Non-strict Interventionism: The Case Of Right-Nested Counterfactuals

Schulz, K.; Smets, S.; Velázquez-Quesada, F.R.; Xie, K.

DOI
10.1007/s10849-022-09358-x

Publication date
2022

Document Version
Final published version

Published in
Journal of Logic, Language and Information

Journal of Logic, Language and Information (2022) 31:235–260
https://doi.org/10.1007/s10849-022-09358-x

Non-strict Interventionism: The Case Of Right-Nested Counterfactuals

Katrin Schulz[1] · Sonja Smets[1,2] · Fernando R. Velázquez-Quesada[1] · Kaibo Xie[1]

Accepted: 17 February 2022 / Published online: 26 April 2022

Abstract
The paper focuses on a recent challenge brought forward against the interventionist approach to the meaning of counterfactual conditionals. According to this objection, interventionism cannot account for the interpretation of right-nested counterfactuals, the problem being its strict interventionism. We will report on the results of an empirical study supporting the objection. Furthermore, we will extend the well-known logic of intervention with a new operator expressing an alternative notion of intervention that does away with strict interventionism (and thus can account for some critical examples). This new notion of intervention operates on the valuation of the variables in a causal model, and not on their functional dependencies.

Keywords Counterfactual · Causal model · Intervention

1 Introduction

The meaning of counterfactual conditionals bears an intrinsic relation to a number of central scientific problems, like the nature of reasoning, the possibility of knowledge, and the status of laws of nature. Therefore, this topic has fascinated many thinkers from various disciplines, like philosophy, logic, psychology and others. But, despite

✉ Kaibo Xie
xkb@mail.tsinghua.edu.cn

Katrin Schulz
K.Schulz@uva.nl

Sonja Smets
S.J.L.Smets@uva.nl

Fernando R. Velázquez-Quesada
F.R.VelazquezQuesada@uva.nl

1 ILLC, Universiteit van Amsterdam, Amsterdam, Netherlands

2 Department of Information Science and Media Studies, University of Bergen, Bergen, Norway

a lot of effort, no consensus has been reached yet about how the meaning of these sentences needs to be approached.

One way to conceptualize the evaluation of counterfactuals, very common in the literature, goes as follows. When evaluating a counterfactual, we select, given the antecedent A and the context of evaluation, certain (hypothetical) situations in which the antecedent is true, and then check whether they make the consequent B true as well. The challenge of accounting for counterfactuals then becomes to define the relevant *selection function* correctly. Following the approach of Lewis (1973), Stalnaker (1968), which still is the dominant approach in the philosophical literature, the selection is based on similarity: we take those hypothetical situations that are most similar to the actual world. But this proposal is known to be problematic: among other things, it appears to be too flexible. In recent years, the interventionist approach to counterfactuals became very popular (Pearl 2013; Schulz 2011; Kaufmann 2013; Halpern 2016; Ciardelli et al. 2018 and others). This approach describes the truth conditions of counterfactuals with respect to a representation of the relevant causal dependencies, building on Causal Models as introduced in Pearl (2000), Spirtes et al. (2000). The approach got its name from the way it describes the selection function. In the selected hypothetical scenarios, the antecedent has been made true by intervention on the actual causal dependencies: it is cut off from its causal parents and stipulated to be true by law. Then, one checks in the resulting model whether the consequent holds.[1] A consequence of this line of approach is that counterfactuals cannot express abductive counter-to-fact reasoning. The underlying intuition is that counterfactuals do not concern epistemic inferences about what certain hypothetical observations would teach us, but essentially are about how actual changes of the facts would affect the world.

