Argumentation-based Probabilistic Causal Reasoning

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

[1] Artificial Intelligence Group, University in Hagen, Germany
{firstname.lastname}@fernuni-hagen.de
[2] Dep. of Advanced Computing Sciences, Maastricht University, The Netherlands
t.rienstra@maastrichtuniversity.nl

Abstract. We introduce an argumentation-based approach for conducting probabilistic causal reasoning. For that, we consider Pearl's causal models where causal relations are modelled via structural equations and a probability distribution over background atoms. The probability that some causal statement holds is then computed by constructing a probabilistic argumentation framework and determining its extensions. This framework can then be used to generate argumentative explanations for the (non-)acceptance of the causal statement. Furthermore, we present an argumentation-based version of the twin network method for dealing with counterfactuals. Finally, we show that our approach yields the same results for causal and counterfactual queries as Pearl's model.

Keywords: causality · argumentation · counterfactuals.

1 Introduction

A recent work [17] presents a machine learning model capable of predicting the mortality within the next 24 hours of the in-patients of a hospital with an accuracy of 95%. This impressive example of the recent advances in AI research is also an excellent example of the limits of machine learning approaches. While it is of course helpful to know which patients need immediate treatment to prevent them from dying, the model leaves us completely in the dark regarding the kind of treatment they need. Imagine this kind of algorithm to be used during a major incident where triage is necessary. Using this model to decide who receives treatment could do more harm than it helps, because patients that could be saved with simple and fast methods would be excluded from treatment. This is one of many potential applications for AI where an explanation of the output of the model is needed. Due to this issue, Explainable Artificial Intelligence (XAI) has become an important research area, which is a very productive but also challenging area of research [13].

A major contribution towards a formal theory of causality is the work on causal graphs by Pearl [14]. He models causal relationships with a double-layered formalism. On the one hand, there are the structural equations which are used to compute the value of an observable variable from a given set of values for a

fixed number of unobservable background variables. On the other hand, there is a directed acyclic graph, which represents the causal dependencies between observable and background variables. A causal explanation is then formalized as a set of logical statements on causal dependencies. His approach has been widely recognized and, in particular, has been adopted in recent work on XAI [11,16].

For verifying a given causal explanation one needs a reasoning formalism which can process causal statements. We propose to use abstract argumentation frameworks as introduced by Dung in [6]. An abstract argumentation framework consists of a set of arguments—in our case causal statements—and a binary attack relation between them. An argumentation semantics is applied to this structure to determine sets of collectively acceptable arguments—so called extensions—which we use to represent consistent sets of causal statements. As a non-monotonic formalism, it can handle inconsistent input, which makes it well-suited for causal reasoning, where additional information can falsify a previously inferred causal dependency. In our approach, causal statements are interpreted as arguments in an abstract argumentation framework and the attack relation represents contradicting causal inferences. This allows us to question the reasoning process during a query. A representation of causal inferences with an argumentation framework offers an intuitive and well-researched access to all maximal consistent causal theories fitting some given facts.

We present two methods for integrating uncertainty into our causal argumentation frameworks. Our first approach makes use of default reasoning to accommodate inconsistent assumptions to reason from. This allows us to reason while staying ambiguous with regard to some background variables. We presented a preliminary discussion of this method in a recent workshop paper [1]. In the second approach we refine our causal argumentation frameworks by bringing probabilities into play. In order to represent Pearl's causal theory to the full extent with argumentation, we introduce probabilistic causal argumentation frameworks, which are based on the probabilistic argumentation frameworks by Hunter [10]. To summarise, our contributions are:

- We demonstrate how causal argumentation frameworks can be used to conduct defeasible reasoning on causal statements (Section 3.1), following up on our work [1].
- We introduce an enhanced version, probabilistic causal argumentation framework and show that it captures Pearl's probabiblistic causal reasoning adequately (Section 3.2).
- We employ probabilistic argumentation frameworks for reasoning with interventional and counterfactual statements and show they produce the same results as Pearl's three-step-method and twin model approach (Section 4).

Moreover, Section 2 introduces the necessary formal context, Section 5 discusses related works, and Section 7 concludes the paper. Proofs of technical results are omitted due to space restrictions and can be found in an online appendix.[1]

[1] http://mthimm.de/misc/bbrt_ratio24.pdf

2 Preliminaries

We set $\mathcal{L}$ to be the language of propositional logic over a finite set of atoms At with the usual connectives $\{\wedge, \vee, \neg, \rightarrow, \leftrightarrow\}$ and $\vdash$ is the standard entailment operator. A *valuation* $val : \mathsf{At} \rightarrow \{true, false\}$ is an assignment of truth values to propositional variables. Our causal reasoning framework builds on a well-known form of default reasoning based on maximal consistent subsets [12]. We define a knowledge base Δ as a pair (K, A) where we assume that $K \subseteq \mathcal{L}$ is a set of *facts* and $A \subseteq \mathcal{L}$ is a set of *assumptions*. Facts are true, thus we assume that K is consistent while assumptions are statements that we are willing to assume true unless we have evidence to the contrary.

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.*

The argumentative part of our causal reasoning method relies on the notion of the *argumentation framework* (AF for short) as introduced by Dung [6].

Definition 2. *An argumentation framework is a pair $\mathsf{AF} = (\mathsf{Arg}, \mathsf{R})$ where Arg is a set of* arguments *and where $\mathsf{R} \subseteq A \times A$ is called the* attack relation.

