PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:

xxxxx, xxxxx, xxxx

Keywords:

explanation, argumentation, explainable AI, argument generation

Author for correspondence:

Ulrike Hahn

e-mail: u.hahn@bbk.ac.uk

THE ROYAL SOCIETY PUBLISHING

Argument and Explanation

Ulrike Hahn[1], Marko Tešić[2]

Birkbeck, University of London, UK

[1]u.hahn@bbk.ac.uk

[2]m.tesic@bbk.ac.uk

In this paper, we bring together two closely related, but distinct, notions: argument and explanation. We clarify their relationship. We then provide an integrative review of relevant research on these notions, drawn both from the cognitive science and the AI literatures. We then use this material to identify key directions for future research, indicating areas where bringing together cognitive science and AI perspectives would be mutually beneficial.

1. Introduction

Arguments and explanations are invaluable elements of our everyday lives. Arguments help us establish support for claims and play a role in changing people's beliefs about these claims, while explanations provide us with an understanding of the world around us. Due to their pervasiveness and practical importance in our lives, argument and explanation have became the focus of extensive research within philosophy, psychology, and artificial intelligence (AI). However, that research has not seen the degree of mutual integration it deserves.

The concepts of 'argument' and 'explanation' are closely intertwined and they are multiply interrelated within cognitive science and AI. The aim of the present paper is to provide broad overviews of research areas that, by their content, should be deeply connected, but, presently, remain almost wholly separate. In order to bridge those divides, we highlight the central issues within research on argument and explanation in both cognitive science and AI, respectively. Specifically, we proceed as follows: We (i) first set the stage with a brief discussion of the general concepts of argument and explanation; we then (ii) go through the respective literatures on argumentation in both AI and cognitive science, in particular psychology; we then do the same for (iii) explanation research. Finally, (iv) we bring argument and explanation together in order to suggest ways in which insights from argumentation research might inform explanation research, and vice versa, both within and across AI and cognitive science.

2. Argument vs. Explanation

What is an argument and what an explanation? We all know what they are intuitively. Beyond that, however, providing more explicit understanding of both notions and their relationship is not entirely trivial –precisely because both concepts seem so closely linked. Both argument and explanation have a common point of departure. First and foremost, both are answers to a *why question*: why is something the case? And both involve the provision of reasons in response. This parallel is so compelling that argument and explanation, in fact, coincide on some formal accounts.

Chief among these is the classic model of explanation in the philosophy of science, Hempel's covering law model of explanation (also known as the deductive-nomological model) [1]. On this account, an explanation is a deductive derivation from a general law of nature. For example, we might try to explain why the sun is in a particular place today. This is our "explanandum", the phenomenon to be explained. On Hempel's account of scientific explanation we avail ourselves of one or more general laws—say, the laws of planetary motion—plus some particular facts—the position of the sun at a previous point in time. Putting these together, one shows that the explanandum (the sun's current position) is derivable as a deductive derivation from that general law and the particular facts. That is, one shows that the phenomenon one is trying to explain follows logically from general law and particular facts.

In short, explanations—on this account—literally *are* arguments. It is at this point useful to clarify the ways in which the term 'argument' is itself multiply ambiguous [2, 3]. The first sense of the term 'argument' is that of an argument as a 'reason'. Giving an argument for something is providing a reason for it. Here, the strongest possible reasons are those from which a claim or conclusion follows by necessity.

This leads to the second, closely related, sense. Here, the term 'argument' is used not just to refer to the reason but to the unit comprising reason(s) and claim. That is how the term 'logical argument' is understood, and the way 'explanation' on Hempel's account *is* an argument: the explanation is a structured unit comprising one or more premises and a conclusion.

Third, the term 'argument' is used to refer not just to information content, but to a social activity. Here, it is not (just) a single premise and conclusion pair or a sequence of inter-connected claims and counter-claims that is in focus, but also the dialogical, social, activity that is giving rise to these claims. For this sense of argument as a dialogical activity, the argumentation literature distinguishes multiple forms, contrasting, for example, a quarrel with a rational debate. The latter involves the exchange of reasons that are aimed at 'convincing a reasonable critic' [4]. It is this latter type of exchange that is the focus of this paper.

Explanation, as a term, shares a corresponding ambiguity: it can refer to the reason, and to the social activity of providing that reason in a particular context. That, too, reflects a commonality across the two terms that we return to below. We assume in the following that it will be clear from the context, for both argument and explanation, what specific meanings are intended.

Despite the notable overlap between argument and explanation just outlined, further reflection reveals that, however closely linked, the two are nevertheless distinct [5]. For example, in the context of rational argument, we are typically trying to advance reasons/arguments that seek to change others' beliefs in an as yet uncertain claim [4, 6]. By contrast explanations can be provided for claims and events we already know to be true [7].

This can be illustrated with inference to the best explanation (IBE) [8, 9, 10]. A canonical example of inference to the best explanation is the following: we go into the kitchen and see that our cheese has been nibbled. The best hypothesis in this situation is that there was a mouse in the kitchen [11]. IBE proponents maintain that the very fact that the mouse hypothesis constitutes the best hypothesis vis a vis our nibbled cheese confers additional epistemic support to the mouse hypothesis being true. One may or may not subscribe to this theoretical position, but, clearly, the cheese-eating mouse is an explanation for the nibbled cheese. Equally clearly, the presence of a mouse is not (in this context) an argument for believing that the cheese has been nibbled. We

already know that the cheese has been nibbled. Rather, nibbled cheese is an argument for the presence of a mouse and mouse is an explanation for nibbled cheese.

