Explanation in AI systems

Marko Tešić and Ulrike Hahn
Department of Psychological Sciences, Birkbeck, University of London

OXFORD
UNIVERSITY PRESS

Acknowledgements

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) as part of the Bayesian Argumentation via Delphi (BARD) project, and by the Humboldt Foundation.

1
Explanation in AI systems

In this chapter, we consider recent work aimed at guiding the design of algorithmically generated explanations. The chapter proceeds in four parts. Firstly, we introduce the general problem of machine-generated explanation and illustrate different notions of explanation with the help of Bayesian belief networks. Secondly, we introduce key theoretical perspectives on what constitutes an explanation, and more specifically a 'good' explanation, from the philosophy literature. We compare these theoretical perspectives and the criteria they propose with a case study on explaining reasoning in Bayesian belief networks and present implications for AI. Thirdly, we consider the pragmatic nature of explanation with the focus on its communicative aspects that are manifested in considerations of trust. Finally, we present conclusions.

1.1 Machine-generated explanation

Recent years have seen a groundswell of interest in machine-generated explanation for AI systems (DARPA, 2016; Doshi-Velez and Kim, 2017; Montavon *et al.*, 2018; Rieger *et al.*, 2018; Samek *et al.*, 2017). Multiple factors exert pressure for supplementing AI systems with explanations of their outputs. Explanations provide transparency for what are often black-box procedures. Hence transparency is viewed as critical for fostering the acceptance of AI systems in real-world practice (Bansal *et al.*, 2014; Chen *et al.*, 2014; Fallon and Blaha, 2018; Hayes and Shah, 2017; Mercado *et al.*, 2016; Wachter *et al.*, 2017*b*), last but not least, because transparency might be a necessary ingredient for dealing with legal liability (Felzmann *et al.*, 2019; Doshi-Velez *et al.*, 2017; Goodman and Flaxman, 2016; Wachter *et al.*, 2017*a*). At the same time, decades of research in AI make plausible the claim that AI systems genuinely able to navigate real-world challenges are likely to involve joint human-system decision making, at least for the foreseeable future. This however, requires AI systems to communicate outputs in such a way as to allow humans to make informed decisions.

The challenge of developing adequate, machine-generated explanation is a formidable one. For one, it requires an accessible model of how AI system's outputs or conclusions were arrived at. This poses non-trivial challenges for many of the presently most successful types of AI systems, such as convolutional, deep learning networks (Collobert *et al.*, 2011; Goodfellow *et al.*, 2016; Graves and Schmidhuber, 2005; Krizhevsky *et al.*, 2012) which are already notoriously opaque to external observers, let alone offering up representations of a system's internal processes that could be used to drive explanation generation. As such, the automated generation of explanation is arguably easier to achieve with AI systems that operate with models that are formulated at acceptable user levels of engagement; at least here, the step of translating low-level representations into a suitable higher-level representations accessible to us is, in a large number of cases, already taken care of.

Bayesian Belief Networks (BNs) are an AI technique that has been viewed as significantly more interpretable and transparent than deep neural networks (Gunning and Aha, 2019), while still possessing a notable predictive power and being applied to various contexts ranging from defence and military (Falzon, 2006; Laskey and Mahoney, 1997; Lippmann *et al.*, 2006) and cyber security (Chockalingam *et al.*, 2017; Xie *et al.*, 2010), over medicine (Agrahari *et al.*, 2018; Wiegerinck *et al.*, 2013), and law and forensics (Lagnado *et al.*, 2013; Fenton *et al.*, 2013), to agriculture (Drury *et al.*, 2017) as well as psychology, philosophy, and economics (see below). As such, BNs seem to serve well one of the goals of this chapter which is to bring and overlay insights on explanations from different areas of research: they are a promising meeting point connecting the research on machine-generated explanation in AI and the research on human understanding of explanation in psychology and philosophy. We thus use BNs as the focal point of our analysis in this chapter. Given the increasing popularity of BNs within AI (Friedman *et al.*, 1997; Pernkopf and Bilmes, 2005; Roos *et al.*, 2005; Ng and Jordan, 2002), including their relation to deep neural networks (Choi *et al.*, 2019; Rohekar *et al.*, 2018; Wang and Yeung, 2016) and efforts to explain deep neural networks via BNs (Harradon *et al.*, 2018), this should be intrinsically interesting. Furthermore, we take the kinds of issues we identify here to be indicative of the kinds of problems and distinctions that will likely emerge in *any* attempt at machine-generated explanation.

1.1.1 Bayesian belief networks: a brief introduction

Bayesian belief networks provide a simple graphical formalism for summarising and simplifying joint probability distributions in such a way as to facilitate Bayesian computations (Neapolitan, 2003; Pearl, 1988). Specifically, BNs use independence relations between variables to simplify the computation of joint probability distributions in cases of multivariate problems. As Bayesian models, they have a clear normative foundation: “being Bayesian”, that is, assigning degrees of belief in line with the axioms of probability (the Dutch Book argument, Ramsey 2016; Vineberg 2016), and the use of Bayes’ rule to update believes in light of new evidence (also known as ‘conditionalization’) which uniquely minimises the inaccuracy of an agent’s beliefs across all possible worlds (that is, regardless of how the world turns out), on the condition that inaccuracy is measured with the Brier score and those worlds are finite (see, e.g., the formal results outlined in Pettigrew 2016). In other words, Bayesian computations specify how agents *should* change their beliefs, if they wish those beliefs to be accurate. As a result, the Bayesian framework has seen widespread application not just in AI, but also in philosophy, economics, and psychology (Bovens and Hartmann, 2003; Dardashti *et al.*, 2019; Dewitt *et al.*, 2018; Dizadji-Bahmani *et al.*, 2011; Fenton *et al.*, 2013; Hahn and Oaksford, 2006; Hahn and Oaksford, 2007; Hahn and Hornikx, 2016; Harris *et al.*, 2016; Howson and Urbach, 2006; Liefgreen *et al.*, 2018; Madsen *et al.*, 2018; Neil *et al.*, 2008; Phillips *et al.*, 2018; Pilditch *et al.*, 2018; Rehder, 2014; Rottman and Hastie, 2014; Spiegler, 2016; Tešić, 2019; Tešić and Hahn, 2019; Tešić *et al.*, 2020).

In particular, BNs are helpful in spelling out the implications of less intuitive interactions between variables. This is readily illustrated with the example of "explaining away", a phenomenon that has received widespread psychological investigation (Davis and Rehder, 2017; Fernbach and Rehder, 2013; Liefgreen *et al.*, 2018; Morris and Larrick, 1995; Pilditch *et al.*, 2019; Rehder, 2014; Rehder and Waldmann, 2017; Rottman and Hastie, 2014; Rottman and Hastie, 2016; Sussman and Oppenheimer, 2011; Tešić *et al.*, 2020). Figure 1.1 illustrates a simple example of explaining away. There are two potential causes, a physical abuse and haemophilia (a genetic bleeding disorder), of a single effect, bruises on a child's body. Before finding out anything about whether there are bruises on a body, the two causes are independent: learning that a child is suffering from haemophilia will not change our beliefs about whether the child is physically abused. However, if we learn that the child has bruises on its body, then the two causes become dependent: additionally learning that the child is suffering from haemophilia will change (decrease) the probability that it has been physically abused since haemophilia alone is sufficient to explain away the bruises.

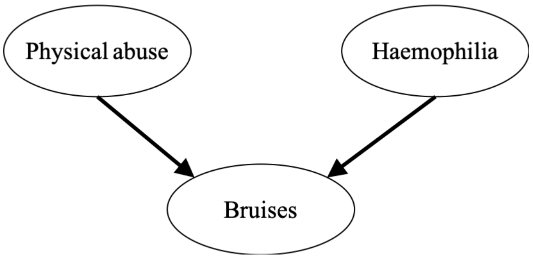


Fig. 1.1 A BN model of explaining away.

