Merging Beliefs and Counterfactuals

Emiliano Lorini[1] , **Dmitry Rozplokhas**[2]
[1]IRIT, CNRS, Toulouse University
[2]TU Wien
Emiliano.Lorini@irit.fr, dmitry@logic.at

Abstract

We present a unified framework for modeling agents' epistemic states and counterfactual conditionals. The novelty of our approach lies at the semantic level. We introduce a computationally grounded semantics in which an agent's doxastic accessibility relation, needed to model the standard notion of deductively closed belief, is derived from the agent's belief base, and in which the notion of comparative similarity between states, needed to interpret counterfactual conditionals, is computed from both the atomic formulas describing the agents' belief bases and those describing the environment. We use this semantics to interpret a language of beliefs and counterfactuals. We demonstrate the expressiveness of our language and the flexibility of the semantics by formalizing a wide range of notions: counterfactual dependence, counterfactual notions of reason and knowledge, conditional belief. Furthermore, we show that our semantics illuminates the subtle interplay between belief change and counterfactuals. On the computational side, we provide a succinct formulation of model checking for our language and establish a PSPACE complexity result.

1 Introduction

Epistemic logic and the logic of counterfactual conditionals have largely developed independently since the seminal work of Hintikka [Hintikka, 1962] on the logic of knowledge and belief, and of Lewis [Lewis, 1973] and Stalnaker [Stalnaker, 1968] on the logic of counterfactual conditionals. Epistemic logic has been extensively studied in computer science and artificial intelligence (AI) as a framework for modeling agents' uncertainty in distributed and multi-agent systems [Meyer and van der Hoek, 1995; Fagin *et al.*, 1995]. By contrast, counterfactuals and their associated logics have recently attracted renewed attention due to their relevance to causal reasoning [Galles and Pearl, 1998; von Kügelgen *et al.*, 2023; Aguilera-Ventura *et al.*, 2025; Bochman, 2018] and explainability [Mittelstadt *et al.*, 2019; Mothilal *et al.*, 2020; Sokol and Flach, 2019; Kenny and Keane, 2021], two central topics in AI.

While epistemic logic is primarily concerned with supporting reasoning about agents' beliefs, including first-order beliefs about the environment and higher-order beliefs about other agents' beliefs, the logic of counterfactual conditionals focuses on reasoning about ontic hypothetical necessity and possibility, that is, how the world would/might be if a given condition were satisfied.

The aim of the present work is to merge these two approaches by providing a unified framework for representing and reasoning about agents' beliefs and counterfactuals. This unification is motivated not only by conceptual considerations but also by practical ones. In particular, we seek to enhance the reasoning capabilities of artificial agents. Intelligent agents possess not only beliefs that represent their views of the environment and of other agents' mental states, but also the ability to reason counterfactually. This capacity is crucial for explaining natural events, as well as one's own and others' beliefs, decisions, and actions, and for attributing responsibility to oneself and to others, all of which are especially relevant to machine ethics and human–machine interaction.

At the language level, the integration we propose consists of combining doxastic modalities of the form $\Box_i$, representing what agent i implicitly believes (i.e., what the agent can deduce from its belief base) with counterfactual conditional modalities in the style of Lewis and Stalnaker, of the form $\Box\!\!\rightarrow$. To interpret this language, we rely on a computationally grounded semantics, inspired by [Lorini, 2020], in which the accessibility relations over the states of a model are not taken as primitives but are computed from the information contained in the states. In particular, an agent's doxastic accessibility relation, required to model the agent's implicit belief modality, is computed from information about the agent's belief base (i.e., what the agent explicitly believes), whereas the comparative similarity relation between states, needed to interpret the counterfactual conditional modality, is computed from both information about the agents' belief bases and information about the environment.

We show that our framework provides a unified setting for modeling a rich variety of epistemic concepts studied in formal epistemology and for capturing the subtle interplay between beliefs, counterfactuals, and belief change. In particular, we show that our semantics enables belief change operations on belief bases, including forgetting, contraction, revision, and update, to be represented directly as formulas in the

language of beliefs and counterfactual conditionals. Furthermore, we show that counterfactual notions of reason, knowledge, and conditional belief can also be naturally expressed within this language. On the computational side, we exploit our semantics to provide a succinct formulation of model checking for the language of beliefs and counterfactual conditionals and establish a PSPACE complexity result.

The paper is organized as follows. In Section 2, we discuss related work. In Section 3, we present our framework, including the semantics and the interpretation of the language of beliefs and counterfactuals. Section 4 presents some formal properties of beliefs, counterfactuals, and their interrelations. In Section 5, we investigate the connection between our framework and the theory of belief change, while in Section 6 we formalize various epistemic concepts. Section 7 addresses model checking and its computational complexity. Proofs are provided in the extended version of this paper available at `https://hal.science/hal-05622406`.

2 Related Work

A natural connection between epistemic states and counterfactuals is provided by the notion of conditional belief. Its logic, as well as its relationship with belief revision, has been widely studied in recent years [van Benthem, 2007; Baltag and Smets, 2006]. At the model-theoretic level, an important difference between counterfactual conditionals, at least from a Lewisian perspective, and conditional belief is that the former relies on a notion of (ontic) comparative similarity between states, whereas the latter relies on a notion of subjective plausibility between states. There are two key differences between these two notions. The first concerns the so-called *absoluteness* property: unlike the comparative similarity relation, which is state-dependent, the plausibility ordering is state-independent, remaining fixed across all states in the model or, more restrictively, within an agent's epistemic state. The second concerns the so-called *centering* property: whereas the actual state belongs to the set of states that are closest to itself (and, in the case of strong centering, is its unique element), the actual state is not necessarily included in the set of states that are most plausible according to the agent. These differences give rise to distinct logics, as shown in [Friedman and Halpern, 1994]. In the present paper, we start from a language of counterfactual conditionals interpreted using a notion of comparative similarity and augment it with belief modalities. We show that a plausibility-based notion of conditional belief can be naturally expressed in the resulting language.

A close work to ours is [Holliday, 2015], where the notion of a counterfactual belief (CB) model is introduced. A CB model combines a comparative similarity relation for counterfactual reasoning with a doxastic accessibility relation for belief. In contrast to our computationally grounded semantics, where comparative similarity reflects differences in agents' belief bases, thereby interrelating the counterfactual and epistemic dimensions, the two dimensions in a CB model are treated independently. A further difference is that Holliday employs CB models to interpret a range of epistemic and doxastic modalities, including Nozick's counterfactual notion of knowledge (which we consider in Section 6), without introducing an explicit counterfactual conditional modality. By contrast, we integrate counterfactual conditional modalities with doxastic modalities and use the resulting language to express a variety of doxastic and epistemic concepts.

