Human-Centric Intelligent Systems
https://doi.org/10.1007/s44230-024-00066-2

RESEARCH ARTICLE

Evaluating the Usefulness of Counterfactual Explanations from Bayesian Networks

Raphaela Butz[1] · Arjen Hommersom[1,2] · Renée Schulz[3] · Hans van Ditmarsch[4]

Received: 18 October 2023 / Accepted: 8 March 2024

Abstract
Bayesian networks are commonly used for learning with uncertainty and incorporating expert knowledge. However, they are hard to interpret, especially when the network structure is complex. Methods used to explain Bayesian networks operate under certain assumptions about what constitutes the best explanation, without actually verifying these assumptions. One such common assumption is that a shorter length of the causal chain of one variable to another enhances its explanatory strength. Counterfactual explanations gained popularity in artificial intelligence over the last years. It is well-known that it is possible to generate counterfactuals from causal Bayesian networks, but there is no indication which of them are useful for explanatory purposes. In this paper, we examine how to apply findings from psychology to search for counterfactuals that are perceived as more useful explanations for the end user. For this purpose, we have conducted a questionnaire to test whether counterfactuals that change an actionable cause are considered more useful than counterfactuals that change a direct cause. The results of the questionnaire indicate that actionable counterfactuals are preferred regardless of being the direct cause or having a longer causal chain.

Keywords Bayesian networks · XAI · Counterfactuals · Do-calculus

Abbreviations

AI	Artificial Intelligence
BN	Bayesian network
CBN	Causal Bayesian network
XAI	EXplainable Artificial Intelligence

Raphaela Butz, Arjen Hommersom, Renée Schulz and Hans van Ditmarsch contributed equally to this work.

✉ Raphaela Butz
raphaela.butz@ou.nl

Arjen Hommersom
arjen.hommersom@ou.nl

Renée Schulz
renee@jrcs.co.jp

Hans van Ditmarsch
hans.van-ditmarsch@irit.fr

[1] Department of Computer Science, Open University of the Netherlands, Valkenburgerweg 177, Heerlen 6419 AT, The Netherlands

[2] ICIS, Radboud University, Toernooiveld 200, Nijmegen 6525EC, The Netherlands

[3] Digital Innovation Lab at Just Right Customer Solution, JRCS, 1 Chome-2-14 Higashiyamatomachi, Shimonoseki 750-0066, Yamaguchi, Japan

[4] CNRS, IRIT, Université Paul Sabatier, 118 Route de Narbonne, Toulouse Cedex 9 31062, France

1 Introduction

Bayesian networks (BNs) [1] are popular tools for representation, reasoning, and learning with uncertainty in artificial intelligence (AI). However, while BNs provide a graph structure of the direct dependencies between random variables, they are in practice hard to reason with for domain experts. For example, two random variables that are unconditionally independent may become dependent if a third variable is observed (a process that is called *explaining away*). This makes the representation and reasoning with BNs sometimes counter-intuitive and the interpretation of the results difficult in practice. Explaining Bayesian networks has therefore been a topic in literature for quite some time (see e.g., [2] for early work).

With the General Data Protection Regulation in place stating that everyone has the right to know how their data is processed, explainable artificial intelligence (XAI) is getting more important. The European Commission for AI published ethics guidelines to gain trustworthiness [3].

Published online: 04 April 2024

However, these guidelines are formulated in imprecise language and lack explicit and clearly defined rights and guarantees. The AI Act proposed by the European Commission is building on the groundwork of the General Data Protection Regulation by determining more specific regulations [4]. The AI Act defines different levels of risk for AI systems, and imposes corresponding obligations and requirements. E.g. AI systems used in critical infrastructure, would have to undergo a conformity assessment. Given these regulations it is more important than ever to make XAI methods available to end users.

XAI algorithms can broadly be divided into two subcategories of explanations: (1) explanations that enhance the understanding of a decision or prediction of a model and (2) explanations that enhance the understanding of the model itself [5]. Research is currently being conducted on a relatively new type of explanation, called counterfactual explanations, that can help the user to understand the decision of a model. These explanations indicate which circumstance, represented by a random variable, could have been changed to obtain the desired outcome [6]. Research from the field of psychology shows that with counterfactuals one can ask 'what would have been' which may guide the user to future possibilities for change [7]. Furthermore, actionable counterfactuals, meaning counterfactuals which alternate a condition that is perceived as changeable, are preferred as explanations over those that alternate a condition that cannot be changed, such as the current age.

Counterfactual explanations make use of causal relationships and chains of causal relationships to identify how the outcome could be different. These causal chains may be intelligible to the user, as suggested by Lewis [8], and therefore provide a solid basis for explanations. Another advantage of counterfactuals is that they can help to show whether a machine learning algorithm is fair or unbiased. Designing fair classifiers is difficult, and counterfactuals can be used to evaluate whether the algorithm would give the same prediction regardless of an individual's age, race, sex, or other fairness attributes [9]. In this paper, we derive these counterfactuals and causal chains from causal Bayesian networks (CBNs), which are Bayesian networks where the associated graph constitutes a causal diagram thereby enabling causal reasoning [10].

Using the length of causal chains is a common approach in literature to measure the usefulness of explanations in a Bayesian network [11, 12], as it is inherent in a causal Bayesian network. This is supported through the work by Lewis [8] who suggests that shorter chains may be better explanations than longer causal chains, where a causal chain is the path of reasoning from one to another variable. Studies form psychology [7] suggest that, it is more useful to give an explanation that offers a future course of action than to explain with a direct cause that is not controllable.

Counterfactual explanations from causal Bayesian networks can be computed using Pearl's do-calculus [10]. However, this method does not provide information as to how valuable they are as an explanation. For example, someone could ask the question 'What had to be different to not get heartburn.' Possible counterfactuals could be: 'You wouldn't have heartburn if you had less stomach acid' or 'You wouldn't have heartburn if you ate a banana instead of fried chicken'. The second counterfactual seems to be a more useful explanation, because it is actionable but one cannot directly control stomach acid, even though both counterfactuals are true.

In this paper we evaluate how and whether the do-calculus can be applied taking into account insights gained from psychology and philosophy on the use of counterfactuals. In particular, we investigate the hypothesis that an actionable explanation, such as 'eating a banana to reduce heartburn', is perceived as a more useful explanation than a direct cause with a shorter causal chain, such as 'having less stomach acid would reduce heartburn'. Therefore, we used a questionnaire to test the usefulness of acquiring additional information from actionability to enhance the quality of explanations instead of using shorter chains, as this could be used to enhance existing methods for explaining Bayesian networks.

Previous research [13] has shown that BNs themselves are perceived as an understandable representation of a case or situation. Participants felt that they were able to understand the explanation that was given directly from the BN. The graph structure of a BN allows users to ask questions for further interpretations. In addition, subjects of the study perceived short additional explanation sentences as useful.