The example illustrates not just BNs ability to model explaining away situations and provide us with both qualitative and quantitative normative answers, but also their advantage over classical logic and rule-based expert systems. A rule-based expert system consisting of a set of IF-THEN rules and a set of facts (see Grosan and Abraham 2011) may carry out an incorrect chaining in situations representing explaining away. For instance, a rule-based system may combine plausibly-looking rules "If the child is suffering from haemophilia, then it is likely the child has bruises" with "If the child has bruises, then it is likely the child is physically abused" to get "If the child is suffering from haemophilia, then it is likely the child is physically abused". However, we know that actually the opposite is true: learning about haemophilia makes physical abuse less likely (Pearl, 1988). The application of rule-based expert systems to legal and medical contexts (Grosan and Abraham, 2011) where explaining away and other causal-probabilistic relationships can be found highlights the importance of accurately capturing these relationships in computational terms.

1.1.2 Bayesian belief networks: explaining evidence

Another important feature of BNs is that they can be used for both predictive reasoning and diagnostic reasoning (often referred to as an abduction, see, e.g., Korb and

Nicholson 2010). Consider the BN in Figure 1.1. An example of predictive reasoning would be inferring a probability that the child has bruises given that it's suffering from haemophilia (i.e. inference from causes (h) to effects (e)); whereas diagnostic reasoning would be inferring a probability of physical abuse from learning that the child has bruises (i.e. inference from effects (e) to causes (h)). Often, diagnostic reasoning (abduction) is used to find the most probable explanations (causes) of observed evidence (effects), that is to find the configuration h with the maximum $p(\mathrm{h} \mid \mathrm{e})$ (Pearl, 1988). Similarly, Shimony's (1991) partial abduction approach first marginalizes out variables that are not part of explanations (x) and then searches for the most probable h: that is, find h with the maximum $\sum_x p(\mathrm{h}, \mathrm{x} \mid \mathrm{e})$. More recently, Yuan *et al.* (2011) introduced a method they call 'Most Relevant Explanation' (MRE) which chooses the explanation that has the highest likelihood ratio compared to all other explanations: that is, find h with the maximum $p(\mathrm{e} \mid \mathrm{h})/p(\mathrm{e} \mid \overline{\mathrm{h}})$, where $\overline{\mathrm{h}}$ denotes all other alternative explanations to h. Nielsen *et al.* (2008) introduced a 'Causal Explanation Tree' (CET) method which uses the post-intervention distribution of variables (Pearl, 2000) in selecting explanations, which is in contrast to all previous methods since they use non-interventional distribution of variables in a BN. Drawing on their definition of causation, Halpern and Pearl (2005*b*) develop a definition of explanation to address a question of why certain evidence holds given users epistemic state. Their definition of explanation states that (i) a user should consider evidence to hold, (ii) an explanation (h) is a sufficient cause of evidence, (iii) h is minimal (i.e. it does not contain irrelevant or redundant elements), and (iv) h is not known at the beginning, but it is considered as a possibility. This is an improvement to other accounts. However, their account of causation has as an output again a set of variables in a BN model which are deemed as causes of evidence in the model. Yap *et al.* (2008) employ Markov blanket to determine which variables should feature in an explanation. A Markov blanket is of a node X includes all nodes that are direct parents, children, or children's parents of node X. A powerful property of Markov blanket is that knowing the sates of all the variables in a Markov blanket of X would uniquely determine the probability distribution of X: additionally learning the states of other variables outside the Markov blanket of X would not affect the probability distribution of X. Yap et al.'s 'Explaining BN Inferences' procedure identifies Markov nodes of evidence (i.e. nodes in a Markov blanket of the evidence node) and learns context specific independences in Markov nodes with a decision tree to exclude irrelevant nodes in an explanation of the evidence.

Despite the difference among the methods, they all share at least one commonality: explanation of evidence is exhausted by a set of variables in a BN that these methods have pointed to. In other words, evidence was provided a justification in terms of a set of variables. This is undoubtedly useful, but in certain contexts (e.g. high-cost domains, see Herlocker *et al.* 2000) it is arguably not enough to meet the demands of user transparency. In contrast to the notion of explanation as a justification of evidence is that of explanation of *reasoning processes* in BNs, and expert systems in general (Lacave and Díez, 2002; Sørmo *et al.*, 2005; Wick and Thompson, 1992). Here one is interested in how evidence propagates in a BN rather than in selecting a set of variables that would account for evidence. Explaining reasoning processes in BNs been a research focus amongst researchers for some time (see Lacave and Díez 2002 for an overview). We next describe one more recent attempt in the context of the Bayesian Argumentation via Delphi (BARD) project.

1.1.3 Bayesian belief networks: explaining reasoning processes

The BARD project (Cruz *et al.*, 2020; Dewitt *et al.*, 2018; Liefgreen *et al.*, 2018; Nicholson *et al.*, 2020; Phillips *et al.*, 2018; Pilditch *et al.*, 2018; Pilditch *et al.*, 2019) set as its goal the development of assistive technology that could facilitate group decision-making in an intelligence context. To this end, BARD provides a graphical user interface enabling intelligence analysts to represent arguments as BNs and allowing them to examine the impact of different pieces of evidence on arguments as well as to bring groups of analysts to a consensus via an automated Delphi method. An essential component of the system is the algorithm for generating natural language explanations of inference in a BN, or more specifically, an explanation of evidence propagation in a BN. This algorithm builds on earlier work by Zukerman and colleagues that have sought to use BNs to generate arguments (Zukerman *et al.*, 1998; Zukerman *et al.*, 1999). The algorithm uses an evidence-to-goal approach to generate explanations for a BN. An explanation starts with the given pieces of evidence and traces paths that describe their influence on intervening nodes until the goal is reached. In essence, the algorithm adopts a causal interpretation of the links between the connected nodes, finds a set of rules that describe causal relations in a BN, and calculates all paths between evidence nodes and target nodes and builds corresponding trees in order to determine the impact of evidence on target nodes. Figure 1.2 provides an example. There we have four pieces of evidence: *Emerson report*, *Quinns report*, and *AitF Sawyer Report* all stating that 'The Spider' is in the facility and *Comms Analyst Winter Report* stating that 'The Spider' is not in the facility. The goal is to explain the impact of these four pieces of evidence on two target variables, namely *Is The Spider in the facility?* and *Are logs true? (Are Emerson & Quinn spies?).* The algorithm first finds all relevant paths between evidence and target nodes, builds a corresponding tree and calculates the impact of evidence on the target, which is simply a difference between the probability of the target node *before* learning particular piece(s) of evidence and *after* learning particular piece(s) of evidence. This way the algorithm can find *HighImpSet*—nodes that have the highest impact on the target node, *NoImpSet*—nodes that, in light of the other evidence nodes, have no impact on the target node, and *OppImpSet*—nodes that have the opposite impact to that of *HighImpSet*. Finally, the algorithm realizes the explanations in English language using sentences, clauses and phrases devised and combined by means of a semantic grammar (Burton, 1976). The output of the algorithm is presented in Figure 1.3.

