Argumentation-based Causal and Counterfactual Reasoning

Lars Bengel[1], Lydia Blümel[1], Tjitze Rienstra[2] and Matthias Thimm[1]

[1]Artificial Intelligence Group, FernUniversität in Hagen, Germany
[2]Department of Advanced Computing Sciences, Maastricht University, The Netherlands

Abstract
In this paper we present a model for argumentative causal and counterfactual reasoning in a logical setting. Causal knowledge is represented in this system using Pearl's causal model of a set of structural equations and a set of assumptions expressed in propositional logic. Queries concerning observations or actions can be answered by constructing an argumentation framework and determining its extensions. For counterfactual queries we propose an argumentation-based implementation of the twin network method and analyse its expressiveness.

Keywords
Causality, Counterfactuals, Abstract Argumentation

1. Introduction

The area of explainable AI (XAI) has received increasing attention in recent years and many different approaches have been proposed (for an overview see e. g., [1, 2]). A central aspect of XAI is to provide human-understandable explanations for the output of some AI model. To provide these kinds of explanations, it seems only natural to utilise the techniques that humans employ to explain events and actions to each other. One such technique is *formal argumentation* [3]. In the past, many works have already suggested argumentation-based approaches for the purpose of providing explanations [4, 5]. More recently, we have seen different argumentative approaches to explain the output of various machine learning and other AI methods [6, 7].

Another field of research that is of increasing relevance is *causality* [8]. Especially interesting is how humans use causal connections for their explanations [9]. In the context of XAI, there are many advocates in the literature for using causal explanations [10, 11, 12]. Causality is furthermore closely linked with the notion of counterfactuals, which are statements that express what would have been true if something that is the case had been different. Counterfactual explanations have received much attention in recent research on XAI [13]. Relevant is also [14], in which the authors provide a logical representation of Pearl's causal models in the framework of the causal calculus [15]. They show that the non-monotonic semantics of this causal calculus

1st International Workshop on Argumentation for eXplainable AI (ArgXAI, co-located with COMMA '22), September 12, 2022, Cardiff, UK

lars.bengel@fernuni-hagen.de (L. Bengel); lydia.bluemel@fernuni-hagen.de (L. Blümel);
t.rienstra@maastrichtuniversity.nl (T. Rienstra); matthias.thimm@fernuni-hagen.de (M. Thimm)

corresponds directly to the solutions of causal models. In addition to that, this representation is also adequate for answering interventional queries for the causal model.

In this paper we contribute to these lines of research by investigating an explainable, argumentation-based account of reasoning with causal models. The starting point in our approach is a propositional knowledge base that encodes a causal model in the sense of Pearl [8] restricted to boolean variables and default assumptions about background variables. The approach makes it possible to answer *interventional* and *counterfactual* queries by constructing an argumentation framework whose stable extensions provide the answer to the queries. Our method of answering counterfactual queries is a reformulation of the dual approach due to Pearl, where a counterfactual query can be answered equivalently using two different methods: a three-step *abduction-action-prediction* procedure, and the so called *twin network* method [8]. We prove that, in our setting, the two methods are also equivalent.

Our work can be understood as an initial investigation into the possibility of causal reasoning within Dung's model of argumentation.

To summarise, the contributions of this paper are the following.

1. We introduce the notion of causal knowledge bases (Section 3).
2. We provide a novel approach for argumentation-based explanations for counterfactual reasoning implementing the twin network method (Section 4).
3. We prove the standard three-step procedure and the twin network method for evaluating counterfactuals by Pearl are equivalent under our notion (Section 4).
4. We discuss our approach and related work (Section 5).

In Section 2 we present the necessary background on knowledge bases and formal argumentation and Section 6 concludes the paper.

2. Preliminaries

The causal reasoning framework we develop builds on a simple and well-known form of default reasoning based on maximal consistent subsets [16]. Let $\mathcal{L}$ be a set of propositional formulas generated by a finite set of atoms. We assume there is a set $K \subseteq \mathcal{L}$ of *facts* and a set $A \subseteq \mathcal{L}$ of *assumptions*. We call a pair $\Delta = (K, A)$ a *knowledge base*. Facts are true and thus we assume that K is consistent. Assumptions are statements that we are willing to assume true unless we have evidence to the contrary. The consequences of a knowledge base $\Delta = (K, A)$ are determined by K together with the *maximal K-consistent* subsets of A.