The framework presented in Section 3 extends the belief-base approach to epistemic logic introduced in [Lorini, 2020], both at the semantic and at the language level, by incorporating counterfactual conditionals.

3 Logical Framework

In this section, we present our framework. We begin by introducing the semantics and then turn to the language. In our semantics, a state specifies the truth of ontic facts describing the environment, as well as the agents' explicit beliefs, that is, the information stored in their belief bases. From the information associated with a state, we derive an agent's doxastic accessibility relation, which represents the states the agent considers possible, as well as a comparative similarity relation for reasoning about counterfactual situations. We use this semantics to interpret a language of implicit belief and counterfactual conditionals. The implicit belief modality captures what an agent can deduce from its explicit beliefs.

3.1 Semantics

Assume a countably infinite set of atomic propositions $\mathbb{P}_0 = \{p, q, \ldots\}$ and a finite set of agents $\mathbb{A} = \{1, \ldots, n\}$. We define the language $\mathcal{L}_0$ for explicit belief, as follows:

$$\mathcal{L}_0 \stackrel{\text{def}}{=} \quad \alpha \quad ::= \quad p \mid \neg\alpha \mid \alpha \wedge \alpha \mid \triangle_i \alpha,$$

where p ranges over $\mathbb{P}_0$ and i ranges over $\mathbb{A}$. The formula $\triangle_i \alpha$ reads "agent i has the explicit belief that α".

We define the set of atomic formulas as follows:

$$\mathbb{P} = \mathbb{P}_0 \cup \{\triangle_i \alpha \mid i \in \mathbb{A} \text{ and } \alpha \in \mathcal{L}_0\}.$$

The following definition introduces the notion of state.

Definition 1 (State)**.** *A* state *is a finite set of atomic formulas from $\mathbb{P}$. We denote by* $\mathbf{S}$ *the set of states. Elements of* $\mathbf{S}$ *are denoted by $S, S', \ldots$*

The truth conditions for the formulas of $\mathcal{L}_0$ are defined with respect to states, as follows.

Definition 2 ($\mathcal{L}_0$-truth conditions)**.** *Let $S \in \mathbf{S}$. Then,*

$$\begin{aligned} S \models p &\Leftrightarrow p \in S, \\ S \models \triangle_i \alpha &\Leftrightarrow \triangle_i \alpha \in S, \\ S \models \neg\alpha &\Leftrightarrow S \not\models \alpha, \\ S \models \alpha_1 \wedge \alpha_2 &\Leftrightarrow S \models \alpha_1 \text{ and } S \models \alpha_2. \end{aligned}$$

As the previous definition highlights, the language $\mathcal{L}_0$ can seen as a propositional language built from the set of atomic formulas $\mathbb{P}$ and a state S is nothing but a propositional valuation for the propositional language $\mathcal{L}_0$.

We define the agent i's belief base at a state S, as the set of i's explicit beliefs at S.

Definition 3 (Belief base)**.** *Let $i \in \mathbb{A}$ and $S \in \mathbf{S}$. Then,*

$$B_i(S) = \{\alpha \in \mathcal{L}_0 \mid \triangle_i \alpha \in S\}.$$

The following definition introduces an agent's doxastic accessibility relation. It is computed from its belief base.

Definition 4 (Doxastic accessibility relation)**.** *Let $i \in \mathbb{A}$ and $S, S' \in \mathbf{S}$. Then, $S\mathcal{R}_i S'$ iff $\forall \alpha \in B_i(S), S' \models \alpha$.*

According to the previous definition, at state S the agent i considers the state S' possible if S' satisfies all facts that i explicitly believes at S. The second component of our semantics is the comparative similarity relation between states.

Definition 5 (Comparative similarity relation)**.** *Let $S, S', S'' \in \mathbf{S}$. State S' is at least as similar to state S as state S'' is, denoted by $S'' \preceq_S^{\mathsf{sim}} S'$, if $(S \Delta S') \subseteq (S \Delta S'')$, where Δ stands for symmetric difference. We write $S'' \prec_S^{\mathsf{sim}} S'$ if $S'' \preceq_S^{\mathsf{sim}} S'$ and $S' \not\preceq_S^{\mathsf{sim}} S''$.*

According to the previous definition, state S' is at least as similar to state S as state S'' is if S'' differs from S at least as much as S' differs from S. In other words, the similarity between two states is fully determined by the set of atomic formulas whose truth values are the same in the two states. Note that $\preceq_S^{\mathsf{sim}}$ is a partial preorder and, moreover, satisfies the following strong centering property:

$$\forall S, S' \in \mathbf{S}, \text{ if } S \neq S' \text{ the } S' \prec_S^{\mathsf{sim}} S. \tag{1}$$

We conclude this section by defining the notion of model as a state supplemented with a set of states also called *context* or *universe*. The latter includes all states that are compatible with the agents' common ground [Stalnaker, 2002].

Definition 6 (Model)**.** *A model is a pair (S, U) with $S \in U \subseteq \mathbf{S}$. The class of models is denoted by $\mathbf{M}$.*

It is possible to specify both a context and a model succinctly by restricting states to be defined from a finite vocabulary, together with an integrity constraint in $\mathcal{L}_0$.

Definition 7 (Context and model specification)**.** *A context specification is a pair (Γ, α), where Γ is finite set of atomic formulas from $\mathbb{P}$ and $\alpha \in \mathcal{L}_0$ is an integrity constraint.*

A model specification is a triple (S, Γ, α), where (Γ, α) is a context specification and $S \in \mathbf{S}_{(\Gamma,\alpha)}$, with

$$\mathbf{S}_{(\Gamma,\alpha)} = \big\{ S \in \mathbf{S} \mid S \models \alpha \text{ and } S \subseteq \Gamma \big\}.$$

The set $\mathbf{S}_{(\Gamma,\alpha)}$ is the context induced by the integrity constraint α and the the finite vocabulary Γ.

3.2 Language

In this section, we introduce a language for implicit belief and counterfactual conditionals on the top of the language $\mathcal{L}_0$ defined above. It is noted $\mathcal{L}$ and defined, as follows:

$$\mathcal{L} \stackrel{\text{def}}{=} \quad \varphi \quad ::= \quad \alpha \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_i \varphi \mid \varphi \boxright \varphi,$$

where α ranges over $\mathcal{L}_0$ and i ranges over Agt. The other Boolean constructions $\top$, $\bot$, $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined from α, $\neg$ and $\wedge$ in the standard way. The formula $\Box_i\varphi$ is read "agent i implicitly believes that φ" in the sense that "φ is deducible from agent i's explicit beliefs". Its dual, $\Diamond_i\varphi \stackrel{\text{def}}{=} \neg\Box_i\neg\varphi$, is read as "according to agent i, it is possible that φ". The conditional $\varphi \boxright \psi$ is read "if φ were true, ψ would be true". Its dual $\varphi \diamondright \psi \stackrel{\text{def}}{=} \neg(\varphi \boxright \neg\psi)$ is read "if φ were true, ψ might be true".