Recently, this approach has been criticized by Fisher (2017). He claims that interventionism makes incorrect predictions for right-nested counterfactuals. According to Fisher, the problem is a particular property of the interventionist approach, *strict interventionism*, which he argues needs to be dropped in a proper account. We will argue, using the results of an empirical study, that Fisher is right in his critique. But this does not mean that the interventionist approach needs to be given up in general. We will propose a variation of the approach that drops strict interventionism and thus can account for Fisher's core-observations. We will also make precise how this new proposal relates to the classical interventionist approach as spelled out in Halpern (2016). We will do so by providing an axiomatization of the new operator for counterfactual reasoning that we introduce. As it will turn out, this new operator can be already defined in terms of the classical intervention operator. Furthermore, to a large extent they both make the same counterfactuals true. So, even though we formalise intervention in a slightly different way, in terms of logical properties we propose a conservative modification of the original interventionist approach.

[1] It turns out that for recursive causal models the interventionist selection function can be understood as just one particular way to make similarity precise (Halpern 2013; Marti and Pinosio 2014).

2 The Interventionist Approach to Counterfactuals

Our presentation of the interventionist approach to counterfactuals is based on Briggs (2012). Still, we will only introduce the parts that are relevant for the discussion at hand. The two central ingredients are (i) the causal model, which contains information about the relevant causal dependencies, and (ii) the operation of intervention, which defines the selection function by mapping a given causal model onto a class of models that make the antecedent of the relevant counterfactual true.

Causal models represent the causal dependencies between a given *finite* set of variables. The variables are sorted into the set $\mathcal{U} = \{U_1, \ldots, U_m\}$ of *exogenous* variables (those whose value is causally independent from the value of other variables in the system) and the set $\mathcal{V} = \{V_1, \ldots, V_n\}$ of *endogenous* variables (those whose value causally depends on the value of other variables in the system). Each variable X has a fixed range $\mathcal{R}(X)$, the *finite* set of possible values it can take. Given this basic information $\langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ (the model's *signature*), a causal model is defined as follows.

Definition 1 (Causal model) A *causal model* for the signature $\langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ is a tuple $\langle \mathcal{S}, \mathcal{A} \rangle$ where

- $\mathcal{S} = \{F_{V_j} \mid V_j \in \mathcal{V}\}$ is a set assigning to each endogenous variable V_j a function

$$F_{V_j} : \big(\mathcal{R}(U_1) \times \cdots \times \mathcal{R}(U_m) \times \mathcal{R}(V_1) \times \cdots \times \mathcal{R}(V_{j-1}) \\ \times \mathcal{R}(V_{j+1}) \times \cdots \times \mathcal{R}(V_n)\big) \to \mathcal{R}(V_j).$$

- $\mathcal{A}$ is the *valuation* function, assigning to every $X \in (\mathcal{U} \cup \mathcal{V})$ a value $\mathcal{A}(X) \in \mathcal{R}(X)$ that *complies with the functions in* $\mathcal{S}$: for all endogenous variables $V_j \in \mathcal{V}$,

$$\mathcal{A}(V_j) = F_{V_j}\big(\mathcal{A}(U_1), \ldots, \mathcal{A}(U_m), \mathcal{A}(V_1), \ldots, \mathcal{A}(V_{j-1}), \mathcal{A}(V_{j+1}), \ldots, \mathcal{A}(V_n)\big).$$

In a causal model, the set $\mathcal{S}$ fixes the causal dependencies among the variables. Each function F_V, called V's *structural function*, describes how the value of V causally depends on that of other variables. The function $\mathcal{A}$ defines the value of all variables in the model, doing it in a way that is consistent with the causal laws fixed by $\mathcal{S}$. Here are two causality-related concepts that will be important through the text.

Definition 2 (Dependency) Let $\langle \mathcal{S}, \mathcal{A} \rangle$ be a causal model for $\langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ (recall: $\mathcal{U} = \{U_1, \ldots, U_m\}$ and $\mathcal{V} = \{V_1, \ldots, V_n\}$). Given an endogenous variable $V_j \in \mathcal{V}$, let $\langle X_1, \ldots, X_{m+n-1} \rangle$ be the (m+n-1)-tuple $\langle U_1, \ldots, U_m, V_1, \ldots, V_{j-1}, V_{j+1}, \ldots, V_n \rangle$.