Definition 1. *Let $\Delta = (K, A)$ be a knowledge base and $\phi, \psi \in \mathcal{L}$. A set $\Sigma \subseteq A$ is a* maximal K-consistent subset of A *whenever $\Sigma \cup K$ is consistent and $\Sigma' \cup K$ is inconsistent for all $\Sigma' \subseteq A$ such that $\Sigma \subset \Sigma'$. We say that:*

- *Δ entails ψ (written $\Delta \mid\!\sim \psi$) whenever $\Sigma \cup K \vdash \psi$ for every maximal K-consistent subset of A.*
- *ϕ Δ-entails ψ (written $\phi \mid\!\sim_\Delta \psi$) whenever $(K \cup \{\phi\}, A)$ entails ψ.*

In the next section we define a special class of knowledge bases used to represent causal knowledge. Our goal is to provide an argumentative method for causal and counterfactual reasoning based on such knowledge bases. The argumentative aspect of this method relies on a characterisation of reasoning with maximal consistent subsets based on the notion of *argumentation framework* or *AF*, for short [3].

Definition 2. *An argumentation framework is a pair $F = (\mathbf{A}, \Rightarrow)$ where A is a set whose elements are called* arguments *and where $\Rightarrow \subseteq A \times A$ is called the* attack relation.

We mostly use infix notation for attacks and write $a \Rightarrow b$ for $(a, b) \in \Rightarrow$ and $a \not\Rightarrow b$ for $(a, b) \notin \Rightarrow$. If $a \Rightarrow b$ we also say that a *attacks* b. Given an AF, a *semantics* determines sets of jointly acceptable arguments called *extensions*. The *admissible*, *preferred* and *stable* semantics are defined as follows [3].

Definition 3. *Let $F = (\mathbf{A}, \Rightarrow)$ be an AF. A set $E \subseteq \mathbf{A}$ is:*

- conflict-free *if for all $a, b \in E$ we have $a \not\Rightarrow b$.*
- self-defending *if for all $a \in E$ and $b \in \mathbf{A} \setminus E$ such that $b \Rightarrow a$, there is a $c \in E$ such that $c \Rightarrow b$.*
- admissible *if E is conflict-free and self-defending.*
- preferred *if E is admissible and there is no admissible set $E' \subseteq \mathbf{A}$ such that $E \subset E'$.*
- stable *if E is conflict-free and for every $a \in \mathbf{A} \setminus E$ there is a $b \in E$ such that $b \Rightarrow a$.*

Following Cayrol [17] we define an argument induced by a knowledge base $\Delta = (K, A)$ to be a pair (Φ, ψ) where $\Phi \subseteq A$ is a minimal set of assumptions (called the *premises* of the argument) that, together with K, consistently entails ψ (called the *conclusion* of the argument). One argument *undercuts* another if the conclusion of the former is the negation of a premise of the latter. The set of all arguments induced by Δ, together with undercut as the attack relation, forms the AF induced by Δ [1]. Formally:

Definition 4. *Let $\Delta = (K, A)$ be a knowledge base.*

- *An* argument induced by Δ *is a pair (Φ, ψ) such that*
  - $\Phi \subseteq A$,
  - $\Phi \cup K \nvdash \bot$,
  - $\Phi \cup K \vdash \psi$, *and if $\Psi \subset \Phi$ then $\Psi \cup K \nvdash \psi$.*
- *An argument (Φ, ψ)* undercuts *an argument (Φ', ψ') if for some $\phi' \in \Phi'$ we have $\phi' \equiv \neg\psi$.*
- *The* AF induced by Δ *is the AF $(\mathbf{A}, \Rightarrow)$ where $\mathbf{A}$ consists of all arguments induced by Δ and $a \Rightarrow b$ holds whenever a undercuts b. We denote the AF induced by Δ by $F(\Delta)$.*

Cayrol et al. [17] showed that there is a one-to-one correspondence between maximal K-consistent subsets of a knowledge base and the stable extensions of the induced AF. Thus, given a knowledge base $\Delta = (K, A)$, the question of whether ϕ Δ-entails ψ can be determined by constructing the AF induced by $(K \cup \{\phi\}, A)$ and checking whether every stable extension contains an argument with conclusion ψ.

[1] Note that this construction method allows arguments to have syntactically different, but semantically equivalent conclusions.

Proposition 1. *Let $\Delta = (K, A)$ be a knowledge base. Then $\phi \mid\!\sim_{\Delta} \psi$ if and only if every stable extension E of $F(K \cup \{\phi\}, A)$ contains an argument with conclusion ψ.*

We conclude this section by noting that the form of argumentation we consider here can also be captured by structured argumentation formalisms such as ASPIC and ABA [18].

