework also provides a measure of argument quality that indicates not how beliefs actually change, but how they should change if we were fully rational. Like logic, it sets out ideals that reasoners may or may not live up to. But what is the basis of that "should"? Why should reasoners accept classical logic or the Bayesian framework as a normative standard?

Probability theory ensures "coherence" among beliefs. In effect, it says "if you have these beliefs, then you cannot have these other beliefs" or your beliefs will be inconsistent. In this way, it is just like classical logic. Both evaluate the extent to which a conclusion follows from a particular set of premises. Neither logic nor probability will tell you whether the premises themselves are true, so neither can tell you that the conclusion actually is true (unless premises or conclusion are a logical truth). On reflection, asking anything more from a measure of argument quality is unrealistic. In particular, argument quality can't hinge on having to await the final verdict on the truth or falsity of particular propositions, given that these are typically themselves part of a wider network of beliefs [68, 69].

Nevertheless, both logic and probability have implications beyond coherence to correspondence, that is, the degree of match between beliefs and world. Consider first *in*consistency and *in*coherence: inconsistent claims, that is, logical contradictions cannot be true. As noted above, inconsistency is just a special case of (probabilistic) incoherence [70]. So-called Dutch book theorems demonstrate that if one's beliefs are incoherent, then, if one bets in line with one's beliefs, there will be bets where a loss is guaranteed, regardless of how the world turns out (see e.g., [71]). In effect, we place bets whenever we choose between actions with uncertain outcomes. Incoherent

probabilities will mean placing bets against nature we cannot win. But the probability calculus does more than rule out these types of sure losses. The Bayesian framework links to the accuracy of our beliefs. A common measure of (in)accuracy is the squared error of a forecast to true value, also known as the Brier score [72,73], proposed originally to evaluate the prediction accuracy of meteorologists. Formal results show that if error is measured via the Brier score then being Bayesian (coherent probabilities plus updating via Bayes' rule) will minimise the inaccuracy of an agents' beliefs across "all possible worlds" that can be entertained by that agent, that is, regardless of how the world actually turns out [74]. In fact, measuring (in)accuracy this way uniquely implies "being Bayesian".

The normative 'bite' of the Bayesian framework, then, stems from the fact that coherence has implications for correspondence. There is a normative basis in instrumental rationality such that if an agent wants to minimise inaccuracy (in this sense), then that agent should be Bayesian. Of course, we are entirely free to choose not to, but if we do, then the Bayesian framework provides a normative standard.

Beyond Bayesian Argumentation?

Our discussion of Bayesian argumentation illustrates what the ingredients of an independently justified normative framework for argument quality are, and what benefits such a framework provides. The ultimate goal, of course, are standards that would apply to all of everyday argument. Known argument schemes for informal argument capture significant proportions of real-world argument, but what does that leave unaccounted for, and how could norms for the remainder be advanced? Examining this is the aim of the "Critical Analysis Project", an applied project that both seeks to further develop norms for rational argument and to provide an accessible way of communicating argument quality to the general public. This web-based project (http://critical-analysis.org/) conducts critical scrutiny of published newspaper opinion pieces. It sheds light not just on what aspects of argument quality go beyond known schemes, it also provides insight into future directions. Crucially, the role of the analysts in the project is that of a referee, not a (further) party in the debate. This restricts analysis to "objective" criteria in the sense of criteria that anyone with accuracy goals could subscribe to. This, it turns out, once again, seems to give a central role for (in)consistency. Everyday discourse, on inspection, shows many looser, types of inconsistency beyond those directly captured by logic and probability theory. These may involve actions (attacking someone for interrupting and then interrupting oneself) or relations between beliefs and actions (a refusal to 'put one's money where one's mouth is').

Speculatively, this suggests that 'consistency' is the fundamental notion underpinning rational argument. This opens up a new, general, perspective on relevance. Relevance and inconsistency are two sides of the same coin: a reason is relevant to a claim precisely to the extent that it could not be accepted without adjusting our views on the claim, if we are to avoid 'inconsistency'. An irrelevant reason, by contrast, is one we could readily accept without anything having to change. Logic and probability merely spell out particular aspects of this more general, fundamental notion. This makes clear why the role of classical logic in everyday argument is limited. The relevance constraints imposed by logical contradiction are weak: they only rule out states of affairs that are strictly impossible. But most everyday argument concerns whether or not things that are possible actually happen to be true, so these constraints leave most arguments in the space of the allowed. The Bayesian framework provides further consistency, and hence relevance, constraints. The key challenge for future research will be to develop similarly clearly articulated notions of (in)consistency that go beyond them.

