# Causal Modeling Semantics for Counterfactuals with Disjunctive Antecedents

Giuliano Rosella* Jan Sprenger†

February 12, 2022

**Abstract**

This paper applies Causal Modeling Semantics (CMS, e.g., Galles and Pearl 1998; Pearl 2000; Halpern 2000) to the evaluation of the probability of counterfactuals with disjunctive antecedents. Standard CMS is limited to evaluating (the probability of) counterfactuals whose antecedent is a conjunction of atomic formulas. We extend this framework to disjunctive antecedents, and more generally, to any Boolean combinations of atomic formulas. Our main idea is to assign a probability to a counterfactual $(A \vee B) \boxdot\!\!\rightarrow C$ at a causal model $\mathcal{M}$ by looking at the probability of $C$ in those submodels that *truthmake* $A \vee B$ (Briggs 2012; Fine 2016, 2017). The probability of $p((A \vee B) \boxdot\!\!\rightarrow C)$ is then calculated as the average of the probability of $C$ in the truthmaking submodels, weighted by the inverse distance to the original model $\mathcal{M}$. The latter is calculated on the basis of a proposal by Eva et al. (2019). Apart from solving a major problem in the research on counterfactuals, our paper shows how work in semantics, causal inference and formal epistemology can be fruitfully combined.

**Keywords**: Counterfactuals; Causal Modeling Semantics; Similarity Distance; Probability of Counterfactuals.

*Corresponding author. Center for Logic, Language and Cognition (LLC), Department of Philosophy and Education, Palazzo Nuovo, Via Sant'Ottavio 20, 10124 Torino, Italy. giuliano.rosella@unito.it

†Center for Logic, Language and Cognition (LLC), Department of Philosophy and Education, Palazzo Nuovo, Via Sant'Ottavio 20, 10124 Torino, Italy. jan.sprenger@unito.it

# 1 Introduction

How should we evaluate the probability of a counterfactual like "If it had been sunny, Mary would have attended the football match"? We are inclined to say that it should be in line with the probability of the consequent "Mary attends the football match" in a hypothetical situation where the sun was actually shining. The debate is about how to determine this probability. The conditional probability $p(C|A) = p(A \wedge C)/p(A)$, or equivalently, the probability of $C$ after conditionalizing on $A$, has been proposed by Adams (1965, 1975) for evaluating *indicative conditionals*, but arguably, conditionalization is no adequate tool for evaluating counterfactuals (e.g., the Oswald/Kennedy example in Adams 1970; see also Lewis 1973b, p. 72).

Which mechanism should replace conditionalization for calculating the probability of a counterfactual $A \boxrightarrow C$? David Lewis's similarity semantics (LSS) for counterfactuals gives a precise account of their truth conditions (Lewis 1973a,b), but is silent on their probability. A couple of years later, Lewis (1976) closed that gap by means of suggesting the equation $p(A \boxrightarrow C) = p_A(C)$: we should *image* the probability distribution $p$ on the $A$-worlds and evaluate $C$ relative to that distribution (see also Gärdenfors 1982; Günther 2022).

On the other hand, computer scientists and formal epistemologists have proposed *causal modeling semantics (CMS)*: a counterfactual $A \boxrightarrow C$ needs to be evaluated by looking at the truth value of its consequent $C$ after a suitable *intervention* on $A$ (Skyrms 1980; Pearl 2000, 2017). This proposal, which relies on causal models as a graphical tool for reasoning and inference, is elaborated in Galles and Pearl (1998) and developed further in Briggs (2012). Transferring this approach to probability, one obtains $p(A \boxrightarrow C) = p(C|do(A))$, i.e., the probability of the counterfactual is the probability of $C$ after intervening on $A$.[1]

The divergences and convergences of CMS and LSS have been studied from various angles. Pearl (2000, pp. 72-73) shows that a certain type of imaging is equivalent to an intervention on $A$ that is represented by the *do*-operator. It is agreed, however, that standard CMS and LSS are different in (at least) one crucial respect (Briggs 2012; Halpern 2013; Pearl 2017): they assign truth conditions to different classes of counterfactuals. Standard CMS, as developed in Galles and Pearl 1998, cannot account for the truth conditions of counterfactuals with disjunctive antecedents of the form $(A \vee B) \boxrightarrow C$. That is, we cannot assign truth conditions to a sentence such as "if it had been sunny *or* the tickets had been discounted, Mary would have attended the football match", and the same holds for the assignment of probabilities of

[1] In general, $p(C|do(A)) \neq p(C|A)$, unless $C$ is causally downstream of $A$. Both CMS and LSS thus avoid the well-known triviality results for the probability of conditionals (Lewis 1976).

the form $p(C|do(A \vee B))$. It is simply not clear what it means to intervene as to satisfy the logical *disjunction* of two propositions. In other words, while CMS has a very strong theoretical motivation, it has limited expressive power.

By contrast, the LSS framework assigns truth values to counterfactuals with *arbitrary* antecedents. For the disjunctive case, they are determined by the truth value of $C$ in the closest possible $(A \vee B)$-world(s). The probability of the counterfactual $(A \vee B) \boxright C$ is given by the probability of $C$ after *imaging* the probability distribution on $A \vee B$. However, the interpretation and logical properties of counterfactuals with disjunctive antecedents are the subject of substantive debate (e.g., Nute 1975; Loewer 1976; McKay and Van Inwagen 1977), and LSS does not determine a canonical algorithm for calculating $p((A \vee B) \boxright C)$.

Our paper develops a novel proposal for evaluating the probability of such counterfactuals. Building on Briggs' 2012 pioneer work, which combines CMS with truthmaker semantics (Fine 2016, 2017), we propose a CMS-based account for evaluating the probability of counterfactuals with disjunctive antecedents. We work in a propositional language allowing for simple (i.e., non-nested) counterfactuals. Specifically, we propose to evaluate the probability of $(A \vee B) \boxright C$ as the weighted probability of $C$ in all submodels that *truthmake* $A \vee B$. Their weights are determined by the algorithm developed in Eva et al. (2019). This procedure extends to calculating the probability of counterfactuals with arbitrary Boolean compounds of atomic formulas in the antecedent.

Our proposal illustrates how work in semantics, formal epistemology and causal modeling can join forces in order to solve a longstanding conceptual problem. It synthesizes truthmaker semantics with ideas from LSS and CMS, yielding more convincing results than LSS, and more general results than standard CMS alone. At the same time, our account preserves some elements of LSS, by weighting the contributions of causal submodels as a function of their similarity to the original model.

The paper is structured as follows. In Section 2 and 3, respectively, we recapitulate the basics of causal modeling semantics and explain how truthmaker semantics can serve to establish a logic of counterfactuals. Section 4 introduces probabilistic causal models, Section 5 outlines our account and Section 6 compares it with the LSS treatment of the probability of counterfactuals. Section 7 wraps up our results and suggests future work.

## 2 Causal Modeling Semantics (CMS)

This section recaps the causal modeling framework for the semantics of counterfactuals (CMS, e.g., Galles and Pearl 1998; Pearl 2000; Halpern 2000), as presented by Briggs (2012). First, we need to introduce causal models, us-

ing a running example (simplified from Pearl 2000) that will accompany us throughout the paper. It involves four Boolean variables, whose values are represented by the numbers zero and one.

> A prisoner is condemned to death and led to the execution court. He stands in front of two soldiers, who will fire at the captain's signal. If at least one of the soldiers fires, the prisoner dies. The captain gives the signal ($C = 1$), the two soldiers fire ($X = 1$, $Y = 1$), and the prisoner dies ($D = 1$).

The main ingredients of this causal model are a set of variables $\mathcal{V} = \{C, X, Y, D\}$, and the set of structural equations that describe their causal dependencies: $\mathcal{S} = \{X = C, Y = C, D = max(X, Y)\}$. This means that the executioners fire if the captain gives the signal and the prisoner dies if one of the two executioners fires. The dependencies can also represented graphically, as in Figure 1 below.