From this simple example it becomes apparent that arguments and explanations are two different, separable concepts. We can (but need not only) have explanations for events that are certain, but we typically do not (though we sometimes can) consider arguments for things we already believe to be certain. Consideration of the simple mouse example illustrates cases of clear difference. However, there will also be cases where the distinction is blurred.

In summary, there are multiple, close links between the notions of argument and explanation, and the degree of conceptual overlap is such that there may be occasions where it is hard to clearly decide whether one is looking at one or the other, or even occasions where what is being advanced might be both. This strongly suggests that the two notions might be usefully studied together.

The main goal of this paper is to enable more integrated research into argument and explanation in future. Specifically, we are interested in bringing together the study of the two notions in psychology/cognitive science and AI.

3. Argument in AI and Psychology

(a) Argument in AI

We start with a brief overview of argumentation research within AI. This is itself a rather disparate field, where many of the areas to be mentioned have little connection with one another. One of the reasons for that, we suspect, is the historic accident by which much of the early literature on 'argumentation' in AI was not actually concerned with everyday natural language argument (in the third sense outlined above, i.e., as a dialectical activity involving the exchange of multiple, inter-related reasons). The goal of early work on argumentation in AI, arguably, did not view this dialectical activity as a target phenomenon that it wanted to understand in its own right (and consequently build systems to execute or, at least, support). Rather it was interested in argumentation as a tool for accomplishing something else. Some of the most foundational work on argumentation in AI, such as Dung style semantics, for example, has roots as a means of trying to elucidate logic programming [12, 13, 14, 15].

Subsequently this field developed a plethora of non-classical logics and argumentation frameworks as tools for dealing with uncertainty, in particular, tools for non-monotonic reasoning [16, 17]. Much of this work was conceived, either explicitly or implicitly, as an alternative to using probability theory for coping with uncertainty [18] (and one lesson learned was that alternatives to probability could turn out to be 'probability in disguise' [19, 20]).

As a result of its tool-based focus, this body of research in AI is often only rather loosely connected with research that has concerned itself more directly with everyday argument, in particular with natural language text.

The following strands seem worth highlighting in this latter context.

(i) The Argument Interchange Format

The first involves the argument interchange format (AIF) –a canonical machine readable format for representing natural language arguments. A prime use for this format has been argument mapping, see Fig. 1. The sample map in the figure was drawn with the software tools of OVA (for 'Online Visualisation of Argument') developed by Reed and colleagues [21]. The particular example is from a recent project examining different ways of representing scientific knowledge, in particular where there is scientific disagreement that might be important to communicate to policy makers in order to reflect accurately extant uncertainty [22]. OVA allows one to take a PDF, highlight text in that PDF, read that into a text box that OVA converts into a node in the map, and then, via pull-down menu, select different types of inferential relationships that connect it with other parts of an argument in order to form an overall map of the dialectical exchange. The utility

of this is that it facilitates the creation of argument maps (including creation at scale, in multi-contributor projects) by allowing one to aggregate automatically different maps. It also supports navigation of such maps, and the use of a variety of computational processes defined over these.

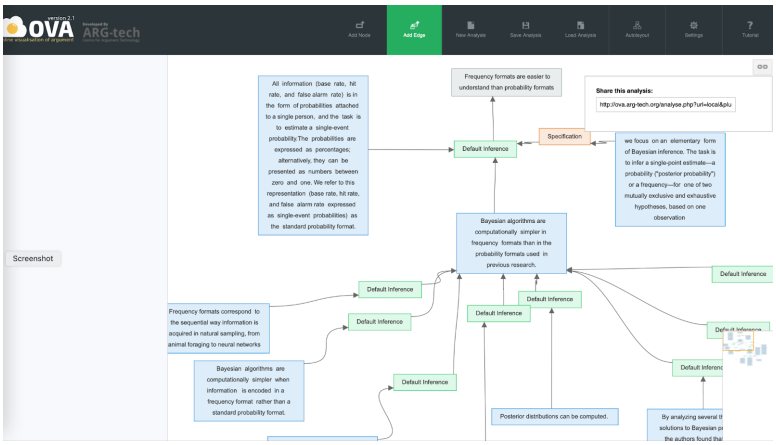

Figure 1. A screenshot of a section of an argument map created with the help of the OVA software.

Finally, anything annotated in OVA may be added to a large, openly accessible, database with many thousands of argument maps based on the AIF format which will continue to grow as long as people are using these types of tools. OVA is one of multiple tools for argument mapping (e.g., [23, 24]), and one of many systems drawing on the AIF. Argument mapping remains popular for a wide range of tasks from large scale computer-aided discourse visualisation [25], through to critical reasoning [26].

For the map in Fig. 1, it was human analysts going through the text, identifying arguments and identifying appropriate argumentative relationships. In recent years, however, much research has gone into trying to automate such activities.

(ii) Argument Mining

Automation of these elements is the focus of argument mining research (for reviews see e.g., [23] and [27]). The goal here is to take the steps just outlined with respect to Fig. 1 –extraction of natural language arguments and their relations from text, and the subsequent generation of machine-processable representations for computational models of argument– and have these be conducted by machine. As a field, argument mining has developed rapidly from a niche interest into a focal topic in AI [28] that now commands significant resources both in academia and the corporate sector. Argument mining research has itself brought together researchers from multiple areas such as natural language processing (NLP) and knowledge representation and reasoning. Lawrence and Reed (2020) [23] highlight three historic routes to argument mining research: sentiment analysis [29], controversy detection [30], and argumentative zoning [31].

Argumentative zoning seeks to take scientific documents and identify relevant argumentative components. This involved a standard computational linguistics process of researcher-developed annotation tools, which were then used to create corpora that serve as training materials for automated classification (e.g., [32]).

The overall goal of summarising scholarly articles, however, has recently also entered firmly into the sights of transformer based NLP tools: BERT [33] and the rapidly expanding list of Large Language Models (LLMs).