Concluding Remarks

Assessing argument quality is a central theoretical and practical concern, yet, it has been difficult to provide non-trivial accounts of what makes arguments 'good'. While critical thinking books and courses are helpful in getting people to identify arguments, the promise of providing meaningful tools for critical evaluation has been hard to fulfil. The Bayesian framework offers a tool with independent normative justification. It deals by design with the problem of aggregation across multiple pieces of evidence, both for and against a claim [75]. And though it accords well with fundamental intuitions, it also corrects where intuitions go badly wrong. This, in principle, allows the Bayesian framework to provide normative guidance for any area concerned with argument quality and evidence evaluation whether this be education, the courts [76,77,78,79], or intelligence analysts [80,81]. Consequently, the Bayesian framework has enabled new descriptive projects assessing lay understanding of argument quality. For this, the framework has both provided new, theoretically motivated questions for inquiry and new methodological tools [82,19] (see Outstanding Questions). In particular, it becomes possible to compare argument evaluation across very different domains (for example, everyday contexts versus socio-scientific communications) because very different contents can be compared against the same normative standard. The challenge for the future will be developing equally rigorous standards for the remainder of everyday, informal argument. This is clearly an ongoing project, but the discernible shape of what complete account would look like, should now be clear.

Acknowledgements

The author was supported by the Humboldt Foundation's Anneliese Maier Research Award.

References

[1] Petty, R. E. (2018). *Attitudes and persuasion: Classic and contemporary approaches*. New York, N.Y.: Routledge.

[2] O'Keefe, D. J. (2006). Pragma-dialectics and persuasion effects research. P. Houtlosser & A. van Rees (Eds.) *Considering Pragma-Dialectics*. New York, N.Y.: Routledge. pp- 235-243.

[3] Hamblin, C. L. (1970). *Fallacies*. London: Methuen.

[4] Pirie, M. (2015). *How to win every argument: The use and abuse of logic*. Bloomsbury Publishing.

[5] Weinberg, J., Nichols, S., & Stich, S. (2001). Normativity and Epistemic Intuitions. *Philosophical Topics, 29*(1/2), 429-460.

[6] Elqayam S., Evans J. S. (2011). Subtracting "ought" from "is": Descriptivism *versus* normativism is the study of human thinking. Behav. Brain Sci. 34, 233–248 10.1017/S0140525X1100001X

[7] Corner, A.J. & Hahn, U. (2013) Normative theories of argumentation: Are some norms better than others? Synthese, 190, 3579-3610.

[8] Kuhn, D. (1991). *The skills of argument.* Cambridge, England: Cambridge University Press.

[9] Corner, A., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory and Language*, *64*(2), 133-152.

[10] Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The Appeal to Expert Opinion: Quantitative support for a Bayesian Network Approach. Cognitive Science, 40*,* 1496-1533.

[11] Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and brain sciences*, *23*(5), 645-665.

[12] Kahneman, D. (2011). *Thinking, fast and slow* (Vol. 1). New York: Farrar, Straus and Giroux.

[13] Papathomas, L., & Kuhn, D. (2017). Learning to argue via apprenticeship. *Journal of experimental child psychology*, *159*, 129-139.

[14] Macagno, F. (2016). Argument relevance and structure. Assessing and developing students' uses of evidence. *International Journal of Educational Research*, *79*, 180-194.

[15] Hopgood, A. A. (2016). *Intelligent systems for engineers and scientists*. Boca Raton, FL: CRC press, Taylor & Francis Group.

[16] Neapolitan, R. E. (2012). *Probabilistic reasoning in expert systems: theory and algorithms*. CreateSpace Independent Publishing Platform.

[17] Woods, J., Irvine, A., & Walton, D. N. (2004). *Argument*: *Critical thinking, logic and the fallacies* (Rev. ed.). Toronto, Ontario, Canada: Prentice Hall.

