A Non-Interventionist Approach to Causal Reasoning based on Lewisian Counterfactuals

Carlos Aguilera-Ventura[1], **Xinghan Liu**[2], **Emiliano Lorini**[1] and **Dmitry Rozplokhas**[2]
[1]IRIT, CNRS, Toulouse University, France
[2]TU Wien, Austria

Abstract

We present a computationally grounded semantics for counterfactual conditionals in which i) the state in a model is decomposed into two elements: a propositional valuation and a causal base in propositional form that represents the causal information available at the state; and ii) the comparative similarity relation between states is computed from the states' two components. We show that, by means of our semantics, we can elegantly formalize the notion of actual cause without recurring to the primitive notion of intervention. Furthermore, we provide a succinct formulation of the model checking problem for a language of counterfactual conditionals in our semantics. We show that this problem is PSPACE-complete and provide a reduction of it into QBF that can be used for automatic verification of causal properties.

1 Introduction

The theory of counterfactual conditionals is one of the cornerstones of modern analytic philosophy since the seminal works of Lewis [Lewis, 1973] and Stalnaker [Stalnaker, 1968]. It has been recently applied in the field of explainable AI to explain the decisions and predictions of artificial intelligent systems [Mittelstadt *et al.*, 2019; Mothilal *et al.*, 2020; Sokol and Flach, 2019; Kenny and Keane, 2021]. The theory of counterfactual conditionals is intimaly related to the theory of causation, and the logic of counterfactual reasoning to the logic of causal reasoning.

As an alternative to Lewis' logic of counterfactual conditionals, Halpern and Pearl [Galles and Pearl, 1998; Halpern, 2000; Halpern and Pearl, 2005a] have introduced a logic of interventionist conditionals as a special kind of counterfactual conditionals in which the antecedent of the conditional is an intervention on a causal model. Unlike Lewis who interprets his logic of counterfactual conditionals by means of abstract comparative similarity relations between possible worlds, Halpern and Pearl interpret their logic by means of a structural equation model (SEM) semantics. The fact that counterfactual conditionals are more general than interventionist conditionals is emphasized by Pearl [Pearl, 2009] who introduced the notion of the three-layer 'causal ladder' (or hierarchy) in which counterfactuals are at the top of of the hierarchy, interventions are in the middle and mere associations are at the bottom layer. Counterfactuals are placed at the top of the ladder since they subsume interventions, in the sense that an interventional question can be formulated as a special kind of counterfactual question but not vice versa.

The theory and the corresponding logic of interventionist conditionals have become the dominant paradigm in the field of formal causal reasoning in AI in the recent years, while Lewisian conditionals are much less prominent. A variety of causal concepts have been formalized using interventionist conditionals including actual cause [Halpern, 2016; Beckers, 2021; Halpern, 2008; Beckers and Vennekens, 2017], NESS (Necessary Element of a Sufficient Set) cause [Beckers, 2021; Halpern, 2008], contrastive cause [Miller, 2021], explanation [Halpern and Pearl, 2005b; Woodward, 2003; Woodward and Hitchcock, 2003], responsibility and blame [Chockler and Halpern, 2004; Halpern and Kleiman-Weiner, 2018; Alechina *et al.*, 2017], discrimination [Chockler and Halpern, 2022] and harm [Beckers *et al.*, 2022]. Thus, the general impression we get from these works is that the primitive notion of intervention is necessary to define and formalize such causal concepts. In this paper, we show that this impression is not well-founded. In particular, we prove that the notion of actual cause, one of the central pillars in the modern theory of causality, can be naturally and elegantly formalized in a language of counterfactual conditionals in Lewis' style without recurring to the notion of intervention.

To obtain our result, we rely on the computationally grounded semantics for causal reasoning recently proposed in [Lorini, 2023; de Lima and Lorini, 2024]. There is a crucial difference between Lewis' original semantics for counterfactual conditionals and the semantics on which we rely. In the former, the notion of possible state (or world) in a model is undecomposed and the comparative similarity relation between states used to interpret counterfactuals is abstract. In the latter, a state is decomposed into two elements: i) a propositional valuation, and ii) a causal base in propositional form that represents the causal information available at the state. Moreover, the comparative similarity relation is grounded in and computed from the states' two components. In this sense, it is a two-dimensional semantics for counterfactual conditionals. Specifically, according to this semantics, a state S' is considered at least as similar to a state S as a state S'' if i) the causal information the state S' shares with the state S is at least as much as the causal information the state S'' shares with the state S, and ii) S'' differs from S with respect to the truth values of propositional atoms at least as much as S' differs from S.

Our semantics offers greater flexibility than the abstract Lewisian semantics and allows us to give a precise interpretation of Lewis' vague concept of a 'small miracle' [Lewis, 1979]. Lewis uses this concept to distinguish backtracking from non-backtracking counterfactuals. Roughly speaking, according to Lewis, in a backtracking counterfactual only the propositional atoms representing the initial conditions can be changed to satisfy the antecedent of the conditional, while the causal laws are kept fixed. On the contrary, in a non-backtracking counterfactual, the causal laws can be changed by imagining 'small miracles'. According to the two-dimensional semantics we use, a 'small miracle' is nothing but a *minimal change of a causal base that can possibly occur to satisfy the antecedent of a conditional*.

The paper is structured as follows. In Section 2 we discuss some work that is directly related to our work. In Section 3 we present the formal framework: the two-dimensional semantics, the language of counterfactual conditionals and its interpretation over it, and a list of interesting validities for this language. Section 4 presents the main conceptual result of the paper. After some formal preliminaries introducing the notion of equational state, we prove a theorem highlighting that the notion of actual cause, as defined in [Halpern, 2015] using the notion of intervention, can be equivalently defined in our language of Lewisian counterfactuals without recurring to interventions. Section 5 is devoted to the computational aspects of our novel semantic approach to counterfactual conditionals. We provide a succinct formulation of the model checking problem for the language of counterfactual conditionals in our semantics. With 'succinct' we mean that the model with respect to which a formula has to be checked is not given explicitly with its set of possible worlds and its comparative similarity relations, but it is given in a compact form. We show that this problem is PSPACE-complete and provide a reduction of it into QBF that can be used for automatic verification of causal properties. As far as we know, nobody before us provided a succinct formulation of the model checking problem for Lewis' logic of counterfactual conditionals and a tight complexity result for this problem.

The extended version of this paper with all proofs is available in https://arxiv.org/pdf/2505.12972.

