One counterfactual does not make an explanation

Raphaela Butz[1], Arjen Hommersom[1,2], Marco Barenkamp[3], and Hans van Ditmarsch[1]

[1] Department of Computer Science, Open University of the Netherlands
[2] ICIS, Radboud University, The Netherlands
[3] European Polytechnical University, Bulgaria

Abstract. 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 what type of counterfactuals 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 showed that actionable counterfactuals are preferred regardless of being the direct cause or having a longer causal chain.

1 Introduction

Bayesian networks (BNs) [14] are popular tools for representation, reasoning, and learning with uncertainty in 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 BN 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. [11] for early work).

With the General Data Protection Regulation (GDPR) 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 [8]. However, these guidelines are formulated in imprecise language and lack explicit and clearly defined rights and guarantees. This leaves the question open to interpretation what a well-explained algorithm is, and how to implement these guidelines.

XAI algorithms can broadly be divided into two subcategories of explanations: explanations that enhance the understanding of a decision or prediction of a model and explanations that enhance the understanding of the model itself [2]. 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 be changed to obtain the desired outcome [25].

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 the philosopher Lewis [12], and therefore provide a solid basis for explanations. Furthermore, 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 [5]. Another advantage of counterfactuals is, that they can help to show whether a machine learning algorithm is fair or unbiased. Designing fair classifiers is sometimes difficult, and counterfactuals can determine whether the algorithm would give the same prediction if an individual person had a different age, race, sex, or other fairness attributes [18].

Counterfactuals from causal Bayesian networks can be computed using Pearl's do-calculus [15]. 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 not to 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 since one cannot directly control stomach acid, even though both counterfactuals are true. In this paper, we investigate the hypothesis that actionable variables, such as eating a banana, is perceived as a more useful explanation than a direct cause with a shorter causal chain, such as excess stomach acid.

In the next section we discuss the background of this work more broadly and introduce preliminaries for causal Bayesian networks and on computing counterfactuals. Section 3 gives a general overview of methods used explaining Bayesian networks. In Section 4 we describe our case study, which is evaluated in Section 5.

2 Preliminaries

The philosophical concept of counterfactuals is discussed first. Then we summarise the effects of counterfactuals described in psychology and we focus on those which have a positive effect on further actions. Finally, we discuss the computation of counterfactuals from a causal Bayesian network.

2.1 Philosophical background

Counterfactuals have long been discussed in philosophy, for example, in the work of Lewis [12]. 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 [6]. 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 [12]. For example 'If fried chicken had been sold out I would have eaten pizza buns instead and gotten heartburn.' Independent on 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 [12]. 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 Subsection 2.2. Lewis leaves it open for interpretation what intelligible implies. Thankfully, research has been done on this topic in the psychological field.

2.2 Relevance 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. [5], where literature is categorised that is relevant in context of 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 [13]. 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 [7].

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.' [7]. It gives us a solution for the future: 'Eat less fried chicken' [19]. However, these counterfactuals have the disadvantage of reinforcing negative feelings such as regret [22], whereas imagining a worse outcome helps us to feel better. People like to think how an outcome could have been better [16]. 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 [3]. 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 [17] have conducted studies that investigate which counterfactuals are more intelligible. In [17], 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 outcomes [17]. Their results show that people stated that the other component might still operate more often when the question included the phrase 'not operating' instead of 'failed'.

2.3 Computing counterfactuals

One currently prominent approach to computing counterfactuals is based on Pearl's do-calculus [15]. This approach does not rest upon Lewis' approach of similarity between possible worlds, but rather is based on causal relationships between variables. One possible representation where such counterfactuals can be evaluated are causal Bayesian networks (CBNs), 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.

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 using the do-calculus [1]. Counterfactuals used in the questionnaire for this paper were computed in this manner.

3 Related Work

Several approaches for explaining Bayesian networks have been proposed that framed the problem in terms of argumentation theory. The approach of Vreeswijk et al. [24] uses a multi-agent system to decide if an inference rule is supporting a logical argument. Williams et al. [26] use argumentation theory to decide which arguments are justified for a particular patient in order to explain predictions of the Bayesian network. In Timmer et al. [20] the approach from Williams et al. was further refined.