We say that an argument $a \in \mathsf{Arg}$ *attacks* another argument $b \in \mathsf{Arg}$ iff we have that $(a, b) \in \mathsf{R}$. We may also use infix notation for attacks and write $a\mathsf{R}b$ for $(a, b) \in \mathsf{R}$. Given an AF, a *semantics* determines sets of jointly acceptable arguments called *extensions*. In this work, we only make use of the *stable semantics*, for other semantics see [6].

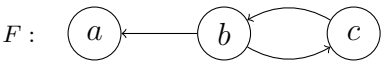
Definition 3. *Let $\mathsf{AF} = (\mathsf{Arg}, \mathsf{R})$ be an AF. A set $E \subseteq \mathsf{Arg}$ is:*

- conflict-free *iff for all $a, b \in E$ we have $(a, b) \notin \mathsf{R}$.*
- stable *iff E is conflict-free and for every $a \in \mathsf{Arg} \setminus E$ there is a $b \in E$ such that $(b, a) \in \mathsf{R}$.*

With $\mathbf{stb}(\mathsf{AF})$ we denote the set of stable extensions of an AF. For the argumentative part of our approach to reasoning with a probabilistic causal model, we use the notion of *probabilistic argumentation framework* (PAF for short) [9]. In this framework, probabilities are assigned to sets of arguments $S \subseteq \mathsf{Arg}$, called *framework states*, which implies that the existence of arguments is not independent of each other. Whenever an argument a is part of some framework state S, i. e., we have that $a \in S$, we say that a is active in S.

Definition 4. *A* probabilistic argumentation framework *is a pair $\mathsf{PAF} = (\mathsf{AF}, P_{\mathsf{AF}})$ where $\mathsf{AF} = (\mathsf{Arg}, \mathsf{R})$ is an argumentation framework and $P_{\mathsf{AF}} : 2^{\mathsf{Arg}} \rightarrow [0, 1]$ is a function with $\sum_{S \in 2^{\mathsf{Arg}}} P_{\mathsf{AF}}(S) = 1$.*

Example 1. Consider the PAF in Figure 1. We evaluate the framework by considering the different framework states and their respective extensions. For instance, the framework state $S_1 = \{a, b\}$ has a probability of 0.4 and only one stable extension $\{b\}$. On the other hand, the framework state $S_3 = \{a, b, c\}$ with probability 0.2 has two stable extensions $\{a, c\}$ and $\{b\}$.



$P_{\mathsf{AF}}(S)$	a	b	c	$\mathbf{stb}(S)$
0.4	✓	✓		$\{\{b\}\}$
0.4		✓	✓	$\{\{b\}, \{c\}\}$
0.2	✓	✓	✓	$\{\{a, c\}, \{b\}\}$

Fig. 1. The PAF $(\mathrm{F}, P_{\mathsf{AF}})$ with three frameworks states as depicted in the table.

3 Causal Reasoning

In the following, we will introduce an argumentation-based approach to perform reasoning with a causal model. The main advantage of this approach is the ability to not only determine whether some causal statement holds, but also provide an argumentative explanation on why it holds or not.

In Section 3.1, we introduce our approach for qualitative causal reasoning from [1], based on a modified version of Pearl's causal model [14], where we only consider Boolean-valued variables. In this scenario, we model the uncertainty via defeasibility which allows us to qualitatively answer queries directly in an argumentation framework. On the other hand, quantitative causal reasoning means computing the exact probability that the conclusion holds under the given observation. For this type of reasoning, we consider probabilistic causal models [14] and define a novel approach for answering queries with the help of a probabilistic argumentation framework (Section 3.2).

3.1 Defeasible Causal Reasoning

To model defeasible causal reasoning, we essentially use the causal model of Pearl [14] except that we restrict our attention to Boolean-valued variables. As described in Definition 5 below, a causal model[2] K is a set of formulas which we call *Boolean structural equations* (terminology adopted from [2]). We distinguish between two types of atoms in these equations: the *background* atoms $U(K)$ and *explainable* atoms $V(K)$. Variables that are determined outside of the model are represented as background atoms $u \in U(K)$ and are considered unobservable and uncontrollable. An explainable atom $v \in V(K)$ is functionally dependent on

[2] Here, we deviate from Pearl's notation for causal models which are defined as the triple (U, V, K), explicitly listing background and explainable atoms [14]. However, with $(U(K), V(K), K)$ we recover Pearls notation of a causal model.

other atoms of the model. We specify this dependency in the form of Boolean structural equations of the form $v \leftrightarrow \phi$, where ϕ is a logical formula over the set of atoms that v is dependent on. Intuitively, a structural equation for some explainable atom v 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 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. A* causal model *K is a set of* Boolean structural equations*, exactly one equation κ_v for each atom $v \in V(K)$. With $U(K)$ we denote the set of background atoms appearing in K and with $V(K)$ we denote the set of explainable atoms appearing in K.*

Furthermore, a causal model induces a *causal graph* G whose vertices are the explainable atoms of the model [14]. Background atoms of the model are represented as a different type of vertex. Given a Boolean structural equation $v \leftrightarrow \phi$, we call an atom appearing in ϕ a *parent* of v. The causal graph G contains an edge from atom $v \in U \cup V$ to atom $v' \in V$ whenever v is a parent of v'. We say a causal model K is Semi-Markovian if the causal graph induced is acyclic [14].