(iii) Large Language Models

LLMs are models with hundreds of billions of parameters that estimate the probability distribution over word sequences [33, 34]. Crucially, state of the art LLMs are able to provide reasons for their solutions to problems (e.g. [35, 36]) and provide evidence for their claims (e.g. [37]). They have also very rapidly become such a focal point of current discussion, not just within the academic literature, that they arguably need no further introduction. At the same time, the recent pace of developments has been such that any evaluations are likely to be superseded at the time of print (see e.g., [38]). This makes more detailed analysis of current capabilities rather futile. There are, however, interesting questions about the relationship between AI and Cognitive Science that are posed by these models, and we refer the reader to two papers in this special issue that pursue these further (XXcite Goodman; cite Pavlick).

(iv) Project Debater

LLMs also seem poised to soon challenge the quantum leap provided by IBM's "Project Debater" [39]. This system can be given a novel claim or proposition and then finds arguments in support of that claim and does so in an interactive debate with an opponent. Project debater is capable of generating arguments that seemed convincing to the audience of a debating contest against a human debating champion. This very recently represented not just a wholly new level of automation (and performance) in the context of argument, but one that was hard imagine a mere decade ago at the advent of argument mining research [40]. Project Debater rests on a combination of some of the aforementioned approaches and technologies. One obvious question for the future is the extent to which 'generalist' LLMs will be able to match (or exceed) such performance.

(v) Bayesian Argumentation in AI

Finally, it is worth mentioning a small pocket of research articles that concern themselves with Bayesian argumentation. These include early [41, 42, 43, 44] and more recent [45, 46] attempts to generate arguments from Bayesian Belief Networks (BBN, on these generally, see [47, 48, 49]). This work merits mention here not because it reflects a sizeable community or body of research within AI, but by virtue of constituting one of the comparatively few, potential points of connection between AI and cognitive science: by virtue of its use of the Bayesian framework, this AI research on argumentation links up with research on argumentation within psychology. We turn to that work next.

(b) Psychology of Argumentation

By contrast to argumentation research within AI, the psychology of argumentation is a tiny field (for an introduction see e.g., [2]). This seems at odds with the central role of argumentation across many real-world contexts. Possibly even more surprising is that a significant proportion of that work has not been conducted by psychologists. Much of what one could class as part of the psychology of argumentation was conducted either in education studies or in communication sciences (and we return to some of the reasons for this below).

One focal point within the psychology of argumentation is the body of work that has concerned itself with critical thinking. This research has sought to understand how one can foster critical thinking, and how good people are at evaluating certain types of arguments [50]. Critical thinking research within education studies (and, relatedly, within developmental psychology) has made wide-spread use of the Toulmin framework developed by Stephen Toulmin in the 1950s [51]. The basic components of Toulmin's framework are illustrated in the argument map shown in Fig. 2. Specifically, the Toulmin framework introduces a number of very general distinctions in terms of types of relations that obtain between different components of an overall more complex argument. The inferential relationship between a reason and a claim, for example, rests on the 'warrant', and that warrant may itself receive further support ('backing'). In effect, the warrant explicates why the reason is relevant to the claim.

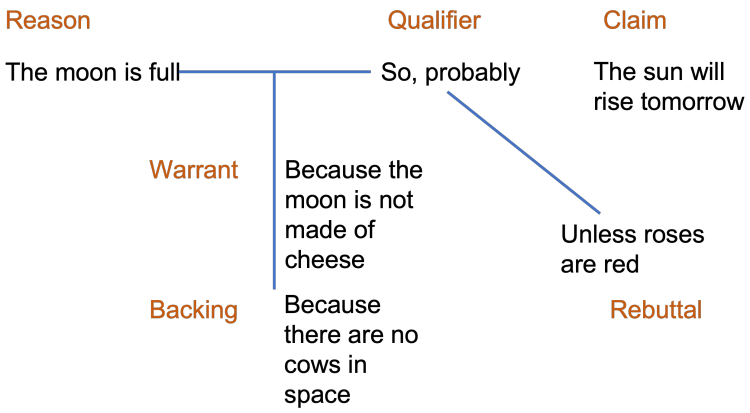

Figure 2. An example of an argument scheme developed using Toulmin's framework. Figure adapted from [52].

The nonsense content in Figure 2 is chosen to make salient the fact that this scheme captures little about content: the fact that something is classed as a reason for a claim is ultimately based on the fact that somebody *advanced it* as a reason for a claim. There is nothing in the Toulmin framework that tells one whether it is actually a good, sensible or cogent reason and hence one that *should* change one's belief in the claim at issue (for discussion of this point see [52]). This severely limits the scheme's utility for the chosen purpose of understanding or fostering critical thinking.

In response, one attempt to move from the descriptive perspective of 'simply given as a reason' to the normative perspective of 'constitutes a good reason' lies in so-called scheme-based approaches to argumentation [2, 53]. These have become popular within the critical thinking literature, in the informal argument literature within philosophy, and within the AI argumentation literature.

The argument mapping software OVA (Fig. 1) (as described above) offers schemes identified in that research literature as inbuilt components: a user can select an 'argument from expertise' or an 'appeal to popular opinion' or particular types of causal argument to represent the inferential relationship (in effect, the warrant) between reason and claim. These argument schemes represent defeasible argument types that are putatively good, but that might be overturned by further evidence. In addition to identifying schemes that represent recurring patterns in everyday argument, the scheme-based literature has sought also to identify so-called 'critical questions' that assist with evaluation [53]. These questions offer standard considerations that might help identify a particular instance of this scheme as weak or strong, good argument or bad. This, in turn, has prompted an empirical literature examining how people try to reason with these [54]. The fact that these schemes are also used in a variety of computational systems creates a further point of overlap in argumentation research across AI and cognitive science (beyond OVA see, e.g., [55]).