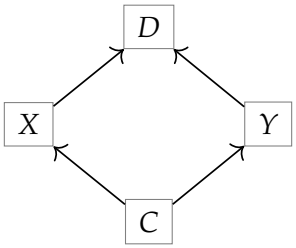


Figure 1: Causal graph for the prisoner execution story. $C$ stands for the captain (not) firing, $X, Y$ for the soldiers (not) shooting, $D$ for the prisoner dying/living.

The *parents* $PA(V)$ of a variable $V$ are simply the variables from which there is an arrow into $V$. For example, $C$ is the only parent of $X$ and $Y$, and $X$ and $Y$ are the parents of $D$. Structural equations describe the value of a variable as a function of the value of its parents. In general, a **causal model** $\mathcal{M}$ is a triple $\mathcal{M} = \langle \mathcal{V}, \mathcal{S}, a \rangle$ where:

- $\mathcal{V}$ is a non-empty finite set of variables $\mathcal{V} = \{V_1, V_2, ..., V_n\}$;
- $\mathcal{S}$ is a set of structural equations, where each element has the form $V = f_V(V_{i_1}, V_{i_2}, \ldots, V_{i_n})$ and $PA(V) = \{V_{i_1}, \ldots, V_{i_n}\}$ (i.e., each structural equation defines the value of $V$ uniquely by the value of its parents; no cycles are allowed);
- $a : \mathcal{V} \to \mathcal{R}(\mathcal{V})$ is a function assigning an *actual value* to each variable $V$, in a way that is consistent with the range of $V$ and the structural equations.

The last part, the assignment of actual values, is not necessarily required for making predictions with causal models, but it is crucial when we want to use them for counterfactual reasoning.

Some additional terminology will be useful: when a variable $V_1$ is connected to another variable $V_1$ via a sequence of directed arrows from $V_1$ into $V_2$, we say that $V_2$ is a *descendant* of $V_1$. For instance, in Figure 1, $D$ is a descendant of $C$, $X$ and $Y$. As in (Briggs 2012), we will restrict our attention to only those models not containing any loop, namely models in which there is no sequence of arrows connecting a variable to itself. Moreover, in a causal model, we say that a variable is *exogenous* when it has no parents (e.g., $C$ in Figure 1) and *endogenous* when it is not exogenous, so that its value can be determined by the value of other variables in the model (e.g., $X$, $Y$ and $D$ in Figure 1).

Now, we need to introduce the notion of an *intervention* on a causal model. An atomic formula in our language has the form $V = v$, expressing the fact that the variable $V$ takes a certain value $v$. The intervention $do(V = v)$ on a causal model $\mathcal{M}$ breaks the dependency of $V$ on its parents via the structural equations (i.e., it eliminates all arrows into $V$) and assigns the value $V = v$ to it. The intervention generates a causal submodel $\mathcal{M}'$ where the formula $V = v$ is true and the structural equation $f_V$ is no longer part of the causal model: the variable $V$ now depends on the intervention, but not any more on its parents.

We can generalize this idea to conjunctions of interventions. For a causal model $\mathcal{M} = \langle \mathcal{V}, \mathcal{S}, a \rangle$, the intervention $do(V_1 = v_1, V_2 = v_2, \ldots, V_n = v_n)$ generates a **submodel** $\mathcal{M}' = \langle \mathcal{V}', \mathcal{S}', a' \rangle$ of $\mathcal{M}$ such that:

- $\mathcal{V}' = \mathcal{V}$, i. e. $\mathcal{M}'$ has the same variables as $\mathcal{M}$;
- $\mathcal{S}' = \mathcal{S} \setminus \{f_{V_1}, \ldots, f_{V_n}\}$;
- $a' : \mathcal{V} \setminus \{V_1, V_2, \ldots, V_n\} \to \mathcal{R}(\mathcal{V})$ assigns actual values to the variables not affected by the intervention, in line with the structural equations in $\mathcal{S}'$.

Conceptually, an intervention on a causal model manipulates some variables, forces them to take a certain value and breaks the causal mechanism between them and their parents. For an example, consider the causal model of the execution story depicted above; we want to know what would have happened if the two executioners had not fired $(X = 0 \wedge Y = 0)$. The answer is given by the intervention $do(X = 0, Y = 0)$ which would generate the model in Figure 2.

Our intervention has broken the causal mechanism that links $C$ to $X$ and $Y$, and we have forced $X$ and $Y$ to value zero. What happens to $D$ now? It continues to be determined by the structural equation $D = max(X, Y)$, but $X = 0$ and $Y = 0$ as a result of our intervention, hence $D = max(0, 0) = 0$. And so the prisoner will live.

The concept of intervention in a causal model explicates our intuitive counterfactual reasoning: in order to know *what would have happened* to the

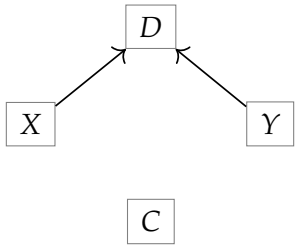


Figure 2: Causal graph for the prisoner execution story, where we intervene on $X$ and $Y$ and break the dependency on the captain's signal $C$.

prisoner *had the executioners not fired*, we perform an intervention on the latter and see how it would have affected the prisoner, according to the known causal mechanisms.

More generally, in standard CMS, counterfactuals and interventions are connected in the following way: a counterfactual of the form $(A_1 \wedge A_2 \wedge ... \wedge A_n) \boxdot\!\!\rightarrow B$ is true at a causal model $\mathcal{M}$ that contains $A_1, \ldots, A_n$ and $B$ as variables if and only if $B = 1$ holds at the causal model $\mathcal{M}'$ generated by the intervention $do(A_1 = 1, A_2 = 1, \ldots, A_n = 1)$ on $\mathcal{M}$. (As before, we use $A_1 = 1$ for expressing that the Boolean variable $A_1$ takes the value "true".) For instance, the counterfactual "if the two executioners hadn't fired, then the prisoner would not have died" is true at the causal model of the execution story since, as we have seen above, after performing the intervention $do(X = 0, Y = 0)$, $D = 0$ holds in the new submodel.

Notice that an intervention of the form $do(A)$ is only defined when $A$ is an atomic formula or a conjunction of atomic formulas. This imposes a restriction on the class of counterfactuals that standard CMS can account for: only counterfactuals of the form $(A_1 \wedge A_2 \wedge ... \wedge A_n) \boxdot\!\!\rightarrow B$ can assume a truth value. CMS does not provide truth conditions for counterfactuals with logically complex antecedents. For instance, we cannot say whether the counterfactual "if one of the two executioners hadn't fired, then the prisoner would not have died" ($(X = 0 \vee Y = 0) \boxdot\!\!\rightarrow D = 0$) is true or false at the causal model of the execution story. This limitation is due to the fact that the *disjunctive intervention* $do(X = 0 \vee Y = 0)$ is not defined (see also Pearl 2017). Intuitively, there is more than one possible realization of $do(X = 0 \vee Y = 0)$: we could manipulate $X$, $Y$, or both variables at the same time (compare Sartorio 2006; Briggs 2012; Günther 2017). Each of the three interventions $do(X = 0)$, $do(Y = 0)$ and $do(X = 0, Y = 0)$ would be a good candidate for an intervention that brings about the state "$X = 0$ or $Y = 0$". But their effects on $D = max(X, Y)$ differ. For the intervention $do(X = 0)$ and $do(Y = 0)$, the prisoner would still die (since the other soldier fires) but for the intervention $do(X = 0, Y = 0)$, he would live. Thus, if *just one* executioner hadn't fired, the

prisoner would have died anyway; if *both* hadn't fired, he would live. So, in the end, standard CMS as presented in Galles and Pearl 1998 and Pearl 2000 does not provide a unique answer to the question of evaluating counterfactuals with disjunctive antecedents. And this is arguably a disadvantage of CMS with respect to LSS, where Lewisian spheres or Stalnaker's selection functions provide definite answers to the question of in which worlds we need to evaluate counterfactuals, and how the results need to be combined (e.g., Lewis demands that the consequent holds in all nearest possible worlds). In order to overcome this shortcoming, Briggs (2012) has proposed an extension of CMS that we present in the next section.