3. Causal Knowledge Bases

Our definition of a causal model is essentially that of Pearl [8] except that we restrict our attention to Boolean-valued variables. A causal model consists of a set U of *background* atoms, a set V of *explainable* atoms, and a set K of formulas which we call *Boolean structural equations* (this terminology was adopted from [14]). Background atoms represent variables that are determined outside of the model. They are typically unobservable and uncontrollable. Every explainable atom v is functionally dependent on other atoms of the model. This dependency is specified by a Boolean structural equation of the form $v \leftrightarrow \phi$. Intuitively, this equation represents the causal mechanism by which v is determined by the other atoms in the model. We use bi-implication because the represented causal mechanism determines not only when v is true, but also when v is false.

Definition 5. *A* causal model *is a triple (U, V, K) where U and V partition the set of atoms into, respectively, a set of background and explainable atoms. K consists of a set of* Boolean structural equations*, one for each atom $v \in V$. A Boolean structural equation for v is a formula of the form*

$$v \leftrightarrow \phi$$

where ϕ is a propositional formula that does not contain v.

Given a causal model (U, V, K) we can reconstruct U and V from K as follows; (1) for every formula $v \leftrightarrow \phi$ in K, add v to V, and (2) add all the remaining atoms that appear in K but not in V to U. Therefore we can refer to a causal model as K. For every Boolean structural equation $v \leftrightarrow \phi$, we call an atom appearing in ϕ a *parent* (background or explainable) of v. Like in the standard case, a causal model induces a *causal graph G* whose vertices are the explainable atoms of the model [8]. This graph contains an edge from atom v to atom v' whenever v is an explainable parent of v'.

Example 1. *Let $U = \{\mathit{corona}, \mathit{influenza}, \mathit{at\text{-}risk}\}$ and $V = \{\mathit{covid}, \mathit{flu}, \mathit{short\text{-}of\text{-}breath}, \mathit{fever}, \mathit{chills}\}$. The set K consisting of the following equations represents a causal model for a patient diagnosis scenario.*

$$\begin{aligned}
\mathit{covid} &\leftrightarrow \mathit{corona}\\
\mathit{flu} &\leftrightarrow \mathit{influenza}\\
\mathit{fever} &\leftrightarrow \mathit{covid} \vee \mathit{flu}\\
\mathit{chills} &\leftrightarrow \mathit{fever}\\
\mathit{short\text{-}of\text{-}breath} &\leftrightarrow \mathit{covid} \wedge \mathit{at\text{-}risk}
\end{aligned}$$

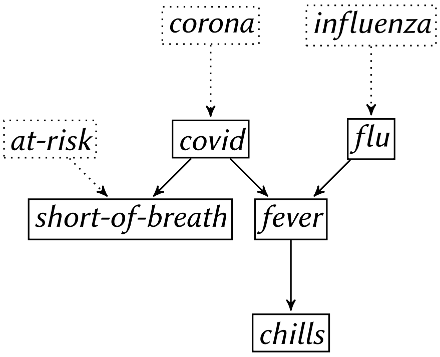


Figure 1: Causal graph for Example 1.

In words: corona virus causes covid, influenza virus causes flu, covid and flu cause fever, fever causes chills, and covid causes short-of-breath, but only if the patient is at risk for this condition (represented by the background atom at-risk). Note that corona, influenza and at-risk are background atoms and thus assumed unobservable.

Figure 1 depicts the causal graph for this model. This figure also includes the background atoms of the model, drawn using dotted lines.

We define a *causal knowledge base* to be a knowledge base where the set of facts is a causal model and where the set of assumptions is restricted to contain assumptions about background atoms. Here we depart from the probabilistic approach to causal modelling, where belief about background atoms is represented using probability distributions over their values [8].

Definition 6. *A* causal knowledge base *is a knowledge base $\Delta = (K, A)$ where K is a causal model and where A is a set of* background assumptions*, at least one for each background atom. A* background assumption *for an atom u is a literal $l \in \{u, \neg u\}$.*

By restricting background assumptions to be literals we prevent the ability to express dependencies among background atoms. There are three possible attitudes towards a background atom u, since we can assume just u, just $\neg u$, or both. Assuming just u or $\neg u$ amounts to assuming that u is true or false, unless we have evidence to the contrary. Assuming both u and $\neg u$ represents a state of uncertainty where u may, depending on the evidence, be true as well as false. If $\Delta = (K, A)$ is a causal knowledge base then Δ-entailment represents a relation between observations and predictions. These predictions include causes as well as effects of the observation in accordance with the causal model K and background assumptions A.