Given a formula $\varphi \in \mathcal{L}$, we denote by $\mathbb{P}(\varphi)$ the set of atomic formulas occurring in φ (including nesting occurences under $\triangle_i$). The definition generalizes to a set of formulas $X \subseteq \mathcal{L}$ in a straightforward way: $\mathbb{P}(X) = \bigcup_{\varphi \in X} \mathbb{P}(\varphi)$.

Formulas in the language $\mathcal{L}$ are interpreted relative to a model as follows. (Boolean cases are omitted since they are defined in the usual way.)

Definition 8 ($\mathcal{L}$-truth conditions)**.** *Let $(S, U) \in \mathbf{M}$. Then,*

$$(S,U) \models \alpha \;\Leftrightarrow\; S \models \alpha,$$
$$(S,U) \models \Box_i\varphi \;\Leftrightarrow\; \forall S' \in U, \text{if } S\mathcal{R}_i S' \text{ then } (S',U) \models \varphi,$$
$$(S,U) \models \varphi \boxright \psi \;\Leftrightarrow\; \forall S' \in \mathit{Closest}(\varphi,S,U), (S',U) \models \psi,$$

where $\mathit{Closest}(\varphi,S,U) = \big\{ S' \in U \mid (S',U) \models \varphi \text{ and } \nexists S'' \in U \text{ such that } (S'',U) \models \varphi \text{ and } S' \prec_S^{\mathsf{sim}} S'' \big\}$.

In the previous definition, $\mathit{Closest}(\varphi,S,U)$ is the set of φ-states in U that are the closest (or most similar) ones to the state S. We call $\mathit{Closest}(\varphi,S,U)$ the φ-neighborhood of state S within U. In agreement with Lewis' interpretation of counterfactual conditionals, at model (S, U) it is the case that "if φ were true, ψ would be true" if every state in the φ-neighborhood of state S within U satisfies ψ. The implicit belief modality $\Box_i$ is interpreted through the doxastic relation $\mathcal{R}_i$: the agent i implicitly believes that φ if φ is true at all states in U that agent i considers possible.

For notational convenience we sometimes write $S \models \varphi$ instead of $(S, \mathbf{S}) \models \varphi$. Given a formula φ of the language $\mathcal{L}$, we say that φ is valid (denoted $\models \varphi$) if $(S, U) \models \varphi$ for every $(S, U) \in \mathbf{M}$. The following example illustrates the language $\mathcal{L}$ and its semantic interpretation.

Example 1. *We consider an example involving two agents, $\mathbb{A} = \{1, 2\}$, representing two robots in a building. It is common knowledge among the robots that: (i) if the fire alarm rings (al), then everyone will explicitly believe this, and if someone explicitly believes that the fire alarm rings, then the alarm indeed rings; (ii) if there is a fire in the building, then the building must be evacuated (ev). Furthermore, each robot explicitly believes that if the fire alarm rings, then there is a fire in the building (fi). We assume that this information resides at the level of the robots' private belief bases and is not common knowledge, since the fire alarm is defective (even though the robot are unaware of this), so the alarm may ring even in the absence of a fire. Finally, robot 2 privately and wrongly believes that robot 1 does not explicitly believe that if the fire alarm rings, then there is a fire in the building. We are in a situation in which there is a fire in the building, the alarm correctly rings and there is need to evacuate the building. Our scenario is represented by the following actual state:*

$$S_0 = \big\{ al, fi, ev, \triangle_1 al, \triangle_2 al, \triangle_1(al \rightarrow fi), \triangle_2(al \rightarrow fi), \triangle_2\neg\triangle_1(al \rightarrow fi) \big\},$$

together with the following integrity constraint representing the agents' common ground:

$$\alpha_0 \stackrel{def}{=} \bigwedge_{i \in \mathbb{A}} (al \leftrightarrow \triangle_i al) \wedge (fi \rightarrow ev).$$

We assume the vocabulary from which the context is built includes all relevant atomic formulas occurring in the state

S_0 and in the integrity constraint α_0. That is, we consider the model specification $(S_0, \Gamma_0, \alpha_0)$ such that $\Gamma_0 = \mathbb{P}(S_0 \cup \{\alpha_0\})$. It is routine to verify that for every agent $i \in \{1, 2\}$:

$$(S_0, \mathbf{S}_{(\Gamma_0,\alpha_0)}) \models \Box_i ev \wedge (\neg al \boxright \neg \Box_i ev) \wedge (\neg \triangle_i al \boxright \neg \Box_i ev).$$

This means that at the actual state each robot can infer from its belief base that the building should be evacuated. Moreover, if the alarm did not ring, the robot would not be able to infer that the building should be evacuated. Finally, if a robot did not believe that the alarm rings, it would not be able to infer that the building should be evacuated. We also have:

$$(S_0, \mathbf{S}_{(\Gamma_0,\alpha_0)}) \models \Box_2\Big(\neg\triangle_1(al \to fi) \wedge \big(\triangle_1(al \to fi) \boxright \Box_1 ev\big)\Big).$$

This means that robot 2 can deduce from its belief base that robot 1 does not explicitly believe that the alarm ringing implies a fire in the building, and that, if it did believe so, it would infer that the building must be evacuated.

4 Properties

A central property of counterfactuals is the so-called limit closure: if a formula φ is possible, there should exist a closest state to the actual one in which it holds. As the following theorem shows, our semantics satisfies this property.

Theorem 1. *Let (S, U) be a model. Then, for any $\varphi \in \mathcal{L}$, if there is $S' \in U$ s.t. $(S', U) \models \varphi$ then $Closest(\varphi, S, U) \neq \emptyset$.*

The following proposition lists some validities that illustrate interesting properties of counterfactuals and beliefs, as well as their interaction.