As can be seen, the output in Figure 1.3 provides significantly more information to the user than just a single verdict on whether or not the variable *Is The Spider in the facility?* is part of the explanation of the evidence, as would be the output of methods looking for a justification of evidence. In addition to the impact sets, it provides a natural language explanation on how different pieces of evidence influence the probability of the target variable. Nevertheless, there remain continued challenges with this approach. First, the algorithm retains difficulties in coping adequately with soft evidence, namely evidence that we do not learn with probability 1. For instance,

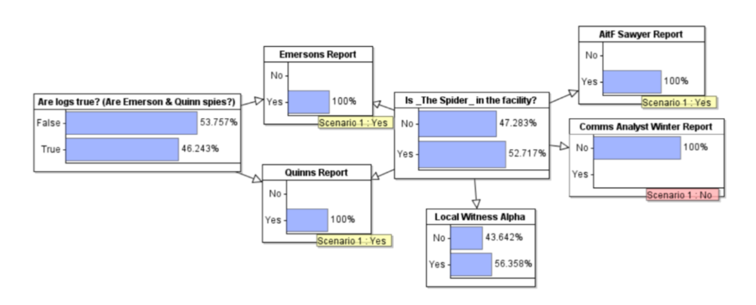
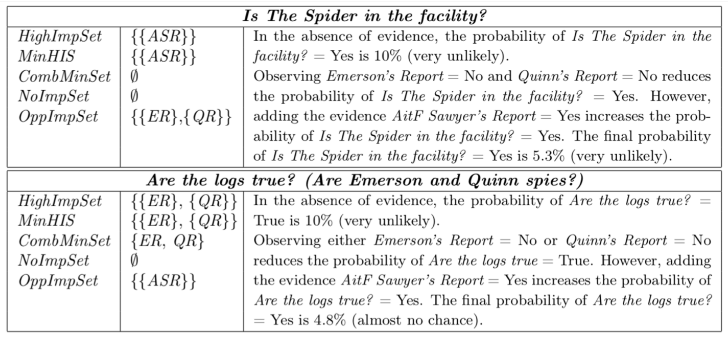


Fig. 1.2 A BN of a fictional scenario used in BARD testing phase. Four pieces of evidence are available: *Emerson Report*=Yes, *Quinns Report*=Yes, *AitF Sawyer Report*=Yes, and *Comms Analyst Winter Report*=No.



Is The Spider in the facility?		

HighImpSet
MinHIS
CombMinSet
NoImpSet
OppImpSet
	{{*ASR*}}
	{{*ASR*}}
	∅
	∅
	{{*ER*},{*QR*}}
		In the absence of evidence, the probability of *Is The Spider in the facility?* = Yes is 10% (very unlikely).
		Observing *Emerson's Report* = No and *Quinn's Report* = No reduces the probability of *Is The Spider in the facility?* = Yes. However, adding the evidence *AitF Sawyer's Report* = Yes increases the probability of *Is The Spider in the facility?* = Yes. The final probability of *Is The Spider in the facility?* = Yes is 5.3% (very unlikely).

Are the logs true? (Are Emerson and Quinn spies?)		

HighImpSet
MinHIS
CombMinSet
NoImpSet
OppImpSet
	{{*ER*}, {*QR*}}
	{{*ER*}, {*QR*}}
	{*ER*, *QR*}
	∅
	{{*ASR*}}
		In the absence of evidence, the probability of *Are the logs true?* = True is 10% (very unlikely).
		Observing either *Emerson's Report* = No or *Quinn's Report* = No reduces the probability of *Are the logs true* = True. However, adding the evidence *AitF Sawyer's Report* = Yes increases the probability of *Are the logs true?* = Yes. The final probability of *Are the logs true?* = Yes is 4.8% (almost no chance).

Fig. 1.3 A summary report generated by the BARD algorithm applied on the BN from Figure 1.2. In addition to the natural language explanation, it provides sets with nodes that are *HighImpSet*, *NoImpSet* and *OppImpSet*. For the purposes of this chapter we can ignore *MinHIS* and *CombMinSet*.

imagine that in the BN in Figure 1.2 we additionally learn that the logs are most likely true, but we are not absolutely convinced. To reflect that we would set *p(Are logs true? (Are Emerson & Quinn spies)* = True) to equal 0.95 for instance. Thus the probability of *Are logs true? (Are Emerson & Quinn spies)* = True has changed from 0.46243 to 0.95, but it didn't go all the way to 1. The current version of the algorithm is not able to calculate the impact of such change. Second, the explanations generated by the algorithm are not aimed specifically at what a human user *might find hard to understand*. To make matters worse, it is arguably the interactions between variables and the often counterintuitive effects of these, that users will most struggle with (for psychological evidence to this effect see for example Dewitt *et al.* 2018; Liefgreen *et al.* 2018; Phillips *et al.* 2018; Pilditch *et al.* 2018; Pilditch *et al.* 2019; Tešić *et al.* 2020). In other words, the system generates an (accurate) explanation, but not necessarily a good explanation. For further guidance on what might count as a good explanation we consult research on this topic within the philosophy of science and epistemology, where the topic has raised decades of interest.

1.2 Good explanation

1.2.1 A brief overview of models of explanation

The historic point of departure for thinking about the nature of explanation in philosophy is the covering law model (Hempel and Oppenheim, 1948), also known as the "deductive-nomological model" of scientific explanation (where nomological means pertaining to the laws of nature). This model construes explanation as a deductive argument with true premises that has the phenomenon to be explained (the so-called explanandum) as its conclusion. Specifically, this conclusion is derived from general laws and particular facts. For example, an explanation of a position of a planet at a point in time consists of a derivation of that position from the Newtonian laws governing gravity (general law), and information about the mass of the sun, the mass of the planet, and position at a particular time and velocity of each (particular facts) (Woodward, 2017). A key feature of this model is that it views explanation and prediction as essentially two sides of the same coin. In the same way that Newtonian laws and information about the mass of the sun and the planet etc. can be used to predict the position of the planet at some future time the inference can also be used to explain the position of the planet after we observe it. In other words, we see here the same tight coupling between diagnostic reasoning and predictive reasoning that we mentioned earlier in the context of BNs. However, while this coupling works in BN's across the range of possible probabilities, it becomes forced in the covering law model when dealing with probabilistic explanations, in particular, when dealing with cases where the probability of observing the conclusion is low. Not only do probabilistic contexts move the inference from deduction to an ampliative inference where the conclusion is no longer certain, the symmetry between explanation and prediction also becomes forced. We might for example readily explain someone being struck by the lightning by appealing to stormy weather conditions and the fact that they were out in the open. But we would nevertheless hasten to predict that someone will be struck by the lightning even if they are out in the open and there is a storm as it is a low probability event. This limits the utility of the covering law model within the social sciences where deduction is not commonplace and where low probability events are often found. Hempel himself was aware of these difficulties to the extent that he proposed two versions of the model the deductive-nomological model and an inductive-statistical one, and himself thought that the inductive statistical model applied only when the explanatory theory involves high probabilities. Even this restriction, however, does not deal appropriately with the asymmetries involved in explanation. These can be observed even in purely deductive context as is illustrated by the following example from Salmon (1992). Imagine there is a flagpole with a shadow of 20m and someone asks why that shadow is 20m long. In this context, it seems appropriate to explain the length of the shadow by appealing to the height of the flagpole, the position of the sun, and the laws of trigonometry. These together adequately explain the shadows length. But note that this inference can be reversed: when can also use the sun's position, the laws of trigonometry, and the length of the shadow to explain the height of the flagpole. This, however, seems wrong; an adequate explanation of the height of that flagpole presumably involves an appeal to the maker of the flagpole in some form or other. Examples such as these serve to illustrate not just the limits of Hempel's account but of the limits of deductive approaches in the context of explanation more generally.

The asymmetric relations involved in explanation prompted alternative accounts of scientific explanation within the subsequent literature. Chief among these are causal accounts which assert that to explain something is to give a specification of its causes. The standard explication of cause in this context is that of factors without which something could not be the case (i.e. *conditio sine qua non*). This deals readily even with low probability events, and causes can be identified through a process of "screening off". If one finds that $p(M \mid N, L) = p(M \mid N)$, then N screens off L from M and that M is causally irrelevant to L. For example, a reading of a barometer (B) and whether there is a storm (S) are correlated. However, knowing the atmospheric pressure (A) will make these two independent: $p(B \mid A, S) = p(B \mid A)$, suggesting no causal relationship between B and S. However, the notion of cause in itself is notoriously fraught as is evidenced by J. L. Mackie's convoluted (Mackie, 1965) definition whereby a cause is defined as an "insufficient but necessary part of an unnecessary but sufficient condition". This rather tortured definition reflects the difficulties with the notion of causation when multiple causes are present which is giving rise to overdetermination (for example, decapitation and arsenic in the blood stream can both be the causes of death), the difficulties created by causal chains (for example, tipping over the bottle which hits the floor which releases the toxic liquid) and the impact of background conditions (for example, putting yeast in the dough causes it to rise, but only if it is actually put in the oven, the oven works, the electrical bills have been paid, and so on). It is a matter of ongoing research to what extent causal Bayes nets, that is BN's supplemented with the *do*-calculus (Pearl, 2000), provide a fully satisfactory account of causality and these difficulties (see also Halpern and Pearl 2005*a*). At the same time, the difficulty of picking out a single one out of multiple potential causes points to the second main alternative to Hempel's covering law model, namely so-called pragmatic accounts of explanation.