## 3 Truthmaker Semantics for Causal Modeling

Briggs' extension relies on truthmaker semantics (TMS), a semantic framework developed in a series of recent publications by Kit Fine (2016, 2017). The idea underlying TMS is that of an *exact truthmaker* of a sentence $A$, namely something in the world which *truthmakes* $A$ and is *wholly relevant* for the truth of $A$.

This intuitive idea can be fruitfully combined with CMS. An intervention $do(A)$ is *admissible* on a causal model $\mathcal{M}$ when it does not perform two inconsistent value assignments to the same variable, like $do(V_1 = 0 \wedge V_1 = 1)$. For a causal model $\mathcal{M} = \langle \mathcal{V}, \mathcal{S}, a \rangle$, we can define the *set of submodels* of $\mathcal{M}$ generated by any intervention $do(A)$ as $S(\mathcal{M}) = \langle S, \sqcup \rangle$ where

- $S$ is the set of submodels of $\mathcal{M}$ generated by any admissible intervention $do(A)$;
- $\mathcal{M}[A]$ indicates the submodel generated by performing $do(A)$ on $\mathcal{M}$;
- $\sqcup$ is an operation of *fusion* among the models in $S$ defined by $\mathcal{M}[A] \sqcup \mathcal{M}[B] := \mathcal{M}[A \wedge B]$.

In other words, the fusion of the two submodels $\mathcal{M}[A]$ and $\mathcal{M}[B]$, defined by the interventions $do(A)$ and $do(B)$, corresponds to the submodel defined by the fusion of the two interventions. We assume that only consistent fusions are allowed, in the sense that $do(A \wedge B)$ is an admissibbile intervention on $\mathcal{M}$.

Now, consider a language $\mathcal{L}$ where atomic formulas have the form $V = v$ and complex formulas are obtained from Boolean combinations of atomic formulas. For a model $\mathcal{M}$, consider its space of proper submodels $S(\mathcal{M}) = \langle S, \sqcup \rangle$ where $\mathcal{M} \notin S$. We can inductively define relations of *truthmaking* $\Vdash \subseteq S \times \mathcal{L}$ and and falsemaking $\dashv\Vdash \subseteq S \times \mathcal{L}$ between any member $s$ of $S$ and formulas

in the language as follows:

$$
\begin{array}{lcl}
s \Vdash V = v & \Leftrightarrow & s = \mathcal{M}[V = v] \\
s \dashv\Vdash V = v & \Leftrightarrow & s = \mathcal{M}[V = v'] \text{ for some } v \neq v' \\
s \Vdash \neg A & \Leftrightarrow & s \dashv\Vdash A \\
s \Vdash A \wedge B & \Leftrightarrow & \textit{for some } t, u \ (t \Vdash A, u \Vdash B \textit{ and } s = t \sqcup u) \\
s \Vdash A \vee B & \Leftrightarrow & s \Vdash A, \ s \Vdash B, \textit{ or } s \Vdash A \wedge B
\end{array}
$$

where $s \Vdash A$ means that $s$ *truthmakes* (=is a truthmaker of) $A$. State $s$ is a truthmaker of $V = v$ if and only if it corresponds to the submodel defined by the intervention $do(V = v)$, and a falsemaker of $V = v$ if and only if it corresponds to the submodel defined by an intervention that sets $V$ to a value different from $v$. Since states in $S(\mathcal{M})$ can be identified with interventions, we can say, for simplicity, that an intervention $do(V_1 = v_1, ..., V_n = v_n)$ on $\mathcal{M}$ truthmakes a formula $A$ if and only if $\mathcal{M}[V_1 = v_1, ..., V_n = v_n]$ is a truthmaker of $A$.

Evidently, $s$ falsemakes $A$ iff $s$ is a truthmaker of $\neg A$. State $s$ truthmakes a *conjunction* of variable assignments iff it is the fusion of two states that truthmake the two individual assignments—in other words, iff $s$ is the causal submodel defined by the intervention that assigns the right values to both variables. Finally, $s$ is truthmaker of a *disjunction* of variable assignments iff it truthmakes one of the two assignments, or its conjunction. This interpretation of truthmaking a disjunction is also at the center of Briggs' (and our own) proposal for expanding CMS.

Consider a propositional language $\mathcal{L}$, which we extend to a language $\mathcal{L}^{\rightarrow}$ with a simple, non-nested counterfactual operator: for any formulas $A, B \in \mathcal{L}$, let $A \ \square\!\!\rightarrow \ B \in \mathcal{L}^{\rightarrow}$. We can now give inductively defined truth conditions for formulas of $\mathcal{L}^{\rightarrow}$, including simple counterfactuals.

**Truth Conditions for Formulas of $\mathcal{L}^{\rightarrow}$ (Briggs)** A $\mathcal{L}^{\rightarrow}$-formula is true at a causal model $\mathcal{M} = \langle \mathcal{V}, \mathcal{S}, a \rangle$ in the following conditions:

$$
\begin{array}{lcl}
\mathcal{M} \vDash V = v & \Leftrightarrow & a(V) = v \\
\mathcal{M} \vDash \neg A & \Leftrightarrow & \mathcal{M} \nvDash A \\
\mathcal{M} \vDash A \wedge B & \Leftrightarrow & \mathcal{M} \vDash A \textit{ and } \mathcal{M} \vDash B \\
\mathcal{M} \vDash A \vee B & \Leftrightarrow & \mathcal{M} \vDash A \textit{ or } \mathcal{M} \vDash B \\
\mathcal{M} \vDash A \ \square\!\!\rightarrow \ B & \Leftrightarrow & \textit{for every } s \textit{ in } S(\mathcal{M}) \textit{ such that } s \Vdash A, s \vDash B
\end{array}
$$

Thus, a counterfactual $A \ \square\!\!\rightarrow \ B$ is true at a causal model $\mathcal{M}$ if and only if $B$ is true at all the members of $S(\mathcal{M})$ that truthmake $A$. Consider again the execution example and the counterfactual "if one of the two executioners had not fired, then the prisoner would not have died". We can formalize this counterfactual as $(X = 0 \vee Y = 0) \ \square\!\!\rightarrow \ D = 0$. The truthmakers of $X = 0 \vee Y = 0$ are the submodels $\mathcal{M}[X = 0]$, $\mathcal{M}[Y = 0]$ and $\mathcal{M}[X = 0 \wedge Y = 0]$.

The first two submodels validate $D = max(X, Y) = 1$ since the second soldier is not affected by the intervention, and so $(X = 0 \vee Y = 0) \boxdot\!\!\rightarrow D = 0$ is false at $\mathcal{M}$.

Briggs' extension of CMS allows us to assign a truth value to counterfactuals with disjunctive antecedents—in fact, to counterfactuals with arbitrary Boolean compounds of atomic formulas in the antecedent. The main innovation to CMS consists in evaluating counterfactuals in the submodels that truthmake the antecedent. Implicit in Briggs' approach is a relevance principle for the truth conditions of counterfactuals, which we will also use later when defining their probability:

**Relevance Principle (Truth Conditions)** The truth value of a counterfactual $A \boxdot\!\!\rightarrow B$ at a causal model $\mathcal{M}$ depends exclusively on the truth value of $B$ in the submodels $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ generated by the interventions on the variables in $A$ that truthmake $A$.

## 4 Probabilistic Causal Models

In this section, we introduce probabilistic causal models in order to assign a probability to a counterfactual. We will also see how the problem of the limited expressive power of CMS re-emerges at the probabilistic level: causal modeling semantics does not allow to assign a probability to counterfactuals with disjunctive antecedent.

A **probabilistic causal model** is a tuple $\mathcal{M} = \langle \mathcal{V}, \mathcal{G}, p \rangle$ where

- $\mathcal{V}$ is a set of variables;
- $\mathcal{G} \subset \mathcal{V} \times \mathcal{V}$ is a set of directed edges between the variables in $\mathcal{V}$, defining the parents and descendants of each variable;
- $p$ is a probability distribution on $\mathcal{V}$ subject to the *Markov condition*, that is, each variable $V$ is probabilistically independent of its non-descendants, conditional on its parents.

The probability distribution $p$ fulfils the role of a structural equation (i.e., it describes how variables depend on their parents), but without the assumption of determinism. Consider again the execution scenario from Section 2 with the probability distribution $p$ described in Table 1. Thanks to the Markov condition, it is sufficient to specify the probability of the exogenous variables, and the conditional probability of the endogenous variables, given the values of their parents.