Theorem 2. *Let $\varphi, \psi \in \mathcal{L}$, $\alpha \in \mathcal{L}_0$ and $p \in \mathbb{P}_0$. We have the following validities:*

$$\models \big(\Box_i\varphi \wedge \Box_i(\varphi \to \psi)\big) \to \Box_i\psi \quad (2)$$

If $\models \varphi$ *then* $\models \Box_i\varphi$ (3)

$$\models \triangle_i\alpha \to \Box_i\alpha \quad (4)$$

$$\models \varphi \boxright \varphi \quad (5)$$

$$\models (\varphi \wedge \psi) \to (\varphi \boxright \psi) \quad (6)$$

$$\models \big((\varphi \boxright \chi) \wedge (\psi \boxright \chi)\big) \to (\varphi \vee \psi) \boxright \chi \quad (7)$$

$$\models \big(p \wedge (\varphi \boxright \psi)\big) \to (\varphi \wedge p) \boxright \psi \quad (8)$$

$$\models \big(\neg p \wedge (\varphi \boxright \psi)\big) \to (\varphi \wedge \neg p) \boxright \psi \quad (9)$$

$$\models \big(\triangle_i\alpha \wedge (\varphi \boxright \psi)\big) \to (\varphi \wedge \triangle_i\alpha) \boxright \psi \quad (10)$$

$$\models \big(\neg\triangle_i\alpha \wedge (\varphi \boxright \psi)\big) \to (\varphi \wedge \neg\triangle_i\alpha) \boxright \psi \quad (11)$$

Validity (2), together with property (3), constitutes the two fundamental principles of the normal modal logic K for the implicit belief modality. Validity (4) captures the interaction between explicit and implicit belief: explicitly believing α entails implicitly believing α.

Validities (5), (6), and (7) are standard properties of counterfactual conditionals. In particular, validity (5) expresses the *reflexivity* property: all closest φ-states to the actual state must satisfy φ. This follows directly from the semantic interpretation of counterfactuals provided in Definition 8.

Validity (6) corresponds to the so-called *strong centering* property, namely property (1) in Section 3.1. Finally, validity (7) follows from the fact that the relation $\preceq_S^{\mathsf{sim}}$ is a partial preorder.

The validities (8), (9), (10), and (11) illustrate the interaction between counterfactual conditionals, propositional atoms, and agents' explicit beliefs. Specifically, they show that the comparative similarity used to interpret counterfactual conditionals (Definition 5) depends entirely on information about atomic propositions and agents' explicit beliefs at the state level. They capture a form of monotonicity under the accumulation of true atomic formulas and their negations. Note that the following formula is not valid:

$$\big((\varphi \boxright \psi) \wedge (\varphi \diamondright \chi)\big) \to (\varphi \wedge \chi) \boxright \psi \quad (12)$$

If $\preceq_S^{\mathsf{sim}}$ were a total preorder, then the earlier formula (12) would be valid. This formula is an axiom of Lewis' V-logics and corresponds to several well-known axioms and postulates in other areas, e.g., the final postulate in AGM theory [Alchourrón *et al.*, 1985] and rational monotonicity (RM) in non-monotonic reasoning [Kraus *et al.*, 1990]. Since $\preceq_S^{\mathsf{sim}}$ is not total, RM does not hold in our setting.

5 Belief Change Operations

In our framework, we can capture the subtle interplay between counterfactual conditionals and beliefs, on the one hand, and belief change, on the other. Specifically, we can express standard implementations of belief change operations, including forgetting, contraction, revision, and update, by combining counterfactual conditionals and belief modalities of the language $\mathcal{L}$ in various ways.

We focus on the belief base approach to belief change, where a belief base is represented as a finite set of formulas representing an agent's explicit beliefs, and the set of the agent's implicit beliefs is given by its classical deductive closure, denoted by the operator $Cn(\cdot)$ [Hansson, 1999]. In our framework, we can work with the formulas from $\mathcal{L}_0$, thereby accomodating higher-order explicit beliefs as well. The notions of forgetting, contraction, and revision can be defined in this setting in different ways. We start with the simplest notion of forgetting.[1]

Definition 9 (Forgetting)**.** *Let $X \subseteq \mathcal{L}_0$ and $\alpha \in \mathcal{L}_0$. The deductive closure of X after forgetting α is defined as:*

$$forg(X, \alpha) = Cn\big(X \setminus \{\alpha\}\big).$$

As the following theorem highlights, forgetting is expressible in the language $\mathcal{L}$, assuming that the model includes all states. (We remind that $S \models \varphi$ stands for $(S, \mathbf{S}) \models \varphi$). Specifically, α_2 is in the deductive closure of an agent's belief base after forgetting α_1 if and only if, if the agent did not explicitly believe α_1, it would be able to infer α_2.

[1] Here, forgetting is understood as the removal of a formula from the belief base [Lorini, 2020], rather than as contraction of the vocabulary [Lin and Reiter, 1994; Delgrande and Wang, 2015].

Theorem 3. *Let $\alpha_1, \alpha_2 \in \mathcal{L}_0$ and $S \in \mathbf{S}$. Then*

$$\alpha_2 \in \mathit{forg}\big(B_i(S),\alpha_1\big) \text{ iff } S \models (\neg\triangle_i\alpha_1) \boxdot\!\!\rightarrow (\Box_i\alpha_2).$$

Belief contraction is a more involved operation, which requires that a formula is not only removed from the belief base, but also cannot be deduced from it. At its core is the following notion of an α-remainder, borrowed from formal models of theory change [Alchourrón *et al.*, 1985] and belief base change [Hansson, 1993].

Definition 10 (α-remainder)**.** *Let $X \subseteq \mathcal{L}_0$, $\alpha \in \mathcal{L}_0$ and $X' \subseteq X$. We say that X' is a α-remainder of X if $\alpha \notin Cn(X')$, and for all $X'' \subseteq X$, if $X' \subset X''$ then $\alpha \in Cn(X'')$. We denote by $X\perp\alpha$ the set of all α-remainders of X.*

According to the previous definition, an α-remainder of a set of $\mathcal{L}_0$-formulas X is a maximal subset of X from which α is not deducible. One standard definition for contraction, defined via α-remainders, is full meet base contraction (also known as simple base contraction) [Ginsberg, 1986].

Definition 11 (Full meet base contraction)**.** *Let $X \subseteq \mathcal{L}_0$ and $\alpha \in \mathcal{L}_0$. The full meet contraction of X by α is defined as:*

$$\mathit{contr}(X,\alpha) = \bigcap_{X' \in X\perp\alpha} Cn\big(X'\big).$$

Full meet contraction can be expressed in a similar way to forgetting through a counterfactual, this time with the antecedent involving the implicit belief modality rather than the explicit belief one.

Theorem 4. *Let $\alpha_1, \alpha_2 \in \mathcal{L}_0$ and $S \in \mathbf{S}$. Then*

$$\alpha_2 \in \mathit{contr}\big(B_i(S),\alpha_1\big) \text{ iff } S \models (\neg\Box_i\alpha_1) \boxdot\!\!\rightarrow (\Box_i\alpha_2).$$

Full meet base revision (also known as simple base revision [Nebel, 1989]) is then defined via the well-known Levi identity [Levi, 1977]. This is equivalent to the definitions in [Fagin *et al.*, 1983; Ginsberg, 1986].