According to van Fraassen (1977) an explanation always has a pragmatic component: specifically what counts as an explanation in any given context depends on the possible contrasts the questioner has in mind. For example, consider the question "why did the dog bury the bone?". Different answers are required for different prosodic contours: "why did the *dog* (i.e., not some other animal) bury the bone?"; why did the dog *bury* the bone? (say, rather than eat it); why did the dog bury the *bone*? (say, rather than the ball). In short, pragmatic accounts bring into the picture the recipient of an explanation while rejecting a fundamental connection between explanation and inference assumed by Hempel's model.

1.2.2 Explanatory virtues

Philosophy has not only tried to characterise the nature of explanation, it has also sought to identify the so-called "explanatory virtues". Of the many things that might count as an explanation according to a particular theoretical account of explanation, not all may seem equally good or compelling. Among 'explanations', we might ask what distinguishes better ones from poorer ones. In search of explanatory virtues that characterise good explanation, a number of factors have been identified: explanatory power, unification, coherence, and simplicity are chief among these. Explanatory power often relates to the ability of an explanation to decrease the degree to which we find the explanandum surprising; the less surprising the explanadum in light of an explanation the more powerful the explanation. For instance, a geologist may find a prehistoric earthquake as explanatory of deformation in layers of bedrock to the extent that these deformations would be less surprising given the occurrence of such an earthquake (Schupbach and Sprenger, 2011). Unification refers to explanations' ability to provide a unified account of a wide range of phenomena. For example, Maxwell's theory (explanation) managed to unify electricity and magnetism (phenomena). Coherence renders explanations that better fit our already established beliefs to be preferred to those that do not (Thagard, 1989). Explanations can also have internal coherence, namely how parts of an explanation fit together. An often motioned explanatory virtue is simplicity. According to Thagard (1978), simplicity is related to the size and nature of auxiliary assumptions needed by an explanation to explain evidence. For instance, the phlogiston theory of combustion needed a number of auxiliary assumptions to explain facts that are easily explained by Lavoisier's theory: it assumed existence of a fire-like element 'phlogiston' that's given away in combustion and that had 'negative weight' since bodies undergoing combustion increase in weight. Others operationalise simplicity as a number of causes invoked in an explanation: the more causes the less simple an explanation (Lombrozo, 2007).

While all of these factors seem intuitive, debate persists about their normative basis. In particular, there is ongoing debate within the philosophy of science about whether these factors admit of adequate probabilistic reconstruction (Glymour, 2014). At the same time, there is now a sizeable program within psychology that seeks to examine the application of these virtues to every day lay explanation. This body of work probes the extent to which lay reasoners endorse these criteria when distinguishing better from worse explanations (Bechlivanidis *et al.*, 2017; Bonawitz and Lombrozo, 2012; Johnson *et al.*, 2014*a*; Johnson *et al.*, 2014*b*; Lombrozo, 2007; Lombrozo, 2016; Pennington and Hastie, 1992; Sloman, 1994; Williams and Lombrozo, 2010; Zemla *et al.*, 2017). To date, researchers found some degree of support for these factors, but also seeming deviations in practice.

Finally, there is a renewed interest in both philosophy and psychology in the notion of inference to the best explanation (Harman, 1965; Lipton, 2003). Debate here centres around the question of whether the fact that an explanation seems in some purely non-evidential way better than its rivals should provide grounds for thinking that explanation is more probable. In other words, the issue is whether an explanation exhibiting certain explanatory considerations that other explanations do not should be considered more likely to be true (Douven, 2013; Harman, 1967; Henderson, 2013; Lipton, 2003; Thagard, 1978). Likewise this has prompted psychological research into whether such probability boosts can actually be observed in reasoning contexts (Douven and Schupbach, 2015). The research on explanatory virtues in both philosophy and psychology is still very much active.

1.2.3 Implications

What if anything can be inferred for the project of machine-generated explanation from this body of research? First, the notion of explanation that emerges is a potentially very different one across different parts of this literature. An explanation is variously a hypothesis or a variable, and inference, or an answer to a question. From afar, in the context of inference, we may distinguish explanation as a product from explanation as a the process (Lombrozo, 2012). From a product perspective, an explanation is a hypothesis or a claim that accounts for evidence when prompted to do so. In contrast, explanation can also be viewed as a cognitive activity (process) that has as its goal to generate explanation 'products'. This distinction nicely corresponds to what we have found in the literature on explanations in BNs reported in Section 1.1. There, explanations are also viewed as products that consist of nodes in a BN that aim to account for other nodes in a BN (i.e. evidence nodes) as well as reasoning processes that include not just the final products but also how one arrives at these products. This nicely illustrates how different disciplines can come to similar conceptual distinctions without these distinctions being communicated from one to another. Establishing the necessary communication channels will arguably help connect the research areas working on closely related questions thus bringing different perspectives and inputs into these areas.

This brings us to our second point. It seems clear that BNs provide a potential tool that is compatible with present thinking about the explanation at least in principle. They can capture the asymmetry in explanation as arcs are directed and can have a causal interpretation (Pearl, 2000), whilst at the same time being able to make predictions. This is in contrast to, for instance, a rule-based expert system with IF-THEN rules and a set of facts which would be susceptible to the symmetry 'error' in explanation illustrated by the flagpole example from Section 1.2.1. A BN on the other hand would be able to account for the asymmetry given a causal interpretation and directional representation of arrows. However, it is neither clear how explanations in BNs can capture the pragmatic component that van Fraassen raises nor how to operationalise explanatory virtues in the context of BNs. These are all potential avenues for further research.

Another point that can be drawn is that the debates about the nature of explanation and explanatory virtues have been conducted at very high levels of abstraction. They have also typically focused on philosophy of science and issues tightly related to it. This is true even for psychological research on explanation, to the extent that it has tried to model psychological investigations more or less directly on philosophical distinctions. However, for the purposes of developing suitable AI algorithms, it does also seems important to work in the opposite direction, as it were from the bottom up. In other words, it seems important to simultaneously start with simple applications of BN's to multiple variable problems, and consider what kinds of explanations a human (expert) would produce. This would shed light on the kinds of explanations that seem natural and appropriate human users as well as provide guidelines on possible theories of explanation. A similar point has been made in an AI literature with additionally emphasising the importance of human-generate explanations serving as a baseline for comparison with machine-generated explanations (Doshi-Velez and Kim, 2017). To explore these ideas further, we conducted a case study on explanation in BNs which we describe next.

1.2.4 A brief case study on human-generated explanation

The main motivation of the study was to find out what kinds of explanations a human (expert) would produce upon being presented with evidence in a BN. This is interesting from both the psychological and the AI perspective as, on the one hand, it could give us further inputs into human explanatory intuitions and preferences and, on the other hand, it could inform the AI researcher that aims to build algorithms for an automated generation of explanations.

In the study we used four BNs of different complexity found on a publicly available BN repository `https://www.norsys.com/netlibrary/index.htm`. The number of nodes in the BNs ranged from 4 to 18 and the number of arcs raged from 4 to 20. Figure 1.4 includes a BN used in the study.