Analogously to the non-probabilistic case, probabilistic causal models provide an excellent tool for reasoning about counterfactuals. Again, the notion of an intervention is crucial. Pearl (2000) proposes that the probability of a

| C | |
|---|---|
| 1 | 0.5 |
| 0 | 0.5 |

| C | X | |
|---|---|---|
| | 1 | 0 |
| 1 | 0.9 | 0.1 |
| 0 | 0.1 | 0.9 |

| C | Y | |
|---|---|---|
| | 1 | 0 |
| 1 | 0.9 | 0.1 |
| 0 | 0.1 | 0.9 |

| X | Y | D | |
|---|---|---|---|
| | | 0 | 1 |
| 1 | 0 | 0.5 | 0.5 |
| 0 | 1 | 0.5 | 0.5 |
| 0 | 0 | 0.9 | 0.1 |
| 1 | 1 | 0.1 | 0.9 |

Table 1: Probability distribution for the variables in the execution example, as a function of the values of their parents.

counterfactual $A \boxdot\!\!\rightarrow C$ at a probabilistic causal model $\mathcal{P}$, given a certain evidence $E$, amounts to the probability of $B$ in the submodel generated by the intervention $do(A)$, where $A$ is an atomic formula or a conjunction of atomic formulas. In other words, $p(A \boxdot\!\!\rightarrow C|E) = p(C|do(A), E)$. This corresponds to the following procedure:

1. Update the probability $p(U = u)$ of each exogenous variable $U$ on the evidence $E$, via Bayesian conditionalization, to the new probability $p'(U' = u) = p(U = u|E)$, *without changing the conditional dependencies among the variables*. This is because the evidence should not change the structure of the causal relationships between the variables: it just informs us which context we are likely to be in (see Pearl 2000, pp. 33-38). So $p'$ induces a new probability distribution on the (endogenous) variables, too.
2. Perform the intervention $do(A)$ on $\mathcal{M}$ to obtain a new submodel $\mathcal{M}'$ of $\mathcal{M}$; accordingly, change the probability distribution so that variables involved in the intervention do not depend on their parents anymore.
3. Use the new submodel $\mathcal{M}' = \langle \mathcal{V}, \mathcal{G}', p' \rangle$ with post-intervention graph $\mathcal{G}' \subseteq \mathcal{G}$ and probability distribution $p'(\circ|do(A))$ to calculate the probability of $B$ at $\mathcal{M}'$ (i.e., $p'(B|do(A))$).

For example, consider now the probabilistic execution model with the numbers from Table 1. Assume that we have learned about the death of the prisoner, without knowing whether the captain has given the signal, or whether the executioners have fired. We have thus learnt the evidence $E = \{D = 1\}$. By the procedure specified above, we need to update the probability of the *exogenous variables*, i.e., $p'(C) = p(C = 1|D = 1) = 0.82$, which induces a new probability distribution $p'$ on the endogenous variables.[2] Now, we want to compute the probability of $D = 0$ under the counterfactual assumption that $X$ has not fired, $X = 0$, or in other words, we assign a probability to

[2]Henceforth, unless otherwise stated, we will use $p'$ to refer to the probability distribution induced by $p'(C) = p(C = 1|E) = 0.82$.

the counterfactual "if executioner $X$ hadn't fired, then the prisoner would not have died" ($X = 0 \boxdot\!\!\rightarrow D = 0$). Following the above procedure, we obtain that

$$\begin{aligned} & p'(D = 0|do(X = 0)) \\ = & \sum_{x,c\in\{0,1\}} p(D = 0|X = 0, Y = y) \times p(Y = y|C = c) \times p(C = c|D = 1) \\ = & \ 0.598. \end{aligned}$$

In other words, it is 59.8% probable that the prisoner would not have died under the counterfactual supposition that the executioner $X$ hadn't fired. This is, by the way, much less than the conditional probability $p'(D = 0|X = 0) = 0.752$ because *updating on* $X = 0$ (with all other variables being unknown) would suggest an inference to the best explanation, i.e., that the captain did not give the signal. Hence, also the probability of $Y = 0$ goes up sharply when we learn $X = 0$, and so does the probability of $D = 0$.

Like deterministic CMS, the probabilistic framework does not account for the probability of counterfactuals with disjunctive antecedents since interventions are only defined for atomic formulas, and their conjunctions. We will now develop a proposal that expands probabilistic CMS to arbitrary Boolean compounds of atomic formulas in the antecedent, similar to what Briggs has achieved for deterministic CMS.

## 5 CMS with Similarity Metrics

Suppose now that we want to use probabilistic CMS in order to calculate the probability of a counterfactual with disjunctive antecedents, or any Boolean compound of formulas that is more complex than a conjunction of elementary interventions. When we apply Pearl's procedure described in the previous section, steps (2) and (3) fail because the model generated by the intervention $do(X = 0 \vee Y = 0)$ is not well defined and consequently we cannot compute $p'(D = 0)$.

We now try to solve this problem using the Relevance Principle from Section 3, transferring Briggs' idea to evaluate counterfactuals with disjunctive antecedents by evaluating the consequent on the submodels that truthmake the antecedent:

> **Relevance Principle (Probability)** The probability of a counterfactual $A \boxdot\!\!\rightarrow B$ at a causal model $\mathcal{M}$ depends exclusively on the probability of $B$ in the submodels $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ generated by the interventions on the variables in $A$ that truthmake $A$.

Thus, we obtain three submodels respectively generated by $do(X = 0)$, $do(Y = 0)$ and $do(X = 0 \wedge Y = 0)$. See Table 2. Step (2) is working now: performing the

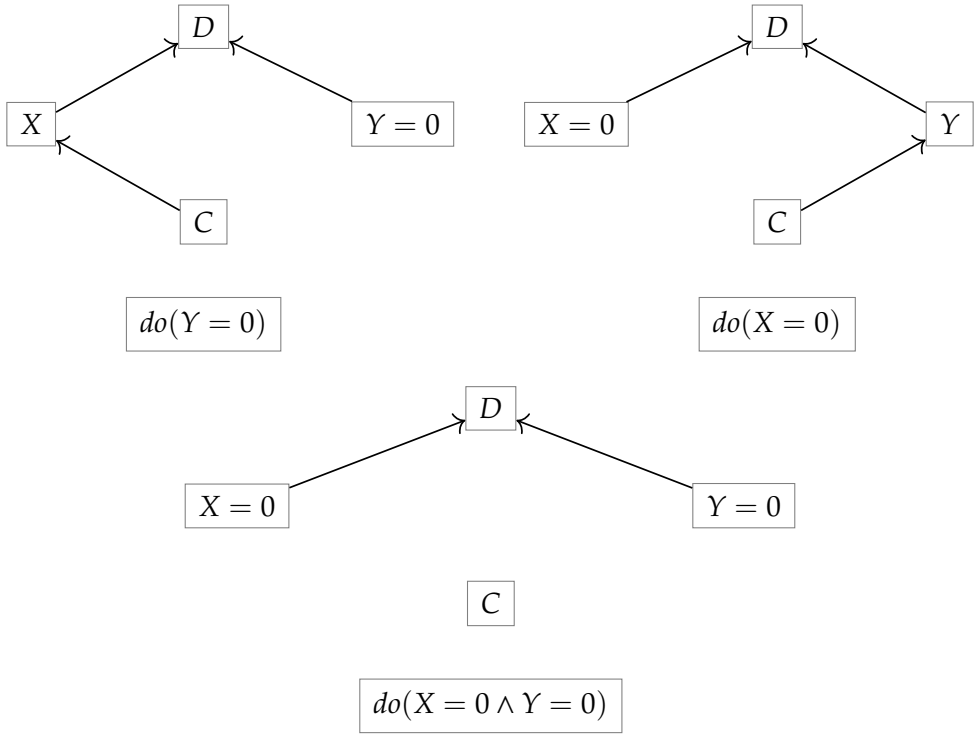


Table 2: The three submodels that truthmake the proposition $X = 0 \vee Y = 0$ in the execution example, with the interventions used to generate them.

intervention $do(X = 0 \vee Y = 0)$ amounts to selecting three *specific* submodels. However, step (3) is still problematic: we have now three possible submodels with respect to which we can compute the probability of $p'(D = 0)$, and it is not clear how these probabilities should be combined. In fact, for the models generated by $do(X = 0)$ and $do(Y = 0)$, $p'(D = 0) = 0.598$, whereas $p'(D = 0) = 0.9$ in the model generated by $do(X = 0 \wedge Y = 0)$.