We say that the endogenous variable $V_j \in \mathcal{V}$ is *directly dependent* on a variable $X_i \in (\mathcal{U} \cup \mathcal{V}) \setminus \{V_j\}$ (in symbols, $X_i \rightarrowtail_{\mathcal{S}} V_j$) if and only if there are $x_1 \in \mathcal{R}(X_1), \ldots, x_{i-1} \in \mathcal{R}(X_{i-1}), x_{i+1} \in \mathcal{R}(X_{i+1}), \ldots, x_{m+n-1} \in \mathcal{R}(X_{m+n-1}))$ and there are $x_i' \neq x_i'' \in \mathcal{R}(X_i)$ such that $F_{V_j}(x_1, \ldots, x_i', \ldots, x_{m+n-1}) \neq F_{V_j}(x_1, \ldots, x_i'', \ldots, x_{m+n-1})$. When $X_i \rightarrowtail_{\mathcal{S}} V_j$, we will also say that X_i is a *parent* of V_j.

We say that $V_j \in \mathcal{V}$ is *causally dependent* on $X_i \in (\mathcal{U} \cup \mathcal{V}) \setminus \{V_j\}$ if and only if $X_i \rightarrowtail_{\mathcal{S}}^{+} V_j$, with $\rightarrowtail_{\mathcal{S}}^{+}$ the transitive closure of $\rightarrowtail_{\mathcal{S}}$.

Definition 3 (Recursive causal models) A causal model is said to be *recursive* if and only if it contains no circular dependencies between the variables (i.e., if and only if $\rightarrowtail$'s transitive closure, $\rightarrowtail^+_{\mathcal{S}}$, is irreflexive [and thus a strict partial order]).

In a recursive model $\langle \mathcal{S}, \mathcal{A} \rangle$, no circular causal dependencies occur. Thus, if the values of all exogenous variables are fixed, the value of every endogenous variable V is uniquely determined (from the values of the exogenous variables and the causal dependencies as described by $\mathcal{S}$). In the rest of the paper we will only consider recursive causal models, which from now on will be called simply *causal models*.

Causal models can be described by different languages; here is our choice.

Definition 4 (Language $\mathcal{L}_{[\,]}$) Formulas ϕ of the language $\mathcal{L}_{[\,]}$ over the signature $\langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ are given by

$$\phi ::= X{=}x \mid \neg\phi \mid \phi \wedge \phi \mid [\overrightarrow{X}{=}\overrightarrow{x}\,]\phi$$

for $X \in \mathcal{U} \cup \mathcal{V}$, $x \in \mathcal{R}(X)$, $\overrightarrow{X} = (X_1, \ldots, X_k) \in (\mathcal{U} \cup \mathcal{V})^k$, $k \in \mathbb{N}$, $X_i \neq X_j$ for $i \neq j$ and $\overrightarrow{x} = (x_1, \ldots, x_k)$ with $x_i \in \mathcal{R}(X_i)$. Sentences of the form $[\overrightarrow{X}{=}\overrightarrow{x}\,]\phi$ represent counterfactual conditionals, with $\overrightarrow{X}{=}\overrightarrow{x}$ being the antecedent and ϕ being the consequent.

A counterfactual conditional $[\overrightarrow{X}{=}\overrightarrow{x}\,]\phi$ states what would hold *if* certain variables ($\overrightarrow{X}$) were set to particular values ($\overrightarrow{x}$). This formulation already exposes the approach introduced here as an interventionist approach: it talks about *setting* variables to a particular values. Thus, counterfactuals are taken to reason about the consequences of *manipulating* the actual situation in certain ways. As mentioned in the introduction, this excludes abductive reasoning in counterfactuals. Changing the facts will not affect what happened before, but only what still is to come. The exact details of how this idea is formalized are spelled out in the semantics provided below.