[18] Hahn, U. & Oaksford, M. (2007) Induction, deduction, and argument strength in human reasoning and argumentation. In: Feeney, A. and Heit, E. (eds.) *Inductive Reasoning: Experimental, Developmental, and Computational Approaches.* Cambridge, UK: Cambridge University Press, pp. 269-301. ISBN 9780521672443.

[19] Corner, A.J. & Hahn, U. (2009) Evaluating Science Arguments: Evidence, Uncertainty & Argument Strength. Journal of Experimental Psychology: Applied, 15, 199-212.

[20] Toulmin, S. E. (1958; 2003). *The Uses of Argument* (update edition). Cambridge: Cambridge University Press.

[21] Hastings, A. C. (1962). A Reformulation of the Modes of Reasoning in Argumentation. Unpublished doctoral dissertation, Northwestern University.

[22] Klieger, A., & Rochsar, A. (2017). Impartation of argumentation skills: impact of scaffolds on the quality of arguments. *Journal of Advances in Education Research*, *2*(3), 183-190.

[23] Hahn, U., Blum, R., & Zenker, F. (2017) Causal Argument. In, M. Waldmann (ed.) The Oxford Handbook of Causal Cognition. Oxford, UK: Oxford University Press.

[24] Hahn, U. & Oaksford, M. (2007) The burden of proof and its role in argumentation. Argumentation, 21, 39-61.

[25] Eeemeren, F. H. van & R. Grootendorst (2004) A Systematic Theory of Argumentation. The Pragma-Dialectical Approach, Cambridge University Press, Cambridge.

[26] Walton, D. N. (1995). *A pragmatic theory of fallacy.* Tuscaloosa, AL: The University of Alabama Press.

[27] Dunne, P. E., Hunter, A., McBurney, P., Parsons, S., & Wooldridge, M. (2011). Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, *175*(2), 457-486.

[28] Eva, B., & Hartmann, S. (2018). Bayesian argumentation and the value of logical validity. Psychological Review, 125(5), 806.

[29] Hahn, U. & Oaksford, M. (2007) The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. Psychological Review, 114, 704-732.

[30] Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, *103*(1), 107-119.

[31] Hahn, U. & Hornikx, J. (2016). A normative framework for argument quality: Argumentation schemes with a Bayesian foundation. Synthese, 193, 1833-1873.

[32] Walton, D. N. (1998). *The new dialectic: Conversational contexts of argument.* Toronto, Ontario, Canada: University of Toronto Press.

[33] Walton, D. N. (2004). *Relevance in argumentation.* Mahwah, NJ: Erlbaum

[34] Hahn, U., & Oaksford, M. (2006). A Bayesian approach to informal argument fallacies. *Synthese, 152,* 207–236.

[35] Collins, P.J. & Hahn, U. (2017) Fallacies of argumentation. In. Thompson, V. and Ball, L. (eds.) International Handbook of Thinking and Reasoning. New York, N.Y. Routledge.

[36] Oaksford, Mike, and Nick Chater. "A rational analysis of the selection task as optimal data selection." *Psychological Review* 101.4 (1994): 608.

[37] Rumain, B., Connell, J., & Braine, M. D. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults: If is not the biconditional. *Developmental psychology*, *19*(4), 471.

[38] Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2014). The negations of conjunctions, conditionals, and disjunctions. *Acta Psychologica*, *151*, 1-7.

[39] Kaye, D. H., & Koehler, J. J. (1991). Can jurors understand probabilistic evidence? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *154*(1), 75-81.

[40] Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive psychology*, *38*(2), 191-258.

[41] Over, D., Douven, I., & Verbrugge, S. (2013). Scope ambiguities and conditionals. *Thinking & Reasoning*, *19*(3-4), 284-307.

[42] Hoeken, H., Timmers, R., & Schellens, P. J. (2012). Arguing about desirable consequences: What constitutes a convincing argument? *Thinking & reasoning*, *18*(3), 394-416.

[43] Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216.

[44] Adams, R. C., Sumner, P., Vivian-Griffiths, S., Barrington, A., Williams, A., Boivin, J., ... & Bott, L. (2017). How readers understand causal and correlational expressions used in news headlines. *Journal of experimental psychology: applied*, *23*(1), 1.

[45] Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451.