2 Related Work

The connection between the logic of interventionist conditionals and Lewis' logic of counterfactual conditionals was studied in [Galles and Pearl, 1998] and more recently in [Zhang, 2013]. Galles & Pearl show how a comparative similarity relation between possible worlds can be computed by a means of interventions: a first world is more similar to a second world than a third world is if it takes less local interventions to transform the first world into the second world than to transform the third world into the second world. As noticed by Zhang, the semantics of counterfactual conditionals based on selection functions in Stalnaker's style can also be reconstructed by means of interventions: the function selects for each intervention the solutions of the underlying causal model produced by it, as the closest worlds to the actual one relative to the intervention. Zhang studies the subclass of causal models, the so-called *solution-conservative* causal models, for which the principles of the logic of interventionist conditionals that correspond to the axioms of Lewis' logic of counterfactual conditionals are valid. However, Galles & Pearl's and Zhang's approach is fundamentally different from our approach. They focus on the logic of interventionist conditionals and aim to elucidate the relation with Lewis' logic. We focus on counterfactual conditionals and get rid of interventions. We show that the notion of actual cause has a natural and elegant interpretation in the logic of counterfactual conditionals that do not require the notion of intervention.

A recent analysis of the distinction between backtracking and non-backtracking counterfactuals in an interventionist setting was given in [von Kügelgen *et al.*, 2023]. This semantic account of non-backtracking counterfactuals is fundamentally different from ours. Following Pearl [Pearl, 2009], they make the concept of non-backtracking counterfactual conditional coincide with the concept of interventionist conditional and the concept of 'small miracle' with the concept of intervention. As pointed out above, our interpretation of Lewis' concept of a 'small miracle' does not rely on the concept of intervention but rather on the concept of minimal change of a causal base.

Alternative semantics for actual causality based on the situation calculus (SC) have also been proposed. Batusov and Soutchanski [Batusov and Soutchanski, 2018] formalize actual causality using atemporal SC action theories with sequential actions. Khan and Lespérance [Khan and Lespérance, 2021] extend causal reasoning to epistemic contexts involving incomplete information and multiagent settings, analyzing how agents acquire knowledge of actual causes through actions and sensing.

Last but not least, it is worth mentioning the work on the connection between counterfactuals and causal rules in the framework of causal calculus presented in [Bochman, 2018; Bochman, 2021]. We share with Bochman and previous work in [Lorini, 2023; de Lima and Lorini, 2024] the idea of expressing causal information through causal rules expressed in propositional form, as an alternative to the SEM semantics of Halpern and Pearl and to the causal team semantics introduced in [Barbero and Sandu, 2021].

3 Formal Framework

In this section, we first present the two-dimensional semantics for counterfactual conditionals. Then, we introduce a language that supports reasoning about propositional facts, information in a causal base and counterfactuals. We show how the language can be interpreted using the two-dimensional semantics. Finally, we discuss some of its formal properties in relation to Lewis' logic.

3.1 Semantics

In [Lorini, 2023] a rule-based semantics for causal reasoning is presented. The main feature of the semantics is its two-dimensional nature: one dimension representing the actual environment, and the other dimension representing the causal information. In this section, we extend this semantics with comparative similarity relations to be able to interpret counterfactual conditionals.

Let $\mathbb{P}$ be an infinite countable set of atomic propositions whose elements are denoted $p, q, \ldots$ We note $\mathcal{L}_{\mathsf{PROP}}(\mathbb{P})$, or simply $\mathcal{L}_{\mathsf{PROP}}$, the propositional language built from $\mathbb{P}$. Elements of $\mathcal{L}_{\mathsf{PROP}}$ are denoted $\omega, \omega', \ldots$ Given $\omega \in \mathcal{L}_{\mathsf{PROP}}$, we note with $\mathbb{P}(\omega)$ the set of atomic propositions occurring in ω. Moreover, if $X \subseteq \mathcal{L}_{\mathsf{PROP}}$ then $\mathbb{P}(X) = \bigcup_{\omega \in X} \mathbb{P}(\omega)$.

The following definition introduces the concept of state, namely, a causal base supplemented with a propositional valuation that is compatible with it.

Definition 1 (State)**.** *A state is a pair $S = (C, V)$, where $C \subseteq \mathcal{L}_{\mathsf{PROP}}$ is a causal base, and $V \subseteq \mathbb{P}$ is a valuation s.t. $\forall \omega \in C, V \models \omega$. The set of all states is denoted by $\mathbf{S}$. A state $S = (C, V)$ is said to be finite if both C and V are finite.*

The propositional valuation V represents the actual environment, while C represents the base of causal information (viz. the causal base). It is assumed that the former is compatible with the latter, that is, if ω is included in the actual causal base (i.e., $\omega \in C$) then it should be true in the actual environment (i.e., $V \models \omega$). We let super- and subscripts to be inherited, e.g., S^* always stands for (C^*, V^*).

A model is nothing but a state supplemented with a set of states that includes it.

Definition 2 (Model)**.** *A model is a pair (S, U) such that $S \in U \subseteq \mathbf{S}$. The set of models is denoted $\mathbf{M}$.*

The component U is called *context* (or *universe*) of interpretation. We call $(S, \mathbf{S})$ a universal model (i.e., a model including all possible states). For notational convenience, we simply write S instead $(S, \mathbf{S})$ to denote a universal model.

Let us illustrate the previous notion of model with the help of an example.

Example 1 (Videogame)**.** *Consider a virtual character controlled by a video gamer using three keyboard keys. Each configuration of these keys corresponds to a specific causal base, which determines the action the virtual character will perform depending on which key is activated by the gamer. Assume that three actions are possible: 'move forward' (fo), 'move backward' (ba), and 'jump' (ju). Suppose that:*

i) the controls are configured such that activating key 1 (ac_1) causes the character to move forward; activating key 2 (ac_2) causes it to move backward; and activating key 3 (ac_3) causes it to jump;

ii) in the actual situation, no key is activated and the character remains stationary;

iii) a hard constraint in the game prevents the gamer from activating more than one key at the same time.

So, according to hypotheses i), ii) and iii), we are in a model (S_0, U_0) with $S_0 = (C_0, V_0)$ such that

$$C_0 = \{ac_1 \to fo, ac_2 \to ba, ac_3 \to ju\},$$
$$V_0 = \emptyset,$$
$$U_0 = \{(C', V') \in \mathbf{S} \;:\; V' \models \bigwedge_{\substack{x,y \in \{1,2,3\}\,: \\ x \neq y}} (ac_x \to \neg ac_y)\}.$$

We define the following comparative similarity relation between states.

Definition 3 (Similarity relation between states)**.** *Let $S = (C, V), S' = (C', V'), S'' = (C'', V'') \in \mathbf{S}$. We say that state S' is at least as similar to state S as state S'' is, denoted $S'' \preceq_S S'$, if*

$$(C \cap C'') \subseteq (C \cap C') \text{ and } (V \Delta V') \subseteq (V \Delta V''),$$

where Δ stands for symmetric difference.

According to the previous definition, state S' is at least as similar to state S as state S'' is if i) the causal information that S'' shares with S is included in the causal information that S' shares with S, and ii) the environment of S'' differs from the environment of S at least as much as the environment of S' differs from the environment of S. The reason why the similarity relation uses 'set-inclusion' for the causal part and 'symmetric difference' for the propositional part is that the *causal* similarity between two states is determined by the information that is shared by their causal bases, while their *propositional* similarity is determined by the set of atomic propositions whose truth values are the same in their propositional valuations.