With that in mind, we developed the idea that short sentence explanations in form of counterfactuals could be perceived as understandable and useful for human interpreters structural data. Aligning to what was supported through literature and our previous work, the following hypotheses were formulated:

[**H0**] There is no difference in the perceived usefulness of counterfactuals with non-actionable short causal chains to actionable counterfactuals in explanations.

[**H1**] Actionable counterfactuals are perceived as more useful explanations than non-actionable shorter causal chains.

According to the hypotheses, short sentence explanations in the form of counterfactuals were formulated to explore the usefulness for supporting understandability for human participants. This method is particularly useful when a user wants to interact with the causal Bayesian network (CBN) to receive feedback, identify opportunities for improvement,

and answer open questions. For instance, you can use this method to determine how the CBN evaluates the quality of a therapy, what aspects can be improved, and how different scenarios would affect the outcome. The method allows for constructive and solution-oriented communication with the CBN based on facts and data.

This paper is structured as follows. In Sect. 3 we provide the background for causal BNs and computing counterfactuals. Section 2 gives a general overview of methods used explaining BNs. In Sect. 4, we explore how the insights from psychology can be applied to BNs. In Sect. 5 the methodology and the case study used to conduct our questionnaire are discussed. The case study is evaluated in Sect. 6. We discuss the applicability of the results in Sect. 7 and conclude the research in Sect. 8.

2 Background

In this chapter, first, we discuss XAI and its relevance for BNs, and, second, we introduce the philosophical notion and introduction to counterfactuals. Third, we summarise their effects on human behaviour as described in the field of psychology. Finally, we discuss the contribution of our work compared to the current state of the art.

2.1 Explainable Artificial Intelligence

It is not surprising that the development of well explained AI algorithms, or XAI methods, is closely linked to the development of AI methods. Already in the 1950 s, the beginning of AI research, there were approaches using symbolic logic, which offered some explanatory sentences. These systems used problem-solving algorithms that operated on data represented in the given formal language. [14]. For example, McCarthy described such an AI method in 1958 which has a transparent systems behaviour by design and could be understood by users [15].

With the increasing popularity of machine learning in the 1980 s, the paradigm of AI research changed. Instead of programming rules and facts, AI systems learned from data and adapted on their own. This led to major advances in areas such as image recognition and language processing. However, this was accompanied by a loss of explainability. AI systems became so called black boxes whose inner workings were no longer comprehensible [16]. This led to new challenges and risks for the application of AI systems. How to ensure that AI systems are fair, trustworthy, and accountable? How to detect and correct errors or biases? How to inform and empower users of AI systems? These questions motivated the development of XAI as a distinct research area.

XAI aims to develop methods and tools that improve or enable the explainability of AI systems. There are different approaches and dimensions of XAI, two of them are are divided into post-hoc and ante-hoc [17] algorithms. The goal of ante-hoc algorithms is to create models that are self-explanatory, their application therefore depends on the selected AI model and the underlying data. Post-hoc algorithms are applied after the model has been trained. Ante-hoc algorithms are incorporated into the model design and training process. Post-hoc algorithms are divided into model-specific and model-agnostic techniques. Model-specific techniques use the internal structure of the given algorithm, whilst model-agnostic techniques can be applied to any type of model. Post-hoc techniques can be further divided into global and local algorithms. Global techniques describe the overall behaviour of the model, while local techniques focus on a single prediction [17].

The dimensions of XAI also depend on the context and the addressee of the explanation. Depending on the objective, different aspects of the explanation may be important. For example: cause-and-effect relationships (Why?), alternatives (What if?), generalisability (What else?), or reliability (How sure?). Depending on the addressee, different formats of explanation may be appropriate such as, visual (graphics), verbal (text), or interactive (dialogue). XAI is a dynamic and interdisciplinary field of research. As AI continues to progress, XAI reflects the challenges of AI research. In this paper we want to examine explanations through alternatives (What if?) which are presented in text format. By providing short counterfactuals in response to questions, we address the format of interactive dialog as well.

2.2 XAI and Bayesian Networks

Research conducted between 1988 and 1999 aimed to explain Bayesian networks (BNs) by focusing on how BN models were generated, and how inferences could be made. This research has been summarised in [2]. More recent research has shifted its focus to the dynamic behavior of BNs, explaining posteriors and certain variables of interest. These methods are typically post-hoc, local, and model-specific.

In literature, several methods have been proposed for combining formal argumentation with Bayesian networks [12, 18–20]. These methods are based on building inference rules from variable-value pairs. Vreeswijk et al. [19] use a multi-agent system to determine if an inference rule supports a logical argument. Williams et al. [20] use argumentation theory to decide which arguments are justified for a particular patient in order to explain predictions of the Bayesian network. Timmer et al. [18] refine the approach from Williams et al., and in [12], they introduce a support graph. This support graph reduces the number of rules extracted from

the BN by only considering variables that are not conditionally independent, given the variable of interest, i.e., the variables which are in the Markov blanket of the variable of interest. Furthermore, Timmer et al. do not only show one argument to explain a variable of interest, but show arguments derived from different non-blocking paths in the network (so-called support chains). Therefore, the user can decide which argument is best for explaining the variable of interest.

There are also other approaches for explaining a variable of interest given evidence in Bayesian networks. Yap et al. [21] introduced a method that explains the variable of interest by capturing how variable interactions in a BN lead to inferences, independently of the evidence, just using variables needed to predict the behavior of the variable of interest. Vlek et al. [22] provide a text form report for different scenarios, consistent with the evidence, regarding a case in legal evidence. The report estimates the probability of each chosen scenario being likely, to present a global perspective on the case. In Kyrimi et al. [11], variables of interest are not explained by all variables, but only from variables having a significant impact on it. To achieve this, Kyrimi et al. compute the impact of the evidence and all variables in the Markov blanket of the variable of interest.

2.3 Philosophical Background of Counterfactuals

Counterfactuals have long been discussed in philosophy, for example, in the work of Lewis [8]. The sentence structure of a counterfactual consists of a false antecedent followed by a conclusion that is true in the form 'If A had been the case, then B would be the case', for example: 'If I hadn't eaten fried chicken I wouldn't have heartburn'. The conclusion can be stated in a negative or positive form. The truth condition of the conclusion of those counterfactuals is difficult to determine. Usually in logical reasoning an argument is constructed by using one or several premises to come to a conclusion, which is either true or false. The antecedent of a counterfactual however never happened but just could have been, which is hard to reason with. To cope with this logical clash, Lewis makes use of Carnap's ontology of possible worlds [23]. With this method, it is evaluated how far a possible world is away from the actual situation.