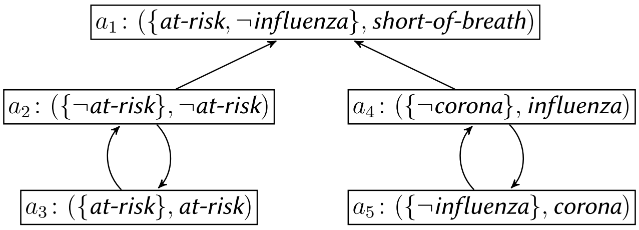
Example 2. *Consider the causal knowledge base $\Delta = (K, A)$ where K is the causal model defined in Example 1 and where $A = \{\textit{at-risk}, \neg\textit{at-risk}, \neg\textit{corona}, \neg\textit{influenza}\}$. In words, these assumptions mean that we are uncertain whether the patient is at risk to develop shortness of breath, and a corona or influenza infection is assumed false unless there is evidence to the contrary. Now consider the question of whether observing fever entails shortness of breath, i.e., whether*

$$\textit{fever} \mathrel{|\!\!\sim_{\Delta}} \textit{short-of-breath}. \tag{1}$$

Flu and covid cause fever. Thus, fever is evidence for flu or covid. Of these two possible causes, covid may cause shortness of breath, but only if the patient is at risk, which may or may not be true. Hence, fever may or may not be evidence for shortness of breath. This reasoning is depicted by the AF $F = F((K \cup \{\mathit{fever}\}, A))$ *shown in Figure 2 (we only depict arguments relevant to the conclusion of short-of-breath). This AF has four stable extensions* $\{a_2, a_4\}$*,* $\{a_3, a_4\}$*,* $\{a_2, a_5\}$*, and* $\{a_3, a_5, a_1\}$*. The argument* a_1 *with conclusion short-of-breath is included in some but not all of these extensions. We thus have that* (1) *is false, but note that*

$$\mathit{fever} \mid\!\sim_{\Delta} \neg \mathit{short\text{-}of\text{-}breath}$$

is also false. Thus, given fever, shortness of breath is possible but not necessary.



a_1: ({at-risk, ¬influenza}, short-of-breath)
a_2: ({¬at-risk}, ¬at-risk)
a_4: ({¬corona}, influenza)
a_3: ({at-risk}, at-risk)
a_5: ({¬influenza}, corona)

Figure 2: The AF $F(K \cup \{\mathit{fever}\}, A)$.

4. Counterfactuals

In this section we show how causal knowledge bases provide a means to evaluate counterfactuals. Pearl introduces two methods for evaluating counterfactuals in causal models [8], a three-step procedure and the twin network method. We now demonstrate these two processes are equivalent under our notion of causal knowledge bases.

Before dealing with counterfactual statements we start with *interventional statements* of the form

$$\text{if } v \text{ would be } x \text{ then } \psi \text{ would be true} \tag{2}$$

where v is a variable we intervene on and x a truth value (we use $\top$ for true and $\bot$ for false). The antecedent of this statement represents the *action* of setting v to x. In our running example, a patient might be given Ibuprofen to treat fever. This is an action that amounts to setting *fever* to $\bot$. Note that this is different from observing $\neg$*fever*. While the observation $\neg$*fever* would, in our model, imply $\neg$*covid* and $\neg$*flu*, the action of setting *fever* to $\bot$ amounts to overriding the causal mechanism that determines *fever*, which does not affect *covid* or *flu*. Like Pearl, given a causal model K, we denote by $K_{[v=x]}$ the new causal model where the equation $v \leftrightarrow \phi$ is replaced with $v \leftrightarrow x$.

Definition 7. *Let* K *be a causal model, let* v *be an explainable atom, and let* $x \in \{\top, \bot\}$*. We denote by* $K_{[v=x]}$ *the causal model defined by*

$$K_{[v=x]} = \{(v' \leftrightarrow \phi) \in K \mid v' \neq v\} \cup \{(v \leftrightarrow x)\}.$$

Example 3. *Let $\Delta = (K, A)$ be the causal knowledge base from Example 2. Since short-of-breath is caused by covid, covid causes fever, and fever causes chills, we have that shortness of breath leads to the prediction of chills*

$$\textit{short-of-breath} \mathrel{|\!\sim_{\Delta}} \textit{chills}. \tag{3}$$

But what if we observe short-of-breath in a patient that has been given Ibuprofen? The answer is given by the causal knowledge base $\Delta' = (K_{[\textit{fever}=\bot]}, A)$. Here, the action of setting fever to $\bot$ blocks the effect of covid on chills. We now have that shortness of breath leads to the prediction of no chills

$$\textit{short-of-breath} \mathrel{|\!\sim_{\Delta'}} \neg\textit{chills}. \tag{4}$$