3.2 Language

The following definition introduces our modal language for causal reasoning.

Definition 4 (Language)**.** *We structure the language in two layers:*

$$\mathcal{L}_0 \quad \stackrel{def}{=} \quad \alpha ::= p \mid \top \mid \neg\alpha \mid \alpha \wedge \alpha \mid \triangle\omega,$$
$$\mathcal{L} \quad \stackrel{def}{=} \quad \varphi ::= \alpha \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \mid \varphi \mathrel{\Box\!\!\rightarrow} \varphi,$$

where p ranges over $\mathbb{P}$ and ω over $\mathcal{L}_{\mathsf{PROP}}$. The boolean constructs $\bot$, $\vee$, $\to$, and $\leftrightarrow$ are defined in the standard way as abbreviations.

We call $\mathcal{L}_0$ the language of causal information and $\mathcal{L}$ the language of causal counterfactual conditionals. Formula $\triangle\omega$ is read "it is causally relevant that ω" or "ω is a causal information of the actual state". Formula $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ is read "if φ were true, ψ would be true". Its dual $\varphi \mathrel{\Diamond\!\!\rightarrow} \psi := \neg(\varphi \mathrel{\Box\!\!\rightarrow} \neg\psi)$ is read "if φ were true, ψ might be false".

The following definition introduces the satisfaction relation between models and formulas of the language $\mathcal{L}$. (We omit semantic interpretations for the boolean connectives $\neg$, $\wedge$ and for $\top$ since they are defined in the usual way.)

Definition 5 (Satisfaction relation)**.** *Let $(S, U) \in \mathbf{M}$ with $S = (C, V)$. Then, (boolean cases are omitted)*

$(S, U) \models p$ *iff* $p \in V$,

$(S, U) \models \triangle\omega$ *iff* $\omega \in C$,

$(S, U) \models \varphi \mathrel{\Box\!\!\rightarrow} \psi$ *iff* *for all* $S' \in \texttt{Closest}(\varphi, S, U)$, $(S', U) \models \psi$, *where*

$$\texttt{Closest}(\varphi,S,U) = \Big\{ S' \in U : (S',U) \models \varphi \text{ and } \nexists S'' \in U \text{ such that } (S'',U) \models \varphi \text{ and } S' \prec_S S'' \Big\},$$

and $S' \prec_S S''$ *iff* $S' \preceq_S S''$ *and* $S'' \not\preceq_S S'$.

The set $\texttt{Closest}(\varphi,S,U)$ is the set of φ-closest states to state S relative to the context U.

The formula $\triangle\omega$ has the expected set-theoretic interpretation: it is causally relevant that ω iff the propositional formula ω is included in the actual causal base. The counterfactual conditional $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ holds at model (S,U) if φ were true, ψ would be true iff all φ-closest states to state S relative to the context U satisfy ψ.

Let us go back to Example 1 of the videogame to illustrate the semantic interpretation of formulas.

Example 2 (Videogame continued)**.** *It is easy to verify that at model* (S_0,U_0) *i) if the gamer activated key* 3*, the virtual character might jump, and ii) if the gamer activated key* 3 *without changing the causal rule relating key* 3 *to the jumping action, the virtual character would jump, that is,*

$$(S_0,U_0) \models (ac_3 \mathrel{\Diamond\!\!\rightarrow} ju) \wedge \Big(\big(ac_3 \wedge \triangle(ac_3 \rightarrow ju) \big) \mathrel{\Box\!\!\rightarrow} ju \Big).$$

Recall that we write $S \models \varphi$ instead of $(S,\mathbf{S}) \models \varphi$ for notational convenience. We say that a formula $\varphi \in \mathcal{L}(\mathbb{P})$ is valid, denoted by $\models \varphi$, if $(S,U) \models \varphi$ for every model $(S,U) \in \mathbf{M}$. We say φ is satisfiable if $\neg\varphi$ is not valid.

3.3 Some Properties

The following proposition highlights some interesting properties of our counterfactual conditionals.

Proposition 1. *Let* $\varphi,\psi \in \mathcal{L}$, $\omega \in \mathcal{L}_{\mathsf{PROP}}$ *and* $p \in \mathbb{P}$. *We have the following validities:*

$$\models \varphi \mathrel{\Box\!\!\rightarrow} \varphi \tag{1}$$

$$\models (\varphi \mathrel{\Box\!\!\rightarrow} \psi) \rightarrow (\varphi \rightarrow \psi) \tag{2}$$

$$\models \big(\varphi \mathrel{\Box\!\!\rightarrow} \chi \wedge \psi \mathrel{\Box\!\!\rightarrow} \chi\big) \rightarrow (\varphi \vee \psi) \mathrel{\Box\!\!\rightarrow} \chi \tag{3}$$

$$\models \big(p \wedge (\varphi \mathrel{\Box\!\!\rightarrow} \psi)\big) \rightarrow (\varphi \wedge p) \mathrel{\Box\!\!\rightarrow} \psi \tag{4}$$

$$\models \big(\neg p \wedge (\varphi \mathrel{\Box\!\!\rightarrow} \psi)\big) \rightarrow (\varphi \wedge \neg p) \mathrel{\Box\!\!\rightarrow} \psi \tag{5}$$

$$\models \big(\triangle\omega \wedge (\varphi \mathrel{\Box\!\!\rightarrow} \psi)\big) \rightarrow (\varphi \wedge \triangle\omega) \mathrel{\Box\!\!\rightarrow} \psi \tag{6}$$

$$\models \triangle\omega \mathrel{\Box\!\!\rightarrow} \omega \tag{7}$$

The first three validities can be proven straightforwardly. The validity (1) is standard in conditional logics. The validity (2) is called *weak centering* in the literature [Lewis, 1973]. The name comes from its semantic condition, namely if $(S,U) \models \varphi$, then $S \in \texttt{Closest}(\varphi,S,U)$. However, the property *strong centering*, i.e., if $(S,U) \models \varphi$ then $\{S\} = \texttt{Closest}(\varphi,S,U)$, does not hold. A counterexample would be

$$U = \{S,S'\} \text{ with } S = \{\emptyset,\{p\}\} \text{ and } S' = \{\{p\},\{p\}\}.$$

We have $\{S\} \neq \texttt{Closest}(p,S,U) = U$, albeit $(S,U) \models p$. The validity (3) comes from the fact that the comparative similarity relation $\preceq_S$ of Definition 3 is a partial preorder.