Lewis argued that two events can be causally related without being counterfactually dependent on each other, thus counterfactual dependence is not a requirement for causation [8]. For example 'If fried chicken had been sold out I would have eaten pizza buns instead and still have heartburn.' From either fried chicken or pizza buns I would have had heartburn, hence the pizza buns are the cause of my stomach ache but not counterfactually dependent on the result. Lewis used the possible world semantics to model this counterfactual dependence by determining the similarity of possible worlds. An event B is counterfactually dependent on A if and only if, if A would not occur B would not occur. Lewis later refined his definition as chains of counterfactual dependence where A is the cause of B if and only if there is a causal chain of counterfactual dependence leading from A to B.

According to Lewis, we must distinguish between causation and explanation. Causation is a dependency that exists without any subjective interpretation. An explanation depends on identifying a causal chain that is intelligible to the user [8]. If an apple falls from a tree, the cause is gravity, but the ripeness of the apple is also the cause. How useful one of these causes is as an explanation depends on each person, but still follows some general rules, which are discussed in Sect. 2.4. Lewis leaves it open for interpretation what intelligible implies. Thankfully, research has been done on this topic in the psychological field.

2.4 Relevance of Counterfactuals in Psychology

People use counterfactuals in their daily life to consider what might have been, in order to draw conclusions for future actions. They tend to design counterfactuals that add a new piece of information to the situation and allow new conclusions to be drawn. Several papers are discussed below that address the question of what heuristically constitutes a good counterfactual explanation. We follow the work of Byrne et al. [7], where literature is categorised that is relevant for XAI.

Counterfactuals can be created by either adding or deleting information from a set of evidence. Adding information is mostly used to determine how a result could have been better, and aids creative problem solving [24]. For example, we could argue: 'If I took supplements earlier I wouldn't have heartburn after eating the fried chicken.' Counterfactuals can be used to remove information as well. This leads us to our first example: 'If I hadn't eaten fried chicken I wouldn't have heartburn'. This subtractive form of reasoning is less often used than the additive form [25].

Another method to categorise counterfactuals is whether an outcome could have been better or worse. Thinking of a better outcome helps to change our behaviour in future, for example: 'If I had eaten half as much fried chicken, I would be feeling better now.' [25]. It gives us a solution for the future: 'Eat less fried chicken' [26]. However, these counterfactuals have the disadvantage of reinforcing negative feelings such as regret [27], whereas imagining a worse outcome helps us to feel better. People like to think how an outcome could have been better [28]. For example: 'If I would have eaten ice cream as well I would feel way worse'. They will use a counterfactual with a worse outcome, if there is less chance for future preventive action and want to deflect negative emotions, especially after large losses [29]. By appreciating what is still there, negative emotions do not tend to feel

so overwhelming, e.g., 'If they didn't take my legs I would be dead'. Hence, by considering the worse outcome we shift our focus to still being alive instead of the loss of our legs.

Rips and Edwards [30] have conducted studies that investigate which counterfactuals are more intelligible. In [30], people answered questions about simple machines of the form 'If component A had not operated/failed, would component B have operated?'. They discovered that people tend to do causal backtracking, which can be described as following an (allegedly) causal chain of events to its source. For example, given that A operating always causes B to operate, participants tended to answer the question 'If B didn't operate, did A operate?' with 'No' whilst answering 'If someone prevented B from operating, would A operate?' with 'Yes'. Hence in the former case, participants *causally backtracked*: they explained B not operating by its cause. They also discovered that the wording of counterfactuals is crucial. Using the word 'failed' instead of 'not operating' leads to different results [30]. Participants were more likely to believe that the other component may still function when the phrase 'not operating' was used instead of 'failed' even though the scenario described was equivalent.

2.5 Related Work

Several studies have been conducted on the usefulness of counterfactuals in general [24–30] (see Sect. 2.4). However, to our knowledge, there are no studies to date that evaluate the usefulness of counterfactuals for BN explanations based on user feedback. Additionally, to our knowledge, there are no papers that have evaluated which counterfactuals should be extracted from a BN from the user perspective or have developed methodologies for this. All papers published so far are choosing counterfactuals in an arbitrary way.

Keane [31] has recognised this problem as well for XAI-techniques based on classifiers (e.g. decision tree, k-NN, deep learning algorithms). They criticised that most methods do not guarantee that the generated counterfactuals are useful. Therefore Keane [31] did define what good counterfactuals are and suggested a novel approach for generating good counterfactuals. They defined good counterfactuals as follows: (*i*) counterfactuals have a different class label and are close in the feature space of the query; (*ii*) counterfactuals have no more than two feature differences with the query, and (*iii*) counterfactuals are valid data points in the domain and do not suggest impossible or unrealistic feature changes. In [31] no empirical results are reported on how useful their method would be for users. Instead, they acknowledged the need for more user testing to evaluate their notion of good counterfactuals and the effectiveness of their approach.

Miller [32] proposed an extension of structural causal models to add contrastive explanations for different types of questions, one of them being counterfactuals. They start with a definition of a *contrastive cause* and build a model of explanations with it. A contrastive cause for a counterfactual question is a pair of events that cause the fact and the foil respectively. The fact is the event that happened, and the foil is the event that did not happen but could have happened. The explanation is then given by the path in the structural causal model that leads from the actual cause and the foil to the contrastive cause. Miller discusses how the model can be applied to explainable artificial intelligence, where contrastive questions are often considered but contrastive explanations are rarely given. However, Miller did not investigate how useful such explanations are perceived by end-users or made any suggestions what counterfactuals to choose.

A study comparing some explanatory methods of BNs, even if they do not include counterfactuals, was conducted by Butz et al. [13]. The purpose of the study is to evaluate the user experience of four different explanation approaches for Bayesian network inference in the medical domain. They surveyed a group of participants on their perceived understanding of the explanations and found that Bayesian networks were easier to interpret than their associated XAI methods. They also found that working with scenarios and natural language sentences helped the understandability of Bayesian networks graph structure. Natural language can hinder the quick and precise comprehension of information in a written format, that is why it was suggested to use BN graphs compared with short sentences, which could also be provided via audio. In continuation of this work this paper aims to find first reference points for useful sentences to explain a BN's behaviour. Counterfactuals were chosen as first reference point because they are inherently interactive. To generate counterfactuals, a what-if question against the BN must be asked first. The approach in this paper defines and evaluates which counterfactual answer, out of several, is useful for the end-user.

3 Preliminaries

This section contains technical preliminaries on BNs including counterfactual computation from causal BNs.

3.1 Bayesian Networks

BNs are a type of probabilistic graphical models, which represent the probabilistic independence relationship in the form of a directed acyclic graph [33]. The nodes represent random variables and the arcs model the absence of probabilistic independences. The joint probability distribution of the graph is defined by the conditional probability of every node given their parents:

$$P(V) = \prod_{i=1}^{n} P(V_i \mid \text{pa}(V_i)) \tag{1}$$

where $V = \{V_1, \ldots, V_n\}$, and $\text{pa}(V_i)$ represent the parents of V_i in the graph. In the following, we will assume that each random variable is discrete.