We now turn to counterfactual statements. These are statements of the form

$$\text{given } \phi, \text{ if } v \text{ had been } x \text{ then } \psi \text{ would be true} \tag{5}$$

where ϕ is an observed piece of evidence and ψ is the counterfactual conclusion. An example of a counterfactual in our running example would be: given that the patient has fever, would the patient have had fever if we had administered a covid vaccine? This amounts to the statement *given fever, if covid had been $\bot$ then fever would be true.* The standard causal modelling approach of evaluating a counterfactual statement of the form (5) is based on a three-step procedure [8] that we adapt for our setting as follows.

Definition 8. *Given a causal knowledge base $\Delta = (K, A)$, the truth of a counterfactual statement* (5) *is determined by:*

- *Step 1 (abduction) Determine the maximal $K \cup \{\phi\}$-consistent subsets $\Sigma_1, \ldots, \Sigma_n$ of A. These sets represent the possible configurations of the background atoms consistent with evidence ϕ.*
- *Step 2 (action) Set v to x by updating the causal model K to $K_{[v=x]}$.*
- *Step 3 (prediction) For each Σ_i obtained in step 1, determine whether $\Sigma_i \cup K_{[v=x]} \vdash \psi$. If the answer is yes for all $i \in 0, \ldots n$, then the counterfactual statement* (5) *is true.*

The difficulty with this procedure is that it requires the computation of the maximal $K \cup \{\phi\}$-consistent subsets of A. The same difficulty arises in probabilistic causal models, where the abduction step requires the computation of a probability distribution over configurations of the background atoms. The *twin network* method overcomes this difficulty [8]. It is based on constructing a so called *twin model* whose causal network consists of two parts, one to represent the actual world, and one to represent the counterfactual world. While the two networks share the same background variables, the counterfactual part consists of a “counterfactual copy” of all the explainable variables. In what follows we use v^* to denote a unique new atom representing the counterfactual copy of v, and use ϕ^* to denote the formula ϕ with each occurrence of an explainable atom v replaced with v^*.

Definition 9. *The* twin model *for a causal model K is the causal model K^* defined by*

$$K^* = K \cup \{(v^* \leftrightarrow \phi^*) \mid (v \leftrightarrow \phi) \in K\}$$

.

While the abduction step takes place in the actual network, the action and prediction steps take place in the counterfactual network. Let us now prove this intuition of separated areas in the twin model holds up formally during the evaluation, i. e., the three step procedure is equivalent to the evaluation of the twin model.

Proposition 2. *Let $\Delta = (K, A)$ be a causal knowledge base. The counterfactual statement*

$$\textit{given } \phi, \ \textit{if } v \textit{ had been } x \textit{ then } \psi \textit{ would be true}$$

is true in Δ if and only if $\phi \mathrel{|\!\sim}_{\Delta^*_{[v^*=x]}} \psi^*$.

Proof. We prove that the *twin model* method is equivalent to the three-step procedure for evaluating counterfactuals from Pearl as given in Definition 8.

Let $\Delta = (K, A)$ be a causal knowledge base. First, we recall that $\phi \mathrel{|\!\sim}_\Delta \psi$ if and only if $(K \cup \{\phi\}, A)$ entails ψ. $(K \cup \{\phi\}, A)$ entails ψ whenever for every maximal K-consistent subset Σ of A, we have that $\Sigma \cup K \vdash \psi$.

($\Leftarrow$) Now, for the given counterfactual statement, consider $\Delta^* = (K^*, A)$ with K^* as given in Definition 9. Furthermore, we modify Δ^* by setting $v^* = x$ in K^*, i.e., we have $\Delta^*_{[v^*=x]}$. That means, K^* fully contains K as well as an exact copy of K (each copy of a formula ϕ of K is denoted with ϕ^*). In this copy, we also have set $v^* = x$. Therefore, we can essentially also consider K^* as the union of K and $K_{[v=x]}$.

With the twin model method, we have that $\phi \mathrel{|\!\sim}_{\Delta^*_{[v^*=x]}} \psi^*$. That means, according to Definition 1 we determine the $K^* \cup \{\phi\}$-consistent subsets $\Sigma^*_1, ..., \Sigma^*_n$ of A. Now, since $K \subset K^*$, it follows that $\Sigma^*_1 \cap K, ..., \Sigma^*_n \cap K$ are maximal consistent with $K \cup \{\phi\}$ (as required by step 1 abduction).