The validities (4), (5) and (6) are of particular interest since they highlight the interaction between counterfactual conditionals, propositional atoms and causal information. If $\preceq_S$ were a total preorder, the formula

$$\big((\varphi \mathrel{\Box\!\!\rightarrow} \psi) \wedge (\varphi \mathrel{\Diamond\!\!\rightarrow} \chi)\big) \rightarrow (\varphi \wedge \chi) \mathrel{\Box\!\!\rightarrow} \psi$$

would be valid. This formula is an axiom of Lewis' V-logics that relates to many axioms/postulates in other fields, e.g., the last postulate in AGM theory [Alchourrón *et al.*, 1985], and *rational monotonicity* (RM) in non-monotonic reasoning [Kraus *et al.*, 1990]. Since $\preceq_S$ is not total, RM is not valid here. Nevertheless, the validities (4), (5) and (6) indicate that our semantics is monotonic under cumulation of true propositional atoms and their negation, and of actual causal information. Finally, the validity (7) highlights the interaction between causal information and counterfactual conditionals, and comes from the validity (1) and the validity of $\triangle\omega \rightarrow \omega$.

4 Actual Cause

In this section, we turn to actual cause. We first provide some preliminary notions, the notion of equational state and the notion of intervention, that are needed to define actual cause in Halpern & Pearl's sense. We focus on the most recent interventionist definition of actual cause given in [Halpern, 2008]. The section culminates with a theorem showing that Halpern's notion of actual cause can be equivalently formulated in our language of counterfactual conditionals without interventions.

4.1 Equational States

We consider a subclass of states in which, in line with the structural equational modeling (SEM) approach to causality, causal information is represented in equational form.

An equational formula for a proposition p is a propositional formula of the form $p \leftrightarrow \omega$ which unambiguously specifies the truth value of p using a propositional formula ω made of propositions other than p, with $\leftrightarrow$ the usual biconditional boolean connective "if and only if". We note $\mathcal{L}_{\mathsf{EQ}}$ the corresponding set of equational formulas:

$$\mathcal{L}_{\mathsf{EQ}} = \Big\{ p \leftrightarrow \omega : p \in \mathbb{P}, \omega \in \mathcal{L}_{\mathsf{PROP}}; \text{ and } p \notin \mathbb{P}(\omega) \Big\}.$$

For every $p \in \mathbb{P}$, $\mathcal{L}_{\mathsf{EQ}}(p)$ is the set of equational formulas for p. For notational convenience, elements of $\mathcal{L}_{\mathsf{EQ}}$ are also denoted $\epsilon, \epsilon', \ldots$

An equational state is a special kind of state whose causal base is a finite set of equational formulas.

Definition 6 (Equational state)**.** *An equational state is a state* $S = (C,V)$*, with* $C \subseteq \mathcal{L}_{\mathsf{EQ}}$ *finite, and such that*

$$\forall p \in \mathbb{P}, \forall p \leftrightarrow \omega, p \leftrightarrow \omega' \in C, \omega = \omega'.$$

The set of equational states is denoted by $\mathbf{S}_{Eq}$.

According to the previous definition, the causal base of an equational state should contain at most one equational formula for each atomic proposition.

From an equational state, it is straightforward to extract a a set of endogenous variables and a set of exogenous ones. A variable is endogenous if there is an equational formula for it in the actual causal base, it is exogenous if it appears in the actual causal base but there is no equational formula for it.

Definition 7 (Exogenous and endogenous variables)**.** *Let $S = (C, V)$ be an equational state. Its set of exogenous variables $exo(S)$ and its set of endogenous variables $end(S)$ are defined, as follows:*

$$
\begin{aligned}
end(S) =& \big\{p \in \mathbb{P}(C) : \exists\omega \in \mathcal{L}_{\mathsf{PROP}}\big(\mathbb{P} \setminus \{p\}\big) \textit{ such that} \\
& \quad p \leftrightarrow \omega \in C\big\}, \\
exo(S) =& \mathbb{P}(C) \setminus end(S).
\end{aligned}
$$

From an equational state it is also possible to extract its graphical counterpart. Specifically, given an equational state $S = (C, V) \in \mathbf{S}_{Eq}$, we can extract the causal graph $G_S = \big(N_S, \mathcal{P}_S\big)$ with $N_S = \mathbb{P}^+(C)$ and where the *causal parent* function $\mathcal{P}_S \colon N_S \longrightarrow 2^N$ is defined as follows, for every $p \in N_S$:

$$
\begin{aligned}
&(i)\ \mathcal{P}_S(p) = \mathbb{P}^+(\omega) \text{ if } p \leftrightarrow \omega \in C, \\
&(ii)\ \mathcal{P}_S(p) = \emptyset \text{ if } \mathcal{L}_{\mathsf{EQ}}(p) \cap C = \emptyset,
\end{aligned}
$$

where $\mathbb{P}^+(\omega) = \mathbb{P}(\omega) \cup \{\top\}$ if $\top$ occurs in ω, $\mathbb{P}^+(\omega) = \mathbb{P}(\omega)$ if $\top$ does not occur in ω, and $\mathbb{P}^+(C) = \bigcup_{\omega \in C} \mathbb{P}^+(\omega)$. Note that if $p \in N_S$ then, $\mathcal{P}_S(p) = \emptyset$ iff $p \in exo(S)$.

The following example is a classic in the literature on formal models of causality. We use it to illustrate the previous definition.

Example 3. *Suzy and Billy decide to throw a rock simultaneously, aiming at the bottle. Suzy is a bit faster, so her rock breaks the bottle, not Billy's. However, Billy is just as accurate as Suzy: had she not thrown, Billy's rock would have shattered the bottle shortly after. This leads to the following causal structure: i) Suzy throws her rock (st) iff she decides to do so (sd), ii) Billy throws his rock (bt) iff he decides to do so (bd), iii) Suzy hits the bottle (sh) if and only if she throws her rock (st), iv) Billy hits the bottle (bh) if and only if he throws his rock (bt) while Suzy does not hit the bottle (¬sh), v) the bottle is shattered (bs) if and only if either Billy or Suzy hits it. The actual state $S_0 = (C_0, V_0)$ is described as follows:*

$$
\begin{aligned}
C_0 =& \big\{st \leftrightarrow sd, bt \leftrightarrow bd, \\
& \quad sh \leftrightarrow st, bh \leftrightarrow (bt \wedge \neg sh), bs \leftrightarrow (sh \vee bh)\big\}, \\
V_0 =& \{sd, bd, st, bt, sh, bs\}.
\end{aligned}
$$

The causal graph extracted from it is given in Figure 1.

$$
\begin{tikzcd}
sd \arrow[r] & st \arrow[r] & sh \arrow[dd] \arrow[dr] & \\
& & & bs \\
bd \arrow[r] & bt \arrow[r] & bh \arrow[ur] &
\end{tikzcd}
$$

Figure 1: Causal graph

4.2 Interventions

We conceive an intervention as a possibly empty finite set of equational formulas of type $p \leftrightarrow \top$ or $p \leftrightarrow \bot$ with at most one equational formula for each variable. We define the set of interventions as follows:

$$
\begin{aligned}
Int =& \big\{\{p_1 \leftrightarrow \tau_1, \ldots, p_k \leftrightarrow \tau_k\} : \forall 1 \le k', k'' \le k, \\
& \quad \text{if } k' \neq k'' \text{ then } p_{k'} \neq p_{k''} \text{ and } \tau_1, \ldots, \tau_k \in \{\top, \bot\}\big\}.
\end{aligned}
$$