There are two other research traditions with normative, philosophical orientation, that have prompted psychological research. First is research on the procedural rules that govern rational discourse (e.g., and under the header of 'pragma-dialectics' [56, 57] , or 'fairness rules' [58, 59, 60]). Second is psychological research on reasoning [61, 62], and more recently, Bayesian argumentation [52, 63]. This research is explicitly concerned with ways to measure the degree of support that an argument actually conveys for a claim.

As just outlined, and despite its popularity in the critical thinking literature, the Toulmin framework offers no real normative component. The critical questions of the scheme-based literature improves on that, but the normative foundation of those questions themselves very much remains unexplained. It is in order to move beyond that, toward more fine-grained

evaluation of argument quality, that the Bayesian framework has been employed. For example, it has been used both to provide a normative treatment of so-called argument fallacies (examining the extent to which such arguments *should* be viewed as persuasive), and to then look, descriptively, at how people actually evaluate them relative to this normative standard.[1] We return to the implications of these interwoven normative and descriptive concerns at the end.

Finally, there is research relevant to a psychology of argumentation under the banner of 'persuasion' or 'attitude change' as studied within social psychology [65]. While some of that research involves reasons that might be classed as aimed at a 'reasonable critic' (and thus overlaps with the type of argument considered in this paper), other aspects of this literature pursue concerns that might be more appropriately classed as 'marketing'. We likewise return to the persuasion literature in the final section of this paper.

4. Explanation in AI and Psychology

(a) Explanation in AI

In turning our attention to research on explanation within AI, we move back to a large body of research with significant heterogeneity, paralleling the diversity seen within AI research on argumentation. For one, it spans different notions of the term explanation as outlined in Section 2 above.

One body of research treats explanation simply as the most probable cause (i.e., literally 'the mouse' in the earlier cheese example). This is exemplified by a research tradition that has used Bayesian belief networks (BBNs) to identify the most probable cause through abductive reasoning [47, 66, 67].

However, researchers have also been interested not just in identifying a single most probable cause, but in explanation as explicating a process of reasoning: a BBN might tell us that a body of evidence should raise our posterior degree of belief (say, in there being a mouse) to a certain degree, but a user might want to know also how and why one can infer that as a function of the probabilities involved [68].

As an example of the latter, the BARD project aimed to build assistive technologies that would allow a group of people to collaboratively build a BBN, then perform inference over that BBN, and receive computer generated explanations [69].

By far the largest literature on explanation in AI, however, pertains to explaining black box machine learning (ML) models. In such models, the lack of transparency regarding how outputs were generated poses multiple challenges to the user, last but not least challenges with respect to trust. There are presently two main (at times overlapping) strands of this research in the literature: global and local explainability methods. Global methods aim to explain the behaviour of the whole ML model, whereas local methods aim to explain the specific predictions of a model.

Methods for explainable AI are further divided into model-agnostic methods, that would apply to any ML model, on the one hand, and, on the other, model-specific methods, that can only be applied to certain types of models such as, for example, tree-based models or neural networks [70].

The techniques for explaining AI systems are diverse and range from example-based methods [71], feature importance [72, 73] and saliency maps [74], through to counterfactual explanations [75, 76, 77, 78]. For a review on ML explainability techniques see [70].

These different techniques assume different definitions of what, precisely, constitutes an explanation, driven partly by the fact that different techniques suit different data modalities. For example, saliency maps are almost exclusively applied to ML model that process image data. On

[1]This probabilistic framework has also been used not just to assess the strength of arguments about facts, but through the inclusion of utilities, the strength of practical arguments as well, e.g., [64]. Relatedly, Bayesian decision theory has the potential to support future development of a meta-framework that elucidates questions about when argument or explanation might be worthwhile. We thank a reviewer for highlighting this issue.

the other hand, counterfactual explanations are often applied to ML models dealing with tabular data.

The main goal of explainable AI (XAI) methods for machine learning models is to increase understanding of model behaviour. Such explanation of ML models can be used to expose strengths and weaknesses of an ML model and thus used to calibrate trust in ML models [79] –both for researchers and end users.

In effect, research on XAI spans the range of tools that might be used for a concrete decision or recommender system. The nature of the decision or recommender system in question will shape the definition of what would constitute, for that system, an explanation (e.g., information about feature contributions or explanation by example), and shape the computational methods for deriving those explanations. Increasingly this will also prompt empirical investigation of how users actually perceive and understand those explanations.

The importance of user testing in XAI shifts the field in the direction of psychology. Moreover, recent reviews of work on the psychology of explanation (e.g. [80]) have had significant impact on the evaluation and creation of explainable AI methods. XAI is thus one area that is already forging closer links between AI and cognitive science/psychology. We next explore in more detail psychological research on explanation.

(b) Psychology of explanation

The psychology of explanation is, arguably, a larger, and more well-defined, field than the psychology of argumentation. It too, however, is still a comparatively small field relative to other areas of psychology and, in that, remains somewhat at odds with the centrality of explanation to human cognition. As a field it is currently also largely separate from the psychology of argumentation.

One hallmark of relative maturity within psychology is that an area can lay claim to some form of 'classic', hallmark finding. The so-called "illusion of explanatory depth" is not only a contender for such a finding, it is also of both theoretical and practical interest to anyone interested in an explainable AI. The illusion of explanatory depth refers to the rather pervasive finding that people struggle to give meaningful causal or mechanistic explanations for all manner of real world systems that they competently deal with on a daily basis [81], and that the kind of explanations that they do produce often seem more convincing to them than they merit.