The action step of Definition 8 is also executed in the twin model by setting $v^* = x$ in K^*.

In the prediction step, we have to determine whether $\Sigma_i \cup K_{[v=x]} \vdash \psi$ for all K-consistent Σ_i. In the twin model method, we consider the $K^* \cup \{\phi\}$-consistent subsets. As already explained above, $K_{[v=x]}$ is fully contained in K^* and therefore, similarly to above, if we have that $\Sigma^*_i \cup K^* \vdash \psi^*$ it follows that $\Sigma^*_i \cup K_{[v=x]} \vdash \psi$. Thus, the twin model method is equivalent to the three-step procedure for evaluating counterfactual statements.

($\Rightarrow$) Suppose the counterfactual statement is true according to the three step procedure. In the twin model it holds that K and $\{(v^* \leftrightarrow \phi^*) \mid (v \leftrightarrow \phi) \in K\}$ have no variables in common. Thus, for any Σ maximal consistent wrt. $K^* \cup \{\phi\}$ it holds that $\Sigma' = \Sigma \cap \{(v^* \leftrightarrow \phi^*) \mid (v \leftrightarrow \phi) \in K\}$ is maximal consistent wrt. $K_{[v=x]} \cup \{\phi\}$. By the assumption we have $\Sigma' \cup K_{[v=x]} \cup \{\phi\} \vdash \psi$ for any such Σ'. By monotony of classic entailment it follows that $\Sigma \cup K^*_{[v^*=x]} \cup \{\phi\} \vdash \psi$. □

We have now proven that the standard three-step procedure can be reduced to a single evaluation step on the twin model. To conclude, let us give some examples of counterfactual reasoning with the AFs build from a twin model.

Example 4. *Let $\Delta = (K, A)$ be the causal knowledge base defined in Example 2. Suppose we have evidence that the patient has fever. Would the patient have had fever if we had administered a covid vaccine (i.e., if covid had been false)? This depends on whether the fever is caused by covid or flu. If fever was caused by covid then the vaccine would indeed have prevented fever. If, on the*

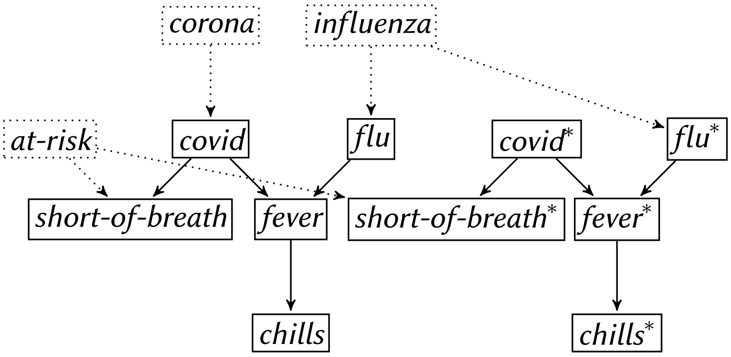


Figure 3: Twin causal diagram for Example 4.

other hand, fever was caused by flu, then the vaccine would not have had an effect, and the patient would still have had fever. To evaluate this counterfactual statement we check whether

$$\textit{fever} \mid\!\sim_{\Delta^*_{[\textit{covid}^*=\bot]}} \textit{fever}^*. \tag{6}$$

Figure 3 shows the causal graph of the twin model $K^*_{\textit{covid}^*=\bot}$ and Figure 4 shows the AF $F = (K^*_{[\textit{covid}^*=\bot]} \cup \{\textit{fever}\}, A)$. This AF has two stable extensions $\{a_1, a_3\}$ and $\{a_2\}$. While the first entails fever, the second does not. Thus, (6) is false, but note that

$$\textit{fever} \mid\!\sim_{\Delta^*_{[\textit{covid}^*=\bot]}} \neg\textit{fever}^* \tag{7}$$

is also false. Thus, given that the patient has fever, the patient may or may not have had fever, had we administered a covid vaccine.

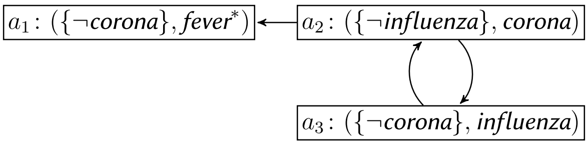

a_1: ({¬corona}, fever*)
a_2: ({¬influenza}, corona)
a_3: ({¬corona}, influenza)

Figure 4: The AF $F((K^*_{[\textit{covid}^*=\bot]} \cup \{\textit{fever}\}, A))$.