Definition 12 (Full meet revision)**.** *Let $X \subseteq \mathcal{L}_0$ and $\alpha \in \mathcal{L}_0$. The full meet revision of X by α is defined as:*

$$\mathit{rev}(X,\alpha) = Cn\big(\mathit{contr}(X,\neg\alpha) \cup \{\alpha\}\big).$$

As the following theorem highlights, α_2 is in the deductive closure of an agent's belief base after full meet revision with α_1 if and only if, if the agent explicitly believed α_1 without falling into a contradiction, it would be able to infer α_2.

Theorem 5. *Let $\alpha_1, \alpha_2 \in \mathcal{L}_0$ and $S \in \mathbf{S}$. Then*

$$\alpha_2 \in \mathit{rev}\big(B_i(S),\alpha_1\big) \text{ iff } S \models (\neg\Box_i\bot \wedge \triangle_i\alpha_1) \boxdot\!\!\rightarrow (\Box_i\alpha_2).$$

Another important notion in the theory of belief change is belief update. It is fundamentally different from belief revision in that it is intended to reflect changes in a dynamic world, rather than accommodating new information about a static one. The standard approach to belief update, due to Katsuno and Mendelzon [Katsuno and Mendelzon, 1991], uses a different representation of an agent's beliefs, namely Winslett's Possible Models Approach [Winslett, 1988], which represents beliefs as the set of models considered possible by the agent (in our case, a subset of states from $\mathbf{S}$). Updating by the fact α replaces each state S with the set of states satisfying α that are closest to S, using symmetric difference as the distance measure.

Definition 13 (Katsuno-Mendelzon update)**.** *Let $X \subseteq \mathcal{L}_0$ and $\alpha \in \mathcal{L}_0$. Moreover, let $\mathbf{S}(X) = \{S \in \mathbf{S} \mid \forall\alpha' \in X, S \models \alpha'\}$ and $\mathbf{S}(\alpha) = \{S \in \mathbf{S} \mid S \models \alpha\}$. The set of possible states after updating the base X with α is defined as:*

$$\mathbf{S}(X \diamond \alpha) = \bigcup_{S \in \mathbf{S}(X)} \mathit{Closest}(\alpha,S,\mathbf{S}(\alpha)).$$

The set of deducible formulas after updating X with α is:

$$\mathit{upd}(X,\alpha) = \{\alpha' \in \mathcal{L}_0 \mid \forall S \in \mathbf{S}(X \diamond \alpha), S \models \alpha'\}.$$

Since our framework is based on the same notion of distance, the fact that α_2 is in the deductive closure of an agent's belief base after its update with α_1 can be represented as the agent's implicit belief that, if α_1 were true, α_2 would be true.

Theorem 6. *Let $\alpha_1, \alpha_2 \in \mathcal{L}_0$ and $S \in \mathbf{S}$. Then*

$$\alpha_2 \in \mathit{upd}\big(B_i(S),\alpha_1\big) \text{ iff } S \models \Box_i(\alpha_1 \boxdot\!\!\rightarrow \alpha_2).$$

6 Epistemic Concepts

In this section, we highlight the expressive power of the language $\mathcal{L}$ and the semantics under which it is interpreted by using it to define a range of epistemic concepts that are widely studied in epistemology. A novel and interesting aspect of our approach, from the standpoint of expressiveness, is that it naturally captures counterfactual relations, including counterfactual dependence, both between the environment and the agents' epistemic states and among the agents' epistemic states themselves.

We start our conceptual analysis by the defining the classical notion of counterfactual dependence between two facts:

$$\mathsf{Dep}(\psi,\varphi) \stackrel{\text{def}}{=} \psi \wedge \varphi \wedge (\neg\varphi \boxdot\!\!\rightarrow \neg\psi).$$

According to the previous definition, ψ counterfactually depends on φ, denoted by $\mathsf{Dep}(\psi,\varphi)$, if both formulas are true and, moreover, if φ were false, then ψ would be false as well.

Counterfactual dependence allows us to define two counterfactual notions of epistemic reason: *explicit counterfactual reasons* and *implicit counterfactual reasons*. These two notions are based on two assumptions: (i) an agent's reason for believing a certain fact is mental (or internal to the agent) insofar as it consists in another belief of the agent that supports the first belief, and (ii) the support relation between a belief and the reason on which that belief is based is counterfactual in nature. In epistemology, this relation is known as the *basing relation*. Different theories analyze it in different ways: causally, counterfactually, dispositionally, or doxastically (see [Carter and Bondy, 2019; Korcz, 2024] for a general overview of competing theories of the basing relation). Following the taxonomy of epistemic reasons proposed in [Sylvan, 2016a; 2016b], we adopt a mentalist (internalist) view of epistemic reasons and a counterfactual interpretation of the basing relation.[2]

[2] Another important distinction in the theory of epistemic reasons is that between normative reasons (i.e., reasons why an agent *ought to* believe something) and operative reasons (i.e., reasons why an agent *holds* a belief) [Sylvan, 2016b]. We remain agnostic about this distinction, since in the case of logic-based artificial agents, which are our focus, the normative and descriptive views coincide: the beliefs an agent ought to have (by applying the normative principles of deductive rationality) coincide with the beliefs the agent holds.

An explicit counterfactual reason is an agent's explicit belief on which one of its implicit beliefs counterfactually depends. That is, the agent explicitly believes α, and can deduce φ from its explicit beliefs, but it would not be able to deduce φ if it did not explicitly believe α. In this sense, explicitly believing α is a necessary condition for the agent to infer that φ is true. In formal terms:

$$\mathsf{EReas}_i(\alpha,\psi) \stackrel{\text{def}}{=} \mathsf{Dep}(\Box_i\psi,\triangle_i\alpha),$$

where $\mathsf{EReas}_i(\alpha,\psi)$ means that the explicitly belief that α is an explicit counterfactual reason for agent i to deduce ψ.

To define an implicit counterfactual reason, we simply replace an explicit belief with an implicit one:

$$\mathsf{IReas}_i(\varphi,\psi) \stackrel{\text{def}}{=} \mathsf{Dep}(\Box_i\psi,\Box_i\varphi),$$

where $\mathsf{IReas}_i(\varphi,\psi)$ means that the implicit belief that φ is an implicit counterfactual reason for agent i to deduce ψ. The difference between an explicit and an implicit reason is that, in the former, what the agent can deduce counterfactually depends on its explicit beliefs, whereas in the latter, it depends on something else the agent can deduce. We now return to our example to illustrate the previous definitions.