[46] Ricco, R. B. (2007). Individual differences in the analysis of informal reasoning fallacies. *Contemporary Educational Psychology*, *32*(3), 459-484.

[47] Schumann J., Zufferey S. & Oswald S. (2019): What makes a straw man acceptable? Three experiments assessing linguistic factors. *Journal of Pragmatics 141*, 1-15. https://doi.org/10.1016/j.pragma.2018.12.009

[48] van Eemeren, F. H., Garssen, B., & Meuffels, B. (2015). The disguised ad baculum fallacy empirically investigated. Strategic maneuvering with threats. In *Reasonableness and Effectiveness in Argumentative Discourse* (pp. 815-826). Heidelberg, Germany: Springer, Cham.

[49] Van Eemeren, F. H., Garssen, B., & Meuffels, B. (2009). *Fallacies and judgments of reasonableness: Empirical research concerning the pragma-dialectical discussion rules*(Vol. 16). Springer Science & Business Media.

[50] Ervas, F., Gola, E., Ledda, A., & Sergioli, G. (2015). Lexical ambiguity in elementary inferences: an experimental study. *Discipline filosofiche*, *22*(1), 149-172.

[51] Bhatia, J. S., & Oaksford, M. (2015). Discounting testimony with the argument ad hominem and a Bayesian congruent prior model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1548.

[52] Hornikx, J., Harris, A. J., & Boekema, J. (2018). How many laypeople holding a popular opinion are needed to counter an expert opinion? *Thinking & Reasoning*, *24*(1), 117-128.

[53] Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In *Advances in neural information processing systems* (pp. 59-68).

[54] Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *58*(2), 75.

[55] Hahn, U. (2011). The problem of circularity in evidence, argument and explanation. Perspectives on Psychological Science, 6, 172-182.

[56] LaPlace, P.S. (1814) *Essai philosophique sur les Probabilités*. *Œuvres complètes de Laplace*, tome VII, p. cliii, Paris: Gauthier-Villars, 1878-1912.

[57] Hahn, U., & Oaksford, M. (2008). Inference from absence in language and thought. *The probabilistic mind: Prospects for Bayesian cognitive science*, 121-42.

[58] Hahn, U., & Oaksford, M. (2012). Rational argument. *The Oxford handbook of thinking and reasoning*, 277-300.

[59] Harris, A.J.L., Corner, A.J. & Hahn, U. (2013) James is polite and punctual (and useless): A Bayesian formalization of faint praise. Thinking & Reasoning, special issue on 'The new paradigm in Reasoning'. Thinking & Reasoning, 19, 414-429.

[60] Hsu, A. S., Horng, A., Griffiths, T. L., & Chater, N. (2017). When absence of evidence is evidence of absence: Rational inferences from absent data. Cognitive science, 41, 1155-1167.

[61] Harris, A.J.L. and Hsu, A.S. and Madsen, Jens K. (2012) Because Hitler did it! quantitative tests of Bayesian argumentation using ad hominem. *Thinking and Reasoning* 18 (3), pp. 311-343. ISSN 1354-6783.

[62] Haigh, M., Wood, J. S., & Stewart, A. J. (2016). Slippery slope arguments imply opposition to change. *Memory & Cognition*, *44*(5), 819-836.

[63] Deak, C., & Saroglou, V. (2017). Terminating a child's life? Religious, moral, cognitive, and emotional factors underlying non-acceptance of child euthanasia. *Psychologica Belgica*, *57*(1), 59.

[64] Mercier, H., Bernard, S., & Clément, F. (2014). Early sensitivity to arguments: How preschoolers weight circular arguments. *Journal of Experimental Child Psychology*, *125*, 102-109.

[65] Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation Schemes*. Cambridge: Cambridge University Press.

[66] Hahn, U., Harris, A. J., & Corner, A. (2016). Public reception of climate science: Coherence, reliability, and independence. *Topics in cognitive science*, *8*(1), 180-195.

[67] Godden, D., & Zenker, F. (2018). A probabilistic analysis of argument cogency. *Synthese*, *195*(4), 1715-1740.

[68] Quine WV: From a Logical Point of View. 1953, Cambridge, MA, Harvard University Press, 2nd Edition.