Relatedly, and of direct interest to this paper's theme of the relationship between argument and explanation, there is also a body of research that has probed the extent to which people can distinguish reliably between evidence or arguments, on the one hand, and explanations on the other. As outlined in Section 2, this is, arguably, not a completely trivial task. As discussed, the two notions are themselves multiply overlapping, interlinked, and connected. It is thus not surprising that there are findings that suggest young children, for example, struggle with this distinction [82, 83].

Beyond that, there is a sizeable body of research in the psychology of explanation that has taken its cue from a literature in philosophy that concerns itself with the so-called explanatory virtues. Explanatory virtues are properties that explanations potentially have, or should have, in order to count as good explanations, particularly in the context of philosophy of science. There have been psychological studies examining the extent to which lay people, in every day contexts, are sensitive to explanatory virtues or signals of explanatory goodness such as the simplicity of a hypothesis [84, 85, 86, 87, 88, 89, 90]

All in all, the psychology of explanation shares some of the breadth of AI research on this topic. Behavioural research on explanatory virtues primarily involves experiments that are about a small number of causes or hypotheses. By contrast, consideration of the illusion of explanatory depth involves much more elaborate linking of explanations. In that, psychological research on explanation reflects some of the range of inter-related meanings of the term explanation distinguished in Section 2 above.

5. Bringing It All Together

In this final section, we offer thoughts on bringing all of this research together. It should now be apparent that there are multiple reasons for why one would want to bring together these currently distinct four fields. The first is that the notions of argument and explanation are not only theoretically closely related notions, they also involve closely related practical applications. If one is interested in building an explainable AI, one should be taking an interest in what researchers are already doing with respect to machine generated arguments. This particular practical connection is obvious and is, at least to some extent, already being pursued.

However, the very fact that the four areas surveyed have all evolved as largely separate fields means also that theoretical unification is desirable in as much as unification is a natural concern of science. Furthermore, unification is also likely to be both theoretically and methodologically productive within the individual fields: it seems highly likely that these fields hold important shareable but as yet un-shared knowledge. Sharing that knowledge would allow these fields to meaningfully refocus some of their research agenda.

One example of a presently un-shared perspective that is likely to be productive concerns potential transfer from argumentation to explanation research. The psychology of explanation has surfaced a number of basic findings about the effects of providing an explanation: explanations increase our confidence in a claim; they increase our confidence that an event will occur when asked to explain a possible future event; and they increase our confidence regarding an event in the past for which we are not sure if it happened or not [91, 92, 93, 94].

All of these are things that arguments do also and this, again, reflects the functional overlap and similarity between argument and explanation. In fact, the literature mentioned above that examined the extent to which people can distinguish faithfully between arguments and explanation has suggested also that people will use explanations to support a claim where evidence or arguments are sparse or missing [82, 83]. Arguments and explanations clearly target some of the same functional space. Hence it is reasonable to expect that things that are functionally relevant for arguments should also play a role for explanations.

Cross-field sharing of perspectives may thus be beneficial inasmuch as there are features of this functional space that have been central (both practically and theoretically) from the perspective of argument, yet have barely begun to come into view in research on explanation.

Coming from the perspective of argumentation, it is salient that an argument is something that a concrete, specific agent (human or other) provides to a concrete other (or group of others), in a specific context. But this of course is also true of explanations [95]. As discussed in Section 2 above, both argument and explanation may be construed as activities, in this case as communicative acts.

One central concern in the communication of arguments (i.e., testimony) is the *reliability of the source*. For one, the above mentioned social psychological literature on persuasion has spent 30 odd years on this issue. In that literature, the core models of persuasion have been so-called dual route models such as the Elaboration Likelihood Model (ELM). These models have sought to identify cognitively distinct routes, or pathways, for convincing people [96, 97, 98]. One of these is taken to be an analytic route that focuses attention on the content of the argument. The other is a peripheral, heuristic, route that pays attention to characteristics of the source. Much research has gone into trying to understand the contexts in which people resort to one or the other, and what kinds of source characteristics are relevant to heuristic processing.

More recently, argumentation researchers have stressed that when coming at the distinction between source and content from a normative, Bayesian, perspective, both features of the source and the content of an argument will matter, and that these should interact in determining how much beliefs change [2, 54, 99, 100, 101]. Behavioural evidence now suggests that they interact in people's intuitive, informal, evaluations of arguments too, and do so in ways that are not well-captured by extant social psychological models of persuasion [102, 103].

Viewed from the perspective of the psychological literature on argumentation it thus seems surprising that there has been so little research on reliability and explanation, by comparison. This absence can be felt not just in the psychology of explanation, but in AI research also (particularly

as the latter has already highlighted trust, i.e., perceptions of source reliability, as a central goal of xAI).

We have been conducting experimental investigations that indicate that source reliability (that is, characteristics of the person providing the explanation) have effects on the impact of the explanation and that the bi-directional dynamics between content and source mirror some of what has been found in argumentation research [104].

It is consequently encouraging that work on explanation in AI has now finally started to take note of the *pragmatics of explanation*, that is, an understanding of how interpretations of utterances are generated in particular contexts [105, 106, 107].

The second set of considerations for re-focusing the research agenda that emerges from trying to learn from the distinct perspectives across our four areas concerns methods. User testing research, specifically in explainable AI, is effectively applied psychology. This means not only that it can benefit, rather obviously, from the experiences of a long tradition of applied psychological research, there are also more specific, topic specific lessons to be learned.