Note that the language $\mathcal{L}_{[\,]}$ is less expressive than the one used in Briggs (2012), as here a counterfactual's antecedent $\overrightarrow{X}{=}\overrightarrow{x}$ is effectively only a conjunction of atomic sentences (set X_1 to x_1 *and* set X_2 to x_2 *and* so on). Yet, it is more expressive than those in Halpern (2000), as here there are no restrictions on a counterfactual's consequent ϕ.[2] Note also how, in contrast to most literature on causal models, $\mathcal{L}_{[\,]}$ allows exogenous variables in a counterfactual's antecedent.

The second important ingredient of the interventionist approach is the notion of intervention used for evaluating counterfactual conditionals $[\overrightarrow{X}{=}\overrightarrow{x}\,]\phi$. Given a causal model $\langle \mathcal{S}, \mathcal{A} \rangle$ and a counterfactual sentence $[\overrightarrow{X}{=}\overrightarrow{x}\,]\phi$, an intervention provides a single model satisfying the antecedent $\overrightarrow{X}{=}\overrightarrow{x}$; this is the model where the consequent ϕ is evaluated. The model is built in two steps. First, the antecedent is made true by forcing the variables in $\overrightarrow{X}$ to the values $\overrightarrow{x}$. In the case of exogenous variables this

[2] Thus, $\mathcal{L}_{[\,]}$ is also more expressive than the languages used in Galles and Pearl (1997), Galles and Pearl (1998).

is done by simply changing their value in $\mathcal{A}$; in the case of endogenous ones this is done by turning the variables' respective structural function in $\mathcal{S}$ into a constant function (thus effectively making the variables exogenous).[3] Second, the values of the endogenous variables are calculated using the new structural functions.

Definition 5 (Intervention) Let $\langle \mathcal{S}, \mathcal{A} \rangle$ be a causal model. When evaluating formulas in $\mathcal{L}_{[\,]}$, Boolean operators are interpreted as usual; for the rest,

$$\begin{array}{lll} \langle \mathcal{S}, \mathcal{A} \rangle \models X{=}x & \text{iff\_}def & \mathcal{A}(X) = x \\ \langle \mathcal{S}, \mathcal{A} \rangle \models [\,\overrightarrow{X} = \overrightarrow{x}\,]\phi & \text{iff\_}def & \langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}} \rangle \models \phi \end{array}$$

with $\langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}} \rangle$ the causal model where

- $\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}$ is as $\mathcal{S}$ except that for each endogenous variable X_i in $\overrightarrow{X}$, the function F_{X_i} is replaced by a 'constant' function F'_{X_i} that assigns to X_i the value x_i regardless of the values of all other variables.
- $\mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}$ is the unique valuation that is identical to $\mathcal{A}$ with respect to the values of exogenous variables not in $\overrightarrow{X}$, assigns to each exogenous variable X_i in $\overrightarrow{X}$ the indicated value x_i, and complies with the causal dependencies in $\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}$ for the endogenous ones.[4]

Thus, according to the interventionist approach, the selection function f discussed in the introduction should be defined as

$$f\big(\langle \mathcal{S}, \mathcal{A} \rangle, \overrightarrow{X} = \overrightarrow{x}\big) := \langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}} \rangle.$$

It is worthwhile to emphasise that, in the intervened model $\langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}} \rangle$, the valuation $\mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}$ *complies* with the model's causal dependencies, $\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}$: for every $V \in \mathcal{V}$, the value $\mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}(V)$ is given by the variable's structural function as provided by $\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}$. So, intervention happens at the level of causal dependencies, and this change affects the valuation $\mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}$. In Sect. 5 we will introduce a notion of intervention that, working on a more general class of models, changes values directly, leaving causal dependencies unaffected.