Given evidence, the independencies between random variables can be read from the BN's graph structure by means of the d-separation criterion. E.g., considering the the structure $V_1 \rightarrow V_2 \rightarrow V_3$, without given evidence, V_1 and V_3 could be dependent, however, if there is evidence on V_2, then this 'blocks' the information between V_1 and V_3, meaning they are independent given V_2. Generally, two sets of nodes X and Y are d-separated by a set of nodes Z if every path from X to Y is blocked by Z [33].

Causal Bayesian networks are Bayesian networks where the arcs can be interpreted as a causal relationship, i.e., if $C \rightarrow E$ is included in the graph, then C is considered a cause of the effect E. The inclusion of this causal knowledge enables causal reasoning, as we will discuss in the next paragraph.

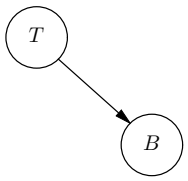
3.2 Computing Counterfactuals

One currently prominent approach to computing counterfactuals is based on Pearl's do-calculus [10]. This approach does not rest upon Lewis' approach of similarity between possible worlds, but is rather based on causal relationships between variables. One possible representation where such counterfactuals can be evaluated are causal Bayesian networks, where the arcs in the graph are interpreted as causal relationships. While conditional probabilities $p(y|x)$ are called observational since it focuses on situations where x is observed to be true, the do-calculus is interventional, and allows one to compute the post-intervention probability $p(y|do(X = x))$, indicating that X is actively set to the value x.

Given a causal model, the joint distribution after intervention can be evaluated by simply removing the conditional probability table of that variable from the factorisation, i.e., given a Bayesian network over variables $V = \{V_1, \ldots, V_n\}$, the interventional distribution is defined by:

$$P(V_1, \ldots, V_{i-1}, V_{i+1}, \ldots, V_n \mid \text{do}(V_i = v_i)) = \prod_{j \neq i} P(V_j \mid \text{pa}(V_j))$$

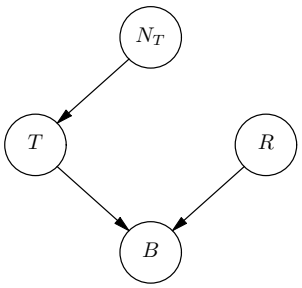
Example 3.1 ([34]) There exists a rather effective treatment for an eye disease. For 99% of all patients, the treatment works and the patient gets cured ($B = 0$); if untreated, these patients turn blind within a day ($B = 1$). For the remaining 1%, the treatment has the opposite effect and they turn blind ($B = 1$) within a day. If untreated, they regain normal vision ($B = 0$). Which category a patient belongs to is controlled by a rare condition that is unknown to the doctor, whose decision whether to administer the treatment ($T = 1$) is thus independent of this condition. The causal assumptions are represented by a causal Bayesian network



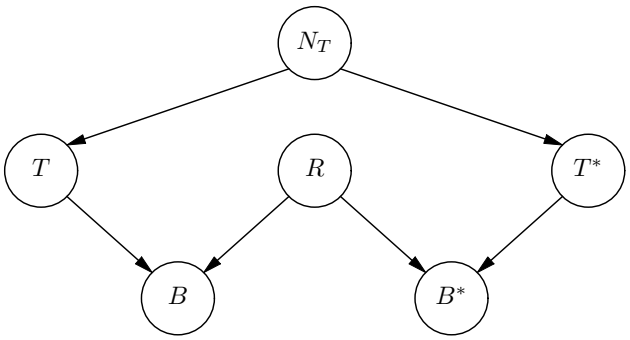
with $P(B = 0 \mid T = 1) = 0.99$ and $P(B = 0 \mid T = 0) = 0.01$. According to this graph it holds, for example, that $P(T \mid \text{do}(B = 1)) = P(T)$, i.e., making someone blind has no effect on the treatment.

Counterfactual questions can be stated in the phrase: what is the probability of y if x would have been true, given that we know u? To compute a counterfactual, we need to take into account both an observational aspect (u) and an interventional aspect, as the part 'if x would have been true' can be seen as a situation where an experimenter controls x. This can be formalised in a CBN by conditioning on u and intervening on variables in a counterfactual situation by constructing a so-called twin network [35]. Counterfactuals used in the questionnaire for this paper were computed in this manner.

Example 3.2 Reconsider Example 3.1. Now suppose we observe a patient that came to the hospital with poor eyesight, received treatment and went blind. A counterfactual question is: 'What would have happened if the doctor had not administered the treatment?'. To construct a twin network from this Bayesian network, the network has to be presented as a structural causal model [36] where noise variables are explicitly represented and each observable variable (B and T) is functionally dependent on other variables. In this case, we introduce a noise variable N_T to represent the probability distribution of the treatment. Blindness is determined by the treatment and an additional variable that represents the rare disease (R). The adapted graph is:



In the twin network construction we use two versions of the observed variables, i.e., B and T in the *actual* world, and B^* and T^* in an 'imaginary' (counterfactual) world. The noise variables link the actual with the counterfactual situation as follows:



In this graph, the counterfactual question can be formalised as:

$$P(B^* \mid T = 1, B = 1, \text{do}(T^* = 0))$$

In this case, it holds that $P(B^* = 0 \mid T = 1, B = 1, \text{do}(T^* = 0)) = 1$, i.e., in case the patient would not have received treatment, the patient would not have turned blind.

4 Applicability to Causal Bayesian Networks

Computing a useful counterfactual gets challenging when a user asks an open-ended question, for example: 'What could have gone differently to improve my situation?'. In this case, Pearl's method [10] can be used to generate several counterfactuals from a causal Bayesian network, satisfying the answer. But which of these counterfactuals offer a good explanation? In order to address this question, insights derived from the field of psychology have been summarised in Sect. 2.4. The following will demonstrate how these insights can be incorporated into the do-calculus. The equations will be used in Sect. 5 to create the questionnaire. We aim to identify some $Z_i = z_i'$ such that:

$$P(x \mid z) < P(x \mid z \setminus z_i, \text{do}(Z_i = z_i')) \tag{2}$$

where $z = \{Z_0 = z_0, \ldots, Z_n = z_n\}$, represents the observed evidence in the graph, i.e., another value of Z_i increases the chance of x. In case this holds for some Z_i, then we call $Z_i = z_i'$ *eligible*. The question remains which of these eligible counterfactuals is a useful explanatory answer. Byrne [7] distinguishes the content of counterfactuals between exceptions, controllability, actions, recent events and probability.