It is clear that Briggs' solution for the truth conditions of a counterfactual with disjunctive antecedents will not help. There, the consequent needed to be true in *all* states that truthmake the antecedent. Briggs (2012, pp. 152-154) recognizes that this is a *choice*, inspired by Lewis' possible world semantics, which requires that a proposition be true in all nearest possible worlds. The conceptual motivation is that there is no convincing argument for identifying a unique best submodel, and so Briggs assumes that the consequent needs to be true in all of them. While this is a reasonable choice in the context of a *logic* of counterfactuals, we cannot apply the same approach to the *probability* of counterfactuals where the output of the submodels are no Boolean values, but real numbers.

However, all truthmaking submodels $s$ should be relevant in the sense that the values of $p'_s(D = 0)$ should *bound* the overall probability of the counterfactual from above and below:

**Convexity Principle** For the probability of a counterfactual $A \boxdot\!\!\rightarrow B$ at a model $\mathcal{M}$, and the set of submodels $|A|_{\mathcal{M}}$ where we intervene on the variables in $A$ as to truthmake $A$,

$$\min(\{p_s(B) : s \in |A|_{\mathcal{M}}\}) \leq p(A \boxdot\!\!\rightarrow B) \leq \max(\{p_s(B) : s \in |A|_{\mathcal{M}}\})$$

where $p_s$ denotes the probability distribution of the variables in submodel $s$, after updating on the available evidence and performing the truthmaking intervention.

In other words, the probability of a counterfactual cannot be greater (smaller) than the maximum (minimum) probability of the consequent in the causal models that truthmake the antecedent (see also Pearl 2017, p. 9). The crucial question is now which *weight* we need to assign to the different truthmaking models.

A natural starting point is the *straight average* of $p'(D = 0)$ in the three submodels generated by the antecedent $do(X = 0 \vee Y = 0)$. In this way, we would obtain $p'(D = 0) = \frac{0.598+0.598+0.9}{3} = 0.698$. However, straight averaging is at best a default assumption. Alternatively, they could be weighted by means of the similarity to the original model. This idea is elaborated in Lewis's (1976) similarity-based semantics (LSS) for counterfactuals and their probability.

The basic ingredients of LSS are a space of possible worlds $W$ together with a similarity order and a probability distribution $p$ on the elements of $W$. A proposition is represented as a set of possible worlds (i.e., the set of possible worlds where it is true). More precisely, $\sum_{w \in W} p(w) = 1$, and the probability of a proposition $A$ is the sum of the probabilities of the worlds where $A$ is true, that is, $p(A) = \sum_{w \vDash A} p(w)$. Suppose now that we want to evaluate the probability of $B$ given the counterfactual assumption that $A$. Lewis defines, for this purpose, the procedure of *imaging on A* as yielding a probability distribution $p_A$ where all $\neg A$-worlds have probability zero. Their weight is transferred to the most similar worlds where $A$ is true. Define, for any proposition $A \subset W$, the function $f_A : W \rightarrow W$ as mapping any world $w$ to the $A$-world that is most similar to $w$. Thus, for any world $w \in W$:

$$p_A(w) = \sum_{v \in W} p(v) \times \begin{cases} 1 & \text{if } w = f_A(v) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

This procedure assumes that each world is most similar to itself, that is, for any $w \in A$, $w = f_A(w)$. Hence any $A$-world preserves at least its original weight

whereas $\neg A$-worlds transfer their probability mass to the closest possible $A$-world.[3] Lewis (1976, p. 310) then defines the probability of a counterfactual $A \boxdot\!\!\rightarrow B$ as the probability of $B$ after imaging on $A$:[4]

$$p(A \boxdot\!\!\rightarrow B) = p_A(B) = \sum_{w \models B} p_A(w)$$

According to Lewis, imaging is a "minimal revision of the probability function to make the antecedent certain" (Lewis 1976, p. 311), and so it is the appropriate way of belief revision for evaluating a counterfactual. We discuss the merits of this proposal and its relation with CMS in detail in Section 6—at this point, we would like to stress a key feature of imaging: a world $w$ gains weight (=probability mass) from a world $v$ *proportionally to its degree of similarity with* $v$. This differs starkly from Bayesian conditionalization where the updating procedure preserves the prior probability ratio between the remaining possible worlds.

We now apply Lewis' idea to the execution story: what should be the weight of each of the three submodels generated by $do(X = 0)$, $do(Y = 0)$ and $do(X = 0 \wedge Y = 0)$? We claim it must be proportional to the degree of similarity to the original execution model. Once we have weights $\alpha_1, \alpha_2, \alpha_3$ for each of them, we can compute the post-intervention probability of $p'(D = 0)$ as $p'(D = 0) = \alpha_1 \times 0.598 + \alpha_2 \times 0.598 + \alpha_3 \times 0.9$.

The question is how to measure this degree of similarity. A possible answer comes from a recent work of Eva et al. (2019) where the authors introduce two notions of similarity distance between causal models: *evidential* similarity distance, based on the shared probabilistic (in)depencies, and *counterfactual* similarity distance, based on shared counterfactual dependencies. In what follows, we restrict our attention to the latter since probabilistic independencies can hide true causal and counterfactual dependencies.[5]

**Counterfactual Dependence between Variables** A variable $V_2$ is counterfactually dependent on another variable $V_1$ when an intervention on $V_1$ affects the probability distribution of $V_2$, i.e., for some $v \in \mathcal{R}(\mathcal{V}_1)$, $p(V_2|do(V_1 = v) \neq p(V_2)$.[6]

**Counterfactual Similarity Distance (Eva et al., 2019)** Two (probabilistic) causal models $\mathcal{M}$ and $\mathcal{M}'$ are more or less similar to each other, the more counterfactual dependencies they agree on. Specifically, the counterfactual distance between $\mathcal{M}$ and $\mathcal{M}'$ is the absolute value of the difference

[3] Equation (1) assumes that the most similar $A$-world is uniquely determined, but generalization to a function $f_A : W \rightarrow \mathcal{P}(W)$ that assigns a set of closest possible worlds is straightforward (e.g., Gärdenfors 1982)—we will get back to this in the next section.

[4] Actually, Lewis refers to "Stalnaker conditionals" and not specifically to counterfactuals.

[5] In the causal modeling literature, this is known as failure of the Faithfulness Condition.

[6] For example, in the execution model, $D$ counterfactually depends on $X$, $Y$ and $C$; while $X$ and $Y$ counterfactually depends on $C$.

of their counterfactual dependencies normalized by the total number of possible counterfactual dependencies:

$$d(\mathcal{M}, \mathcal{M}') = \frac{|C_{\mathcal{M}} - C_{\mathcal{M}'}|}{N_C} \in [0, 1].$$

Recall that a variable $V_2$ is counterfactually dependent on another variable $V_1$ if we can go from $V_1$ to $V_2$ by following a sequence of arrows from $V_1$ to $V_2$: arrows represent the structural equations, i.e., the *mechanisms* or *laws* that connect variables. Hence, if two models disagree on some counterfactual dependencies among the variables, they disagree on the *mechanism* connecting those variables. So, intuitively, the more laws governing the original model are broken in $\mathcal{M}'$, the more counterfactual-distant from $\mathcal{M}$ a causal model $\mathcal{M}'$ is (see also Lewis 1973a).

There are now two principled options for calculating the probability of counterfactuals. First, we could focus on the submodel that is most similar to $\mathcal{M}$ in the above metric, and neglect the contribution of the other submodels. This is feasible, but it would privilege a particular model and a specific way of truthmaking the antecedent. This is especially implausible when the truthmaking models have a similar distance to the original model and express qualitatively different ways of changing the mechanisms to make the antecedent true.