There are many other approaches that aim to explain a prediction about a variable of interest. For example, in [21], a so-called support graph is introduced. This support graph reduces the number of rules extracted from the BN by only considering variables that are not conditionally independent on the variable of interest. Yap et al. [27] introduced a method to explain 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 behaviour of the variable of interest. Vlek et al. [23] 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, to present a global perspective on the case. In Kyrimi et al. [10], variables of interest are not explained by all variables, but only from variables having a significant impact on them. To achieve this, the method by Kyrimi et al. takes into account the impact of the evidence and all variables in the Markov blanket [9] of the variable of interest.

In many of these methods the possible variables used in an explanation are reduced by using the graph structure of the Bayesian network, for example only using the variables in the Markov blanket. In this paper, we investigate as well whether this is the most adept approach, since actionable variables can be located outside the Markov blanket [10, 21, 27].

4 Case Study

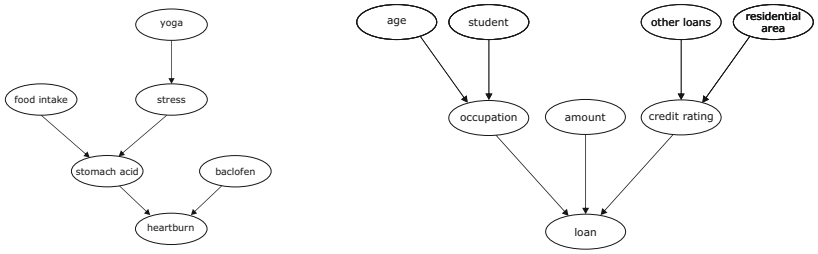
Pearl's method [15] can be used to generate counterfactuals from a causal Bayesian network. But the question remains, which of these counterfactuals offer a good explanation?

As Lewis stated, a counterfactual needs to be intelligible for the user, to be a good explanation [12]. He suggested 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. More recent work [5] suggests that in some situations 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. For example the causal chain from stomach acid to heartburn is shorter than from doing yoga to

heartburn as seen in Figure 1a. We believe that the explanation of doing yoga to reduce heartburn would be more useful than explaining it with too much stomach acid. With these two two concepts we formulated the hypotheses:

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

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



(a) heartburn BN (b) loan BN

Fig. 1: BNs used in the questionnaire

To test our hypothesis, we created three scenarios based on three different CBNs. We used three causal Bayesian networks in our questionnaire. The first CBN is a small network about heartburn as shown in Figure 1a. For this network, we created a scenario with a stressed person who has problems with heartburn and wants to know what they can change to get rid of it. The second network is a medium-sized network about having an accident with your car [4], shown in Figure 2. In this scenario, we designed a person who recently had a minor accident and is wondering what they could do differently to prevent further accidents from happening. The last CBN is a small-sized network about getting a loan. It is shown in Figure 1b. 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, in such a manner that there were more possibilities in which the shortest chain is not at the same time the most actionable variant. We computed all possible counterfactuals for all three scenarios as described in Section 2.3. How we selected the counterfactuals used in the questionnaire is described in the following.

In the questionnaire, we gave two possible counterfactual answers to each question. One that contains the direct cause which is a parent node of our variable of interest. For example, if *accident* is the variable of interest, *antilock* is one parent, another would be *driving quality*. The second answer contained the variable we thought was more actionable but with a longer causal chain. The participants had to choose which of the explanations seemed more useful to them. We asked one question several times but altered the two answers with different