[69] Weisman, K., & Markman, E. M. (2017). Theory-based explanation as intervention. *Psychonomic bulletin & review*, *24*(5), 1555-1562.

[70] Ramsey, F.P. (1926) "Truth and Probability", in Ramsey, 1931, *The Foundations of Mathematics and other Logical Essays*, Ch. VII, p.156-198, edited by R.B. Braithwaite, London: Kegan, Paul, Trench, Trubner & Co., New York: Harcourt, Brace and Company.

[71] Hájek, A., & Hitchcock, C. (2016). The Oxford Handbook of Probability and Philosophy. Oxford: Oxford University Press.

[72] Brier,G.W. (1950).Verification of forecasts expressed in terms of probability.*Mon. Wea.Rev*.78,1–3.doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[73] Siegert, S. (2017). Simplifying and generalising Murphy's Brier score decomposition. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 1178-1183.

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

[75] Kadane, J. B., & Schum, D. A. (2011). *A probabilistic analysis of the Sacco and Vanzetti evidence* (Vol. 773). New York: John Wiley & Sons.

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

[77] Smit, N. M., Lagnado, D. A., Morgan, R. M., & Fenton, N. E. (2016). Using Bayesian networks to guide the assessment of new evidence in an appeal case. *Crime science, 5(*1), 9.

[78] Neil, M., Fenton, N., Lagnado, D., & Gill, R. D. (2018). Modelling competing legal arguments using Bayesian model comparison and averaging. *Artificial Intelligence and Law*, 1-28.

[79] de Zoete, J., Fenton, N., Noguchi, T., & Lagnado, D. (2019). Resolving the so-called "probabilistic paradoxes in legal reasoning" with Bayesian Networks. Science & Justice.

[80] Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M.M., Horowitz, M., Merkle,E. & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world. *Journal of Experimental Psychology: Applied, 21*, 1-14.

[81] Pilditch, T. D., Fenton, N., & Lagnado, D. (2019). The zero-sum fallacy in evidence evaluation. *Psychological science, 30(*2), 250-260.

[82] Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual review of psychology*, *71*.

The Toulmin Framework for Argument Analysis

Figure 1 Hastings' [21] Toulmin diagram (Panel B) of an argument from cause to effect taken from a speech by Dwight Eisenhower (Panel A), see also Hahn et al. [23]. Panel C: A nonsense argument that nevertheless readily fulfils the Toulmin model.

Box 1. Being Bayesian

"Being Bayesian" means two things: 1) assigning probabilities (which represent degrees of belief) in accordance with the laws of probability, and 2) the use of Bayes' rule for belief revision.

1. The laws of probability provide rules for assigning degrees of belief in a way that is "coherent". For example, one's degree of belief in the truth of a claim and one's belief in the truth of its negation must sum to one. This is not merely an arbitrary mathematical convention. A probability of *p=1* represents certainty, and the fact that a claim and its negation sum to one reflects a logical constraint: either it is the case that Donald Trump is president of the United States or it is *not the case* that Donald Trump is president of the United States.

2. Bayes' rule:

$$P(C|e) = \frac{P(e|C) * P(C)}{P(e|C) * P(C) + P(e|\neg C) * P(\neg C)}$$

Bayes' rule details how to combine the prior degree of belief in a claim, *P(C),* with new evidence *e*, to calculate the revised (posterior) degree of belief, *P(C|e)*, that is, the probability of the claim given the evidence.

Evidence will change beliefs more the more *diagnostic* that evidence is, that is, the greater the so-called likelihood ratio:

$$P(e|C)/P(e|\neg C)$$

In other words, evidence will lead to greater belief change the more likely it is to be found only if the claim is true, as opposed to false (the symbol ¬ represents negation). Evidence that is equally likely either way, is entirely non-diagnostic.