As researchers who have studied argument and explanation within psychology, we think it would be fruitful for behavioural testing conducted on XAI to more strongly emphasize, and focus on, normative considerations. What we mean by normative considerations in this context is that one should be thinking about, and using as a way to structure one's research, considerations of what constitutes a *good argument* and a *good explanation*. 'Good', here, is intended not just in the sense of whatever happens to actually convince somebody, but rather what *should* convince somebody, that is, what should convince a rational actor.

This matters for explainable AI because one ultimately wants people's trust to be calibrated to the quality of the system [79]. One cannot just want people to believe or trust a system regardless, as doing so may be dangerous.

However, normative considerations are arguably even more important from a methodological perspective. In our view, researchers will need normative frameworks in order to structure the research in such a way that it becomes *generalisable*. We think this conclusion follows strongly from the history of research both on the psychology of argumentation and the psychology of explanation.

Our above discussion of the psychology of argumentation drew out some of the characteristics of work in that area that give reason to think there are deep substantive reasons for why this topic that is so central to our everyday lives received so little psychological attention. We see as chief among these the limitations of certain tools. Given just the Toulmin framework (Fig. 2), all one can really say, is that there is a claim, some reasons for it, some kind of support relationship and, possibly, a rebuttal. Beyond those crude distinctions, it provides no tools for identifying further categories or objects of study.

In other words, the scheme is too limited to enable meaningful theoretical predictions or identify types of arguments across which one might seek empirical generalizations. From the perspective of that framework, the only questions one can 'see' and hence ask in empirical studies, is whether particular individuals actually offer reasons, and how complex the inter-relationships between those reasons might be. Beyond that, one cannot really distinguish between someone saying "it's raining outside because the pavement is wet " as opposed to "strawberry ice cream is more popular than pistachio partly because humans prefer the colour". These are simply two different arguments with nothing in common other than that both involve a claim and a reason; because they have entirely different content there is no way to form any kind of meaningful generalization over them.

The value of the Bayesian framework as a tool for studying argumentation has been that it allows researchers to ask normative questions about argument strength that attach to the specific content of what is being argued about (see also, [52]). This is made possible because probabilities are intensional and are determined by the content of a proposition [47]. Hence one can ask systematic questions about responses to arguments across different content instantiations. For example, one can ask whether people's argument evaluation is closer to the normative standard

when they are confronted with arguments describing scientific scenarios or with arguments involving familiar, everyday events [64]. Theoretically and practically meaningful questions about how people treat arguments of different content and context thus become possible because one can compare those very different arguments to the same normative standard. In the same vein, we consider it to be more than a coincidence that some of the most succesful work on explanation has been based on normative considerations drawn from philosophy (see [86]).

This leads us to believe that behavioural studies in the context of explainable AI will not generate cumulative insight on the user testing side without use of structuring frameworks. Research will be limited, we suspect, to collecting particulars without deeper insight: in effect, 'this system did this specific thing and this is how convinced people were by it', and then, in a different study 'we did this specific thing with this completely different system and this is how people responded there'. Without a systematizing framework this will produce little in the way of general insight or information gain. And this is precisely why both the psychology of argumentation, and the psychology of explanation have availed themselves of extant normative frameworks.

Adopting tools like the Bayesian framework, for example, to study such questions may provide a theoretical framework that supports general insights. Beyond that, it should be a welcome and exciting prospect, last but not least, because the normative questions raised in the XAI context are themselves interesting; and both argument and explanation involve interesting normative concerns that are unaddressed to date.

6. Conclusions

To conclude, argumentation and explanation are closely related and overlapping, but nevertheless conceptually distinct, notions. Both argumentation and explanation constitute large topics of research in AI and sizable but smaller topics in psychology. We think closer theoretical and practical integration is required both across the argumentation-explanation dimension and the AI-psychology dimension. Such integration will naturally highlight shared constructs such as source reliability, and we suspect others will emerge from those comparisons. Finally, we suggest that without normative considerations to help derive theory to guide experimental work, future research will unlikely meet fully the practical challenges explainable AI is seeking to address, and will remain unlikely to yield robust, generalisable, insight. In short, there is much to gain from closer integration.

Authors' Contributions. Both authors wrote, revised, read and approved the manuscript.

Competing Interests. The author(s) declare that they have no competing interests.

Funding. M.T. was supported by a Royal Academy of Engineering Fellowship.

References

1 Carl G. Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175, 1948.
2 Ulrike Hahn and Jos Hornikx. A normative framework for argument quality: Argumentation schemes with a Bayesian foundation. *Synthese*, 193(6):1833–1873, 2016.
3 Daniel J O'keefe. *Persuasion: Theory and research*. Sage Publications, 2015.
4 Frans H Van Eemeren, Sally Jackson, and Scott Jacobs. Argumentation. In *Argumentation Library*, pages 3–25. Springer, 2015.
5 Ulrike Hahn. The problem of circularity in evidence, argument, and explanation. *Perspectives on Psychological Science*, 6(2):172–182, 2011.
6 Ulrike Hahn and Mike Oaksford. Rational argument. In K. J. Holyoak and R. G. Morrison, editors, *Oxford library of psychology. The Oxford handbook of thinking and reasoning*, pages 277–298. Oxford University Press, Oxford, 2012.

7 Tania Lombrozo. Explanation and abductive inference. In K. J. Holyoak and R. G. Morrison, editors, *Oxford library of psychology. The Oxford handbook of thinking and reasoning*, pages 260–276. Oxford University Press, Oxford, 2012.

8 Gilbert Harman. The inference to the best explanation. *The philosophical review*, 74(1):88–95, 1965.

9 Igor Douven. Inference to the best explanation, dutch books, and inaccuracy minimisation. *The Philosophical Quarterly*, 63(252):428–444, 2013.