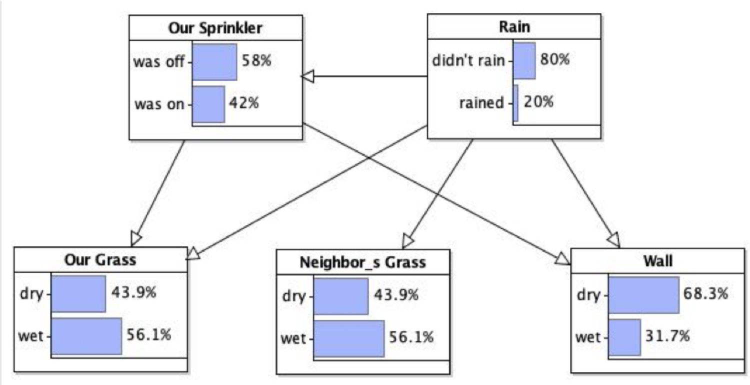


Fig. 1.4 A BN used in the case study.

Three independent raters, all of whom were experts in probabilistic reasoning, were then given access to implementations of the respective model in order to probe the BNs in more detail, and asked to then provide answers to questions that prompted them to consider how learning evidence changed the probabilities of the target nodes. Below are sample questions:

- Given evidence: $\{$*Neighbours grass* = wet$\}$
  Question: How does the probability of '*Our Sprinkler* = was on' change compared to when there was no evidence and why?
- Given Evidence: $\{$*Our grass* = wet, *Wall* = wet$\}$
  Question: How does the probability of '*Rain* = rained' change compared to when the only available evidence was '*Our Grass* = wet' and why?

Subsequently, the three independent sets of answers were subjected to an analysis by a fourth person in order to identify both commonalities and differences across the answers. This then formed the basis of the subsequent evaluation of those answers.

We describe the full set of results elsewhere (Tešić and Hahn, prep*a*), restricting ourselves here to an initial summary of the results. First, we observed high levels of agreement across answers. Differences were typically more presentational than substantive. For example, the following three statements all seek to describe the same state of affairs:

- 'As A is true C is more likely to be true if B is true and less likely to be true if B is false. As we do not know B these alternatives essentially cancel themselves out and leave the probability of C unchanged.'
- 'It does not change. $P(\mathrm{C} \mid \mathrm{A})$ is equal to $P(\mathrm{C})$ if $P(\mathrm{C} \mid \mathrm{A}, \mathrm{B}) = P(\mathrm{C} \mid {\sim}\mathrm{A}, {\sim}\mathrm{B})$ and $P(\mathrm{C} \mid \mathrm{A}, {\sim}\mathrm{B}) = P(\mathrm{C} \mid {\sim}\mathrm{A}, \mathrm{B})$ (assuming $P(\mathrm{B}) = 0.5$), which here is the case.'
- 'According to model parameters: If A and B both true or both false, then C has probability .75. If A true but B false, or vice-versa, then C has probability .25. When we know A is true, and prior for B is 50%, there is a 50% that probability of C is 75% and a 50% that probability of C is 2%, therefore overall probability of C is 50%.'

Second, all appeal to *hypothetical reasoning* as a way of unpacking interactions of evidence variables:

- '*Wall* = wet is a lot more likely if the sprinkler was on than if it rained (as a matter of fact, if it rained, the wall is more likely to be dry than wet). Since, *Our sprinkler* = was on went down, *Wall* = wet went down.'

Third, causal explanations are prevalent, and, where present, typically appeal to the underlying target system being modelled (i.e., sprinkler, wall, rain) as opposed to the model itself:

- 'The probability of rain decreases because, although the sprinkler and rain can both cause our grass to be wet, the wet wall is more likely to happen when the sprinkler is on rather than rain.'

Notably, in appealing to causes, it is the most probable cause that seems to be highlighted as an explanation:

- 'There is a decrease [in probability] because the most likely cause of our grass being wet is the sprinkler and since the wall is dry the sprinkler is unlikely to be on.'

Finally, these data seem to suggest that the structure of the BNs is exploited in order to zero in on 'the explanation' as a subset of all variables described in the problem. Specifically, explanations seemed to make use of the Markov blanket: a set of nodes consisting of the another node's parents, its children, and its children's parents and making that node conditionally independent of the rest of the network (Korb and Nicholson, 2010). In addition to Markov blanket, rater's descriptions mostly followed the direction of evidence propagation, i.e. followed the directed paths in a BN:

- 'The probability of *Battery voltage* = dead increases because failure of the car to start could be explained by the car not cranking and the likely cause of this is a faulty starter system. A dead battery is one possible explanation for a faulty starter system.'

This suggests that the explanatory virtue of 'simplicity' might, in a BN context, be conceptualised in terms terms of a Markov blanket and path direction.

In summary, we see multiple features of the general philosophical literature reflected in these explanations of everyday situations expressed with a BN model: a focus on an inference or a reasoning process; the use of causal explanation for a probabilistic system; the directional nature of explanation (its asymmetry); indications of pragmatic sensitivity in that hypotheticals are used to express relevant 'contrasts'; and, finally, an emerging notion of simplicity in the use of the Markov blanket.

These results are, however, still very much preliminary further research is needed, but hopefully they give us some sense of the kinds of explanations a human (expert) may produce and, potentially, prefer.

1.3 Bringing in the user: bi-directional relationships

One of the themes of this chapter is that an explanation is an explanation for someone. We have already encountered this in pragmatic theories of explanation within the philosophy literature, but it seems particularly pertinent to the AI context. As illustrated by the BARD assistive reasoning tool example, if explanations for AI systems are to be effective, they must take the user into account. On the one hand, this will require detailed research into what it is that the users of a given system do and do not readily understand. These concerns will largely be specific to the context of the AI system in question. But there are also further general considerations, which apply across potential systems, that have not yet been adequately discussed.

1.3.1 Explanations are communicative acts

Only very recently has it been noted that the provision of an explanation, whether from human or machine, is a *communicative act*. In order to understand the impact of explanations, one must consequently consider the pragmatics of (natural language) communication. Pragmatics, that is the part of language that deals with meaning in context, tells us much about how users will come to interpret explanations. The need for AI researchers to consider pragmatics in the context of machine-generated explanation has recently been highlighted by Miller (2019) who very persuasively argues that the research in explainable AI has largely neglected insights from social sciences, one of which is a very important commonplace regarding the social aspect of explanations where it is argued that explanations are often presented relative to the explainer's beliefs about the explainee's beliefs. In particular, Miller argues that explanations go beyond search for and presentation of associations and causes of evidence, but that they are also contextual: explainer and explainee may have different background beliefs regrading certain observed pieces of evidence and an explainable AI should address this.

We briefly consider some further potential implications for the recipients of explanations here. One general feature of communicative acts is that they provide information about the speaker, intended or otherwise. Recent work concerned with the question of trust has highlighted the interplay between culture and its contents and the perceived reliability of the speaker (Olsson and Vallinder, 2013; Bovens and Hartmann, 2003). One upshot of this, is that receiving an explanation is likely to change perceptions of the reliability of the explanation's source. Here, it is not only characteristics of the explanation such as its perceived cogency, how articulately it is framed, or how easily to process it that are likely to influence perceived source reliability, there are also likely to be effects of the specific content. In particular, the extent to which the content of the message fits with our present (uncertain) beliefs about the world has been shown to affect beliefs about the issue at hand as well as beliefs about the perceived reliability of the speaker (Collins *et al.*, 2018; Collins and Hahn, 2019).