Box 2. Types of argument from ignorance

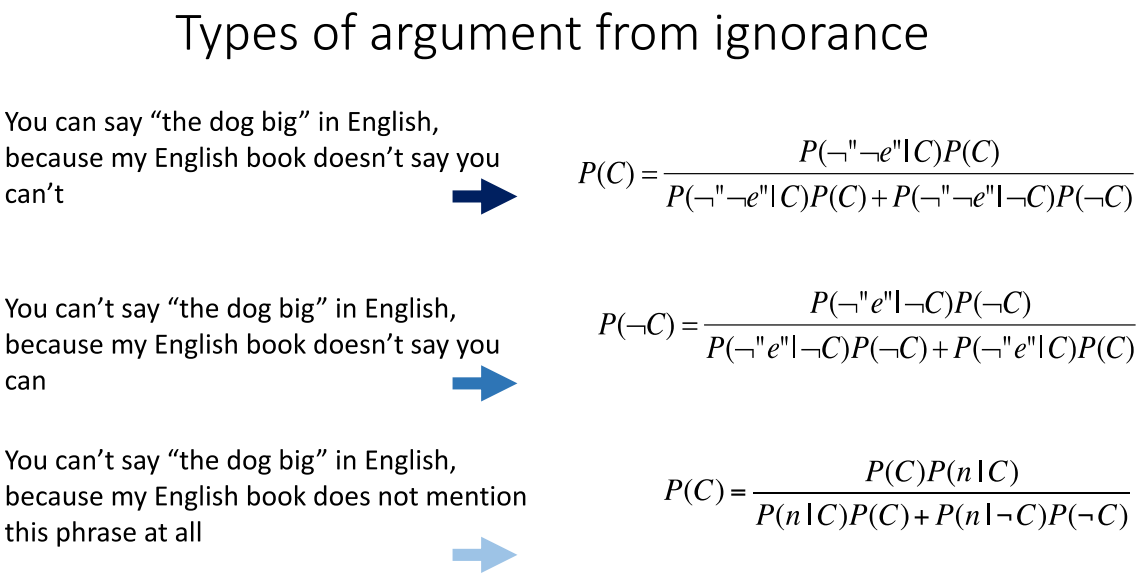

Figure I. Arguments from Ignorance with Different Logical Structures.

The examples in Fig. I mimic potential uses of implicit negative evidence in language acquisition (see e.g., [57]) to illustrate three distinct types of argument from ignorance. These vary both in whether the conclusion of the argument is a positive or a negative, and the types negative evidence or involved.

As the examples make clear, implicit ‘negative evidence’ can consist of the absence of explicit evidence *against* a claim (top example), the absence of explicit evidence *for* a claim (middle example), or the absence of any explicit evidence either way (bottom example).

To capture this, the formalisation (see e.g., [58]) assumes three possible types of evidence: explicit positive evidence “e” (i.e., the textbook explicitly says the linguistic construction exists), explicit negative evidence “¬*e”* (i.e., the textbook explicitly says the construction doesn’t exist), and no explicit evidence, represented as *n* for ‘nothing’ (i.e., the textbook is silent on the issue). On such a scheme, the absence of explicit positive evidence then is ¬ “e” (i.e., the negation of an explicit, positive evidence statement), and absence of explicit negative evidence corresponds to the negation of explicit negative evidence: ¬ “¬*e”*. This gives rise to the respective versions of Bayes’ rule for each of the three cases.

Because the probabilities in all three arguments are systematically related (e.g., the textbook must either say the construction is grammatical, say it is ungrammatical or fail to mention it), a single set of facts about the world will determine the strength of all three. As a result, there will be systematic relations between these types in terms of argument strength. The figure on the left (Fig. II) illustrates this by showing the results of the application for the three versions of Bayes’ rule, and hence the three different argument strengths, calculated over the same sample probability distribution. This illustrates how the Bayesian formalisation not only provides a quantitative treatment of initial informal arguments, but yields novel insights about the structure of different arguments and the relationships between them.