10 Peter Lipton. *Inference to the best explanation*. Routledge, 2003.

11 Bas C Van Fraassen. *The scientific image*. Oxford University Press, 1980.

12 Andrei Bondarenko, Francesca Toni, and Robert A Kowalski. An assumption-based framework for non-monotonic reasoning. In *LPNMR*, volume 93, pages 171–189, 1993.

13 Phan Minh Dung. Negations as hypotheses: An abductive foundation for logic programming. In *ICLP*, volume 91, pages 3–17, 1991.

14 Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77 (2):321–357, 1995.

15 Antonis C Kakas, Robert A. Kowalski, and Francesca Toni. Abductive logic programming. *Journal of logic and computation*, 2(6):719–770, 1992.

16 John L Pollock. Defeasible reasoning. *Cognitive science*, 11(4):481–518, 1987.

17 Henry Prakken and Gerard Vreeswijk. Logics for defeasible argumentation. *Handbook of philosophical logic*, pages 219–318, 2001.

18 Simon Parsons. *Qualitative methods for reasoning under uncertainty*, volume 13. Mit Press, 2001.

19 Eric Horvitz, David Heckerman, and Curtis P Langlotz. A framework for comparing alternative formalisms for plausible reasoning. In *AAAI*, pages 210–214, 1986.

20 Paul Snow. Intuitions about ordered beliefs leading to probabilistic models. In *Uncertainty in Artificial Intelligence*, pages 298–302. Elsevier, 1992.

21 Mathilde Janier, John Lawrence, and Chris Reed. Ova+: An argument analysis interface. In *Computational Models of Argument: Proceedings of COMMA*, volume 266, page 463, 2014.

22 U. Hahn, J.K. Madsen, S. Schubert, and C. Reed. Managing expert disagreement for the policy process and beyond. *Unpublished manuscript*, 2022.

23 John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4): 765–818, 2020.

24 Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce M McLaren. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-supported collaborative learning*, 5(1):43–102, 2010.

25 Chris Reed, Katarzyna Budzynska, John Lawrence, Martın Pereira-Farina, Dominic De Franco, Rory Duthie, Marcin Koszowy, Alison Pease, Brian Pluss, Mark Snaith, et al. Large-scale deployment of argument analytics. In *In Argumentation and Societythe workshop at the 7th International Conference on Computational Models of Argument (COMMA 2018)*, 2018.

26 Jacky Visser, John Lawrence, and Chris Reed. Reason-checking fake news. *Communications of the ACM*, 63(11):38–40, 2020.

27 Manfred Stede and Jodi Schneider. Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191, 2018.

28 Elena Cabrio and Serena Villata. Five years of argument mining: A data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433, 2018.

29 Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.

30 Filip Boltužić and Jan Šnajder. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, 2015.

31 Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1493–1502, 2009.

32 Simone Teufel and Marc Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445, 2002.

33 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

34 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

35 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

36 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

37 Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

38 Samuel R Bowman. Eight things to know about large language models. *arXiv preprint arXiv:2304.00612*, 2023.

39 Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. An autonomous debating system. *Nature*, 591(7850):379–384, 2021.

40 Chris Reed. Argument technology for debating with humans, 2021.

41 Kevin B. Korb, Richard McConachy, and Ingrid Zukerman. A cognitive model of argumentation. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, pages 400–405, 1997.

42 Ingrid Zukerman, Richard McConachy, and Kevin B. Korb. Bayesian reasoning in an abductive mechanism for argument generation and analysis. In *AAAI/IAAI*, pages 833–838, 1998.

43 Ingrid Zukerman, Richard McConachy, Kevin B. Korb, and Deborah Pickett. Exploratory interaction with a Bayesian argumentation system. In *IJCAI*, pages 1294–1299, 1999.

44 Ingrid Zukerman, Richard McConachy, and Sarah George. Using argumentation strategies in automated argument generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 55–62, 2000.

45 Ann E Nicholson, Kevin B Korb, Erik P Nyberg, Michael Wybrow, Ingrid Zukerman, Steven Mascaro, Shreshth Thakur, Abraham Oshni Alvandi, Jeff Riley, Ross Pearson, Shane Morris, Matthieu Herrmann, A.K.M. Azad, Fergus Bolger, Ulrike Hahn, and David Lagnado. BARD: A structured technique for group elicitation of Bayesian networks to support analytic reasoning. *arXiv preprint arXiv:2003.01207*, 2020.

46 Iyad Rahwan and Guillermo R Simari. *Argumentation in artificial intelligence*, volume 47. Springer, 2009.

47 Judea Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kauffman, 1988.

48 Judea Pearl. *Causality*. Cambridge university press, 2009.

49 Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

50 Madsen Pirie. *How to win every argument: the use and abuse of logic*. Bloomsbury Publishing, 2015.

51 Stephen E Toulmin. *The uses of argument*. Cambridge university press, 1958/2003.

52 Ulrike Hahn. Argument quality in real world argumentation. *Trends in Cognitive Sciences*, 2020.

53 Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation schemes*. Cambridge University Press, 2008.

54 Ulrike Hahn, Mike Oaksford, and Adam JL Harris. Testimony and argument: A bayesian perspective. In *Bayesian argumentation*, pages 15–38. Springer, 2013.

55 Thomas F Gordon, Henry Prakken, and Douglas Walton. The carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10-15):875–896, 2007.

56 Frans H Van Eemeren and Peter Houtlosser. The development of the pragma-dialectical approach to argumentation. *Argumentation*, 17(4):387–403, 2003.

57 Frans H van Eemeren. *Argumentation theory: A pragma-dialectical perspective*. Springer, 2018.

58 Ursula Christmann, Christoph Mischo, and Norbert Groeben. Components of the evaluation of integrity violations in argumentative discussions: Relevant factors and their relationships. *Journal of Language and Social Psychology*, 19(3):315–341, 2000.