1.3.2 Explanations and trust

Because message content and source reliability/trustworthiness jointly determine the impact of the communication on our beliefs, these interactions are likely to be consequential for the extent to which the machine conclusion being explained is itself perceived to be true. The literature in AI, in particular recommender systems, has long recognized relationship between trust and explanation (Zhang and Chen, 2020). Majority of research suggests that providing an explanation improves user's trust in an AI system (Herlocker *et al.*, 2000; Sinha and Swearingen, 2002; Symeonidis *et al.*, 2009). However, the situation seems more intricate as more transparent systems do not always lead to increase in trust (Cramer *et al.*, 2008), and sometimes poor explanations can lead to reduced acceptance of the AI systems (Herlocker *et al.*, 2000). To explore the interactions between explanations and trust, in addition to manipulation transparency of AI systems, one would also need to experimentally manipulate a level of trust users have in them. Recently, we have conducted empirical work on the relationship between reliability (which is related to the notion of trust as understood in AI) and explanation in non-AI context (Tešić and Hahn, prep*b*). We have performed three experiments where we used simple dialogues between two people (in the condition where an explanation was provided these were then explainer and explainee) on five different issues to show that (i) providing an explanation for a claim increases not just people's convincingness in the claim but also their reliability of the person providing an explanation compared when there is no such explanation and (ii) providing an explanation has a significantly greater impact on the convincingness and reliability when people's initial (prior) reliability of the source is low compared to when that reliability is high. In the context of AI, these results suggest that providing a (good) explanation of AI system's decisions will arguably increase people's convincingness in/acceptance of these decisions as well as people's perceived the reliability/trust of the system. In particular, the impact of providing an explanation will be greater (and most useful) if people's initial perceived reliability/trust of an AI system is low.

1.3.3 Trust and fidelity

Recent surge of model agnostic post-hoc explanations of black-box deep learning models has significantly pushed the horizons of explainable AI, but at the same time it has also introduced the *fidelity* problems. Namely, unlike explanations of BN where original BN models could be used to generate explanations (either as justification of evidence or as explanation of reasoning processes), deep learning models are not transparent enough for either a lay or an expert human user to be able to explain the models' outputs; rather, one resorts to explanation models that are independent of deep learning models to generate explanations of these black-box models' decisions after these decisions have been made, i.e. post-hoc (Ribeiro *et al.*, 2016; Zhang and Chen, 2020). The explanations models are often model agnostic as they should be able to explain decisions of any (black-box) model. Post-hoc model agnostic explanation models have certainly furthered the work on explanation in AI, but they have also prompted questions regarding the degree to which the explanations generated by models are reflecting the real mechanisms that generated decisions of a deep learning model: i.e. they have raised questions regarding the fidelity of explanation models (Sørmo *et al.*, 2005; Ribeiro *et al.*, 2016). In the literature, the trade-off between fidelity and interpretability of explanations models is often acknowledged: the higher the fidelity of an explanation model to the black-box model the lower the interpretability of that model and its transparency to a human user (Ribeiro *et al.*, 2016). This however brings trust into the consideration. On the one hand, if higher interpretability is to increase trust, then trust may be negatively affected by higher fidelity. On the other hand, if users expect higher fidelity explanation models, then lower fidelity may now negatively affect trust. This potentially interesting relationship between fidelity and trust is another open issue related to the interplay between a user and the system that could be addressed in the future research.

1.3.4 Further research avenues

The effect of the communication on the reliability of the source/trust and possibly fidelity, however, are unlikely to be the only way in which explanations alter beliefs about what it is that is being explained. For example, does providing an explanation constrain and/or make less ambiguous the underlying (causal) structure of the world that the explainee had in mind before receiving the explanation (or in the case of a BN, does providing an explanation restrict the number of potential BN structures that the explainee has mind)? How providing an explanation of an (ab)normal event in a causal chain of events reflects on our perceptions of that explanation? Does a detailed explanation of a usual and obvious succession of events make that explanation less preferred or worse compared to a less detailed explanation? All these questions call for further investigation and can have implications for the explainable AI project.

1.4 Conclusions

We have seen multiple ways to build different notions of what counts as an explanation. One of these involves explanation as identification of the variables that mattered in generating certain outcome. In the context of computational models of explanation in BN's this corresponds to the usual focus on explaining observed evidence via unobserved nodes within the network (Pacer *et al.*, 2013). In other words, the explanation identifies a justification/hypothesis. This is the notion of explanation that has figured prominently in work on computer-generated explanations as well as in psychological and philosophical literature on explanation. The second notion of explanation we considered includes explanation of the inference that links evidence and hypothesis. In the context of BN's this means explaining the inferences that lead to a change (or no change) in the probabilities of the query nodes. In other words, the explanation involves a target hypothesis plus information about the incremental reasoning process that identifies that hypothesis. Finally, we considered the notion of explanation in terms of providing understanding of what is hard to understand and or surprising to the system's user. Explanation in this widest sense requires not just identification of a best hypothesis and explanation of the incremental steps that lead to the identification of that hypothesis but also a user model that tells us what it is that human users find difficult. For this widest sense of explanation, psychological research is essential. Explanation so understood constitutes a fundamental problem of human computer interaction and only empirical research that seeks to understand the human user can lead to fully satisfactory answers.

Choosing among these three different notions of explanation, also directly affects the answer to the question of what counts as a good explanation. As we saw in this chapter, there is some guidance on the notion of good explanation that can be drawn from both the philosophy and psychology literature; but it is also clear that more specific work is required. In particular, most recent research suggests that such work will need to take into account that providing an explanation is a communicative act that changes perceptions of the communicator. In other words, explanation will not merely translate an extant result into a language understood by the user, it will likely affect how the user interprets the output of the system and the reliability of the system itself. This means the provision of explanation will likely affect what the user considers to be the verdict of the system in the first place, which could lead to further intricate relationships between trust and concepts such as fidelity. It is thus essential that future work on explanation within AI engage more fully with the pragmatic consequences of communicating explanation.

References

Agrahari, Rupesh, Foroushani, Amir, Docking, T Roderick, Chang, Linda, Duns, Gerben, Hudoba, Monika, Karsan, Aly, and Zare, Habil (2018). Applications of Bayesian network models in predicting types of hematological malignancies. *Scientific reports*, **8**(1), 6951.

Bansal, Aayush, Farhadi, Ali, and Parikh, Devi (2014). Towards transparent systems: Semantic characterization of failure modes. In *European Conference on Computer Vision*, pp. 366–381. Springer.

Bechlivanidis, Christos, Lagnado, David A, Zemla, Jeffrey C, and Sloman, Steven (2017). Concreteness and abstraction in everyday explanation. *Psychonomic bulletin & review*, **24**(5), 1451–1464.

Bonawitz, Elizabeth Baraff and Lombrozo, Tania (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental psychology*, **48**(4), 1156.

Bovens, Luc and Hartmann, Stephan (2003). *Bayesian epistemology.* Oxford University Press.

Burton, Richard R (1976). Semantic grammar: An engineering technique for constructing natural language understanding systems.

Chen, Jessie Y, Procci, Katelyn, Boyce, Michael, Wright, Julia, Garcia, Andre, and Barnes, Michael (2014). Situation awareness-based agent transparency. Technical report, Army research lab Aberdeen proving ground MD human research and engineering.

Chockalingam, Sabarathinam, Pieters, Wolter, Teixeira, André, and van Gelder, Pieter (2017). Bayesian network models in cyber security: a systematic review. In *Nordic Conference on Secure IT Systems*, pp. 105–122. Springer.

Choi, Arthur, Wang, Ruocheng, and Darwiche, Adnan (2019). On the relative expressiveness of Bayesian and neural networks. *International Journal of Approximate Reasoning*, **113**, 303–323.

Collins, PJ and Hahn, U (2019). We might be wrong, but we think that hedging doesn't protect your reputation. *Journal of experimental psychology. Learning, memory, and cognition.*

Collins, Peter J, Hahn, Ulrike, von Gerber, Ylva, and Olsson, Erik J (2018). The bi-directional relationship between source characteristics and message content. *Frontiers in psychology*, **9**, 18.

Collobert, Ronan, Weston, Jason, Bottou, Léon, Karlen, Michael, Kavukcuoglu, Koray, and Kuksa, Pavel (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, **12**(Aug), 2493–2537.

Cramer, Henriette, Evers, Vanessa, Ramlal, Satyan, Van Someren, Maarten, Rutledge, Lloyd, Stash, Natalia, Aroyo, Lora, and Wielinga, Bob (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, **18**(5), 455.