Example 2 (cont.)**.** *The following holds:*

$$(S_0, \mathbf{S}_{(\Gamma_0,\alpha_0)}) \models \mathsf{EReas}_i(al,ev) \wedge \mathsf{IReas}_i(fi,ev)$$

This means that, for each robot, the ringing of the alarm provides an explicit reason to conclude that the building must be evacuated, whereas the fact that there is a fire in the building provides an implicit reason for reaching the same conclusion.

As the following theorem shows, under the conditions that the model includes all states and beliefs are about formulas of the language $\mathcal{L}_0$, the existence of an explicit counterfactual reason coincides with the fact that the agent would no longer be able to deduce the supported proposition if it were to forget the reason. Thus, under these conditions, counterfactual dependence of the supported belief on its explicit reason can be characterized in purely *internal* terms, namely, only in terms of operations on the agent's belief base.[3]

Theorem 7. *Let $\alpha_1, \alpha_2 \in \mathcal{L}_0$ and $S \in \mathbf{S}$. Then*
$S \models \mathsf{EReas}_i(\alpha_1,\alpha_2)$ *iff* $\alpha_1 \in B_i(S), \alpha_2 \in Cn\big(B_i(S)\big)$, *and*

$$\alpha_2 \notin forg\big(B_i(S), \alpha_1\big).$$

Nozick's counterfactual definition of knowledge [Nozick, 1981] emphasizes that, for a true belief to count as knowledge, there must be a counterfactual dependence between the true proposition and the agent's belief in it. That is:

$$\mathsf{K}_i\varphi \stackrel{\text{def}}{=} \mathsf{Dep}(\Box_i\varphi,\varphi).$$

Here $\mathsf{K}_i\varphi$ denotes that agent i knows that φ, according to Nozick's counterfactual theory of knowledge.[4]

[3]The idea that the basing relation should be internal to the agent is discussed in [Korcz, 2000] as a solution to deviant causal chains and to cases where one belief causes another accidentally, rather than through inference, as described in [Audi, 1986; Plantinga, 1993].

[4]Nozick includes a second condition in his definition, namely that if φ were true, the agent would believe φ, that is, $\varphi \boxdot\!\!\rightarrow \Box_i\varphi$. In our framework, this condition is redundant since, by the strong centering property satisfied by our counterfactual conditionals, it is implied by $\varphi \wedge \Box_i\varphi$, which is part of the definition of $\mathsf{K}_i\varphi$. He includes this second condition as he presupposes weak centering.

The condition $\neg\varphi \boxdot\!\!\rightarrow \neg\Box_i\varphi$ in the definition of $\mathsf{K}_i\varphi$ is known as the *sensitivity condition*: agent i's belief that φ is sensitive to the truth of φ. This condition ensures that the agent does not hold a true belief in φ merely by accident.

Example 3 (cont.)**.** *We have:*

$$(S_0, \mathbf{S}_{(\Gamma_0,\alpha_0)}) \models ev \wedge \Box_1 ev \wedge \neg(\neg ev \boxdot\!\!\rightarrow \neg\Box_1 ev),$$
$$(S_0, \mathbf{S}_{(\Gamma_0,\alpha_0)}) \models \triangle_2 al \wedge \Box_1\triangle_2 al \wedge (\neg\triangle_2 al \boxdot\!\!\rightarrow \neg\Box_1\triangle_2 al).$$

Thus, we have: $(S_0, \mathbf{S}_{(\Gamma_0,\alpha_0)}) \models \neg\mathsf{K}_1 ev \wedge \mathsf{K}_1\triangle_2 al$.

This means that, on the one hand, robot 1 *holds a true belief that the building should be evacuated, but this belief does not amount to knowledge because the sensitivity condition is not satisfied. In this sense, robot* 1*'s true belief is accidental. On the other hand, robot* 1 *knows that robot* 2 *explicitly believes that the alarm is ringing, since it believes this correctly and, moreover, it would not hold this belief if robot* 2 *did not explicitly believe that the alarm is ringing.*

The last concept we consider in this section is conditional belief, where an agent believes that ψ conditional on φ, denoted by $\mathsf{B}_i(\psi \mid \varphi)$, means that if the agent believed φ, it would deduce that ψ holds. As emphasized in Section 2, this concept has been widely employed in epistemic logic and formal epistemology [van Benthem, 2007; Baltag and Smets, 2006; Spohn, 2013]. While following the Lewisian tradition a counterfactual conditional is interpreted through a notion of comparative similarity, the interpretation of a conditional belief relies on a plausibility ordering over possible states. What we are going to show is that in our framework conditional belief can be reduced to a combination of implicit belief and counterfactual conditional, once an appropriate notion of plausibility is provided.

Following [Lorini, 2025, Def. 7] we compute an agent i's plausibility ordering over states from its belief base.

Definition 14 (Plausibility ordering)**.** *Let $i \in Agt$ and $S, S', S'' \in \mathbf{S}$. Then, $S'' \preceq_{i,S}^{\mathsf{plaus}} S'$ iff $Sat_i(S,S'') \subseteq Sat_i(S,S')$, where $Sat_i(S,S') = \{\alpha \in B_i(S) : S' \models \alpha\}$. We write $S'' \prec_{i,S}^{\mathsf{plaus}} S'$ if $S'' \preceq_{i,S}^{\mathsf{plaus}} S'$ and $S' \not\preceq_{i,S}^{\mathsf{plaus}} S''$.*

The fact $S'' \preceq_{i,S}^{\mathsf{plaus}} S'$ should be interpreted as "at state S agent i considers state S' at least as plausible as state S''". According to the previous definition, this means that all explicit beliefs of agent i at state S that are satisfied by state S'' are also satisfied by state S'. Clearly, the relation $\preceq_{i,S}^{\mathsf{plaus}}$ thus defined is a partial preorder. However, unlike the comparative similarity relation $\preceq_S^{\mathsf{sim}}$, it does not satisfy either the strong or the weak centering property, that is, the analogue of property (1) from Section 3.1, nor its weak variant in which strict ordering is replaced by weak ordering.

Let us define the dyadic conditional belief modality as the following abbreviation using counterfacual conditional:

$$\mathsf{B}_i(\psi \,|\, \varphi) \stackrel{\text{def}}{=} \Diamond_i\varphi \boxdot\!\!\rightarrow \Box_i(\varphi \rightarrow \psi).$$

According to the previous definition, a conditional belief coincides with a belief about material implication in all closest states to the actual one at which the condition is considered possible by the agent.

As the following theorem highlights, when the model includes all states, the previous abbreviation captures the standard notion of conditional belief as defined in epistemic logic, according to which believing that ψ conditional on φ means that ψ is true at all most plausible states satisfying the condition φ.