Axiom system. As stated, $\mathcal{L}_{[\,]}$ is different from previous languages used for describing causal models. Thus, it is worthwhile to provide its axiom system. Note how, although the system uses as a parameter the signature $\langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ of both the language and the class of models (in some cases relying on the signature's finiteness), no axiom

[3] If the counterfactual's antecedent were allowed to be an arbitrary Boolean combination of atoms, this step would already produce more than one alternative. This generalisation can be dealt with as in Briggs (2012), which relies on Fine (2012) to provide a procedure for generating all these models. However, the issue of the number of models is orthogonal to this text's main focus. Therefore, we keep the restriction on the form of the antecedent, as done in Pearl (2000) and Halpern (2013).

[4] Note: the valuation is unique, not only because the values of exogenous variables is determined, but also because, if $\langle \mathcal{S}, \mathcal{A} \rangle$ is recursive, so is $\langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}} \rangle$. This is because the intervention operation only removes causal dependencies, and thus no circular dependencies are added.

refers to a particular model. The proof of the theorem below (as well as proofs of all other technical results in this paper) can be found in the Appendix.

Theorem 1 *The axioms and rules in Table 1 define a sound and complete axiom system for $\mathcal{L}_{[]}$ w.r.t. causal models.*

It is useful to have an intuitive understanding of what the axioms state. Axiom **A0** and rule **MP** deal with instances of propositional validities. Axioms **A2** and **A1** simply state, respectively, that every variable has one and only one value. Axiom **A3** indicates, in essence, that if an intervention $\overrightarrow{X}=\overrightarrow{x}$ sets Y to y, then its effects can be replicated by the extended intervention $(\overrightarrow{X}=\overrightarrow{x}, Y=y)$.[5] Axiom **A4** says that an intervention is successful in setting the values of intervened variables. Axiom **A5** states, paraphrasing Halpern (2000), that if the intervention $(\overrightarrow{X}=\overrightarrow{x}, Y=y)$ sets Z to z and the intervention $(\overrightarrow{X}=\overrightarrow{x}, Z=z)$ sets Y to y, then the intervention $\overrightarrow{X}=\overrightarrow{x}$ should be already enough to set Z to z (and, by the commutativity of conjunction, also enough to set Y to y). Axiom **A6** uses the abbreviation $\rightsquigarrow$, a syntactic characterisation of direct dependency $\rightarrowtail$ (Definition 2), to state that direct dependency has no cycles (and thus the model is recursive).[6] Axioms **A7.1** and **A7.2** state that interventions commute with negation and distribute over conjunction (in particular, the former indicates that an intervention is deterministic). **A8** is the generalisation rule, a standard property of modal operators. Axiom **A9** (cf. Briggs (2012); Barbero and Sandu (2019)) deals with nested interventions, stating that a later intervention overwrites an earlier one. Finally, while axiom $\mathbf{A}_{[]}$ states that an intervention with an empty assignment does not affect the given causal model, axiom $\mathbf{A}_{\mathcal{U}}$ states that, for any assignment $\overrightarrow{X}=\overrightarrow{x}$, the valuations before and after an intervention assign the same value to *exogenous* variables not occurring in $\overrightarrow{X}$.

[5] Within defeasible reasoning Koons (2017), **A3** is known as the principle of *weak/cautious monotonicity* (adding conclusions as premises does not invalidate conclusions) Gabbay (1984); Makinson (1988). Together with the *cut* principle Kraus et al. (1990) (removing premises that are also conclusions does not invalidate conclusions, $\left([\overrightarrow{X}=\overrightarrow{x}, Y=y](Z=z) \wedge [\overrightarrow{X}=\overrightarrow{x}](Y=y)\right) \rightarrow [\overrightarrow{X}=\overrightarrow{x}](Z=z)$, also valid), it makes an intervention *cumulative*: if $[\overrightarrow{X}=\overrightarrow{x}]$ sets Y to y, then it has the same effects as $[\overrightarrow{X}=\overrightarrow{x}, Y=y]$.