Elements of Int are denoted $E, E', \ldots$. Given $E \in Int$, let

$$
\widehat{E} =_{def} \bigwedge_{p \leftrightarrow \top \in E} p \wedge \bigwedge_{p \leftrightarrow \bot \in E} \neg p.
$$

For every finite set of atomic propositions $Z \subseteq \mathbb{P}$, we note Int_Z the set of interventions for Z, that is,

$$
\begin{aligned}
Int_Z =& \big\{E \in Int : (\forall p \in Z, p \leftrightarrow \top \in E \text{ or } p \leftrightarrow \bot \in E) \\
& \quad \text{and } (\forall p \notin Z, p \leftrightarrow \top \notin E \text{ and } p \leftrightarrow \bot \notin E)\big\}.
\end{aligned}
$$

From a semantic point of view, an intervention $\{p_1 \leftrightarrow \tau_1, \ldots, p_k \leftrightarrow \tau_k\}$ replaces any equational formula for $p_{k'}$ with $1 \le k' \le k$ in a causal base by the equational formula $p_{k'} \leftrightarrow \tau_{k'}$. Following this idea, the following definition introduces the notion of causal compatibility post intervention.

Definition 8 (Causal compatibility post intervention)**.** *Let $E \in Int$. We define $\Rightarrow^E$ to be the binary relation on the set of states $\mathbf{S}$ such that, for every $S = (C, V), S' = (C', V') \in \mathbf{S}$:*

$$
S \Rightarrow^E S' \textit{ if and only if } C' = \big(C \setminus \bigcup_{p \in \mathbb{P}(E)} \mathcal{L}_{\mathsf{EQ}}(p)\big) \cup E.
$$

$S \Rightarrow^E S'$ means that state $S' = (C', V')$ is compatible with state $S = (C, V)$ after the occurrence of the intervention E. Specifically, the latter is the case if the causal base C' is the result of the following replacement operation applied to the causal base C: first of all remove from C all equational formulas for the propositions on which we intervene through E, and then add to the resulting causal base all equational formulas included in E. Note that if $S \in \mathbf{S}_{Eq}$ and $S \Rightarrow^E S'$ then $S' \in \mathbf{S}_{Eq}$. This means that intervening on an equational state results in an equational state.

4.3 Formalization of Actual Cause

In this section, we recall the definition of actual cause given in [de Lima and Lorini, 2024]. As shown by de Lima & Lorini, under the assumption that the causal graph induced by the underlying equational state is a DAG (directed acyclic graph) their definition is equivalent to Halpern's definition given in [Halpern, 2015].

Before defining actual cause formally, some preliminary notation is needed. A term is a conjunction of literals in which a propositional variable can occur at most once. The set of terms is denoted by $Term$ with elements $\lambda, \lambda', \ldots$ The set $Term_Z$ with $Z \subseteq \mathbb{P}$ denotes the set of terms built from the variables in Z. Given $\lambda, \lambda' \in Term$, with a slight abuse of notation, we write $\lambda' \subseteq \lambda$ (resp. $\lambda' \subset \lambda$) to mean that the set of literals appearing in λ' is a subset (resp. strict subset) of the set of literals appearing in λ. Lastly, $\overline{\lambda}$ denotes the conjunction of the negations of λ's literals. That is,

$$
\overline{\lambda} =_{def} \bigwedge_{p \subseteq \lambda} \neg p \wedge \bigwedge_{\neg p \subseteq \lambda} p.
$$

The definition below introduces the so-called "but" condition. A term λ is a "but" condition for a propositional fact ω at a state S if there exists an intervention on the endogenous variables in λ, along with another intervention that fixes the actual values of some endogenous variables not included in λ, such that, if the values of the exogenous variables remain unchanged, the formula φ will necessarily be false after these interventions.

Definition 9 ("But" condition)**.** *Let $S = (C, V) \in \mathbf{S}_{Eq}$, $\lambda \in Term_{end(S)}$ and $\omega \in \mathcal{L}_{\mathsf{PROP}}$ such that $\mathbb{P}(\omega) \subseteq \mathbb{P}(C)$. We say that λ is a "but" condition for ω at state S, denoted by $\mathsf{But}(S, \lambda, \omega)$, if*

$$\exists E \in Int_{\mathbb{P}(\lambda)}, \exists Z \subseteq end(S), \exists E' \in Int_Z \text{ such that}$$
$$Z \cap \mathbb{P}(\lambda) = \emptyset, S \models \widehat{E'} \text{ and}$$
$$\forall S' \in \mathbf{S}, \text{ if } S \Rightarrow^{E \cup E'} S' \text{ and } S' \models \lambda_S^{exo} \text{ then } S' \models \neg\omega,$$

$$\lambda_S^{exo} =_{def} \bigwedge_{p \in exo(S) \cap V} p \wedge \bigwedge_{p \in exo(S) \setminus V} \neg p.$$

The definition seems complicated, especially the E' part that consists in fixing the actual truth values of some endogenous variables. The existential quantification over such E' is a core aspect of Halpern's definition of actual cause. This quantification is needed to check the absence of causal influence from the other variables on the produced effect. As we will show in Section 4.4, this quantification is not needed when expressing actual cause through counterfactuals.

We use the notion of "but" condition to define the notion of actual cause below. Namely, λ is an actual cause of ω if both λ and ω are true, and λ is a *minimal* "but" condition for ω.

Definition 10 (Actual cause)**.** *Let $S = (C, V) \in \mathbf{S}_{Eq}$, $\lambda \in Term_{end(S)}$ and $\omega \in \mathcal{L}_{\mathsf{PROP}}$ such that $\mathbb{P}(\omega) \subseteq \mathbb{P}(C)$. We say λ is an actual cause of ω at state S if:*

$i)$ $S \models \lambda \wedge \omega$,

$ii)$ $\mathsf{But}(S, \lambda, \omega)$ *holds*,

$iii)$ $\forall \lambda' \subset \lambda, \mathsf{But}(S, \lambda', \omega)$ *does not hold.*

Let us emphasize again that, as shown in [de Lima and Lorini, 2024], the previous definition of actual cause is equivalent to Halpern's definition given in [Halpern, 2015] when the causal graph induced by the equational state S is a DAG.

Let us go back to Billy and Suzy's example.

Example 4 (Billy and Suzy continued)**.** *We have that st is an actual cause of bs at state S_0 in Example 3, while bt is not, for $\mathsf{But}(S_0, bt, bs)$ does not hold.*

4.4 Reduction to Counterfactuals

In this section, we are going to present the central conceptual result of the paper: a theorem highlighting that actual cause can be expressed by means of counterfactual conditionals *without interventions*. The following Lemma 11 is the key to prove it.