Example 5. *Suppose that, in addition to having evidence that the patient has fever, we also have evidence of shortness of breath. Again, we want to know if the patient would have had fever if we had administered a covid vaccine. This time, the evidence of shortness of breath implies that the fever must have been caused by covid. Hence, the vaccine would indeed have prevented fever. We check this formally by checking whether*

$$\textit{fever} \wedge \textit{short-of-breath} \mid\!\sim_{\Delta^*_{[\textit{covid}^*=\bot]}} \textit{fever}^*. \tag{8}$$

Figure 5 shows the AF $F = (K^*_{\textit{covid}^*=\bot} \cup \{\textit{fever} \wedge \textit{short-of-breath}\}, A)$. This AF is the same as the one shown in Figure 4 except that there is an extra argument $(\emptyset, \textit{corona})$ which is based on the evidence short-of-breath. This AF has one stable extension $\{a_4, a_2\}$ which does not entail fever*. Hence, (8) is true, i.e., given that the patient has fever and shortness of breath, the patient would not have had fever, had we administered a covid vaccine.

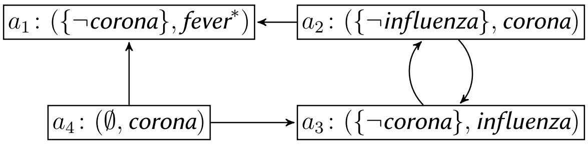

a_1: ({¬*corona*}, *fever**)
a_2: ({¬*influenza*}, *corona*)
a_4: (∅, *corona*)
a_3: ({¬*corona*}, *influenza*)

Figure 5: The AF $F(K^*_{\mathit{covid}^*=\bot} \cup \{\mathit{fever} \wedge \mathit{short\text{-}of\text{-}breath}\}, A)$.

5. Discussion

Our work can be understood as an initial investigation into the possibility of reasoning with causal models using tools of abstract information science and logic. By doing this we contribute to explaining technically advanced causal reasoning procedures with concepts closer to human reasoning. Our account of Pearl's theory was motivated by the non-monotonic interpretation of his causal model in [14]. The need to further explain causal reasoning, e. g., using Bayesian networks, has also been pointed out in [19]. Their contribution towards this is making d-separation explicit in so called support graphs, which eliminate circular causal structures and help to explain interdependent causes. Implementing a similar mechanism within our framework has proven nontrivial. In [20] they illustrate the power of their framework by applying it to legal reasoning. As the authors point out under future work, the support graph is more of a visualization tool while the actual evaluation of arguments is still done in the original Bayesian network formalism. A sophisticated approach towards a framework capable of intrinsic evaluations of causal statements is the method for generating bipolar argumentation frameworks from causal models by Rago et al. [21], which is also based on Pearl's notion of a causal model. For that they create so called explanation moulds, that reinterpret desirable properties of semantics of argumentation frameworks. Rago et. al. point out that, compared to [22], they make better use of the AF representation in their causal reasoning by basing their semantic evaluation on the support and attack relations themselves instead of relying on an additional ontological structure. This criticism does not apply to our framework, we only rely on the causal model. Once the AF is constructed, we do not need the inference apparatus of the underlying knowledge base formalism for causal reasoning in our framework.

There are several major design differences between the two approaches. First Rago et al. use bipolar AFs in contrast to our Dung-style approach. Second, our approach is based on structured argumentation, while theirs uses causal atoms as arguments. On the evaluation level, they apply gradual semantics while we make use of classic stable semantics. For both frameworks an interesting future work direction would be to apply the resp. other type of semantics. An important difference lies in the resp. attack notions. Rago et al. let causes contribute negatively or positively towards the status of an argument via the attack and support relation, respectively. Our focus is on the explanation aspect. Attacks are used to exclude certain premises as causes. This is done by undercut-attacks on the corresponding argument. A current shortcoming of our approach is the restriction to propositional logic in the attack notion, which would have difficulties handling uncertain causal relations. A promising future work direction would therefore be to investigate ways for representing uncertain causal relations in some type of probabilistic argumentation framework, e. g. by adding a probability distribution over possible causal structures [23]. Another option could be to utilize a form of probabilistic instantiation like PABA [24].

