Counterfactuals as Modal Conditionals, and Their Probability

Giuliano Rosella[1] , **Tommaso Flaminio**[2] , **Stefano Bonzio**[2,3]

[1]Department of Philosophy and Education Sciences, University of Turin
[2]Artificial Intelligence Research Institute (IIIA - CSIC), Barcelona
[3]Department of Mathematics and Computer Science, University of Cagliari
giuliano.rosella@unito.it, tommaso@iiia.csic.es, stefano.bonzio@unica.it

Abstract

In this paper we propose a semantic analysis of Lewis' counterfactuals. By exploiting the structural properties of the recently introduced boolean algebras of conditionals, we show that counterfactuals can be expressed as formal combinations of a conditional object and a normal necessity modal operator. Specifically, we introduce a class of algebras that serve as modal expansions of boolean algebras of conditionals, together with their dual relational structures. Moreover, we show that Lewis' semantics based on sphere models can be reconstructed in this framework. As a consequence, we establish the soundness and completeness of a slightly stronger variant of Lewis' logic for counterfactuals with respect to our algebraic models. In the second part of the paper, we present a novel approach to the probability of counterfactuals showing that it aligns with the uncertainty degree assigned by a belief function, as per Dempster-Shafer theory, to its associated conditional formula. Furthermore, we characterize the probability of a counterfactual in terms of Gärdenfors' imaging rule for the probabilistic update.

1 Main Contributions of the Paper

Counterfactuals, statements like "if a were the case, then b would be the case", formally $a \mathrel{\Box\!\!\rightarrow} b$, are crucial for various fields like logic, linguistics, and AI. However, analyzing them often relies on traditional methods. In this work, we introduce a novel framework to analyze Lewis counterfactuals that allows us to combine logical, algebraic, and probabilistic approaches.

Algebraic Models for Lewis Counterfactuals. The best-know approach to counterfactuals and their logic originates from the philosophical literature, in particular from (Lewis 1973) and (Stalnaker 1968). Lewis' logical framework is based on sphere models, which consist of Kripke-style semantic structures. While Lewis' framework and derivatives approaches are still widely used, an algebraic analysis of Lewis counterfactuals is notably missing in the literature. Such investigation would allow us to deepen our understanding of Lewis counterfactuals, their logic and their meaning.

We bridge this gap by introducing a new class of algebras that we call *Lewis Algebras* which consist of a Boolean algebra of conditionals from (Flaminio, Godo, and Hosni 2020) equipped with a normal modal operator $\Box$ satisfying specific properties. This structures contain conditional expressions of the form $(a \mid b)$, read as "a given b", Boolean combinations of those, e.g. $(a \mid b) \wedge (c \mid d)$, $\neg(c \mid d)$, and modal conditionals like $\Box(b \mid a)$ that serve to interpret counterfactuals.

We then delve into an investigation of the properties of Lewis algebras and their dual Kripke frames, which we call *Lewis frames*, through the mirror of Jónnson-Tarski duality (Blackburn, de Rijke, and Venema 2001). By exploiting this duality, we show that every Lewis algebra induces a Lewisian sphere model for counterfactuals, and vice versa every Lewis sphere model for counterfactuals induce a Lewis algebra. More schematically:

- We introduce a new class of algebraic structures, i.e. Lewis algebras, and investigate their mathematical properties;
- We characterize their dual structures and connect them to Lewis models for counterfactuals.

Representation of Lewis Counterfactuals. We leverage Lewis algebras to represent Lewis' logic of counterfactuals C1 in (Lewis 1971). We prove that a slightly stronger version of C1 is sound and complete with respect to Lewis algebras, and consequently Lewis frames. This result goes beyond the technical side: it offers a new conceptual understanding of counterfactuals. In fact, key to this approach is interpreting Lewis counterfactuals $a \mathrel{\Box\!\!\rightarrow} b$ as necessitated modal conditionals $\Box(b \mid a)$ within Lewis algebras. This allows us to put forward a reductionist perspective on counterfactuals and define them by combining a modal necessity operator with probabilistic conditionals. Unlike traditional approaches that regard counterfactuals as primitive operators, Lewis Algebras enable us to re-elaborate the classical truth conditions of counterfactuals. More schematically:

- We interpret Lewis counterfactuals $a \mathrel{\Box\!\!\rightarrow} b$ as necessitated modal conditionals $\Box(b \mid a)$ within Lewis algebras, offering a new conceptual understanding of them;
- We prove soundness and completeness of (a slightly stronger version of) Lewis logic of counterfactuals C1 with respect to our models;