Lemma 11. *Let $S = (C, V) \in \mathbf{S}_{Eq}$ such that its causal graph G_S is a DAG, $\lambda \in Term_{end(S)}$ and $\mathbb{P}(\omega) \subseteq \mathbb{P}(C)$. If $S \models \lambda$, then $\mathsf{But}(S, \lambda, \omega)$ if and only if*

$$S \models \bigvee_{\lambda' \in Term_{\mathbb{P}(\lambda)}} \big((\lambda' \wedge \lambda_S^{exo}) \diamondrightarrow \neg\omega\big).$$

The lemma states that the "but" condition can be captured in terms of a might-conditional under the assumption that the underlying causal graph is a DAG. In particular, under the assumption that the causal graph induced by the state S is a DAG, λ is a "but" condition for the propositional fact ω at S if and only if, at S there exists a term λ' sharing its propositions with λ such that if λ' were true and the exogenous variables had their actual truth values, ω might be false.

We sketch the proof idea of the lemma here. In the first glimpse, we must construct an intervention $E \cup E'$ witnessing $\mathsf{But}(S, \lambda, \omega)$ from some $S' \in \mathtt{Closest}(\lambda' \wedge \lambda_S^{exo}, S, \mathbf{S})$ with $S' \models \neg\omega$ and vice versa. Apparently, a concern would be that the intervention does not necessarily give rise to the closest states to S, as required by S'. Nonetheless, such a concern is unfounded: since ω is a propositional formula, its truth value is only determined by the propositional valuation. The causal base plays the role, together with the fact that the causal graph G_S is a DAG, of ensuring that we can associate some S' with the states resulting from the intervention $E \cup E'$, in such a way that they share the same propositional valuation.

We are now in a position to show the main result of this section, namely the following Theorem 12.

Theorem 12. *Let $S = (C, V) \in \mathbf{S}_{Eq}$ s.t. G_S is a DAG, $\lambda \in Term_{end(S)}$ and $\mathbb{P}(\omega) \subseteq \mathbb{P}(C)$. Then, λ is an actual cause of ω at S, if and only if*

$$S \models \lambda \wedge \big((\overline{\lambda} \wedge \lambda_S^{exo}) \diamondrightarrow \neg\omega\big) \wedge \bigwedge_{\substack{Z \subset \mathbb{P}(\lambda),\\ \lambda' \in Term_Z}} \big((\lambda' \wedge \lambda_S^{exo}) \boxright \omega\big).$$

According to Theorem 12, under the assumption that the underlying causal graph is a DAG, the notion of actual cause can be captured by a combination of conditionals and one might-conditional. In particular, under the assumption that the causal graph induced by the state S is a DAG, λ is an actual cause of ω at S if and only if at S i) λ is true, ii) if the truth values of all variables in λ were changed and the exogenous variables had their actual truth values, ω might be false, and iii) for every term λ' built from a strict subset of the set of propositions in λ, if λ' were true and the exogenous variables had their actual truth values, ω would be true. Theorem 12 highlights the main message of our paper: *actual cause is definable using counterfactual conditionals without having to resort to interventions*.

Example 5 (Billy and Suzy revisited)**.** *We have that st is an actual cause of bs, for $S_0 \models st \wedge (\neg st \wedge \lambda_S^{exo}) \diamondrightarrow \neg bs) \wedge \lambda_S^{exo} \boxright bs$, but bt is not, for $S_0 \models (\neg bt \wedge \lambda_S^{exo}) \boxright bs$.*

5 Model Checking

In this section, we study the model checking problem in the defined framework. To date, satisfiability checking received more attention in the literature on counterfactuals: a seminal paper [Friedman and Halpern, 1994] established PSPACE-completeness for it in general (with a few exceptions for some properties of similarity ordering) and subsequent works proposed various decision procedures [Giordano *et al.*, 2009; Lellmann and Pattinson, 2012; Girlando *et al.*, 2021]. At the same time, using standard methods in model checking [Grädel and Otto, 1999], it is straightforward to verify that model checking can be performed in PTIME if the whole model is given *explicitly* as input, including the set of possible states and the comparative similarity relations. However, explicit models may be extremely large and so unpractical.

In our semantics, model checking can be formulated in a succinct way since the model does not need to be given explicitly: the set of possible states and the comparative similarity relations can be computed *ex post*. Specifically, following [de Lima and Lorini, 2024], we define a succinct "relativized" version of model checking in which three elements are given as input: i) a formula φ of the language $\mathcal{L}$ to be checked, ii) a finite vocabulary Γ of propositional facts from which the context $\mathbf{S}^\Gamma \overset{def}{=} \{S = (C, V) \in \mathbf{S} \,:\, C \subseteq \Gamma\}$ is defined, and iii) a finite state S from $\mathbf{S}^\Gamma$ with respect to which the formula φ is evaluated. The context $\mathbf{S}^\Gamma$ includes all states whose causal bases are constructed from Γ.

Model checking problem.
Input: $\psi \in \mathcal{L}$, finite $\Gamma \subset 2^{\mathcal{L}_{\mathsf{PROP}}}$, finite $S \in \mathbf{S}^\Gamma$.
Output: `true` if $(S, \mathbf{S}^\Gamma) \models \psi$, `false` otherwise.

In the rest of this section, we are going to show that this problem is PSPACE-complete by its polynomial reduction to the quantified Boolean Formula problem (QBF) and vice versa.

Let Γ and ψ be given. The set of relevant atoms is defined as follows: $\Sigma = \mathbb{P}(\Gamma) \cup \mathbb{P}(\psi)$. We can represent each state of $\mathbf{S}^\Gamma$ by $|\Gamma| + |\Sigma|$ bits, defining which facts from Γ are present in the causal base and which relevant atoms are present in the valuation. Accordingly, we use sets of variables $X^i = \{b^i_\omega \mid \omega \in \Gamma\} \cup \{v^i_p \mid p \in \Sigma\}$ for $i \in \mathbb{N}$ to represent the states in the QBF encoding. Then, any state from $\mathbf{S}^\Gamma$ corresponds to some valuation on variables from X^i.