Second, we could propose that the weight of each submodel $\mathcal{M}'$ should be inversely proportional to its distance to the original model $\mathcal{M}$, according to the above distance measure. This is our preferred approach since it takes into account all relevant submodels that truthmake the antecedent (and only them).

For example, consider the execution story and the three submodels generated by $do(X = 0)$, $do(Y = 0)$ and $do(X = 0 \wedge Y = 0)$. The number of total pairwise counterfactual dependencies is $N_C = 12$; the original model $\mathcal{M}$ encodes $C_{\mathcal{M}} = 5$ counterfactual dependencies; each of the models generated by $do(X = 0)$ and $do(Y = 0)$ encodes $C_{\mathcal{M}'} = 4$ counterfactual dependencies and the model generated by $do(X = 0 \wedge Y = 0)$ encodes $C_{\mathcal{M}'} = 2$ counterfactual dependencies. Table 3 describes the counterfactual dependencies of the execution story and its submodels, where $V_1 \boxdot\!\!\rightarrow V_2$ means that $V_2$ counterfactually depends on $V_1$:

| | *Original Model* | $do(X = 0)$ | $do(Y = 0)$ | $do(X = 0 \wedge Y = 0)$ |
|---|---|---|---|---|
| $C \boxdot\!\!\rightarrow X$ | Yes | No | Yes | No |
| $C \boxdot\!\!\rightarrow Y$ | Yes | Yes | No | No |
| $C \boxdot\!\!\rightarrow D$ | Yes | Yes | Yes | No |
| $X \boxdot\!\!\rightarrow D$ | Yes | Yes | Yes | Yes |
| $X \boxdot\!\!\rightarrow Y$ | No | No | No | No |
| $X \boxdot\!\!\rightarrow C$ | No | No | No | No |
| $Y \boxdot\!\!\rightarrow D$ | Yes | Yes | Yes | Yes |
| $Y \boxdot\!\!\rightarrow X$ | No | No | No | No |
| $Y \boxdot\!\!\rightarrow C$ | No | No | No | No |
| $D \boxdot\!\!\rightarrow X$ | No | No | No | No |
| $D \boxdot\!\!\rightarrow Y$ | No | No | No | No |
| $D \boxdot\!\!\rightarrow C$ | No | No | No | No |

Table 3: Counterfactual Dependencies for the Execution Example.

Call $\mathcal{M}$ the original execution model. By looking at the table we can deduce that

$$d(\mathcal{M}, \mathcal{M}[X = 0]) = \frac{1}{12} \qquad d(\mathcal{M}, \mathcal{M}[Y = 0]) = \frac{1}{12}$$
$$d(\mathcal{M}, \mathcal{M}[X = 0 \wedge Y = 0]) = \frac{3}{12}$$

So, $\mathcal{M}[X = 0]$ and $\mathcal{M}[Y = 0]$ are equally similar to $\mathcal{M}$ and $\mathcal{M}[X = 0 \wedge Y = 0]$ is the most distant from $\mathcal{M}$. Hence, $\mathcal{M}[X = 0 \wedge Y = 0]$, which is the most distant submodel, will receive the least weight. Call $|A|_{\mathcal{M}} = \{s | s \Vdash A\}$ the set of truthmakers of $A$, i.e., the submodels generated by the intervention $do(A)$ on $\mathcal{M}$. In the model $\mathcal{M}$ of the execution story,

$$|X = 0 \vee Y = 0|_{\mathcal{M}} = \{\mathcal{M}[X = 0], \mathcal{M}[X = 0], \mathcal{M}[X = 0 \wedge Y = 0]\}.$$

For $s \in |X = 0 \vee Y = 0|_{\mathcal{M}}$, we define its weight as

$$\alpha(s) = \frac{d(\mathcal{M}, s)^{-1}}{\sum_{t \in |X=0 \vee Y=0|_{\mathcal{M}}} d(\mathcal{M}, t)^{-1}},$$

following the rationale that the weight should be inversely proportional to the distance from the original model, normalized by the sum of all weights.

By some computation, we get that

$$\alpha(\mathcal{M}[X = 0]) = \alpha(\mathcal{M}[Y = 0]) = \frac{3}{7} \qquad \alpha(\mathcal{M}[Y = 0 \wedge X = 0]) = \frac{1}{7}$$

Applied to the execution story, we then find that

$$p'((X = 0 \vee Y = 0) \boxdot\!\!\rightarrow D = 0) = \frac{3}{7} \times 0.598 + \frac{3}{7} \times 0.598 + \frac{1}{7} \times 0.9 \approx 0.64,$$

in agreement with the Convexity Principle. We can generalize the weighting procedure as follows: for a causal model $\mathcal{M}$, for an arbitrary formula $A$ in $\mathcal{L}$, for $s \in |A|_{\mathcal{M}}$,

$$\alpha(s) = \frac{d(\mathcal{M}, s)^{-1}}{\sum_{t \in |A|_{\mathcal{M}}} d(\mathcal{M}, t)^{-1}}.$$

Consequently, we calculate the probability of a counterfactual $A \boxdot\!\!\rightarrow B$ with $\mathcal{L}$-sentences $A$ and $B$, relative to a causal model $\mathcal{M}$, as

$$\begin{aligned} p(A \boxdot\!\!\rightarrow B) &= \sum_{s \in |A|_{\mathcal{M}}} \alpha(s) \times p_s(B) \qquad (2) \\ &= \sum_{s \in |A|_{\mathcal{M}}} \frac{d(\mathcal{M}, s)^{-1}}{\sum_{t \in |A|_{\mathcal{M}}} d(\mathcal{M}, t)^{-1}} \times p_s(B) \end{aligned}$$

Equation (2) expresses our main idea in a nutshell: the probability of the counterfactual $p(A \boxdot\!\!\rightarrow B)$, given an evidence $E$, is the probability of the consequent $B$ in all submodels that truthmake the antecedent, weighted inversely by their similarity to the original model, where similarity is measured by the number of shared counterfactual dependencies. Our account thus synthesizes Causal Modeling Semantics with the Relevance Principle (=focusing on models that truthmake the antecedent, as in Briggs (2012)), and Eva et al.'s (2019) proposal for measuring similarity between causal models.

It is easy to see that our definition of the probability of a counterfactual with disjunctive antecedents extends to more complex sentences, too. Fine's truthmaker semantics, already adopted by Briggs (2012) in her development of a general *logic* of counterfactuals, indicates the truthmaking space states of all Boolean compunds of atomic sentences. Thus, for any sentence that we wish to take as the antecedent of a counterfactual, we simply determine the truthmaking states, the interventions on the causal model that correspond to them, and the corresponding counterfactual probabilities. Then we can use the Eva-Stern-Hartmann procedure for weighting the causal models that correspond to the truthmaking states.

For example, if, for binary variables $A$ and $B$, our counterfactual is "if $A = B$, then $C = 1$" (with actual values $A = 1$ and $B = 0$), the antecedent has two truthmakers: the model generated by $do(A = 1, B = 1)$ and the one generated by $do(A = 0, B = 0)$. The two causal models obtained will then have the same weight according to our procedure, since the intervention affects the same variables and yields the same counterfactual dependencies. In other words, the probability of the counterfactual "if $A = B$, then $C = 1$" is

simply the straight average of the probability of $C = 1$ under the interventions $do(A = 1, B = 1)$ and $do(A = 0, B = 0)$.[7]

Taking stock, we have developed a procedure that goes beyond the achievements of Galles and Pearl (1998) and Halpern (2000), who can calculate probabilities of counterfactuals, but only for antecedents representing a (conjunctive) set of interventions. On the other hand, Briggs (2012) has a general logic of counterfactuals, allowing for arbitary Boolean compounds as antecedents, but no extension to probabilistic reasoning. Our contribution provides a probabilistic counterpart of her logic motivated from the very same principles.

## 6 Back to Lewis: Comparison with Imaging

In this section, we compare our account with David Lewis' *imaging procedure* for assigning a probability to a counterfactual with disjunctive antecedents. This is especially interesting since imaging has been proposed as an alternative to Bayesian conditionalization in the context of Causal Decision Theory (Joyce 1999).