Theorem 8. *Let $S \in \mathbf{S}$. Then*

$$S \models \mathsf{B}_i(\psi \,|\, \varphi) \textit{ iff } \forall S' \in Best(i,\varphi,S), S' \models \psi,$$

where $Best(i,\varphi,S) = \{S' \in \mathbf{S} \colon S' \models \varphi$ and $\nexists S'' \in \mathbf{S}$ s.t. $S' \prec_{i,S}^{\mathsf{plaus}} S''$ and $S'' \models \varphi\}$.

7 Model Checking

In Definition 7 we introduced the notion of a model specification. We use this notion to formulate the following succinct model checking problem.

Succinct Model Checking Problem
input: model specification (S, Γ, α), $\varphi \in \mathcal{L}$;
output: **true** if $(S, \mathbf{S}_{(\Gamma,\alpha)}) \models \varphi$; **false** otherwise.

In this section, we study the computational complexity of this problem and study its connection with model checking w.r.t. the universal context $\mathbf{S}$.

Succinct model checking for (S, Γ, α) and φ can be done directly by definition, in polynomial space w.r.t. the size of the input (a detailed algorithm is provided in the supplementary material): (i) if φ is an atom, check $\varphi \in S$; (ii) if φ has a Boolean connective at the top, recursively check its immediate subformula(s); (iii) if $\varphi = \Box_i\varphi'$, check that $(S', \mathbf{S}_{(\Gamma,\alpha)}) \models \varphi'$ for every subset S' of Γ satisfying all formulas in $\{\alpha\} \cup \{\alpha' : \triangle_i\alpha' \in S\}$; (iv) if $\varphi = \varphi_1 \boxdot\!\!\rightarrow \varphi_2$, check that for every subset S' of Γ, either $S' \not\models \alpha$ or $(S', \mathbf{S}_{(\Gamma,\alpha)}) \models \varphi_2$, or $(S', \mathbf{S}_{(\Gamma,\alpha)}) \not\models \varphi_1$, or there exists a subset $S'' \subseteq \Gamma$ with $(S'', \mathbf{S}_{(\Gamma,\alpha)}) \models \varphi_1$ and $S \Delta S'' \subset S \Delta S'$. Although (iii)–(iv) also perform model checking on formulas in $\{\alpha\} \cup \{\alpha' : \triangle_i\alpha' \in S\}$ that are not subformulas of φ, these are all in $\mathcal{L}_0$ and require only linear time for model checking. Apart from these formulas, the recursion depth is bounded by the size of φ, so the procedure uses polynomial space, resulting in a PSPACE upper bound. A matching lower bound follows from the known reduction of TQBF to model checking of similarity-based counterfactuals [Aguilera-Ventura *et al.*, 2025].

Theorem 9. *Succinct model checking is PSPACE-complete.*

Model checking w.r.t. the universal context $\mathbf{S}$ is trickier. While in standard classical and modal logics the vocabulary can always be restricted to the atoms occurring in the input formula, this does not hold in our framework, since additional atoms of the form $\triangle_i\alpha$ may impose additional constraints on occurring atoms, changing the satisfiability of the subformulas in the accessible states. For example, for $\varphi = \Box_i\neg\Box_i p$, $(\emptyset, \mathbf{S}_{(\Gamma,\top)}) \models \varphi$ for $\Gamma = \{p\}$, since $\emptyset \mathcal{R}_i S$ and $S \mathcal{R}_i \emptyset$ for all $S \in \mathbf{S}_{(\Gamma,\top)}$; however, $\emptyset \not\models \varphi$ because $\Box_i p$ is satisfied, e.g., in the state $\{\triangle_i p\}$. Thus, in the general case, model checking w.r.t. the universal context requires considering non-trivial extensions of the set of atoms occurring in the input, and we conjecture that this problem belongs to a higher complexity class than the succinct model checking problem defined above. At the same time, we can isolate a fragment of the language that avoids such problematic instances but still contains many interesting formulas (including representations in Section 5). For this, we need to forbid changes of *polarity* in modal subformulas (such as the occurrence of a modality with existential polarity inside a modality with universal polarity, as in the example above). More precisely, we define a two-layered grammar of formulas in normal form with the connectives $\wedge$ and $\vee$ and without negations. The first layer $\mathcal{L}_\downarrow$ is given by the following grammar, allowing only universal modalities and only formulas from $\mathcal{L}_0$ in the antecedent of the counterfactual:

$$\varphi^\downarrow ::= \alpha \mid \varphi^\downarrow \wedge \varphi^\downarrow \mid \varphi^\downarrow \vee \varphi^\downarrow \mid \Box_i\varphi^\downarrow \mid \alpha \boxdot\!\!\rightarrow \varphi^\downarrow,$$

where α ranges over $\mathcal{L}_0$ and i ranges over $\mathbb{A}$. The second layer $\mathcal{L}_\uparrow$ is given by the following grammar, allowing only formulas from $\mathcal{L}_\downarrow$ under the $\Box_i$ modality and the negation of formulas from $\mathcal{L}_\downarrow$ in the antecedent of the counterfactual:

$$\varphi^\uparrow ::= \alpha \mid \varphi^\uparrow \wedge \varphi^\uparrow \mid \varphi^\uparrow \vee \varphi^\uparrow \mid \Box_i\varphi^\downarrow \mid (\neg\varphi^\downarrow) \boxdot\!\!\rightarrow \varphi^\uparrow,$$

where α ranges over $\mathcal{L}_0$, i ranges over $\mathbb{A}$, and $\varphi^\downarrow$ ranges over $\mathcal{L}_\downarrow$. Thus, $\mathcal{L}_\uparrow$ allows epistemic modalities in the antecedent of a counterfactual, but only with existential polarity and only if no other epistemic modalities occur above. Notice that all the representation theorems in Section 5 comply with the fragment $\mathcal{L}_\uparrow$ (as $(\neg\Box_i\bot \wedge \triangle_i\alpha_1) \boxdot\!\!\rightarrow (\Box_i\alpha_2)$ can be equivalently rewritten as $(\neg(\Box_i\bot \vee \neg\triangle_i\alpha_1)) \boxdot\!\!\rightarrow (\Box_i\alpha_2)$). For $\mathcal{L}_\uparrow$, model checking w.r.t. the universal context can be reduced to succinct model checking.