Example 2. Suppose we are building a causal model to investigate the cause of a surfer's death by drowning at the beach. The explainable variables in this case could be $V_{surf}(K_{surf}) = \{\textit{drowning}, \textit{cramp}, \textit{submersion}, \textit{broken-board}\}$, i. e., the fact itself, two physical conditions leading to it, as well as a side-effect. The background conditions potentially leading to these variables being true are $U_{surf}(K_{surf}) = \{\textit{jellyfish}, \textit{strong-current}, \textit{giant-wave}\}$. We equip these with the structural equations K_{surf}

$$
\begin{aligned}
\kappa_d &: \quad \textit{drowning} \leftrightarrow \textit{cramp} \vee \textit{submersion} \\
\kappa_c &: \quad \textit{cramp} \leftrightarrow \textit{strong-current} \vee \textit{jellyfish} \\
\kappa_s &: \quad \textit{submersion} \leftrightarrow \textit{giant-wave} \wedge \textit{strong-current} \\
\kappa_{bb} &: \quad \textit{broken-board} \leftrightarrow \textit{giant-wave}
\end{aligned}
$$

Figure 2 depicts the causal graph for this model. The background atoms of the model are drawn using dotted lines.

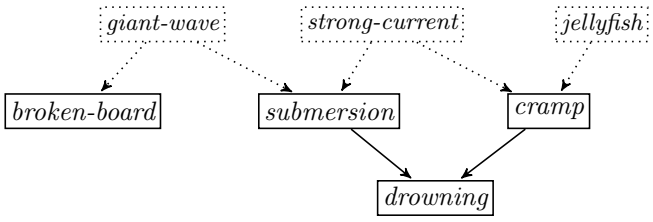


Fig. 2. Causal graph for Example 2.

We now define a *causal knowledge base* as a knowledge base, where the set of facts K is a causal model and the set of assumptions A is limited to assumptions about the background atoms in K.

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 $l \in \{u, \neg u\}$. We denote by $\overline{l}$ the assumption of the opposite, i. e., $\overline{u} = \neg u$ and $\overline{\neg u} = u$.*

Since the background variables are supposed to be independent, we restrict the background assumptions to be literals. This allows us to express three possible stances towards a background atom u: we can assume just u, just $\neg u$, or both. Assuming only u ($\neg u$) amounts to assuming that u is true (false), unless we have evidence to the contrary. On the other hand, if we assume both u and $\neg u$, this represents a state of uncertainty where we are willing to consider u to be true as well as false, depending on the evidence.

Example 3. To continue Example 2 we can now construct a causal KB $\Delta = (K_{surf}, A)$ by combining the causal model K_{surf} with the set of assumptions $A = \{\textit{jellyfish}, \textit{strong-current}, \neg\textit{strong-current}, \textit{giant-wave}\}$. Intuitively, this expresses that we assume a giant wave has happened and that there are dangerous jellyfish present, but are uncertain whether there is a strong current in the area.

Given a causal knowledge base $\Delta = (K, A)$, then Δ-*entailment* can be understood as the relation between observations and predictions, i. e., an observation ϕ Δ-entails some prediction ψ, denoted by $\phi \mathrel{|\!\sim_\Delta} \psi$, if the underlying causal model together with the observation ϕ entails the conclusion ψ. These predictions include causes as well as effects of the observation in accordance with the causal model K and the background assumptions A.

We now describe how we can transform a causal knowledge base into an argumentation framework and how to compute the Δ-entailment in that framework. For that, we adopt the approach by Cayrol et al. [4] to define an argument induced by a knowledge base $\Delta = (K, A)$. An induced argument is 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 some *conclusion* ψ. The attacks between the arguments are given by the undercut relation. We say that an argument *undercuts* another if the conclusion of the former is the negation of a premise of the latter.

Definition 7. *Let $\Delta = (K, A)$ be a causal knowledge base. We define the* AF induced by Δ*, denoted with $F(\Delta) = (\mathsf{Arg}_\Delta, \mathsf{R}_\Delta)$ as follows*

- *The set of* Δ-induced arguments Arg_Δ *is defined as all arguments of the form (Φ, ψ) such that $\psi \in \{u, \neg u \mid U(K) \cup V(K)\}$ and*
  - $\Phi \subseteq A$,
  - $\Phi \cup K \nvdash \bot$,

  - $\Phi \cup K \vdash \psi$, *and if* $\Psi \subset \Phi$ *then* $\Psi \cup K \nvdash \psi$.
- $(\Phi, \psi)\mathsf{R}_\Delta(\Phi', \psi')$, *iff for some* $\phi' \in \Phi'$ *we have* $\overline{\phi'} = \psi$.

As shown by Cayrol et al. [4], there is a one-to-one correspondence between the maximal K-consistent subsets of a knowledge base and the stable extensions of an AF induced according to Definition 7. Given a causal knowledge base $\Delta = (K, A)$, this allows us to answer the question of whether ϕ Δ-entails ψ by constructing the AF induced by $(K \cup \{\phi\}, A)$ and determining whether every stable extension contains at least one argument which concludes ψ.

Proposition 1. *Let* $\Delta = (K, A)$ *be a causal 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.

Example 4. We continue with the causal knowledge base $\Delta = (K_{surf}, A)$ from Example 3. Consider the question whether observing that the surfer has drowned entails that the drowning has been caused by submersion, i. e., consider the statment whether *drowning* $\mid\sim_\Delta$ *submersion*. Submersion and a cramp are the two possible causes of drowning. It depends on the background atoms which one was the actual cause of drowning. We determine the question and the explanation via the induced AF $F = F((K \cup \{\textit{drowning}\}, A))$, shown in Figure 3 (we only depict arguments relevant to the conclusion of submersion). The two stable extensions of this AF are $\{a_1, a_3\}$ and $\{a_2, a_4, a_5\}$. The argument a_4 concludes *submersion*, but is only included in one of the stable extensions. Thus, *drowning* does not entail *submersion*, given the background assumptions A.