When we image on a proposition $A$, and more than one $A$-world is most similar to a $\neg A$-world $w$, we need to generalize Lewisian imaging beyond Equation (1). Günther (2022) shows that there are numerous ways of doing so, depending on how one determines the selection function $f_A : W \to \mathcal{P}(W)$, and how one distributes the mass of $w$ among the selected worlds $f_A(w)$. For the purposes of counterfactual and causal reasoning, the following function proposed by Gärdenfors (1982) is especially attractive:

$$p_A(w) = \sum_{v \in W} p(v) \times \begin{cases} \frac{p(w)}{\sum_{w' \in f_A(v)} p(w')} & \text{if } w \in f_A(v) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In this case, each world $w$ where $A$ is false transfers its probability mass to the closest worlds where $A$ is true, in proportion to the prior probability of these worlds. This type of imaging, which respects the prior probability ratio among the worlds that receive mass from $w$, is called "Bayesianized imaging" by Joyce (1999). Indeed, in the extreme case where $f_A(w) = A$ if $w \notin A$ (i.e., all $A$-worlds are selected), this form of imaging amounts to Bayesian conditionalization on $A$ (Pearl 2000, p. 73; compare also Proposition 1 in Günther 2022).

There is a deep connection between Bayesianized imaging and CMS. Pearl (2017) shows that the probability of a counterfactual $A \boxrightarrow B$, with $A =$

[7] Note that this also holds if it is *actually* the case that $A = B = 1$. Calculating the probability of the counterfactual does not privilege the actual values of variables; all that matters is whether the distance of the truthmaking models from the original model in terms of counterfactual dependencies.

$A_1 \wedge .... \wedge A_n$ being a conjunction of atomic formulas, can be characterized in two equivalent ways: either, by the definition of Causal Modeling Semantics, as

$$p(A \boxdot\!\!\rightarrow B) = p(B|do(A)) \tag{4}$$

or, when we count worlds with equal causal histories as equally similar, and use the function $p_A(w)$ defined according to Equation (3), by

$$p(A \boxdot\!\!\rightarrow B) = p_A(B) = \sum_{w \models B} p_A(w) \tag{5}$$

The first condition ("equal causal history") means that the most similar $A$-worlds to a $\neg A$-world $w$ contain all and only those $A$-worlds that agree with $w$ on the value of the variables that cannot be affected by $do(A)$, i.e., the non-descendants of $A$.

Pearl then shows that these two characterizations are equivalent, i.e.

$$\sum_{w \models B} p_A(w) = p(B|do(A)). \tag{6}$$

In other words, the transformation defined by the *do*-operator can, for atomic interventions or their conjunctions, be interpreted as an imaging-type mass-transfer. This is a significant result showing that generalized imaging and CMS agree for a large class of interventions. Since Bayesianized imaging is the only type of generalized imaging with this property, we put it into the focus the comparison of our own proposal with LSS.

We now extend Bayesianized imaging to the probability of counterfactuals with disjunctive antecedents. Consider the execution model, the probability distribution $p'$, and the counterfactual $(X = 0 \vee Y = 0) \boxdot\!\!\rightarrow D = 0$. We associate a possible world $w$ to each possible realization of the binary variables $C$, $X$, $Y$, $D$; so there are 16 possible worlds in total. The probability of each of them is simply the joint probability of the realizations of the variables in that possible world. In the execution model, we use $p'$ to compute the probability of each of the 16 possible worlds in Table 4; this means that for $c, y, x, d \in \{0, 1\}$,

$$\begin{aligned} & p'(\langle C = c, X = x, Y = y, D = d \rangle) \\ = \; & p'(C = c) \times p(X = x|C = c) \times p(Y = y|C = c) \times p(D = d|X = x, Y = y) \end{aligned}$$

It is clear that after imaging on $(X = 0 \vee Y = 0)$, four worlds will have weight zero in $p'_{X=0 \vee Y=0}$: $w_1$, $w_2$, $w_9$ and $w_{10}$ in Table 4. The question is how their weight should be distributed to the rest; and this depends on what are the closest neighbors to these possible worlds.

The first conceptual obstacle in defining a similarity order is to decide which variables are *not* affected by $do(X = 0 \vee Y = 0)$. Again, we translate the problem into Causal Modeling Semantics. According to Briggs (2012), the

| Worlds | Closest worlds after imaging on $X = 0 \vee Y = 0$ | |
|---|---|---|
| | Option 1: $f(w_i) = \ldots$ | Option 2: $f(w_i) = \ldots$ |
| $w_1 = \langle C = 1, X = 1, Y = 1, D = 1 \rangle$ | $\{w_3, w_4, w_7, w_8\}$ | $\{w_3, w_4, w_5, w_6, w_7, w_8\}$ |
| $w_2 = \langle C = 1, X = 1, Y = 1, D = 0 \rangle$ | $\{w_3, w_4, w_7, w_8\}$ | $\{w_3, w_4, w_5, w_6, w_7, w_8\}$ |
| $w_3 = \langle C = 1, X = 1, Y = 0, D = 1 \rangle$ | $\{w_3\}$ | $\{w_3\}$ |
| $w_4 = \langle C = 1, X = 1, Y = 0, D = 0 \rangle$ | $\{w_4\}$ | $\{w_4\}$ |
| $w_5 = \langle C = 1, X = 0, Y = 0, D = 0 \rangle$ | $\{w_5\}$ | $\{w_5\}$ |
| $w_6 = \langle C = 1, X = 0, Y = 0, D = 1 \rangle$ | $\{w_6\}$ | $\{w_6\}$ |
| $w_7 = \langle C = 1, X = 0, Y = 1, D = 0 \rangle$ | $\{w_7\}$ | $\{w_7\}$ |
| $w_8 = \langle C = 1, X = 0, Y = 1, D = 1 \rangle$ | $\{w_8\}$ | $\{w_8\}$ |
| $w_9 = \langle C = 0, X = 1, Y = 1, D = 1 \rangle$ | $\{w_{11}, w_{12}, w_{15}, w_{16}\}$ | $\{w_{11}, w_{12}, w_{13}, w_{14}w_{15}, w_{16}\}$ |
| $w_{10} = \langle C = 0, X = 1, Y = 1, D = 0 \rangle$ | $\{w_{11}, w_{12}, w_{15}, w_{16}\}$ | $\{w_{11}, w_{12}, w_{13}, w_{14}w_{15}, w_{16}\}$ |
| $w_{11} = \langle C = 0, X = 1, Y = 0, D = 0 \rangle$ | $\{w_{11}\}$ | $\{w_{11}\}$ |
| $w_{12} = \langle C = 0, X = 1, Y = 0, D = 1 \rangle$ | $\{w_{12}\}$ | $\{w_{12}\}$ |
| $w_{13} = \langle C = 0, X = 0, Y = 0, D = 0 \rangle$ | $\{w_{13}\}$ | $\{w_{13}\}$ |
| $w_{14} = \langle C = 0, X = 0, Y = 0, D = 1 \rangle$ | $\{w_{14}\}$ | $\{w_{14}\}$ |
| $w_{15} = \langle C = 0, X = 0, Y = 1, D = 0 \rangle$ | $\{w_{15}\}$ | $\{w_{15}\}$ |
| $w_{16} = \langle C = 0, X = 0, Y = 1, D = 1 \rangle$ | $\{w_{16}\}$ | $\{w_{16}\}$ |

Table 4: Imaging Mass Transfer for the execution example with disjunctive interventions. The two options correspond to two different similarity orders.

disjunctive intervention $do(X = 0 \vee Y = 0)$ can be regarded as encoding three different interventions, $do(X = 0)$, $do(Y = 0)$, and $do(X = 0 \wedge Y = 0)$. The closest worlds to $w_1$ for the first intervention are $w_7$ and $w_8$, for the second, they are $w_3$ and $w_4$, and for the third, $w_5$ and $w_6$. Dependent on how seriously we consider the option of intervening on *both* variables as a way of expressing $do(X = 0 \vee Y = 0)$, this gives us two options for the most similar worlds to $w_1$: $\{w_3, w_4, w_7, w_8\}$ or $\{w_3, w_4, w_5, w_6, w_7, w_8\}$. And vice versa for the other worlds whose weight needs to be cancelled. Both options are represented in the rightmost columns of Table 4.