Theorem 10. *$S \models \varphi^\uparrow$ iff $(S, \mathbf{S}_{(\Gamma,\top)}) \models \varphi^\uparrow$, for $\varphi^\uparrow \in \mathcal{L}_\uparrow$ and $\Gamma = S \cup \mathbb{P}(\varphi^\uparrow)$.*

8 Conclusion

We have presented a logical framework integrating beliefs and counterfactuals, and demonstrated its flexibility and expressiveness by using it to capture a wide range of belief change operations and epistemic concepts studied in epistemic logic and formal epistemology. On the computational side, we have provided a succinct formulation of model checking in our framework and proved its PSPACE-completeness. Moreover, we have identified a non-trivial fragment of our language for which model checking can be formulated more generally, without relying on a notion of finite vocabulary.

Future work will be devoted to studying the proof-theoretic aspects of our framework. In particular, we plan to come up with an axiomatization for the set of validities of the language $\mathcal{L}$. We also plan to extend the language $\mathcal{L}$ with epistemic group modalities, including common belief and distributed belief. Indeed, we plan to generalize to the collective level our analysis of the interplay between epistemic states and counterfactuals. On the practical side, we plan to come up with an implementation of the model checking algorithm used for proving the PSPACE upper bound, or a variant of it relying on a polysize reduction into TQBF.

Acknowledgements

This work is supported the ANR PRCE project DesiRes (ANR-25-CE23-7498-01) and the European Union's Horizon 2020 research and innovation programme under grant agreement No 101034440.

References

[Aguilera-Ventura *et al.*, 2025] Carlos Aguilera-Ventura, Xinghan Liu, Emiliano Lorini, and Dmitry Rozplokhas. A non-interventionist approach to causal reasoning based on Lewisian counterfactuals. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2025)*, pages 4301–4310. ijcai.org, 2025.

[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.

[Audi, 1986] Robert Audi. Belief, reason, and inference. *Philosophical Topics*, 14(1):27–65, 1986.

[Baltag and Smets, 2006] Alexandru Baltag and Sonja Smets. Conditional doxastic models: A qualitative approach to dynamic belief revision. *Electronic Notes in Theoretical Computer Science*, 165:5–21, 2006.

[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.

[Carter and Bondy, 2019] Joseph Adam Carter and Patrick Bondy, editors. *Well-Founded Belief: New Essays on the Epistemic Basing Relation*. Routledge Studies in Epistemology. Routledge, New York, 2019.

[Delgrande and Wang, 2015] James Delgrande and Kewen Wang. A syntax-independent approach to forgetting in disjunctive logic programs. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1482–1488, Austin, Texas, USA, 2015. AAAI Press.

[Fagin *et al.*, 1983] Ronald Fagin, Jeffrey Ullman, and Moshe Vardi. On the semantics of updates in databases. In *Proceedings of the 2nd ACM SIGACT-SIGMOD Symposium on Principles of Database Systems (PODS '83)*, pages 352–365, New York, NY, USA, 1983. Association for Computing Machinery.

[Fagin *et al.*, 1995] Ronald Fagin, Joseph Halpern, Yoram Moses, and Moshe Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, 1995.

[Friedman and Halpern, 1994] Nir Friedman and Joseph 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.

[Ginsberg, 1986] Matthew Ginsberg. Counterfactuals. *Artificial Intelligence*, 30(1):35–79, 1986.

[Hansson, 1993] Sven Ove Hansson. Theory contraction and base contraction unified. *Journal of Symbolic Logic*, 58(2):602–625, 1993.

[Hansson, 1999] Sven Ove Hansson. *A Textbook of Belief Dynamics: Theory Change and Database Updating*. Kluwer, Dordrecht, Netherlands, 1999.

[Hintikka, 1962] Jaakko Hintikka. *Knowledge and Belief*. Cornell University Press, New York, 1962.

[Holliday, 2015] Wesley Holliday. Epistemic closure and epistemic logic I: Relevant alternatives and subjunctivism. *Journal of Philosophical Logic*, 44(1):1–62, 2015.

[Katsuno and Mendelzon, 1991] Hirofumi Katsuno and Alberto Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52(3):263–294, 1991.

[Kenny and Keane, 2021] Eoin Kenny and Mark 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.

[Korcz, 2000] Keith Allen Korcz. The causal-doxastic theory of the basing relation. *Canadian Journal of Philosophy*, 30(4):525–550, 2000.

[Korcz, 2024] Keith Allen Korcz. The epistemic basing relation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2024 edition edition, 2024.

[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.

[Levi, 1977] Isaac Levi. Subjunctives, dispositions and chances. *Synthese*, 34(4):423–455, 1977.

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

[Lin and Reiter, 1994] Fangzhen Lin and Raymond Reiter. Forget it! In *Proceedings of the AAAI Fall Symposium on Relevance*, pages 154–159, 1994.

[Lorini, 2020] Emiliano Lorini. Rethinking epistemic logic with belief bases. *Artificial Intelligence*, 282, 2020.

[Lorini, 2025] Emiliano Lorini. An epistemic theory of deductive arguments. In *Proceedings of the 22nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 450–461, 2025.

[Meyer and van der Hoek, 1995] John-Jules Meyer and Wiebe van der Hoek. *Epistemic logic for AI and computer science*. Cambridge University Press, 1995.

[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 Kommiya 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, pages 607–617, New York, USA, 2020. Association for Computing Machinery.

[Nebel, 1989] Bernhard Nebel. A knowledge level analysis of belief revision. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 301–311, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers.

[Nozick, 1981] Robert Nozick. *Philosophical Explanations*. Harvard University Press, Cambridge, MA, 1981.

[Plantinga, 1993] Alvin Plantinga. *Warrant: The Current Debate*. Oxford University Press, Oxford, UK, 1993.

[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, 2019.

[Spohn, 2013] Wolfgang Spohn. A ranking-theoretic approach to conditionals. *Cognitive Science*, 37(6):1074–1106, 2013.

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

[Stalnaker, 2002] Robert Stalnaker. Common ground. *Linguistics and Philosophy*, 25(5-6):701–721, 2002.

[Sylvan, 2016a] Kurt Sylvan. Epistemic reasons I: Normativity. *Philosophy Compass*, 11(6):339–351, 2016.

[Sylvan, 2016b] Kurt Sylvan. Epistemic reasons II: Basing. *Philosophy Compass*, 11(7):377–389, 2016.

[van Benthem, 2007] Johan van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 17(2):129–155, 2007.

[von Kügelgen *et al.*, 2023] Julius von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In *Conference on Causal Learning and Reasoning (CLeaR 2023)*, volume 213 of *Proceedings of Machine Learning Research*, pages 177–196. PMLR, 2023.

[Winslett, 1988] Marianne Winslett. Reasoning about action using a possible models approach. In *Proceedings of the Seventh AAAI National Conference on Artificial Intelligence (AAAI 1988)*, pages 89–93. AAAI Press, 1988.