Cruz, Nicole, Desai, Saoirse Connor, Dewitt, Stephen, Hahn, Ulrike, Lagnado, David, Liefgreen, Alice, Phillips, Kirsty, Pilditch, Toby, and Tešić, Marko (2020). Widening access to bayesian problem solving. *Frontiers in Psychology*, **11**, 660.

Dardashti, Radin, Hartmann, Stephan, Thébault, Karim, and Winsberg, Eric (2019). Hawking radiation and analogue experiments: A Bayesian analysis. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*.

DARPA (2016). Explainable artificial intelligence (XAI) program. Retrieved from `https://www.darpa.mil/program/xplainable-artificial-intelligence`.

Davis, Zachary and Rehder, Bob (2017). The causal sampler: A sampling approach to causal representation, reasoning, and learning. In *Proceedings of the Cognitive Science Society.*

Dewitt, Stephen, Lagnado, David A, and Fenton, Norman E (2018). Updating prior beliefs based on ambiguous evidence. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society.*

Dizadji-Bahmani, Foad, Frigg, Roman, and Hartmann, Stephan (2011). Confirmation and reduction: A Bayesian account. *Synthese*, **179**(2), 321–338.

Doshi-Velez, Finale and Kim, Been (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Doshi-Velez, Finale, Kortz, Mason, Budish, Ryan, Bavitz, Chris, Gershman, Sam, O'Brien, David, Schieber, Stuart, Waldo, James, Weinberger, David, and Wood, Alexandra (2017). Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.

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

Douven, Igor and Schupbach, Jonah N. (2015). The role of explanatory considerations in updating. *Cognition*, **142**, 299–311.

Drury, Brett, Valverde-Rebaza, Jorge, Moura, Maria-Fernanda, and de Andrade Lopes, Alneu (2017). A survey of the applications of Bayesian networks in agriculture. *Engineering Applications of Artificial Intelligence*, **65**, 29–42.

Fallon, Corey K and Blaha, Leslie M (2018). Improving automation transparency: Addressing some of machine learning's unique challenges. In *International Conference on Augmented Cognition*, pp. 245–254. Springer.

Falzon, Lucia (2006). Using Bayesian network analysis to support centre of gravity analysis in military planning. *European Journal of operational research*, **170**(2), 629–643.

Felzmann, Heike, Villaronga, Eduard Fosch, Lutz, Christoph, and Tamò-Larrieux, Aurelia (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, **6**(1), 2053951719860542.

Fenton, Norman, Neil, Martin, and Lagnado, David A (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive science*, **37**(1), 61–102.

Fernbach, Philip M and Rehder, Bob (2013). Cognitive shortcuts in causal inference. *Argument & Computation*, **4**(1), 64–88.

Friedman, Nir, Geiger, Dan, and Goldszmidt, Moises (1997). Bayesian network classifiers. *Machine learning*, **29**(2-3), 131–163.

Glymour, Clark (2014). Probability and the explanatory virtues. *British Journal for the Philosophy of Science*, **66**(3), 591–604.

Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron (2016). *Deep Learning*. MIT Press.

Goodman, Bryce and Flaxman, Seth (2016). EU regulations on algorithmic decision-making and a “right to explanation”. In *ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. http://arxiv. org/abs/1606.08813 v1.*

Graves, Alex and Schmidhuber, Jürgen (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, **18**(5-6), 602–610.

Grosan, Crina and Abraham, Ajith (2011). Rule-based expert systems. In *Intelligent Systems*, pp. 149–185. Springer.

Gunning, David and Aha, David W (2019). Darpa’s explainable artificial intelligence program. *AI Magazine*, **40**(2), 44–58.

Hahn, Ulrike and Hornikx, Jos (2016). A normative framework for argument quality: argumentation schemes with a Bayesian foundation. *Synthese*, **193**(6), 1833–1873.

Hahn, Ulrike and Oaksford, Mike (2006). A Bayesian approach to informal argument fallacies. *Synthese*, **152**(2), 207–236.

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

Halpern, Joseph Y and Pearl, Judea (2005*a*). Causes and explanations: A structural-model approach. Part I: Causes. *The British journal for the philosophy of science*, **56**(4), 843–887.

Halpern, Joseph Y and Pearl, Judea (2005*b*). Causes and explanations: A structural-model approach. Part II: Explanations. *The British journal for the philosophy of science*, **56**(4), 889–911.

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

Harman, Gilbert (1967). Detachment, probability, and maximum likelihood. *Nous*, 401–411.

Harradon, Michael, Druce, Jeff, and Ruttenberg, Brian (2018). Causal learning and explanation of deep neural networks via autoencoded activations. *arXiv preprint arXiv:1802.00541*.

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

Hayes, Bradley and Shah, Julie A (2017). Improving robot controller transparency through autonomous policy explanation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pp. 303–312. IEEE.

Hempel, Carl G. and Oppenheim, Paul (1948). Studies in the logic of explanation.

Philosophy of science, **15**(2), 135–175.

Henderson, Leah (2013). Bayesianism and inference to the best explanation. *The British Journal for the Philosophy of Science*, **65**(4), 687–715.

Herlocker, Jonathan L, Konstan, Joseph A, and Riedl, John (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pp. 241–250.

Howson, Colin and Urbach, Peter (2006). *Scientific reasoning: the Bayesian approach.* Open Court Publishing.

Johnson, Samuel, Jin, Andy, and Keil, Frank (2014*a*). Simplicity and goodness-of-fit in explanation: The case of intuitive curve-fitting. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 36.

Johnson, Samuel, Johnston, Angie, Toig, Amy, and Keil, Frank (2014*b*). Explanatory scope informs causal strength inferences. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 36.

Korb, Kevin B. and Nicholson, Ann E. (2010). *Bayesian artificial intelligence.* CRC press.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105.

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

Lagnado, David A, Fenton, Norman, and Neil, Martin (2013). Legal idioms: a framework for evidential reasoning. *Argument & Computation*, **4**(1), 46–63.

Laskey, Kathryn Blackmond and Mahoney, Suzanne M (1997). Network fragments: Representing knowledge for constructing probabilistic models. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pp. 334–341. Morgan Kaufmann Publishers Inc.

Liefgreen, Alice, Tešić, Marko, and Lagnado, David (2018). Explaining away: significance of priors, diagnostic reasoning, and structural complexity. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society.*

Lippmann, Richard, Ingols, Kyle, Scott, Chris, Piwowarski, Keith, Kratkiewicz, Kendra, Artz, Mike, and Cunningham, Robert (2006). Validating and restoring defense in depth using attack graphs. In *MILCOM 2006-2006 IEEE Military Communications conference*, pp. 1–10. IEEE.

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

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

Lombrozo, Tania (2012). Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, 260–276.

Lombrozo, Tania (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, **20**(10), 748–759.

Mackie, John L. (1965). Causes and conditions. *American philosophical quarterly*, **2**(4), 245–264.

Madsen, Jens Koed, Hahn, Ulrike, and Pilditch, Toby D (2018). Partial source dependence and reliability revision: the impact of shared backgrounds. In *Proceedings* *of the 40th Annual Conference of the Cognitive Science Society.*

Mercado, Joseph E, Rupp, Michael A, Chen, Jessie YC, Barnes, Michael J, Barber, Daniel, and Procci, Katelyn (2016). Intelligent agent transparency in human–agent teaming for multi-uxv management. *Human factors*, **58**(3), 401–415.

Miller, Tim (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, **267**, 1–38.

Montavon, Grégoire, Samek, Wojciech, and Müller, Klaus-Robert (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, **73**, 1–15.

Morris, Michael W and Larrick, Richard P (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, **102**(2), 331–355.

Neapolitan, R. E. (2003). *Learning Bayesian Networks.* Upper Saddle River, NJ: Prentice Hall.