Probability

People tend to construct plausible counterfactuals, which can be consolidated with their knowledge about the world. For example, people rarely form counterfactuals which let people eat trees. This suggests identifying the eligible factor that has the counterfactually best and therefore most probable desired outcome, which can be formalised as follows:

$$\hat{z}_i = \text{argmax}_{Z_i = z_i'} P(x \mid z \setminus z_i, \text{do}(Z_i = z_i')) \tag{3}$$

Nevertheless people misjudge the likelihood of events, which is why they prefer less probable counterfactuals depending on their focus.

Exceptions

The exceptions category includes counterfactuals which are more likely to have occurred than the actual observed event. This can be formulated as an addition to the probability category in Bayesian networks, instead of using the most probable desired outcome as in Eq. 3. We say that some eligible z_i is the most *surprising* if for each eligible z_j with $z_i \neq z_j$ holds:

$$P(z_i | z \setminus z_i) < P(z_j | z \setminus z_j) \tag{4}$$

Controllability and Actions

The controllability category includes counterfactuals which change variables in peoples control. If tasks seem to be impossible, people tend to change events outside their control, whilst they tend to change events inside of the control of a protagonist for a better outcome. One example Byrne [7] gives is a quiz show in which one envelope contained a difficult math problem that no one could solve in the given time, and the other envelope contained money. The protagonists formed counterfactuals in which they had an easier math problem, while viewers formed the counterfactual, that they would have been better off taking the other envelope. Additionally, people tend to create counterfactuals that are actionable mainly if the previous event did not contain any action. Knowledge of what is actionable and what is not cannot be drawn from a BN. However, it is possible to hand label the variables with the information whether it is changeable for a person:

$$\hat{a}_i = \text{argmax}_{a_i'} P(x \mid z \setminus a_i, \text{do}(A_i = a_i')) \quad (5)$$

where $A \subseteq Z$ are actionable variables. The term actionability summarises the principle that people prefer to choose variables that they themselves have identified as variables they can act upon.

Recent Events

People like to change events with counterfactuals that just happened instead of events further in the past which is described in the recent event category. Modeling this in a BN is not trivial, because it cannot be realised by keeping additional labels for each variable, that indicate which event was observed in order to be able to roll them back until an eligible counterfactual is found. While this could be formalised in context of temporal Bayesian networks, we focus on a-temporal Bayesian networks in this paper.

5 Methodology and Case Study Design

To explore the explainability of counterfactuals in Bayesian networks, a case study was designed. In this case study, the main objective is to find a possible metric to measure the explainability through explicitly exploring shorter versus longer, and actionable versus non-actionable statements. Even though many BNs in the literature feature specific information from knowledge domains, e.g., medical BNs, the objective of this case study was to research the usefulness of actionability statements to enhance the perceived explainability of BNs within the general population without specific or expert knowledge. The questionnaire was sent out via multiple channels, enabling a heterogeneous population to participate, in order to minimise results based on specific knowledge. The method of creating a questionnaire was chosen to include a wide range of international participants, not focusing on a single demographic. An online-based questionnaire enables world wide participation, with the only recorded limits being access to internet and a suitable device (mobile, tablet, or desktop) and the ability to read and understand the questionnaire in English.

To test our hypotheses, mentioned in Sect. 1, we created three scenarios based on three different causal Bayesian networks (CBNs).

The first CBN is a small network about heartburn as shown in Fig. 1. The scenario presented in the CBN is about a stressed person who has problems with heartburn and wants to know what they can change in their life to get rid of the heartburn.

The second network, shown in Fig. 2, is a medium-sized network about a car accident [37]. In this scenario, a person who recently had a minor accident is wondering what they could do differently to prevent further accidents, is described.

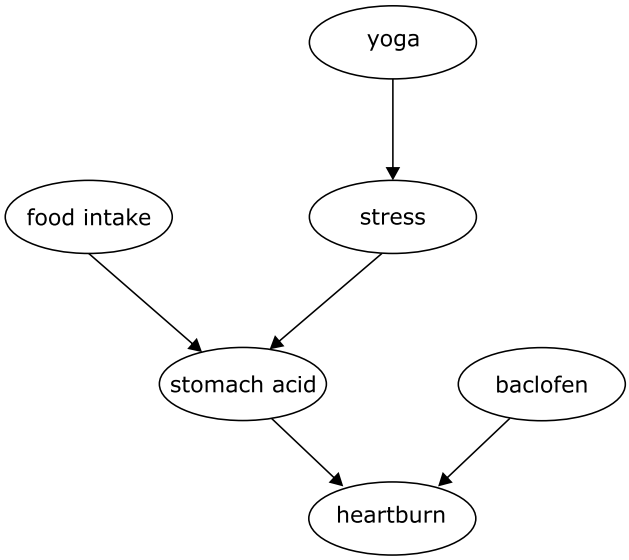


Fig. 1 The BN about heartburn used in the questionnaire

The last CBN is a small-sized network about getting a loan, presented in Fig. 3. The person in this scenario wants to know what they could change to raise their chances of getting a loan. The heartburn CBN and the loan CBN were specifically designed for this study, so that there were more possibilities in which the shortest chain is not at the same time the most actionable variant. In order to compute all possible actionable variables, Eq. 5 was used which was defined in Sect. 4. Counterfactuals with a short causal chain were respectively calculated with the do-calculus described in Sect. 3.2. The following describes how the computed counterfactuals were juxtaposed in the questionnaire.

In the questionnaire, we presented the participants with two possible counterfactual answers and asked them to choose the one that was more useful for them in the given scenario. One of the counterfactual answers contained the direct cause, which is a parent node of our variable of interest. For instance, if the variable of interest is an *accident*, one parent node could be *antilock*, and another could be driving quality (Fig. 2). The second answer contained the variable that we considered to be more actionable but with a longer causal chain. As described in the background section (Sect. 2.2), most methods assume that a direct cause is a good explanation. There were few possible choices in constructing the questions because there are not many counterfactuals in the combination that guarantee a better outcome and also fulfill the criteria of a direct parent node or a longer causal chain. As a first effort, we selected arbitrary actionable variables with chains of different lengths to determine whether they would be preferred in some cases.