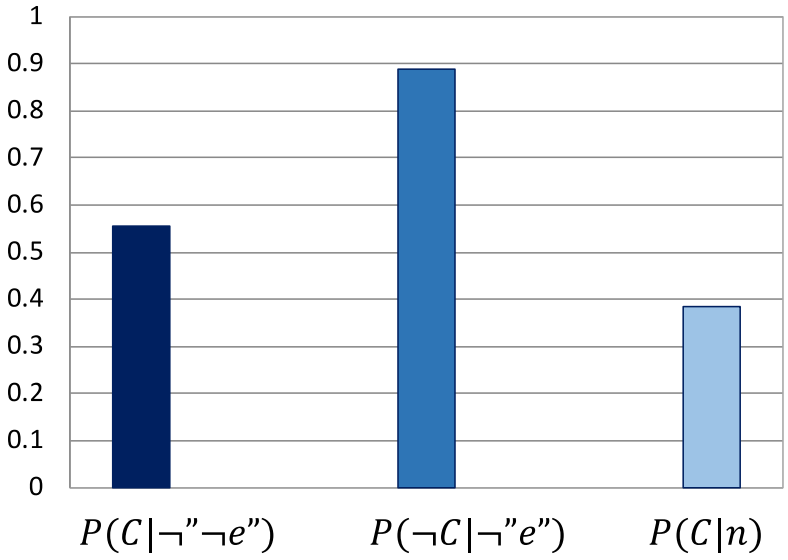

Fig. II. Sample Posteriors Across a Shared Probability Distribution

Table 1. Representative fallacies from the literature, adapted from [35]. All cases, other than those indicated by #, have been explicated with logical/probabilistic norms. Unless otherwise indicated (again by #) the illustrative experimental work assesses lay peoples' performance vis a vis these norms. Finally, the symbol * is used to indicates experimental work that, although not directly on the fallacy in question, is closely related.

Type	Fallacy	Example/Explanation	Exp. study
Invalid inferences of propositional logic, e.g.,	Denial of Antecedent	If p then q; not p; therefore not q.	Oaksford & Chater, 1994 [36]
	Affirmation of Consequent	If p then q; q; therefore p.	Oaksford & Chater, 1994 [36]
	Improper Transposition	If p then q; therefore if not p then not q.	Rumain et al., 1983 [37]
	Denying conjunct	Not both p and q; not p, therefore not q.	Khemlani et al., 2014* [38]
	Commutation of Conditional	If p then q; therefore if q then p.	Kaye & Koehler, 1991 prosecutors fallacy [39]
Invalid syllogisms , e.g.,	Illicit Process - Major	All X are Y; no Z are X; therefore no Z are Y.	Chater & Oaksford, 1999 [40]
Quantifier	Quantifier Shift	For every x there is a cause y; therefore there is some cause y which is the cause of every x	
Modality	Modal Scope	Necessarily, if p then q; therefore, if p, then necessarily q.	Over et al., 2013 * [41]
Causal	Ad consequentiam	Appeal to consequences; belief accepted/rejected in virtue of good/bad consequences	Hoeken et al., 2012 [42]
	Common Cause	X believed to cause Y, when both caused by Z	Waldman & Hagmeyer, 2005 [43]
	Cum hoc, ergo propter hoc	X and Y co-occur; therefore X causes Y	Adams et al., 2017 [44] #
	Post hoc, ergo propter hoc	X precedes Y; therefore X causes Y	Lagnado & Sloman, 2006 [45]
	Slippery Slope	Inoffensive move W leads, by steps X, Y etc, to unacceptable Z	Corner et al., 2011 [9]
Diversion	Irrelevant Conclusion		Ricco, 2007 [46] #

		Ignoratio elenchi; conclusion irrelevant to issue at hand	
	Straw Man #	Argues against a caricature of an opponent's argument	Schumann et al., 2019 [47] #
Emotion	Ad baculum	Threatens force if conclusion not accepted	Van Eemeren et al., 2015 [48] #
	Ad misericordiam	Argues for conclusion using irrelevant appeal to pity	Van Eemeren et al., 2009 [49] #
Language	Ambiguity	Structural or lexical (equivocation); cogency depends on ambiguous meanings	Ervas et al. 2015 [50]
	Complex Question #	Abusive presupposition: When did you stop beating your wife?	
	Composition/Division	Properties of whole attributed to parts and vice versa	
	Vagueness #	Hinges on vagueness of argument's terms.	
Source	Ad hominem	Attacks the source, not content, of argument	Bhatia & Oaksford, 2015 [51]
	Ad populum	Cites popularity of argument, not content	Hornikx et al., 2018 [52]
	Ad verecundiam	Appeals to authority	Harris et al., 2016 [9]
Other	Accident	Hasty generalization from very small sample	Tenenbaum, 1999 [53]
	Ad ignorantiam	Argument from ignorance	Oaksford & Hahn, 2004 [54]
	Question Begging	Petitio Principii; premise assumes conclusion	Hahn & Oaksford, 2007 [29]