Moreover, note that the statement *drowning* $\mid\sim_\Delta \neg$*submersion* does also not hold.

To conclude, we can say if we observe *drowning*, then *submersion* is a possible cause, but not necessary. The explanation for either case is then given by the corresponding stable extension containing the conclusion.

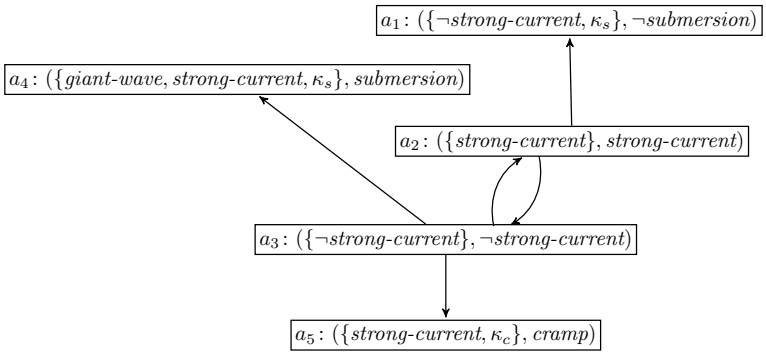


Fig. 3. The AF $F(K \cup \{\textit{drowning}\}, A)$ from Example 4.

3.2 Probabilistic Causal Reasoning

A *probabilistic causal model* [14] is defined as a causal model together with a probability assignment to every background atom. For some causal statement $\phi \mid\!\sim_{\Delta} \psi$, this allows us to determine exactly the probability that ψ holds given ϕ. As implied by Definition 8, we assume that the probabilities of the background atoms are independent, thus the causal model is considered *Markovian*.

Definition 8. *A* probabilistic causal model *is a pair* $\mathcal{C} = (K, \mathsf{P})$ *where* K *is a causal model and* $\mathsf{P} : U \to [0, 1]$ *is a probability assignment.*

Let $\mathcal{C} = (K, \mathsf{P})$ be a probabilistic causal model. A *causal state* $C \in 2^{U(K)}$ is essentially a specific configuration of the background atoms. So, if $u \in C$, then u is considered true in the state C, and otherwise u is false. We then define the *probability distribution* $P_{\mathcal{C}}$ over causal states (which correspond directly to the valuations of $U(K)$) as follows

$$P_{\mathcal{C}}(C) = \prod_{u \in C} \mathsf{P}(u) \prod_{u \in U \setminus C} (1 - \mathsf{P}(u)). \tag{1}$$

Note that the above defined function is indeed well-defined.

Proposition 2. *For any causal model* $\mathcal{C} = (K, \mathsf{P})$*, the probability distribution* $P_{\mathcal{C}}$ *sums up to* 1*.*

Example 5. Consider again the causal model K introduced in Example 2. The background atoms of K are *giant-wave* (g), *strong-current* (s) and *jellyfish* (j). We define the probability assignment P to the background atoms as follows: $\mathsf{P}(\textit{giant-wave}) = 0.8$, $\mathsf{P}(\textit{strong-current}) = 0.5$ and $\mathsf{P}(\textit{jellyfish}) = 0.2$. Then, for the probabilistic causal model $\mathcal{C} = (K, \mathsf{P})$ we compute the probability distribution of the causal states via Equation (1) as follows: $P_{\mathcal{C}}(gs\bar{j}) = P_{\mathcal{C}}(\overline{gsj}) = 0.32$, $P_{\mathcal{C}}(gsj) = P_{\mathcal{C}}(g\bar{s}j) = P_{\mathcal{C}}(\bar{g}s\bar{j}) = P_{\mathcal{C}}(\overline{gsj}) = 0.08$ and $P_{\mathcal{C}}(\bar{g}sj) = P_{\mathcal{C}}(\overline{gs}j) = 0.02$.

For a causal statement $\phi \mid\!\sim_{\mathcal{C}} \psi$ the probability that ψ is predicted to be true, given the observation ϕ is given as the conditional probability $P_{\mathcal{C}}(\psi \mid \phi)$ [14].

Example 6. Consider the causal statement *drowning* $\mid\!\sim_{\mathcal{C}}$ *submersion*. We compute the probability $P_{\mathcal{C}}(\textit{submersion}|\textit{drowning})$ (i. e., probability of submersion given that we observe drowning) using the standard causal model approach. Continuing Example 5, we construct the probability distribution over all valuations of the background atoms, and including all the explainable atoms, whose values are determined by the background atoms, see Table 1. Computing queries based on observations simply amounts to computing a conditional probability based on the probability distribution given above. Using the definition of conditional probability we get $P_{\mathcal{C}}(\textit{submersion}|\textit{drowning}) = P_{\mathcal{C}}(\textit{submersion} \wedge \textit{drowning})/P_{\mathcal{C}}(\textit{drowning}) = 0.4/0.6 = 2/3$. Thus, the probability of submersion given that we observe drowning is 2/3.

gsj	*broken-board*	*submersion*	*cramp*	*drowning*	Prob
000	0	0	0	0	0.08
001	0	0	1	1	0.02
010	0	0	1	1	0.08
011	0	0	1	1	0.02
100	1	0	0	0	0.32
101	1	0	1	1	0.08
110	1	1	1	1	0.32
111	1	1	1	1	0.08

Table 1. Partial probability distribution $P_{\mathcal{C}}$ from Example 6.