The participants had to choose which of the explanations seemed more useful to them. The participants were asked

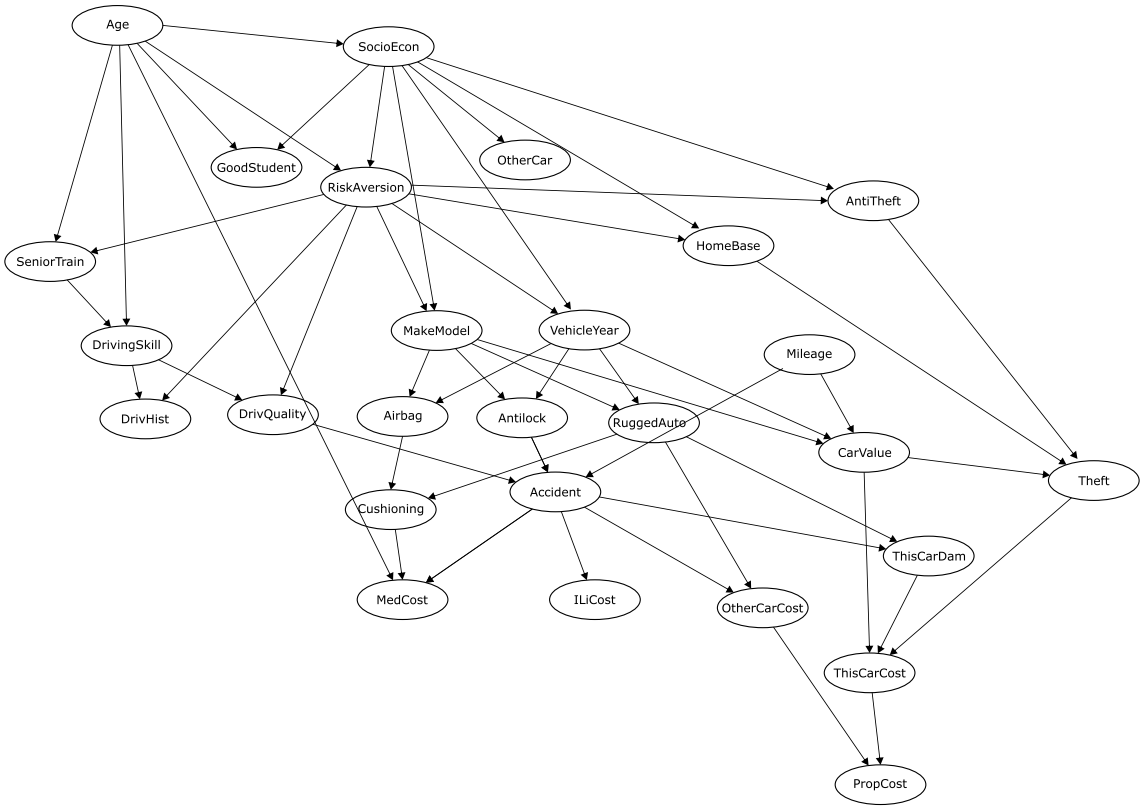


Fig. 2 The BN about car accidents used in the questionnaire

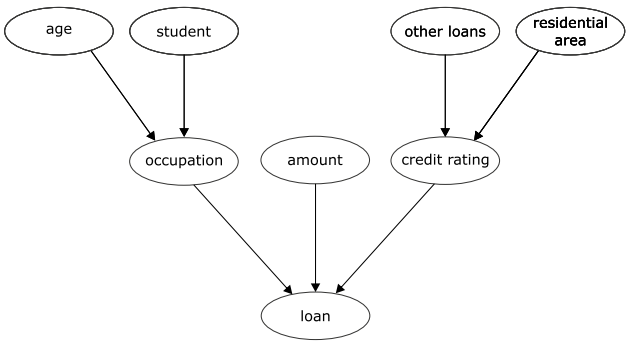


Fig. 3 The BN about loans used in the questionnaire

about their own perceived usefulness of the explanations, because each person perceives an explanation as of varying usefulness. We want to explain to the end user, so we did not choose a scenario to be evaluated as an outsider, because studies have already shown that the evaluation is different if the participant is not affected [7].

We targeted the same question from different perspectives by altering the answers to different pairs of chain lengths. We used the same sentence structure to present the problem to avoid bias by altering the sentences so that one problem description could not be favoured or interpreted differently from the other based on the sentence formulation alone. For example, the question for the first scenario was 'You would not have heartburn if...' and the first pair of counterfactuals was: 'You had less stomach acid' for the shorter chain and 'You ate a banana' for the actionable variable but longer chain. The second pair of counterfactuals to the question was: 'You had taken tablets (baclofen)' for the shorter chain and 'You did yoga' for the presumably actionable variable but longer chain.

At the end of the questionnaire, we asked the participants to rate the variables from easy to change to hardest to change. In addition, we asked which variables are not actionable at all for them. Since actionability might differ to some extent between persons, we were able to measure if the participants selected the variable that is more actionable for them. Note that this means that participants may evaluate the shorter chains as more actionable, for which we correct in the statistical analyses. We calculated probabilities for choosing actionable and shorter-chain explanations by means of a χ^2-test. Finally, we tested whether there is a significant tendency to either choose shorter chains or actionable variables, than what might be expected by chance. As a baseline of comparison, we also consider the probability that the participants chose the most probable counterfactual according to the Bayesian network.

Three different CBNs were chosen, so that a single topic would not have a strong influence on the results. For example, it is possible that people in the health context would prefer actionable answers, while people in the loan context would prefer to have causal answers. We further decided on not showing the CBNs to the participants, because we focused on the question which counterfactual is perceived as a more useful explanation for a question about alternative (counterfactual) situations, and not on how to explain a CBN with it. The participants had no information about the complexity or architecture of the network. Therefore this information was not reflected in the results either. The topics of the CBNs were general because we wanted to ask a heterogeneous selection of people.

6 Evaluation

Fifty-four people participated in the questionnaire. They were acquired by social media posts and circular e-mails at the Open University and at a company focusing on IT solutions. The questionnaire was accessible online. Five questionnaires were inconsistent: they listed variables as not actionable at all, but in their rating the variable was listed as the easiest or one of the easiest variables to change, that is, as actionable. They were therefore excluded. One of the questionnaires listed three variables as not actionable at all but the easiest to change in the rating. This led us to the conclusion that it was intentionally filled out incorrectly, which is why we also decided to excluded it. Four other questionnaires had only one inconsistent variable, which was suspected to be a mistake, so we decided to include them in the analysis. We used a χ^2-test for our analysis.

We asked 13 questions in total, excluding the rating questions. The participants answered five questions in the heartburn scenario four in accident and four in the loan scenario. With 49 valid questionnaires we got a total of 402 answers that preferred a more actionable explanation in contrast to 235 answers preferring the less actionable alternative. In the total of 637 answers, the participants shared our notion about what is more actionable 396 times. An overview of the total answers for each scenario is shown in Fig. 4.

Overall, 64% of the actionable explanations were preferred over less actionable explanations ($p < 10^{-5}$). This is consistent in all three scenarios: in the heartburn scenario 70% preferred the actionable explanations ($p < 10^{-5}$), in the accident scenario 64% preferred the actionable explanations ($p < 10^{-3}$), and in the loan scenario 57% preferred the actionable explanations, though this last one did not reach statistical significance ($p = 0.06$). Not all participants preferred actionable explanations, but 76% of the participants chose more actionable than less actionable explanations throughout the scenarios ($p < 10^{-3}$).