[6] For the syntactic characterisation of direct dependency $\rightarrowtail$, recall that, for $V \in \mathcal{V}$ and $X \in \mathcal{U} \cup \mathcal{V}$, we have $X \rightarrowtail V$ if and only if there are values $\overrightarrow{z}$ of variables in $\overrightarrow{Z} = (\mathcal{U} \cup \mathcal{V}) \setminus \{X, V\}$ and two different values x_1, x_2 of X such that the value V gets by setting $(\overrightarrow{Z}, X)$ to $(\overrightarrow{z}, x_1)$ is different from the value it gets by setting the same variables to $(\overrightarrow{z}, x_2)$. This can be expressed by the formula

$$\bigvee_{\substack{\overrightarrow{z} \in \mathcal{R}((\mathcal{U} \cup \mathcal{V}) \setminus \{X, V\}),\\ \{x_1, x_2\} \subseteq \mathcal{R}(X),\ x_1 \neq x_2,\\ \{v_1, v_2\} \subseteq \mathcal{R}(V),\ v_1 \neq v_2}} [\overrightarrow{Z}=\overrightarrow{z}, X=x_1](V=v_1) \ \wedge\ [\overrightarrow{Z}=\overrightarrow{z}, X=x_2](V=v_2),$$

which is abbreviated as $X \rightsquigarrow V$ (cf. with the syntactic definition of causal dependency in Halpern (2000)).

Table 1 Axiom system for $\mathcal{L}_{[\,]}$ w.r.t. causal models over the given signature

A0	Propositional tautologies	**MP**	From ϕ and $\phi \rightarrow \psi$ infer ψ
A1	$Y{=}y \rightarrow \neg(Y{=}y')$ for $y, y' \in \mathcal{R}(Y)$ with $y \neq y'$	**A2**	$\bigvee_{y \in \mathcal{R}(Y)} Y{=}y$
A3	$\left([\overrightarrow{X}{=}\overrightarrow{x}](Y{=}y) \wedge [\overrightarrow{X}{=}\overrightarrow{x}](Z{=}z)\right) \rightarrow [\overrightarrow{X}{=}\overrightarrow{x}, Y{=}y](Z{=}z)$		
A4	$[\overrightarrow{X}{=}\overrightarrow{x}, Y{=}y](Y{=}y)$		
A5	$\left([\overrightarrow{X}{=}\overrightarrow{x}, Y{=}y](Z{=}z) \wedge [\overrightarrow{X}{=}\overrightarrow{x}, Z{=}z](Y{=}y)\right) \rightarrow [\overrightarrow{X}{=}\overrightarrow{x}](Z{=}z)$ for $Y \neq Z$		
A6	$(X_0 \rightsquigarrow X_1 \wedge \cdots \wedge X_{k-1} \rightsquigarrow X_k) \rightarrow \neg(X_k \rightsquigarrow X_0)$		
A7.1	$[\overrightarrow{X}{=}\overrightarrow{x}]\neg\phi \leftrightarrow \neg[\overrightarrow{X}{=}\overrightarrow{x}]\phi$	**A7.2**	$[\overrightarrow{X}{=}\overrightarrow{x}](\phi \wedge \psi) \leftrightarrow ([\overrightarrow{X}{=}\overrightarrow{x}]\phi \wedge [\overrightarrow{X}{=}\overrightarrow{x}]\psi)$
A8	From ϕ infer $[\overrightarrow{X}{=}\overrightarrow{x}]\phi$		
A9	$[\overrightarrow{X}{=}\overrightarrow{x}][\overrightarrow{Y}{=}\overrightarrow{y}]\phi \leftrightarrow [\overrightarrow{X'}{=}\overrightarrow{x'}, \overrightarrow{Y}{=}\overrightarrow{y}]\phi$ with $\overrightarrow{X'}{=}\overrightarrow{x'}$ the subassignment of $\overrightarrow{X}{=}\overrightarrow{x}$ for $\overrightarrow{X'} = \overrightarrow{X} \setminus \overrightarrow{Y}$		
$\mathbf{A}_{\mathcal{U}}$	$[\,]U{=}u \leftrightarrow [\overrightarrow{X}{=}\overrightarrow{x}]U{=}u$ for $U \in \mathcal{U}$ with $U \notin \overrightarrow{X}$	$\mathbf{A}_{[]}$	$Y{=}y \leftrightarrow [\,]Y{=}y$