In order to determine the probability of a statement $\phi \mid\!\sim_{\mathcal{C}} \psi$, we induce a probabilistic argumentation framework PAF from the probabilistic causal model $\mathcal{C}$. For that we denote with $\mathsf{C}(\phi)$ the set of causal states in which the observation ϕ is true, defined as

$$\mathsf{C}(\phi) = \{C \in 2^{U(K)} \mid K \cup C \cup \{\neg u \mid u \notin C\} \vdash \phi\}.$$

Similar to before, an induced argument is a pair (Φ, ψ) consisting of a set of premises Φ and a conclusion ψ. The set of premises $\Phi \subseteq \{u, \neg u \mid u \in U(K)\}$ must be consistent with some causal state $C \in \mathsf{C}(\phi)$, i. e., the union of C and Φ is not contradictory, and is has to be the minimal K-consistent set to entail the conclusion ψ. The attacks of PAF are again given by the undercut relation.

We define $\mathsf{Arg}_{\mathcal{C}}(C)$ as the set of arguments consistent with a causal state $C \in 2^U$, i. e., $\mathsf{Arg}_{\mathcal{C}}(C) = \{(\Phi, \psi) \in \mathsf{Arg}_{\mathcal{C}} \mid \Phi \cup C \cup \{\neg u \mid u \notin C\} \nvdash \bot\}$, where $\mathsf{Arg}_{\mathcal{C}}$ is the set of induced arguments (see Def. 9). With that, the probability of a framework state S of the PAF is defined as the sum over the probabilities of all causal states C which are consistent with all arguments that are active in S.

Definition 9. *Let $\mathcal{C} = (K, \mathsf{P})$ be a probabilistic causal model. We define the* PAF induced by $\mathcal{C}$*, given the observation ϕ, denoted with $\mathsf{PAF}_{\mathcal{C}} = (F(\mathcal{C}), P_{\mathsf{AF}})$ with $F(\mathcal{C}) = (\mathsf{Arg}_{\mathcal{C}}, \mathsf{R}_{\mathcal{C}})$ as follows:*

- *The set of* $\mathcal{C}$-induced arguments $\mathsf{Arg}_{\mathcal{C}}$ *consists of all arguments (Φ, ψ), with $\Phi \subseteq \{u, \neg u \mid u \in U(K)\}$, such that*
  - *$\Phi \cup C \nvdash \bot$ for some $C \in \mathsf{C}(\phi)$,*
  - *$\Phi \cup K \nvdash \bot$,*
  - *$\Phi \cup K \vdash \psi$, and if $\Psi \subset \Phi$ then $\Psi \cup K \nvdash \psi$.*
- *The set of* $\mathcal{C}$-induced attacks $\mathsf{R}_{\mathcal{C}}$ *is defined via the undercut relation, i. e., an argument (Φ, ψ)* undercuts *an argument (Φ', ψ') iff for some $\phi' \in \Phi'$ we have $\phi' \equiv \overline{\psi}$.*

The probability distribution $P_{\mathsf{AF}} : 2^{\mathsf{Arg}} \to [0, 1]$ over framework states is given as

$$P_{\mathsf{AF}}(S) = \sum_{C \in \mathsf{C}(\phi, S)} P_{\mathcal{C}}(C).$$

$$\textit{where } \mathsf{C}(\phi, S) = \{C \in \mathsf{C}(\phi) \mid S = \mathsf{Arg}_{\mathcal{C}}(C)\}.$$

Note that the above defined probability distribution P_{AF} is indeed well-defined.

Proposition 3. *For any causal model $\mathcal{C} = (K, \mathsf{P})$ and observation ϕ, the probability distribution P_{AF} sums up to* 1.

Example 7. We continue Example 5. To determine the probability of *drowning* $\mid\!\sim_{\mathcal{C}}$ *submersion*, we construct the induced probabilistic argumentation framework $\mathsf{PAF}_{\mathcal{C}} = (F(\mathcal{C}), P_{AF})$, shown in Figure 4 (only arguments relevant to the query are depicted). The framework states with non-zero probability are described in Table 2. Each framework state corresponds to one or more causal state and consists of a subset of arguments for which we can determine whether all stable extensions conclude *submersion*. In this case, only the first framework state satisfies this.

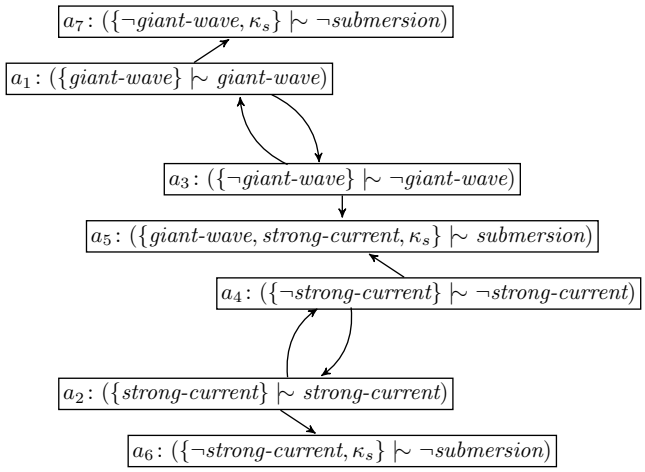


Fig. 4. The AF $F(K \cup \{\mathit{drowning}\})$ from Example 7.