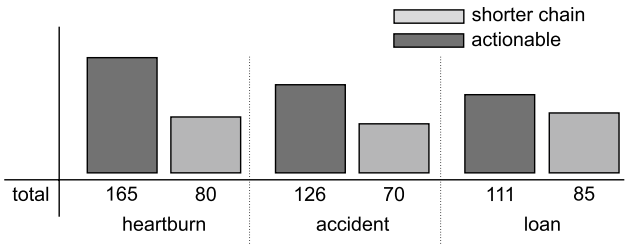


Fig. 4 Overview of all answers split up by scenario

Similarly, 64% of the most probable explanations were preferred over less probable explanations ($p < 10^{-5}$). In this case, the results are not consistent in all scenarios: in the heartburn scenario, the more probable counterfactual was preferred less (44%), whereas in the accident scenario (66%) and loan scenario (87%) the more probable counterfactuals were chosen. Overall, 90% of the participants chose more probable than less probable explanations ($p < 10^{-5}$), though this is primarily explained by the loan scenario.

We did not find a similar trend in the length of the chain: overall items shorter and longer items were chosen equally probable. Similarly, we found no statistical difference between the number of participants that preferred longer or shorter chains more often. Remarkable, in the accident case study the longer-chain explanations were preferred ($p < 10^{-3}$) and in the loan scenario the shorter-chain explanations were preferred ($p < 10^{-6}$). This might indicate that this is highly dependent on the type of application or Bayesian network used.

To test the main hypothesis, we compared whether explanations with actionable long-chain explanations were more likely to be chosen than non-actionable short-chain explanations. We found that this was the case in 60% of the time ($p < 10^{-5}$), which indicates that actionable variables tend to be perceived as a more useful explanations than shorter chains. We did not find an overall difference in a choice between actionable and more probable explanations, though the latter were not consistently preferred across different scenarios.

Another effect that emerged here, is that the interpretation of what is actionable was in many cases not according to the expectations in the design of the questionnaire, i.e., participants rated the variable with the shorter chain as actionable more often than expected. Recall that questions were designed in such a way that answers with short chains were expected to be less actionable. In 73 of the 111 answers for the loan scenario the more actionable counterfactual also had a shorter chain. This was unexpected.

7 Discussion

While actionability is one aspect of plausible counterfactuals (Eq. 5), we formalised several other aspects in Sect. 4, in particular *the most probable desired outcome* (Eq. 3) and *the most surprising outcome* (Eq. 4). The case study presented in this paper was limited to comparing causal chains and actionability given three scenarios. As a baseline, we also considered the most probable explanation to put the results of the actionability and length of causal chains in perspective.

The implication of this limitation is that aspects such as actionability and length of causal chain could not be correlated to other aspects of useful explanations, as discussed in Sect. 4. In particular, the probability of the desired outcome (cf. Equation 3) might provide an alternative explanation of the results. To this end, we computed the probability of participants favouring the more probable counterfactual over the other. This posthoc-analysis shows that this is the case in 66% of the time in the accident and 87% in loan case study. In the heartburn scenario however, the participants choose the more probable counterfactual over the other in only 44% of the times. Moreover, the distances between the probabilities are very small in most cases, e.g., most counterfactuals (five out of seven) for the loan example range between a $43-45\%$ in the probability of a loan being approved.

While we did not find evidence that participants selected the *more* probable counterfactual, we also investigated whether there was a tendency to select the *most* probable one. For the heartburn scenario, this counterfactual was not part of the survey, so the tendency to select the most probable counterfactual could not be analysed. However, the loan scenario offers an interesting result. The most probable desired outcome is 'Your chance getting a loan would be higher if the amount of borrowed money would be lower.' In the scenario, this option was compared to 'Your chance getting a loan would be higher if you were an adult instead of a young adult.' Overall 80% of all people chose lowering the amount of borrowed money over changing the age. However, most participants considered reducing the amount of borrowed money the *most actionable*, whereas becoming an adult was considered the *least actionable*, so the results are consistent with the preference of choosing actionable variables. In the accident scenario the most probable counterfactual is 'Your chance of having an accident would be less severe if your car had antilock'. This counterfactual however had mixed effects in the case study. Only 10% picked a car with antilock over better driving skills, but 64% preferred a car with antilock over a new car.

In summary, while we cannot rule out confounding effects because of limited data, the current results do not indicate that the probability of the desired outcome is a more important factor than the actionability of variables.

8 Conclusion

In this paper, we explored the selection of a useful counterfactual explanation, derived from a BN, when there are multiple to choose from. We have examined this question from several angles. On the one hand, we defined how findings from psychology can be applied to the do-calculus. One example of this application is that humans typically respond better to actionable explanations that lead to positive outcomes. On the other hand, we compared causal chains with actionability in terms of perceived usefulness. The graph structure is mostly used in XAI for BNs, but the results of our questionnaire indicated that actionable variables are preferred while the preferred length of the chain depended on the scenario.

Additionally, we established the more probable counterfactual as the baseline. This evaluation indicated a preferred selection of the more probable counterfactual. However, this preference was not given for every individual scenario. In summary, our finding suggest that actionable counterfactuals appear to be a more robust finding than using the graph structure or the probabilities of counterfactuals.

The implication of this study shows that the approach of most methods that explain Bayesian networks may be sub-optimal, as they rely on the graph structure of the network and use, for example, the Markov blanket of a variable of interest to limit possible variables for an explanation. However, our results suggest that this is most likely not the best method to explain Bayesian networks, because longer-chain actionable variables are generally outside the Markov blanket and would not have been considered. Therefore, this findings have a direct impact on the construction of Bayesian network explanations and can be used to improve these methods.

We acknowledge that our CBNs are relatively small and focus on three specific domains. The consistent results with respect to the usefulness of actionability suggest that these generalise to other domains. On the other hand, we expect that there are differences in the perceived usefulness of shorter versus longer chains depending on domains or potentially the size of the overall BNs, because we observed that preferences differed significantly between the scenarios. This study is restricted to few scenarios. Further research on different scenarios should investigate the circumstances under which shorter chains are preferred over longer chains, even if the latter potentially represent a more actionable variable. By doing so, we can gain a more comprehensive understanding of the factors that influence the effectiveness of chains in different contexts. Similarly

it should be further explored if actionable counterfactuals are superior to the most probable counterfactual.

Another aspect that we would like to investigate further is how to automate the generation of the most useful counterfactual. The results of this paper suggest that we need to label variables according to the extent to which they are actionable or impossible to change. However, a causal Bayesian network provides information about causal relations, not about actionability. It is an open question what the most appropriate and efficient manner is to add knowledge about actionability. One possible approach is to ask the user, which can provide the most useful actionable variables. However, that approach is time consuming.

We suggest that there are further opportunities to investigate the generalisability of the results of this paper. The setting in which we studied actionability compared to causal chains is limited because we focused on a special type of question that can be answered with counterfactuals. Instead of focusing on counterfactuals, another type of explanations that could be useful to investigate are contrastive explanations.