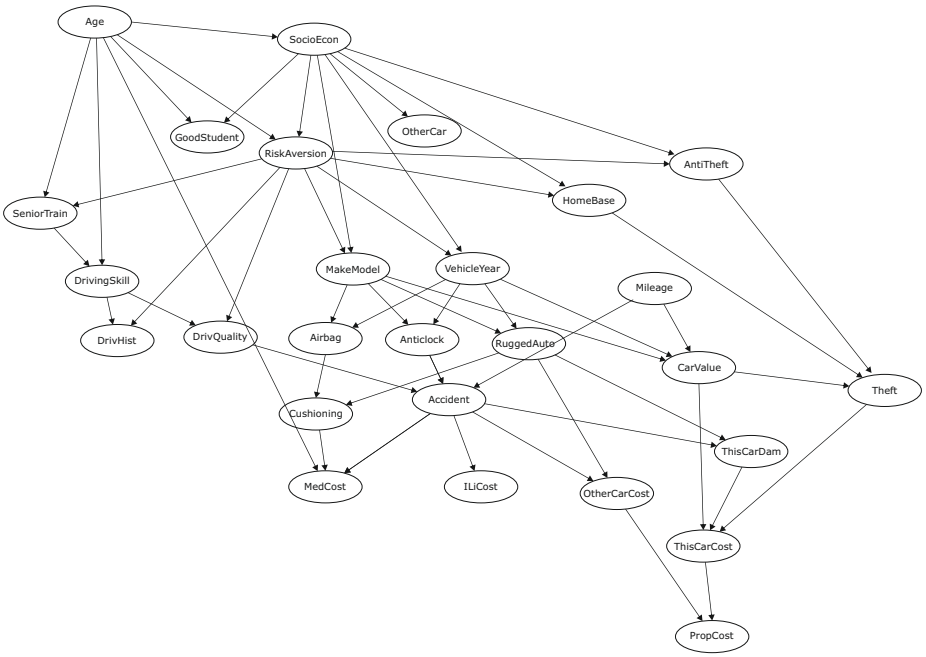


Fig. 2: The BN about car accidents used in the questionnaire.

pairs of shorter chains and actionable variables. 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 same question was: 'You had taken tablets (baclofen)' for the shorter chain and 'You did yoga' for the actionable variable but longer chain.

At the end of the survey we asked the participants to rate the variables from easiest 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.

Three different CBNs were chosen, such 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 counter-

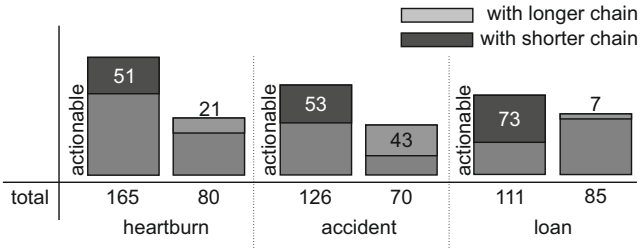


Fig. 3: Overview of all answers split up by scenario. For example, in the heartburn scenario, in a total of 80 answers the less actionable variable was chosen, of which 21 had a longer chain.

factual 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.

5 Evaluation

Fifty-four people participated in the survey, and were acquired by social media posts and circular e-mails at the Open University and one company focusing on IT solutions. The questionnaire was accessible online. Five questionnaires were completely 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. 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 decided to excluded it. Three other questionnaires had only one inconsistent variable, which was suspected to be a mistake, so we decided to include them in the 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 Figure 3.

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 ($p < 10^{-3}$).

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.

Another effect that emerged here, which can be seen in Figure 3, 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. However, for example, in 111 answers people chose the more actionable counterfactual for the loan scenario, of which 73 unexpectedly had a shorter chain.

6 Conclusion

In this paper, we compared two concepts causal chains and actionability. Causal chains are introduced as an important aspect of counterfactual explanations in philosophy. Actionability is a key factor in suitable counterfactual explanations in psychology. We measured which of the two concepts seemed more useful to participants. The results of our questionnaire indicated that actionable variables are preferred over shorter causal chains.

Most methods that explain Bayesian networks 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 all longer-chain actionable variables were outside the Markov blanket and would not have been considered.

There are opportunities to investigate further how well the results of this paper generalise. The setting in which we studied actionability compared to causal chains is limited. We focused on a special type of question that enquires about reasons for particular questions that can be answered with counterfactuals. Another question type are for example contrastive explanations where a common question is 'why t instead of t'?'. We believe that our results should also hold in other types of explanations, i.e., that these type of questions are more useful when considering actionable variables. In addition, our CBNs are relatively small

and focus on three different domains. We believe that the results would not be fundamentally different in other domains when it comes to the usefulness of actionability, we expect that there are differences in the perceived usefulness of shorter versus longer chains: we observed that preferences differed significantly between the two case studies. It should be investigated further in which cases shorter chains are preferred and in which cases longer chains.

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 require a labelling of the variables according to the extent they are actionable or impossible to change. However, a causal Bayesian network provides information about causal relations but not about actionability. The most appropriate and efficient manner to elicit knowledge about actionability from the user in other to provide the most useful explanation is an open question.