$\mathsf{C}(\phi, S)$	$P_{\mathsf{AF}}(S)$	a_1	a_2	a_3	a_4	a_5	a_6	a_7	$S \vdash \phi$
gsj, $gs\bar{j}$	0.4	✓	✓			✓			yes
$g\bar{s}j$	0.08	✓			✓		✓		no
$\bar{g}sj$,$\bar{g}s\bar{j}$	0.1		✓	✓				✓	no
$\bar{g}\bar{s}j$	0.02			✓	✓		✓	✓	no

Table 2. The framework states of the induced $\mathsf{PAF}_{\mathcal{C}} = (F(K \cup \{\mathit{drowning}\}), P_{\mathsf{AF}})$ which correspond to some $C \in \mathsf{C}(\phi)$.

Let $\mathcal{C} = (K, \mathsf{P})$ be a probabilistic causal model and consider some causal statement $\phi \mid\!\sim_{\mathcal{C}} \psi$.

We can compute the probability that ψ holds given that ϕ is true via the induced probabilistic argumentation framework $\mathsf{PAF}_{\mathcal{C}} = (F(K \cup \{\phi\}), P_{\mathsf{AF}})$ as follows. With $\mathsf{S}_{[\psi=true]}$ we denote the set of framework states which entail the conclusion ψ, i. e., for which every stable extension of $\mathsf{PAF}_{\mathcal{C}}$ contains at least one argument with the conclusion ψ. In Pearl's standard causal model approach, the probability $P(\phi \mid\!\sim_{\mathcal{C}} \psi)$ is computed as the conditional probability $P_{\mathcal{C}}(\psi|\phi) = P_{\mathcal{C}}(\psi \wedge \phi)/P_{\mathcal{C}}(\phi)$. Analogously, in our framework the probability $P_{\mathcal{C}}(\psi \wedge \phi)$ amounts to the sum of probabilities over all framework states S that entail ψ, while the probability $P_{\mathcal{C}}(\phi)$ is the sum of probabilities over all causal states in which ϕ is true. Thus, the probability $P(\phi \mid\!\sim_{\mathcal{C}} \psi)$ is then computed as

$$P(\phi \mid\!\sim_{\mathcal{C}} \psi) = \frac{\sum\limits_{S \in \mathsf{S}_{[\psi=true]}} P_{\mathsf{AF}}(S)}{\sum\limits_{C \in \mathsf{C}(\phi)} P_{\mathcal{C}}(C)}. \tag{2}$$

Our main theorem below states that probabilistic argumentative reasoning amounts to the same results as Pearl's classical approach, with the added value of representing causal inference through argumentative reasoning.

Theorem 1. *Let $\mathcal{C} = (K, \mathsf{P})$ be a probabilistic causal model and $\phi \mid\!\sim_{\mathcal{C}} \psi$ is a causal statement. Then $P(\phi \mid\!\sim_{\mathcal{C}} \psi) = P_{\mathcal{C}}(\psi|\phi)$.*

In addition to the probability that the statement is true, the induced PAF also allows us to provide different types of explanations. We can, for example, provide an explanation for the most likely scenario under which the query holds. The same can be done for the situation under which the contrary is most likely to be true. Furthermore, we might also provide an explanation for the scenario in which both outcomes are possible.

Example 8. We continue Example 7. The probability of *drowning* $\mid\!\sim_{\mathcal{C}}$ *submersion* can then be computed via (2). Considering the framework states in Table 2, only one framework state with probability 0.4, corresponding to the causal states gsj and $gs\bar{j}$, entails the conclusion *submersion*. The sum of probabilities over the causal states that are consistent with drowning $\mathsf{C}(\textit{drowning})$ is 0.6. Thus the probability of submersion given that we observe drowning is $P(\textit{drowning} \mid\!\sim_{\mathcal{C}} \textit{submersion}) = 0.4/0.6 = 0.\overline{66}$. In terms of explainability, we have different angles to give an explanation based on the argumentation framework. A positive explanation would be that a giant wave and a strong current cause submersion. On the other hand, we can also say that the most likely reason against submersion is that there is no strong current which means no risk of submersion, as implied by the second framework state.

4 Counterfactual Reasoning

We consider first the *interventional statements* of the form

$$\text{if } v \text{ would be } x \text{ then } \psi \text{ would be true.} \tag{3}$$

The left side of an interventional statement consists of an *action* where the atom v is intervened on, i. e., we set v to the truth value x. It is important to note that this is different from simply observing v or $\neg v$. Performing the action of setting v to x means overriding the causal mechanism that usually determines v. For some causal model K, we denote with $K_{[v=x]}$ the causal model where the structural equation of κ_v is replaced with $v \leftrightarrow x$.

Definition 10. *Let K be a causal model, let $v \in V$ be an explainable atom, and let $x \in \{\top, \bot\}$. We denote by $K_{[v=x]}$ the causal model defined by*

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

Note that we perform the intervention on the causal model K itself, which means we can apply this intervention both to a causal knowledge base $\Delta = (K, A)$ as well as a probabilistic causal model $\mathcal{C} = (K, \mathsf{P})$, depending on whether we want to reason qualitatively or quantitatively. We will then also write $\Delta_{[v=x]}$ and $\mathcal{C}_{[v=x]}$ as a shortcut for $\Delta = (K_{[v=X]}, A)$ or $\mathcal{C} = (K_{[v=x]}, \mathsf{P})$ respectively.