However, if we calculate the probability of the counterfactual $(X = 0 \vee Y = 0) \boxdot\!\!\rightarrow D = 0$, after having learnt the evidence $D = 1$, the result of Bayesianized imaging will, for either of these similarity orders, differ from our proposal. For Option 1, we obtain $p'_{(X=0 \vee Y=0)}(D = 0) \approx 0.56$, and for Option 2, we obtain $p'_{(X=0 \vee Y=0)}(D = 0) \approx 0.57$.[8] This is arguably not a good prediction since it violates the plausible Convexity Principle: the probability of the counterfactual should be bounded from above and below by the (maximal and minimal) probability of the consequent in the causal submodels that

[8] Alessandro Zangrandi's GitHub `https://github.com/zazangra/lewis_imaging` offers a Python program to perform Bayesianized imaging on a causal model.

truthmake the antecedent. To recall:

$$p'(X = 0 \boxdot\!\!\rightarrow D = 0) = 0.598 \qquad p'((X = 0 \wedge Y = 0) \boxdot\!\!\rightarrow D = 0) = 0.9$$
$$p'(Y = 0 \boxdot\!\!\rightarrow D = 0) = 0.598$$

To the extent that the Convexity Principle is plausible and compelling, we should reject any procedure that violates this constraint. It is simply puzzling why the probability of the counterfactual can exceed or fall below the probability of the consequent in all relevant submodels. On an intuitive level, it is puzzling why the death of the prisoner, $D = 1$, is more probable under the hypothetical assumption that *at least one* of the two executioners did not fire ($p'_{X=0 \vee Y=0}(D = 0) \approx 0.56/\approx 0.57$), than under the assumption that *only one* did not fire ($p'_{X=0}(D = 0) = 0.598$).

Primarily, the failure of Convexity in imaging is due to the fact that in calculating $p'_{X=0}(D = 0)$ and $p'_{X=0 \vee Y=0}(D = 0)$, different worlds are involved: there is no systematic connection between these two probabilities, like in our proposal. For instance, when imaging on $X = 0$, part of the mass of $w_3$ is transferred to $w_5$, whose probability mass makes a contribution to $p'_{X=0}(D = 0)$, but not to $p'_{X=0 \vee Y=0}(D = 0)$ (in Option 1). This explains why the latter probability falls below $p'_{X=0}(D = 0)$, i.e., below the bounds resulting from the Convexity Principle. In other words, the violation of the Convexity Principle is due to the fact that Bayesianized imaging does not respect the Relevance Principle: the possible worlds do not contain any information about the causal structure of the model, and hence, the results of Bayesianized imaging can differ substantially from our proposal.

Of course, generalized imaging offers an entire universe of different mass transfer functions. So we do not exclude that the imaging theorist can find a function that complies with the Convexity Principle. However, this must come at the price of choosing a procedure that deviates systematically from CMS for (conjunctions of) atomic interventions. What the imaging theorist *cannot* have is a probability mass transfer function that agrees in regular circumstances with CMS, and that satisfies at the same time the Convexity Principle when applied to more complex interventions. Indeed, Pearl (2017, pp. 6-7) explicitly advises caution when applying imaging to disjunctive interventions, such as the ones that we discussed in this paper. Hence, we conclude that the imaging framework has not yet delivered a convincing response to the problem of evaluating the probability of counterfactuals with disjunctive antecedents.

## 7 Conclusions

The present paper expands Causal Modeling Semantics to the evaluation of the probability of counterfactuals with disjunctive antecedents, and more

generally, any truth-functional compound of atomic sentences. To the best of our knowledge, no other proposal has been advanced in the literature to achieve this goal. Our approach is very natural and based on combining three well-established ideas: (1) Briggs' characterization of disjunctive interventions in a causal modeling framework; (2) Lewis' idea of ordering possible worlds according to their similarity with the actual world; (3) Eva et al.'s definition of similarity distance between causal models by counting shared counterfactual dependencies.

As an alternative to our approach, one can assign probabilities to counterfactuals with disjunctive antecedents by embedding Lewis' process of (generalized) imaging into the causal modeling framework, via Bayesianized imaging. However, this option does not return plausible predictions about the probability of counterfactuals, and what is more, it violates intuitive requirements such as the Convexity Principle and the Relevance Principle.

For future work, it would be worth investigating further applications of this framework and discuss whether it matches our intuitions about the probability of counterfactuals. With respect to the former point, we believe that an application of our framework could shed new lights on the notion of *disjunctive causes* introduced by Sartorio (2006); with respect to the latter point, we think that the question of what constraints one should impose on the probability of counterfactuals is a urgent one. We hope to have started an investigation towards this directions by showing how the probability of a counterfactual $A \boxrightarrow B$ should intuitively have a lower and an upper bound imposed by the *best* and *worst* scenarios for $B$ that we could imagine under the counterfactual supposition that $A$.

### Acknowledgments

Funding: this work was supported by Horizon2020 through ERC Starting Grant No. 640638; and Italian Ministry of University and Research through PRIN grant "From Models to Decision".

## References

Adams, Ernest W. (1965), "The Logic of Conditionals", *Inquiry*, 8, pp. 166-197.

— (1970), "Subjunctive and Indicative Conditionals", *Foundations of Language*, 6, pp. 89-94.

— (1975), *The Logic of Conditionals*, Reidel, Dordrecht.

Briggs, R.A. (2012), "Interventionist Counterfactuals", *Philosophical Studies*, 160, pp. 139-166.

Eva, Benjamin, Reuben Stern, and Stephan Hartmann (2019), "The Similarity of Causal Structure", *Philosophy of Science*, 86, pp. 821-835.

Fine, Kit (2016), "Angellic Content", *Journal of Philosophical Logic*, 45, pp. 199-226.

— (2017), "Truthmaker Semantics", in *A Companion to the Philosophy of Language*, ed. by Bob Hale, Crispin Wright, and Alexander Miller, Wiley, New York, pp. 556-577.

Galles, David and Judea Pearl (1998), "An Axiomatic Characterization of Causal Counterfactuals", *Foundations of Science*, 3, pp. 151-182.

Gärdenfors, Peter (1982), "Imaging and Conditionalization", *Journal of Philosophy*, 79, pp. 747-760.

Günther, Mario (2017), "Disjunctive Antecedents for Causal Models", in *Proceedings of the 21st Amsterdam Colloquium*.

— (2022), "Probabilities of Conditionals and Imaged Probabilities", Draft manuscript.

Halpern, Joseph Y. (2000), "Axiomatizing Causal Reasoning", *Journal of Artificial Intelligence Research*, 12, pp. 317-337.

— (2013), "From Causal Models to Counterfactual Structures", *Review of Symbolic Logic*, 6, pp. 305-322.

Joyce, James (1999), *Foundations of Causal Decision Theory*, Cambridge University Press, Cambridge.

Lewis, David (1973a), "Causation", *Journal of Philosophy*, 70, pp. 556-567.

— (1973b), *Counterfactuals*, Blackwell, Oxford.

— (1976), "Probabilities of Conditionals and Conditional Probabilities", *Philosophical Review*, 85, pp. 297-315.

Loewer, Barry (1976), "Counterfactuals with Disjunctive Antecedents", *Journal of Philosophy*, 73, pp. 531-537.

McKay, Thomas and Peter Van Inwagen (1977), "Counterfactuals with Disjunctive Antecedents", *Philosophical Studies*, 31, pp. 353-356.

Nute, Donald (1975), "Counterfactuals and the Similarity of Worlds", *Journal of Philosophy*, 72, pp. 773-778.

Pearl, Judea (2000), *Causality*, Cambridge University Press, Cambridge.

— (2017), "Physical and Metaphysical Counterfactuals: Evaluating Disjunctive Actions", *Journal of Causal Inference*, 5, pp. 1-10.

Sartorio, Carolina (2006), "Disjunctive Causes", *Journal of Philosophy*, 103, pp. 521-538.

Skyrms, Brian (1980), *Causal Necessity*, Yale University Press, Princeton.