Acknowledgements

We would like to thank reviewers for their suggestions. They have significantly improved the paper.

References

1. Balke, A., Pearl, J.: Probabilistic evaluation of counterfactual queries. In: Probabilistic and Causal Inference: The Works of Judea Pearl, pp. 237–254 (2022)
2. 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 58, 82–115 (2020)
3. Beike, D.R., Markman, K.D., Karadogan, F.: What we regret most are lost opportunities: A theory of regret intensity. Personality & Social Psychology Bulletin 35(3), 385–397 (2009)
4. Binder, J., Koller, D., Russell, S., Kanazawa, K.: Adaptive probabilistic networks with hidden variables. Machine Learning 29(2), 213–244 (1997)
5. Byrne, R.M.J.: Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. pp. 6276–6282 (2019)
6. Carnap, R.: Meaning and Necessity. University of Chicago Press (1947)
7. Epstude, K., Roese, N.J.: The functional theory of counterfactual thinking. Personality & Social Psychology 12(2), 168–192 (2008)
8. European Commission: White paper on artificial intelligence: a European approach to excellence and trust (2020)
9. Korb, K.B., Nicholson, A.E.: Bayesian Artificial Intelligence, Second Edition. CRC Press, Inc., 2nd edn. (2010)
10. Kyrimi, E., Marsh, W.: A progressive explanation of inference in 'hybrid' Bayesian networks for supporting clinical decision making. Proceedings of the Eighth International Conference on Probabilistic Graphical Models (2016)
11. Lacave, C., Díez, F.J.: A review of explanation methods for Bayesian networks. Knowl. Eng. Rev. 17(2), 107–127 (2002)

12. Lewis, D.: Counterfactuals. Blackwell (1973)
13. Markman, K., Lindberg, M., Kray, L., Galinsky, A.: Implications of counterfactual structure for creative generation and analytical problem solving. Personality & Social Psychology Bulletin 33, 312–24 (2007)
14. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann (1988)
15. Pearl, J.: Causality. Cambridge University Press (2009)
16. Rim, S., Summerville, A.: How far to the road not taken? the effect of psychological distance on counterfactual direction. Personality & Social Psychology Bulletin 40 (2013)
17. Rips, L., Edwards, B.: Inference and explanation in counterfactual reasoning. Cognitive science 37 (2013)
18. Russell, C., Kusner, M.J., Loftus, J., Silva, R.: When worlds collide: Integrating different counterfactual assumptions in fairness. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
19. 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 42(3), 383–390 (2012)
20. Timmer, S., Meyer, J., Prakken, H., Renooij, S., Verheij, B.: Inference and attack in bayesian networks. In: Hindriks, K., De Weerdt, M., Van Riemsdijk, B., Warnier, M. (eds.) Proceedings of the 25th Benelux Conference on Artificial Intelligence (BNAIC 2013). pp. 199–206. Delft University Press (2013)
21. Timmer, S.T., Meyer, J.J.C., Prakken, H., Renooij, S., Verheij, B.: A two-phase method for extracting explanatory arguments from Bayesian networks. International Journal of Approximate Reasoning 80, 475–494 (2017)
22. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. In: Wendt, D., Vlek, C. (eds.) Utility, Probability, and Human Decision Making: Selected Proceedings of an Interdisciplinary Research Conference, Rome, 3–6 September, 1973. pp. 141–162. Springer (1975)
23. Vlek, C.S., Prakken, H., Renooij, S., Verheij, B.: A method for explaining Bayesian networks for legal evidence with scenarios. Artificial Intelligence and Law 24(3), 285–324 (2016)
24. Vreeswijk, G.A.W.: Argumentation in bayesian belief networks. In: Rahwan, I., Moraitis, P., Reed, C. (eds.) Argumentation in Multi-Agent Systems: First International Workshop, ArgMAS 2004, New York, NY, USA, July 19, 2004, Revised Selected and Invited Papers. pp. 111–129. Springer (2005)
25. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR (2018)
26. Williams, M., Williamson, J.: Combining argumentation and Bayesian nets for breast cancer prognosis. Journal of Logic, Language and Information 15(1), 155–178 (2006)
27. Yap, G.E., Tan, A.H., Pang, H.H.: Explaining inferences in Bayesian networks. Applied Intelligence 29(3), 263–278 (2008)