Probability of Lewis Counterfactuals. Our work extends to the probability of counterfactuals, a longstanding challenge in philosophy. Notably, it is unclear how to assign a probability to counterfactuals, like $P(a \mathrel{\Box\!\!\rightarrow} b)$ (see (Lewis 1976)). We address this problem by leveraging our finding that Lewis counterfactuals can be interpreted as modal conditionals $a \mathrel{\Box\!\!\rightarrow} b \approx \Box(b \mid a)$. By appealing to classical results connecting modal logic with Dempster-Shafer belief functions, e.g. (Harmanec, Klir, and Resconi 1994) and (Resconi, Klir, and Clair 1992), we prove that the probability of a Lewis counterfactual $a \mathrel{\Box\!\!\rightarrow} b$ corresponds to the belief function of the consequent b, *imaged* on the antecedent a, denoted $Bel_a(b)$.

Imaged belief functions have been introduced in (Dubois and Prade 1994); intuitively Bel_a can be viewed as un *updated* version of an original belief function Bel, filtered through a. In essence, we show that $P(a \mathrel{\Box\!\!\rightarrow} b) = Bel_a(b)$. This result extends to Lewis algebras. Specifically, starting with a probability P over a Boolean algebra $\mathbf{A}$, we can define a *canonically extended probability* μ_P within the Lewis algebra $\langle \mathcal{C}(\mathbf{A}), \Box \rangle$. The probability of a counterfactual under this framework, $\mu_P(\Box(b \mid a))$, coincides with both the original $P(a \mathrel{\Box\!\!\rightarrow} b)$ and $Bel_a(b)$. More schematically:

- We characterize the probability of Lewis counterfactuals using imaged belief functions: $P(a \mathrel{\Box\!\!\rightarrow} b) = Bel_a(b)$;
- We extend the above characterization result to Lewis Algebras using canonically extended probabilities.

References

Blackburn, P.; de Rijke, M.; and Venema, Y. 2001. *Modal Logic*. Cambridge University Press.

Byrne, R. M. J. 2019. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI-2019. International Joint Conferences on Artificial Intelligence Organization.

Dubois, D., and Prade, H. 1994. A survey of belief revision and updating rules in various uncertainty models. *International Journal of Intelligent Systems* 9(1):61–100.

Flaminio, T.; Godo, L.; and Hosni, H. 2020. Boolean algebras of conditionals, probability and logic. *Artificial Intelligence* 286:103347.

Galles, D., and Pearl, J. 1998. An axiomatic characterization of causal counterfactuals. *Foundations of Science* 3:151–182.

Ginsberg, M. L. 1986. Counterfactuals. *Artificial Intelligence* 30(1):35–79.

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

Halpern, J. Y. 2013. From Causal Models to Counterfactual Structures. *Review of Symbolic Logic* 6(2):305–322.

Harmanec, D.; Klir, G. J.; and Resconi, G. 1994. On modal logic interpretation of dempster-shafer theory of evidence. *International Journal of Intelligent Systems* 9(10):941–951.

Kripke, S. A. 1965. Semantical analysis of intuitionistic logic i. In *Formal Systems and Recursive Functions*. Elsevier. 92–130.

Lewis, D. 1971. Completeness and decidability of three logics of counterfactual conditionals1. *Theoria* 37(1):74–85.

Lewis, D. K. 1973. *Counterfactuals*. Cambridge, MA, USA: Blackwell.

Lewis, D. 1976. Probabilities of conditionals and conditional probabilities. *Philosophical Review* 85(3):297–315.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38.

Moravčík, M.; Schmid, M.; Burch, N.; Lisý, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; and Bowling, M. 2017. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* 356(6337):508–513.

Neal, L.; Olson, M.; Fern, X.; Wong, W.-K.; and Li, F. 2018. *Open Set Learning with Counterfactual Images*. Springer International Publishing. 620–635.

Pearl, J. 2000. *Causality*. Cambridge: Cambridge University Press.

Resconi, G.; Klir, G. J.; and Clair, U. S. 1992. Hierarchical uncertainty metatheory based upon modal logic. *International Journal of General Systems* 21(1):23–50.

Rosella, G.; Flaminio, T.; and Bonzio, S. 2023. Counterfactuals as modal conditionals, and their probability. *Artificial Intelligence* 323:103970.

Stalnaker, R. 1968. A Theory of Conditionals. In Rescher, N., ed., *Studies in Logical Theory*, number 2 in American Philosophical Quarterly Monograph Series. Oxford: Blackwell. 98–112.