6. Conclusion

We have proposed a natural way to model causal reasoning with Dung-style AFs by means of causal knowledge bases. The framework introduced integrates given causal knowledge in form of structural equations into an attack relation and a set of structured arguments constructed from sets of assumptions resp. observations from the real world. By doing this we have laid the groundwork for an implementation of the twin network method for counterfactual reasoning. It offers an alternative for computing the truth value of counterfactual statements which, as we proved, is equivalent to the standard approach in our framework. The results we obtain demonstrate that causal reasoning can be done as an instantiation of Dung's model of argumentation. This can be used to provide argumentative explanations for causal and counterfactual queries. That is, the argumentation framework contains not only the arguments used to arrive at a prediction, but also possible counterarguments that lead to alternative predictions.

Acknowledgements The research reported here was partially supported by the Deutsche Forschungsgemeinschaft (grant 375588274).

References

[1] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE access 6 (2018) 52138–52160.

[2] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.

[3] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artificial Intelligence 77 (1995) 321–358.

[4] C. Antaki, I. Leudar, Explaining in conversation: Towards an argument model, European Journal of Social Psychology 22 (1992) 181–194.

[5] B. Moulin, H. Irandoust, M. Bélanger, G. Desbordes, Explanation and argumentation capabilities: Towards the creation of more persuasive agents, Artificial Intelligence Review 17 (2002) 169–222.

[6] K. Čyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative xai: a survey, arXiv preprint arXiv:2105.11266 (2021).

[7] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review 36 (2021) e5.

[8] J. Pearl, Causality: models, reasoning and inference, volume 29, Cambridge University Press, 2000.

[9] M. M. De Graaf, B. F. Malle, How people explain action (and autonomous intelligent systems should too), in: 2017 AAAI Fall Symposium Series, 2017, pp. 19–26.

[10] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.

[11] P. Schwab, W. Karlen, Cxplain: Causal explanations for model interpretation under uncertainty, Advances in Neural Information Processing Systems 32 (2019) 10188–10198.

[12] D. Alvarez-Melis, T. S. Jaakkola, A causal framework for explaining the predictions of black-box sequence-to-sequence models, arXiv preprint arXiv:1707.01943 (2017).

[13] R. M. J. Byrne, Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, ijcai.org, 2019, pp. 6276–6282. doi:10.24963/ijcai.2019/876.

[14] A. Bochman, V. Lifschitz, Pearl's causality in a logical setting, in: B. Bonet, S. Koenig (Eds.), Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA, AAAI Press, 2015, pp. 1446–1452. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9686.

[15] N. McCain, H. Turner, et al., Causal theories of action and change, in: AAAI/IAAI, Citeseer, 1997, pp. 460–465.

[16] N. R. R. Manor, N. Rescher, On inference from inconsistent premises, Theory and Decision 1 (1970) 179–219.

[17] C. Cayrol, On the relation between argumentation and non-monotonic coherence-based entailment, in: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes, Morgan Kaufmann, 1995, pp. 1443–1448.

[18] O. Arieli, A. Borg, J. Heyninck, C. Straßer, Logic-based approaches to formal argumentation, FLAP 8 (2021) 1793–1898. URL: https://collegepublications.co.uk/ifcolog/?00048.

[19] S. T. Timmer, J.-J. C. Meyer, H. Prakken, S. Renooij, B. Verheij, Explaining bayesian networks using argumentation, in: European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty, Springer, 2015, pp. 83–92.

[20] S. T. Timmer, J.-J. C. Meyer, H. Prakken, S. Renooij, B. Verheij, A structure-guided approach to capturing bayesian reasoning about legal evidence in argumentation, in: Proceedings of the 15th International Conference on Artificial Intelligence and Law, 2015, pp. 109–118.

[21] A. Rago, F. Russo, E. Albini, P. Baroni, F. Toni, Forging argumentative explanations from causal models, in: M. D'Agostino, F. A. D'Asaro, C. Larese (Eds.), Proceedings of the 5th Workshop on Advances in Argumentation in Artificial Intelligence 2021 co-located with the 20th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2021), Milan, Italy, November 29th, 2021, volume 3086 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-3086/paper3.pdf.

[22] P. Besnard, M.-O. Cordier, Y. Moinard, Arguments using ontological and causal knowledge, in: International Symposium on Foundations of Information and Knowledge Systems, Springer, 2014, pp. 79–96.

[23] A. Hunter, S. Polberg, N. Potyka, T. Rienstra, M. Thimm, Probabilistic argumentation: A survey, Handbook of Formal Argumentation 2 (2021) 397–441.

[24] K. Čyras, Q. Heinrich, F. Toni, Computational complexity of flat and generic assumption-based argumentation, with and without probabilities, Artificial Intelligence 293 (2021) 103449.