We define an encoding function $\mathsf{Sat}(\varphi, X^i)$, which maps a subformula φ of ψ (or of some formula in Γ) into an open QBF formula satisfiable exactly by valuations on X^i that correspond to states satisfying φ (boolean cases are omitted):

$$\begin{aligned}
&\mathsf{Sat}(p, X^i) = v^i_p\\
&\mathsf{Sat}(\triangle\omega, X^i) = b^i_\omega\\
&\mathsf{Sat}(\varphi_1 \;\Box\!\!\rightarrow \varphi_2, X^i) = \forall X^{i+1}.\, \mathsf{State}(X^{i+1}) \rightarrow\\
&\qquad (\mathsf{Closest}(\varphi_1, X^i, X^{i+1}) \rightarrow \mathsf{Sat}(\varphi_2, X^{i+1}))
\end{aligned}$$

Notice that for encoding quantification over states we need to use a set of variables X^{i+1} different from X^i, and we need to check that causal base and the valuation given by choice of values of X^{i+1} will be compatible (as required by Definition 1). For this, we use the following predicate State:

$$\mathsf{State}(X^i) = \bigwedge_{\omega \in \Gamma} (b^i_\omega \rightarrow \mathsf{Sat}(\omega, X^i)).$$

Predicate Closest encodes the definition of the closest state (Definition 5). However, this definition uses predicate Sat on φ twice: to assert that given state satisfies φ and that no closer state satisfies φ. To keep the encoding polynomial, we need to merge these two instances into one via standard Tseitins Tranformation [Tseitin, 1983] by introducing an extra quantifier:

$$\begin{aligned}
&\mathsf{Closest}(\varphi, X^i, X^j) = \forall X^k.\, \forall r.\, \mathsf{State}(X^k) \rightarrow\\
&\quad (\mathsf{Sat}(\varphi, X^k) \leftrightarrow r) \rightarrow ((\mathsf{Eq}(X^j, X^k) \rightarrow r) \wedge\\
&\qquad\qquad (\mathsf{Closer}(X^i, X^j, X^k) \rightarrow \neg r)).
\end{aligned}$$

Here $k = \max(i, j) + 1$ (to ensure that variables are different). Predicates Eq and Closer encode equality and similarity of states from $\mathbf{S}^\Gamma$ directly by definitions.

Notice that $|\mathsf{State}(X^i)| = \mathcal{O}(\sum_{\omega\in\Gamma} |\omega|)$ since $|\mathsf{Sat}(\omega, X^i)| = \mathcal{O}(|\omega|)$ for $\omega \in \mathcal{L}_{\mathsf{PROP}}$, while $\mathsf{Eq}(X^i, X^j)$ and $\mathsf{Closer}(X^i, X^j, X^k)$ require to do $\mathcal{O}(|\Gamma| + |\Sigma|)$ checks on corresponding variables. So the predicate Sat makes a recursive call for each immediate subformula exactly once with an overhead at each step that is polynomial w.r.t. size of the input. Thus, we have a polynomial-size reduction to QBF, which immediately implies PSPACE-membership of the model checking problem.

For PSPACE-hardness we provide a reverse reduction (from QBF). It is based on the observation that for $p, p' \in \mathbb{P} \setminus V$ there are exactly two states in $\texttt{Closest}(p \vee p', (\emptyset, V), \mathbf{S}^\emptyset)$, one satisfying p and one not satisfying it (but satisfying p'), so we can use a counterfactual with $(p \vee p')$ in the antecedent to emulate boolean quantification over p.

Theorem 13. *The model checking problem is PSPACE-complete.*

Our reduction of actual cause to counterfactuals in Theorem 12 requires model checking with respect to the context $\mathbf{S}$ that contains all states. In general, model checking w.r.t. $\mathbf{S}$ can not be easily reduced to model checking w.r.t. a context $\mathbf{S}^\Gamma$ defined from a finite vocabulary Γ (we believe the former problem belongs to a higher complexity class). However, we can perform such a reduction in the special case when the input formula does not contain nested counterfactuals (which is the case for Theorem 12).

Lemma 14. *If $\psi \in \mathcal{L}$ does not contain nested counterfactuals then $((C, V), \mathbf{S}) \models \psi$ iff $((C, V), \mathbf{S}^\Gamma) \models \psi$ for $\Gamma = C \cup \{\omega : \triangle\omega$ is a subformula of $\psi\}$.*

Due to this reduction, we can employ our QBF encodings to check actual cause via Theorem 12. Moreover, although the last conjunction over $\lambda' \in \mathit{Term}_Z$ includes exponentially many conjuncts, we can obtain a polynomial encoding if we replace this conjunction with quantification over terms (which we can also naturally represent with boolean variables). With this modification we can achieve polynomial QBF encoding with the depth of quantifier alteration equals 2: $\exists\forall$ in the second conjunct and $\forall\exists$ in the third conjunct of the formula in Theorem 12. In this sense, our encoding is "close" to being optimal, since the checking of actual cause was shown to be Σ^P_2-complete in [Eiter and Lukasiewicz, 2002] and so only requiring $\exists\forall$ alternation.

6 Conclusion

Let's take stock. We have shown that the notion of intervention is not essential for the formalization of actual cause, one of the central concepts in the theory of causality. This concept can be captured by Lewisian counterfactual conditionals once a two-dimensional semantics distinguishing the propositional level from the causal level is adopted. We have also shown that model checking for the language of counterfactual conditionals defined in this semantics is PSPACE-complete by means of its reduction into QBF and vice versa.

Our contribution has an impact at both the conceptual and computational level. On the conceptual side, we offer a general framework for unifying counterfactuals and actual cause. On the practical side, we provide a semantics for counterfactuals in which model checking can be formulated succinctly. This is useful in practice for the automatic verification of causal properties.

Directions for future work are manifold. First, we plan to explore the proof-theoretic aspects of our logic of counterfactual conditionals. In Section 3.3, we only presented some interesting validities. We plan to develop a sound and complete axiomatization. Second, we plan to implement the QBF translation given in Section 5, in order to experimentally investigate the automated verification of causal properties—actual cause in particular—in terms of computation time. Third, we plan to extend our analysis based on Lewisian counterfactuals to other notions of cause, with special attention to Wright's notion of NESS cause [Wright, 1988]. Finally, we plan to investigate the relationship between our counterfactual atemporal approach to actual cause and recent work on temporal causal reasoning [Gladyshev *et al.*, 2025]. To this aim, we will extend our framework with an LTL temporal component, in order to account for temporal information in a causal base as well as counterfactual reasoning about temporal facts.

Acknowledgments

This work is supported by the ANR projects EpiRL (grant number ANR-22-CE23-0029), ALoRS (grant number ANR-21-CE23-0018-01), the Austrian Science Fund (FWF) 10.55776/COE12, and European Union's Horizon 2020 research and innovation programme (grant agreement No 101034440).

References

[Alchourrón *et al.*, 1985] Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic*, 50(2):510–530, 1985.

[Alechina *et al.*, 2017] Natasha Alechina, Joseph Y. Halpern, and Brian Logan. Causality, responsibility and blame in team plans. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2017)*, pages 1091–1099. ACM, 2017.

[Barbero and Sandu, 2021] Fausto Barbero and Gabriel Sandu. Team semantics for interventionist counterfactuals: Observations vs. interventions. *Journal of Philosophical Logic*, 50(3):471–521, 2021.

[Batusov and Soutchanski, 2018] Vitaliy Batusov and Mikhail Soutchanski. Situation calculus semantics for actual causality. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 1744–1752. AAAI Press, 2018.

[Beckers and Vennekens, 2017] Sander Beckers and Joost Vennekens. The transitivity and asymmetry of actual causation. *Ergo*, 4(1), 2017.

[Beckers *et al.*, 2022] Sander Beckers, Hana Chockler, and Joseph Y. Halpern. A causal analysis of harm. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[Beckers, 2021] Sander Beckers. The counterfactual NESS definition of causation. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, pages 6210–6217. AAAI Press, 2021.