Neil, Martin, Fenton, Norman et al. (2008). Using Bayesian networks to model the operational risk to information technology infrastructure in financial institutions. *Journal of Financial Transformation*, **22**, 131–138.

Ng, Andrew Y and Jordan, Michael I (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pp. 841–848.

Nicholson, Ann E, Korb, Kevin B, Nyberg, Erik P, Wybrow, Michael, Zukerman, Ingrid, Mascaro, Steven, Thakur, Shreshth, Alvandi, Abraham Oshni, Riley, Jeff, Pearson, Ross et al. (2020). BARD: A structured technique for group elicitation of bayesian networks to support analytic reasoning. *arXiv preprint arXiv:2003.01207*.

Nielsen, Ulf, Pellet, Jean-Philippe, and Elisseeff, André (2008). Explanation trees for causal Bayesian networks. *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, 427–434.

Olsson, Erik J and Vallinder, Aron (2013). Norms of assertion and communication in social networks. *Synthese*, **190**(13), 2557–2571.

Pacer, Michael, Williams, Joseph, Chen, Xi, Lombrozo, Tania, and Griffiths, Thomas (2013). Evaluating computational models of explanation using human judgments. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligenc.*

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

Pearl, Judea (2000). *Causality: models, reasoning and inference.* Volume 29. Springer.

Pennington, Nancy and Hastie, Reid (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of personality and social psychology*, **62**(2), 189.

Pernkopf, Franz and Bilmes, Jeff (2005). Discriminative versus generative parameter and structure learning of Bayesian network classifiers. In *Proceedings of the 22nd international conference on Machine learning*, pp. 657–664. ACM.

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

Phillips, Kirsty, Hahn, Ulrike, and Pilditch, Toby D (2018). Evaluating testimony from multiple witnesses: single cue satisficing or integration? In *Proceedings of the*

40th Annual Conference of the Cognitive Science Society.

Pilditch, Toby D., Fenton, Norman, and Lagnado, David (2019). The zero-sum fallacy in evidence evaluation. *Psychological Science*, **30**(2), 250–260.

Pilditch, Toby D, Hahn, Ulrike, and Lagnado, David A (2018). Integrating dependent evidence: naïve reasoning in the face of complexity. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society.*

Ramsey, Frank P (2016). Truth and probability. In *Readings in Formal Epistemology*, pp. 21–45. Springer.

Rehder, Bob (2014). Independence and dependence in human causal reasoning. *Cognitive psychology*, **72**, 54–107.

Rehder, Bob and Waldmann, Michael R (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & cognition*, **45**(2), 245–260.

Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos (2016). "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*, pp. 1135–1144.

Rieger, Laura, Chormai, Pattarawat, Montavon, Grégoire, Hansen, Lars Kai, and Müller, Klaus-Robert (2018). Structuring neural networks for more explainable predictions. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pp. 115–131. Springer.

Rohekar, Raanan Y, Nisimov, Shami, Gurwicz, Yaniv, Koren, Guy, and Novik, Gal (2018). Constructing deep neural networks by Bayesian network structure learning. In *Advances in Neural Information Processing Systems*, pp. 3047–3058.

Roos, Teemu, Wettig, Hannes, Grünwald, Peter, Myllymäki, Petri, and Tirri, Henry (2005). On discriminative Bayesian network classifiers and logistic regression. *Machine Learning*, **59**(3), 267–296.

Rottman, Benjamin Margolin and Hastie, Reid (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological bulletin*, **140**(1), 109–139.

Rottman, Benjamin M and Hastie, Reid (2016). Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away. *Cognitive psychology*, **87**, 88–134.

Salmon, Wesley C. (1992). Scientific explanation. In *Introduction to the philosophy of science* (ed. M. H. Salmon, J. Earman, C. Glymour, J. Lennox, K. F. Schaffner, W. C. Salmon, J. D. Norton, J. McGuire, P. Machamer, and J. G. Lennox), Chapter 1, pp. 7–41. Englewood Cliffs, New Jersey, Prentice-Hall.

Samek, Wojciech, Wiegand, Thomas, and Müller, Klaus-Robert (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Schupbach, Jonah N and Sprenger, Jan (2011). The logic of explanatory power. *Philosophy of Science*, **78**(1), 105–127.

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

Sinha, Rashmi and Swearingen, Kirsten (2002). The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems*, pp. 830–831.

Sloman, Steven A (1994). When explanations compete: The role of explanatory coherence on judgements of likelihood. *Cognition*, **52**(1), 1–21.

Sørmo, Frode, Cassens, Jörg, and Aamodt, Agnar (2005). Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review*, **24**(2), 109–143.

Spiegler, Ran (2016). Bayesian networks and boundedly rational expectations. *The Quarterly Journal of Economics*, **131**(3), 1243–1290.

Sussman, Abigail B and Oppenheimer, Daniel M (2011). A causal model theory of judgment. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

Symeonidis, Panagiotis, Nanopoulos, Alexandros, and Manolopoulos, Yannis (2009). Moviexplain: a recommender system with explanations. In *Proceedings of the third ACM conference on Recommender systems*, pp. 317–320.

Tešić, Marko (2019). Confirmation and the generalized Nagel–Schaffner model of reduction: a Bayesian analysis. *Synthese*, **196**(3), 1097–1129.

Tešić, Marko and Hahn, Ulrike (2019). Sequential diagnostic reasoning with independent causes. In *Proceedings of the 41th Annual Conference of the Cognitive Science Society*.

Tešić, Marko and Hahn, Ulrike (in prep.*a*). Human-generated explanations of inferences in Bayesian networks: a case study.

Tešić, Marko and Hahn, Ulrike (in prep.*b*). The impact of explanation on explainee's beliefs and explainer's perceived reliability.

Tešić, Marko, Liefgreen, Alice, and Lagnado, David (2020). The propensity interpretation of probability and diagnostic split in explaining away. *Cognitive Psychology*, **121**, 101293.

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

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

Van Fraassen, Bas C. (1977). The pragmatics of explanation. *American Philosophical Quarterly*, **14**(2), 143–150.

Vineberg, Susan (2016). Dutch book arguments. In *The Stanford Encyclopedia of Philosophy* (Spring 2016 edn) (ed. E. N. Zalta). Metaphysics Research Lab, Stanford University.

Wachter, Sandra, Mittelstadt, Brent, and Floridi, Luciano (2017*a*). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, **7**(2), 76–99.

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

Wang, Hao and Yeung, Dit-Yan (2016). Towards Bayesian deep learning: A survey. *arXiv preprint arXiv:1604.01662*.

Wick, Michael R and Thompson, William B (1992). Reconstructive expert system explanation. *Artificial Intelligence*, **54**(1-2), 33–70.

Wiegerinck, Wim, Burgers, Willem, and Kappen, Bert (2013). Bayesian networks, introduction and practical applications. In *Handbook on Neural Information Processing*, pp. 401–431. Springer.

Williams, Joseph J and Lombrozo, Tania (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, **34**(5), 776–806.

Woodward, James (2017). Scientific explanation. In *The Stanford Encyclopedia of Philosophy* (Fall 2017 edn) (ed. E. N. Zalta). Metaphysics Research Lab, Stanford University.

Xie, Peng, Li, Jason H, Ou, Xinming, Liu, Peng, and Levy, Renato (2010). Using Bayesian networks for cyber security analysis. In *2010 IEEE/IFIP International Conference on Dependable Systems & Networks (DSN)*, pp. 211–220. IEEE.

Yap, Ghim-Eng, Tan, Ah-Hwee, and Pang, Hwee-Hwa (2008). Explaining inferences in bayesian networks. *Applied Intelligence*, **29**(3), 263–278.

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

Zemla, Jeffrey C, Sloman, Steven, Bechlivanidis, Christos, and Lagnado, David A (2017). Evaluating everyday explanations. *Psychonomic bulletin & review*, **24**(5), 1488–1500.

Zhang, Yongfeng and Chen, Xu (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval*, **14**(1), 1–101.

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

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