Data availability The authors of this paper are not registered to the HCIN Editorial System as editor or reviewer. Survey data were stored in CVS format. All data is available on request from the corresponding author.

Declarations

Financial interest or Non-financial interest The authors did not receive support from any organisation for the submitted work. All authors certify that they have no affiliations with or involvement in any organisation or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. RB wrote the majority of the manuscript and performed the survey and data analysis. AH contributed to the data analysis. AH, RS and HvD gave feedback and suggestions for the manuscript. All authors contributed to the design of the study and take full responsibility for and have read and approved this final version of this manuscript.

Conflict of interest There are no competing interests.

References

1. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann. 1988. https://doi.org/10.1016/C2009-0-27609-4.
2. Lacave C, Díez FJ. A review of explanation methods for Bayesian networks. Knowl Eng Rev. 2002;17(2):107–27. https://doi.org/10.1017/S026988890200019X.
3. European Commission. White Paper on Artificial Intelligence: a European approach to excellence and trust. European Union. 2020.
4. European Commission. Proposal for a Regulation laying down harmonised rules on artificial intelligence. European Union. 2021.
5. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
6. Wachter S, Mittelstadt B, Russell C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. 2018. https://doi.org/10.2139/ssrn.3063289.
7. Byrne RMJ. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In: Proc. of 28th IJCAI, 2019;pp. 6276–6282. https://doi.org/10.24963/ijcai.2019/876
8. Lewis D. Counterfactuals Blackwell. 1973. https://doi.org/10.2307/2273738.
9. Russell C, Kusner MJ, Loftus J, Silva R. When worlds collide: Integrating different counterfactual assumptions in fairness. In: Advances in Neural Information Processing Systems. 2017;vol. 30.
10. Pearl J. Causality Cambridge University Press. 2009. https://doi.org/10.1017/CBO9780511803161.
11. Kyrimi E, Mossadegh S, Tai N, Marsh W. An incremental explanation of inference in Bayesian networks for increasing model trustworthiness and supporting clinical decision making. Artif Intell Med. 2020. https://doi.org/10.1016/j.artmed.2020.101812.
12. Timmer ST, Meyer J-JC, Prakken H, Renooij S, Verheij B. A two-phase method for extracting explanatory arguments from Bayesian networks. International Journal of Approximate Reasoning. 2017;80:475–494. https://doi.org/10.1016/j.ijar.2016.09.002
13. Butz R, Schulz R, Hommersom A, van Eekelen M. Investigating the understandability of XAI methods for enhanced user experience: When Bayesian network users became detectives. Artificial Intelligence in Medicine 2022;**134**. https://doi.org/10.1016/j.artmed.2022.102438
14. Schwalbe G, Finzel B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery. 2023.
15. McCarthy J. Programs with common sense. In: Proceedings of the Teddington Conference on the Mechanisation of Thought Processes. 1958;pp. 77–84.
16. Saeed W, Omlin C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems. 2023;263:110273. https://doi.org/10.1016/j.knosys.2023.110273
17. Speith T. A review of taxonomies of explainable artificial intelligence (XAI) methods. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT '22, 2022;pp. 2239–2250. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3531146.3534639 .
18. Timmer S, Meyer J, Prakken H, Renooij S, Verheij B. Inference and attack in Bayesian networks. In: Proc. of 25th BNAIC, 2013;pp. 199–206.

19. Vreeswijk GAW. Argumentation in Bayesian belief networks. In: Proc. of ArgMAS, 2005;pp. 111–129. https://doi.org/10.1007/978-3-540-32261-0_8
20. Williams M, Williamson J. Combining argumentation and Bayesian nets for breast cancer prognosis. JoLLI. 2006;15(1):155–78. https://doi.org/10.1007/s10849-005-9010-x.
21. Yap G-E, Tan A-H, Pang H-H. Explaining inferences in Bayesian networks. Applied Intelligence. 2008;29(3):263–78. https://doi.org/10.1172/JCI125014.
22. Vlek CS, Prakken H, Renooij S, Verheij B. A method for explaining Bayesian networks for legal evidence with scenarios. Artificial Intelligence and Law. 2016;24(3):285–324. https://doi.org/10.1007/s10506-016-9183-4.
23. Carnap R. Meaning and Necessity. University of Chicago Press. 1947.
24. Markman K, Lindberg M, Kray L, Galinsky A. Implications of counterfactual structure for creative generation and analytical problem solving. Personality & Social Psychology Bulletin. 2007;33:312–24 https://doi.org/10.1177/0146167206296106
25. Epstude K, Roese NJ. The functional theory of counterfactual thinking. Personality & Social Psychology. 2008;12(2):168–92. https://doi.org/10.1177/1088868308316091.
26. Smallman R, McCulloch K. Learning from yesterday's mistakes to fix tomorrow's problems: when functional counterfactual thinking and psychological distance collide. European Journal of Social Psychology. 2012;42(3):383–90. https://doi.org/10.1002/ejsp.1858.
27. Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. In: Utility, Probability, and Human Decision Making. 1975;pp. 141–162. https://doi.org/10.1126/science.185.4157.1124
28. Rim S, Summerville A. How far to the road not taken? The effect of psychological distance on counterfactual direction. Personality & Social Psychology Bulletin. 2013;**40**. https://doi.org/10.1177/0146167213513304
29. Beike DR, Markman KD, Karadogan F. What we regret most are lost opportunities: A theory of regret intensity. Personality & Social Psychology Bulletin. 2009;35(3):385–97. https://doi.org/10.1177/0146167208328329.
30. Rips L, Edwards B. Inference and explanation in counterfactual reasoning. Cognitive science 2013;**37**. https://doi.org/10.1111/cogs.12024
31. Keane MT, Smyth B. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI(XAI). In: Watson I, Weber R, editors. Case-Based Reasoning Research and Development. Cham: Springer; 2020. p. 163–78.
32. Miller T. Contrastive explanation: a structural-model approach. The Knowledge Engineering Review 2021;**36**. https://doi.org/10.1017/s0269888921000102
33. Korb KB, Nicholson AE. Bayesian Artificial Intelligence CRC Press. 2010. https://doi.org/10.1201/b10391.
34. Peters J, Janzing D, Schölkopf B. Elements of causal inference: foundations and learning algorithms. 2017.
35. Balke A, Pearl J. Probabilistic evaluation of counterfactual queries. In: Probabilistic and Causal Inference: The Works of Judea Pearl. 2022;pp. 237–254. https://doi.org/10.1145/3501714
36. Lauritzen SL. Graphical models. 1996;**17**.
37. Binder J, Koller D, Russell S, Kanazawa K. Adaptive probabilistic networks with hidden variables. Machine Learning. 1997;29(2):213–44. https://doi.org/10.1023/A:1007421730016.