[Bochman, 2018] Alexander Bochman. On laws and counterfactuals in causal reasoning. In *Proceedings of the Sixteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2018)*, pages 494–503. AAAI Press, 2018.

[Bochman, 2021] Alexander Bochman. *A Logical Theory of Causality*. MIT Press, 2021.

[Chockler and Halpern, 2004] Hana Chockler and Joseph Y. Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

[Chockler and Halpern, 2022] Hana Chockler and Joseph Y. Halpern. On testing for discrimination using causal models. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*, pages 5548–5555. AAAI Press, 2022.

[de Lima and Lorini, 2024] Tiago de Lima and Emiliano Lorini. Model checking causality. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*, pages 3324–3332. ijcai.org, 2024.

[Eiter and Lukasiewicz, 2002] Thomas Eiter and Thomas Lukasiewicz. Complexity results for structure-based causality. *Artificial Intelligence*, 142(1):53–89, 2002.

[Friedman and Halpern, 1994] Nir Friedman and Joseph Y. Halpern. On the complexity of conditional logics. In *Principles of Knowledge Representation and Reasoning*, pages 202–213. Elsevier, 1994.

[Galles and Pearl, 1998] David Galles and Judea Pearl. An axiomatic characterization of causal counterfactuals. *Foundation of Science*, 3(1):151–182, 1998.

[Giordano *et al.*, 2009] Laura Giordano, Valentina Gliozzi, Nicola Olivetti, and Camilla Schwind. Tableau calculus for preference-based conditional logics: Pcl and its extensions. *ACM Trans. Comput. Logic*, 10(3), April 2009.

[Girlando *et al.*, 2021] Marianna Girlando, Sara Negri, and Nicola Olivetti. Uniform labelled calculi for preferential conditional logics based on neighbourhood semantics. *Journal of Logic and Computation*, 31(3):947–997, 04 2021.

[Gladyshev *et al.*, 2025] Maksim Gladyshev, Natasha Alechina, Mehdi Dastani, Dragan Doder, and Brian Logan. Temporal causal reasoning with (non-recursive) structural equation models. In *Proceedings of the 39th*

AAAI Conference on Artificial Intelligence (AAAI-25), pages 14949–14957. AAAI Press, 2025.

[Grädel and Otto, 1999] Erich Grädel and Martin Otto. On logics with two variables. *Theoretical Computer Science*, 224:73–113, 1999.

[Halpern and Kleiman-Weiner, 2018] Joseph Y. Halpern and M. Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 1853–1860. AAAI Press, 2018.

[Halpern and Pearl, 2005a] Joseph Y. Halpern and Judea Pearl. Causes and explanations: a structural-model approach. Part I: Causes. *British Journal for Philosophy of Science*, 56(4):843–887, 2005.

[Halpern and Pearl, 2005b] Joseph Y. Halpern and Judea Pearl. Causes and explanations: a structural-model approach. Part II: Explanations. *British Journal for Philosophy of Science*, 56(4):889–911, 2005.

[Halpern, 2000] Joseph Y. Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12:317–337, 2000.

[Halpern, 2008] Joseph Y. Halpern. Defaults and normality in causal structures. In G. Brewka and J. Lang, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference (KR 2008)*, pages 198–208. AAAI Press, 2008.

[Halpern, 2015] Joseph Y. Halpern. A modification of the Halpern-Pearl definition of causality. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 3022–3033. AAAI Press, 2015.

[Halpern, 2016] Joseph Y. Halpern. *Actual causality*. MIT Press, 2016.

[Kenny and Keane, 2021] Eoin M. Kenny and Mark T. Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021)*, pages 11575–11585. AAAI Press, 2021.

[Khan and Lespérance, 2021] Shakil M. Khan and Yves Lespérance. Knowing why — on the dynamics of knowledge about actual causes in the situation calculus. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*, pages 701–709. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2021.

[Kraus *et al.*, 1990] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial intelligence*, 44(1-2):167–207, 1990.

[Lellmann and Pattinson, 2012] Björn Lellmann and Dirk Pattinson. Sequent systems for lewis' conditional logics. In Luis Fariñas del Cerro, Andreas Herzig, and Jérôme Mengin, editors, *Logics in Artificial Intelligence*, pages 320–332, 2012.

[Lewis, 1973] David K. Lewis. *Counterfactuals*. Harvard University Press, 1973.

[Lewis, 1979] David K. Lewis. Counterfactual dependence and time's arrow. *Noûs*, 13(4):455–476, 1979.

[Lorini, 2023] Emiliano Lorini. A rule-based modal view of causal reasoning. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)*, pages 3286–3295. ijcai.org, 2023.

[Miller, 2021] Tim Miller. Contrastive explanation: a structural-model approach. *The Knowledge Engineering Review*, 36:e14, 2021.

[Mittelstadt *et al.*, 2019] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in AI. In *Proceedings of the 2019 conference on Fairness, Accountability, and Transparency*, pages 279–288, 2019.

[Mothilal *et al.*, 2020] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 607–617, New York, NY, USA, 2020. Association for Computing Machinery.

[Pearl, 2009] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

[Sokol and Flach, 2019] Kacper Sokol and Peter Flach. Counterfactual explanations of machine learning predictions: Opportunities and challenges for ai safety. In *Proceedings of the AAAI Workshop on Artificial Intelligence Safety 2019*, volume 2301 of *CEUR Workshop Proceedings*. CEUR Workshop Proceedings, January 2019. 2019 AAAI Workshop on Artificial Intelligence Safety, SafeAI 2019 ; Conference date: 27-01-2019.

[Stalnaker, 1968] Robert Stalnaker. A theory of conditionals. In N. Rescher, editor, *Studies in Logical Theory*, pages 28–45. Oxford University Press, 1968.

[Tseitin, 1983] Grigori Tseitin. On the complexity of derivation in propositional calculus. In Jörg H. Siekmann and Graham Wrightson, editors, *Automation of Reasoning: 2: Classical Papers on Computational Logic 1967–1970*, pages 466–483. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983.

[von Kügelgen *et al.*, 2023] Julius von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In *Conference on Causal Learning and Reasoning, CLeaR 2023, 11-14 April 2023, Amazon Development Center, Tübingen, Germany, April 11-14, 2023*, volume 213 of *Proceedings of Machine Learning Research*, pages 177–196. PMLR, 2023.

[Woodward and Hitchcock, 2003] James Woodward and Christopher Hitchcock. Explanatory generalizations, part i: A counterfactual account. *Noûs*, 37(1):1–24, 2003.

[Woodward, 2003] James Woodward. *Making Things Happen: a Theory of Causal Explanation*. Oxford University Press, 2003.

[Wright, 1988] Richard W. Wright. Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts. *Iowa Law Review*, 73:1001–1077, 1988.

[Zhang, 2013] Jiji Zhang. A Lewisian logic of causal counterfactuals. *Minds and Machines*, 23(1):77–93, 2013.