A *counterfactual statement* is of the form

$$\text{given } \phi, \text{ if } v \text{ had been } x \text{ then } \psi \text{ would be true.} \tag{4}$$

Intuitively this means, if we observe ϕ and if v would have been x, then ψ would have been true. So we reason about a hypothetical or alternative scenario.

In [14], Pearl introduced two approaches to deal with counterfactual statements: a three-step procedure and the twin network method. We base our approach to counterfactual reasoning on the twin network approach. The general idea is to construct a *twin model* which consists of the actual causal model, representing the actual world, and a second model that represents the counterfactual world. Both of these worlds share the same background atoms, i. e., we have $U(K) = U(K^*)$, while for all explainable atoms $v \in V(K)$ we introduce a "counterfactual copy" $v^* \in V(K^*)$ in the counterfactual world.

Definition 11. *The* twin model *for a causal model K is the causal model K^* defined by*

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

Like for the intervention, we may also write Δ^* and $\mathcal{C}^*$ as a shortcut for $\Delta = (K^*, A)$ or $\mathcal{C} = (K^*, \mathsf{P})$ respectively.

First, consider the three-step procedure for evaluating counterfactual statements in a probabilistic causal model as described by Pearl [14].

Definition 12. *Given a probabilistic causal model $\mathcal{C} = (K, \mathsf{P})$, the truth of a counterfactual statement*

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

is determined by:

- *Step 1 (abduction) Update $P_{\mathcal{C}}$ by the evidence ϕ to obtain $P_{\mathcal{C}}(u \mid \phi)$.*

- *Step 2 (action) Modify K by the action $v = x$ to obtain $K_{[v=x]}$.*
- *Step 3 (prediction) Use the modified model $(K_{[v=x]}, P_{\mathcal{C}}(u \mid \phi))$ to compute the probability of ψ, i. e., $P_{\mathcal{C}}(\psi \mid \phi)$.*

The problem of this procedure lies in the abduction step, where we have to compute a probability distribution over configurations of the background atoms. This can be avoided by using the twin network method.

Consider a probabilistic causal model $\mathcal{C} = (K, \mathsf{P})$ and a counterfactual statement (4). Our argumentation-based approach consists of the following steps:

1. Compute the twin model $\mathcal{C}^* \cup \{\phi\}$ which includes the observation ϕ,
2. Perform the intervention $v^*{=}x$ on the counterfactual copy of v to obtain $\mathcal{C}^*_{[v^*=x]} \cup \{\phi\}$,
3. Construct the induced probabilistic AF $\mathsf{PAF}_{\mathcal{C}} = (F(\mathcal{C}^*_{[v^*=x]} \cup \{\phi\}), P_{\mathsf{AF}})$,
4. Determine the probability that ψ^* is true.

Note that the second and fourth step, representing action and prediction step of the standard three-step procedure, take place in the counterfactual world. For the third step we induce the probabilistic argumentation framework from $\mathcal{C}$ as described in Definition 9. The probability that ψ would have been true given ϕ, under the assumption that $v = x$, is calculated as the sum over the probabilities of all framework states $S \in \mathsf{S}_{[\psi^*=true]}$ for which every stable extension of the induced probabilistic argumentation framework of the twin model $\mathsf{PAF}_{\mathcal{C}} = (F((\mathcal{C} \cup \{\phi\})), P_{\mathsf{AF}})$ contains an argument with conclusion ψ^*.

Definition 13. *Let $\mathcal{C} = (K, \mathsf{P})$ be a probabilistic causal model. For the counterfactual statement $\phi \mid\!\sim_{\mathcal{C}^*_{[v^*=x]}} \psi^*$, the probability that ψ would have been true, given ϕ and assuming $v{=}x$, is computed as*

$$P(\phi \mid\!\sim_{\mathcal{C}^*_{[v^*=x]}} \psi^*) = \sum_{S \in \mathsf{S}_{[\psi^*=true]}} P_{\mathsf{AF}}(S).$$

The probabilistic argumentation-based twin network approach is equivalent to Pearl's standard three-step procedure.

Theorem 2. *Let $\mathcal{C} = (K, \mathsf{P})$ be a probabilistic causal model. Given a counterfactual statement $\phi \mid\!\sim_{\mathcal{C}^*_{[v^*=x]}} \psi^*$, we have that $P(\phi \mid\!\sim_{\mathcal{C}^*_{[v^*=x]}} \psi^*) = P_{\mathcal{C}}(\psi \mid \phi)$.*

5 Discussion

In this work, we extended our argumentation-based approach for defeasible causal and counterfactual reasoning from [1] to the probabilistic scenario. The intention of our approach is to bridge the gap from causal reasoning to formal argumentation. Our approach provides an argumentative representation of the causal mechanisms of the model in the context of a specific causal or counterfactual statement. In the literature, approaches for generating explanations for the (non-)acceptance of arguments in an argumentation framework have already

been proposed [5]. The work [8] introduces a new kind of semantics called *related admissibility* which computes sets of arguments that are related to a specific argument. These sets form the basis of different kinds of explanations for the argument. Based on the same idea, they also introduce *dispute forests* that can be used to explain the non-acceptance of an argument. Furthermore, [3] introduce a general framework for explanations in formal and structured argumentation. They define different kinds of explanations, for example, an explanation for or against an argument as well as evidence that supports or is incompatible with an argument. This approach is especially interesting since they also consider the structured argumentation formalism ASPIC$^+$ [15], which is very similar to how we induce argumentation frameworks from causal models in our approach.