59 Ursula Christmann, Christoph Mischo, and Jürgen Flender. Argumentational integrity: a training program for dealing with unfair argumentative contributions. *Argumentation*, 14(4): 339–360, 2000.

60 Christoph Mischo. The role of cognition in reacting to argumentative unfairness. *Pragmatics & cognition*, 11(2):241–266, 2003.

61 Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

62 Keith E Stanovich and Richard F West. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and brain sciences*, 23(5):645–665, 2000.

63 Ulrike Hahn and Mike Oaksford. The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological review*, 114(3):704, 2007.

64 Adam Corner, Ulrike Hahn, and Mike Oaksford. The psychological mechanism of the slippery slope argument. *Journal of Memory and Language*, 64(2):133–152, 2011.

65 Jos Hornikx and Ulrike Hahn. Reasoning and argumentation: Towards an integrated psychology of argumentation. *Thinking & Reasoning*, 18(3):225–243, 2012.

66 Solomon E Shimony. Explanation, irrelevance and statistical independence. In *Proceedings of the ninth National conference on Artificial intelligence-Volume 1*, pages 482–487. AAAI Press, 1991.

67 Changhe Yuan, Heejin Lim, and Tsai-Ching Lu. Most relevant explanation in Bayesian networks. *Journal of Artificial Intelligence Research*, 42:309–352, 2011.

68 Carmen Lacave and Francisco J Díez. A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127, 2002.

69 Erik P Nyberg, Ann E Nicholson, Kevin B Korb, Michael Wybrow, Ingrid Zukerman, Steven Mascaro, Shreshth Thakur, Abraham Oshni Alvandi, Jeff Riley, Ross Pearson, et al. Bard: A structured technique for group elicitation of bayesian networks to support analytic reasoning. *Risk Analysis*, 42(6):1155–1178, 2022.

70 Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

71 Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

72 Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

73 Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

74 Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

75 Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

76 Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 652–663, 2021.

77 Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

78 Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

79 Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*, 2018.

80 Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.

81 Leonid Rozenblit and Frank Keil. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, 26(5):521–562, 2002.

82 Sarah K Brem and Lance J Rips. Explanation and evidence in informal argument. *Cognitive science*, 24(4):573–604, 2000.

83 Deanna Kuhn. How do people know? *Psychological science*, 12(1):1–8, 2001.

84 David Lagnado. *The psychology of explanation: A Bayesian approach*. Unpublished Masters thesis. Schools of Psychology and Computer Science, University of Birmingham, UK, 1994.

85 Tania Lombrozo. Simplicity and probability in causal explanation. *Cognitive psychology*, 55 (3):232–257, 2007.

86 Tania Lombrozo. Explanatory preferences shape learning and inference. *Trends in cognitive sciences*, 20(10):748–759, 2016.

87 Nancy Pennington and Reid Hastie. Reasoning in explanation-based decision making. *Cognition*, 49(1-2):123–163, 1993.

88 Stephen J Read and Amy Marcus-Newhall. Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3):429, 1993.

89 Paul Thagard. The best explanation: Criteria for theory choice. *The journal of philosophy*, 75(2): 76–92, 1978.

90 Paul Thagard. Explanatory coherence. *Behavioral and brain sciences*, 12(3):435–467, 1989.

91 Craig A Anderson, Mark R Lepper, and Lee Ross. Perseverance of social theories: the role of explanation in the persistence of discredited information. *Journal of personality and social psychology*, 39(6):1037, 1980.

92 Craig A Anderson and Elizabeth S Sechler. Effects of explanation and counterexplanation on the development and use of social theories. *Journal of Personality and Social Psychology*, 50(1): 24, 1986.

93 Derek J Koehler. Explanation, imagination, and confidence in judgment. *Psychological bulletin*, 110(3):499, 1991.

94 Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10 (10):464–470, 2006.

95 Denis J Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107 (1):65, 1990.

96 Richard E Petty and John T Cacioppo. Issue involvement as a moderator of the effects on attitude of advertising content and context. *ACR North American Advances*, 1981.

97 Richard E Petty and Pablo Briñol. The elaboration likelihood model. *Handbook of theories of social psychology*, 1:224–245, 2011.

98 Chanthika Pornpitakpan. The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of applied social psychology*, 34(2):243–281, 2004.

99 Ulrike Hahn, Adam JL Harris, and Adam Corner. Argument content and argument source: An exploration. *Informal Logic*, 29(4):337–367, 2009.

100 Adam JL Harris, Ulrike Hahn, Jens K Madsen, and Anne S Hsu. The appeal to expert opinion: Quantitative support for a Bayesian network approach. *Cognitive Science*, 40(6):1496–1533, 2016.

101 Andreas Jarvstad and Ulrike Hahn. Source reliability and the conjunction fallacy. *Cognitive Science*, 35(4):682–711, 2011.

102 Peter J Collins and Ulrike Hahn. Communicating and reasoning with verbal probability expressions. *Psychology of Learning and Motivation*, 69:67–105, 2018.

103 P Collins and U Hahn. We might be wrong, but we think that hedging doesn't protect your reputation. *Journal of experimental psychology. Learning, memory, and cognition*, 2019.
104 Marko Tešić and Ulrike Hahn. The impact of explanation on explainee's beliefs and explainer's perceived reliability, in prep.
105 Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.
106 Marko Tešić and Ulrike Hahn. Explanation in AI systems. In S. Muggleton and N. Chater, editors, *Human-Like Machine Intelligence*. Clarendon Press. Oxford, UK, forthcoming.
107 Marko Tešić and Ulrike Hahn. Can counterfactual explanations of ai systems' predictions skew lay users' causal intuitions about the world? if so, can we correct for that? *Patterns*, 3 (12):100635, 2022.