There also exist other argumentation-based approaches in the literature that highlight the interest in explaining causal reasoning. For instance, the work [18] is concerned with Bayesian networks and introduces the notion of a support graph that makes d-separation explicit, which eliminates circular causal structures and helps to explain interdependent causes.

In a recent work [16], Rago et al. introduce an approach for generating bipolar argumentation frameworks from causal models in the sense of Pearl. They create so called explanation moulds, that reinterpret desirable properties of semantics of argumentation frameworks. In their approach, they interpret causal atoms directly as arguments and causes contribute positively or negatively towards arguments via attack and support relations, respectively.

6 Limitations

In the following we discuss the limitations of the approach introduced in this work. First, our approach is built on classical propositional logic. That means, while being relatively easy to understand, the expressiveness is limited when compared to other higher-order logics.

Our approach is only focused on the actual reasoning with a causal model. That means we consider the underlying causal model to be given and crafted by experts and we assume that the given relations between the variables are indeed causal and not merely correlations.

Furthermore, the computational complexity of this approach to causal reasoning is quite high. Our approach relies on deciding whether some of the arguments are skeptically accepted in the induced argumentation framework. This problem is naturally difficult and in the case of the stable semantics that we use it has been shown to be NP-complete [7]. In addition to that, when considering probabilistic causal reasoning we have to potentially consider exponentially many framework states (wrt. the set of background variables) which increases the complexity of the approach significantly.

Finally, it should also be noted that our approach is to be understood as a groundwork for making causal reasoning explainable. Meaning the induced (probabilistic) argumentation framework can be the basis for crafting human understandable explanations. How exactly these explanations should look like,

is left for future work and some interesting approaches for that matter have already been highlighted in Section 5. Especially in the case of probabilistic causal reasoning this is even more difficult since the probabilistic aspect has to be somehow incorporated into the explanations.

7 Conclusion

We extended our approach for argumentation-based causal reasoning from [1] to deal with probabilistic causal models. For that, we model probabilistic causal reasoning in a probabilistic argumentation framework and compute the probability that the statement is true by reasoning in the framework states. Furthermore, we showed that our approach can also be used for reasoning with counterfactuals, by adapting Pearl's twin network method. Besides computing the probability, the generated probabilistic argumentation framework can be used as the basis for creating explanations of the underlying causal mechanisms of the model in the context of the statement, since it provides both arguments supporting the prediction as well as arguments that refute the prediction.

Future work includes determining structural properties of the generated (probabilistic) AFs and looking into concrete application scenarios to investigate the capabilities of our approach.

Acknowledgement

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

References

1. Bengel, L., Blümel, L., Rienstra, T., Thimm, M.: Argumentation-based causal and counterfactual reasoning. In: 1st International Workshop on Argumentation for eXplainable AI, Cardiff. CEUR Workshop Proceedings, vol. 3209 (2022)
2. Bochman, A., Lifschitz, V.: Pearl's causality in a logical setting. In: Bonet, B., Koenig, S. (eds.) Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015. pp. 1446–1452. AAAI Press (2015)
3. Borg, A., Bex, F.: A basic framework for explanations in argumentation. IEEE Intelligent Systems **36**(2), 25–35 (2021)
4. Cayrol, C.: On the relation between argumentation and non-monotonic coherence-based entailment. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95. pp. 1443–1448 (1995)
5. Čyras, K., Rago, A., Albini, E., Baroni, P., Toni, F.: Argumentative xai: a survey. arXiv preprint arXiv:2105.11266 (2021)
6. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–358 (1995)
7. Dunne, P.E., Wooldridge, M.: Complexity of abstract argumentation. Argumentation in artificial intelligence pp. 85–104 (2009)

8. Fan, X., Toni, F.: On computing explanations in argumentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 29 (2015)
9. Hunter, A.: A probabilistic approach to modelling uncertain logical arguments. International Journal of Approximate Reasoning **54**(1), 47–81 (2013)
10. Hunter, A., Polberg, S., Potyka, N., Rienstra, T., Thimm, M.: Probabilistic argumentation: A survey. Handbook of Formal Argumentation **2**, 397–441 (2021)
11. Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: Explainable reinforcement learning through a causal lens. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 2493–2500 (2020)
12. Manor, N.R.R., Rescher, N.: On inference from inconsistent premises. Theory and Decision **1**, 179–219 (1970)
13. Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N.: Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review pp. 1–66 (2022)
14. Pearl, J.: Causality: models, reasoning and inference, vol. 29. Cambridge University Press (2000)
15. Prakken, H.: An abstract framework for argumentation with structured arguments. Argument & Computation **1**(2), 93–124 (2010)
16. Rago, A., Russo, F., Albini, E., Baroni, P., Toni, F.: Forging argumentative explanations from causal models. In: Proceedings of the 5th Workshop on Advances in Argumentation in Artificial Intelligence 2021). CEUR Workshop Proceedings, vol. 3086. CEUR-WS.org (2021)
17. Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G.E., Irvine, J., Le, Q., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S.L., Chou, K., Pearson, M., Madabushi, S., Shah, N.H., Butte, A.J., Howell, M.D., Cui, C., Corrado, G.S., Dean, J.: Scalable and accurate deep learning with electronic health records. npj Digit. Medicine **1** (2018)
18. Timmer, S.T., Meyer, J.J.C., Prakken, H., Renooij, S., Verheij, B.: Explaining bayesian networks using argumentation. In: European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty. pp. 83–92. Springer (2015)