Fig. 1 A causal model for **Match** and **Headlamp**, before (left side) and after (right side) interpreting the counterfactuals

3 Fisher's Criticism

Fisher (2017) criticizes the approach described above. He claims that it makes incorrect predictions for right-nested counterfactuals. Concretely, he discusses the examples (1) and (2) below.[7]

- **Match.** I hold up a match and strike it, but it does not light. I say

  (1) If the match had lit, then (even) if it had not been struck, it would have lit.

- **Headlamp.** I hold up a headlamp in good working condition. I say

  (2) If the headlamp were emitting light, then if it had had no batteries, the headlamp would be emitting light.

Both examples involve a model of the form shown on the left side of Fig. 1, where A_1 stands for the variable the first antecedent talks about (*"the match lights"* and *"the headlamp emits light"*, respectively) and A_2 for the variable of the second antecedent (*"the match is struck"* and *"the headlamp has batteries"*, respectively).[8] Following the interventionist approach, the evaluation of the first antecedent produces a causal model where A_1 is forced to a particular value, and where the causal connection between A_2 and A_1 has been erased (right side of Fig. 1). Evaluating the second antecedent forces A_2 to a particular value too, but this will no longer affect A_1. Hence, the counterfactuals (1) and (2) are predicted to be true, but intuitively, according to Fisher, they should be false. Fisher traces the problem back to the property of *strict interventionism* (SI).

> (SI) "When a variable V is intervened on so that it is made to take a value v, V remains set to v unless it is intervened upon again per an iterated application of the interventionist recipe." (Fisher 2017: 4939).

Interventionist approaches have this property because their selection function maps a given causal model M and an antecedent A to a new causal model in which a causal variable V occurring in the antecedent A has lost all connections to its causal parents. Any later intervention that might affect V's (former) causal parents will no longer

[7] Fisher also considers an example with the counterfactual "If the match were struck and it lit, then if it hadn't been struck, it would have lit". This is not a good example to make his point, as it contains a conjunction of cause (striking the match) and effect (the match lights) in the antecedent. For the counterexample to work, Fisher needs this conjunction to be interpreted as two independent interventions. However, it could be that *"and"* is interpreted causally in this case: *"If the match were struck and because of that it lit, ..."*. But then the fact that the match lights would be introduced as a causal consequent of the striking of the match and not as an independent intervention.

[8] We ignore other possible variables, as they will not affect the relevant predictions made.

affect V itself. So, as long as ψ does not assign a new value to V, the counterfactual $[\overrightarrow{V} = \overrightarrow{v}][\psi](V_i = v_i)$ will always come out as true.

To solve this problem, Fisher proposes to give up strict interventionism. More concretely, he proposes the following *adequacy condition* for approaches to the meaning of counterfactuals: "A causal model semantics for counterfactuals should admit cases in which the variables implicated in the antecedent of a counterfactual remain causally sensitive to their parents throughout the evaluation procedure." (Fisher (2017):4942). However, he does not propose an alternative approach that has this property.[9] In the rest of the paper we will do the following. First of all, we need to confirm Fishers judgements concerning the target examples (1) and (2) with an actual survey. These are not your every day examples of counterfactuals and we should make sure that the intuitions Fisher reports are generally shared and that they really concern truth conditions, not the assertability of this type of right-nested counterfactuals. This is the subject of Sect. 4. In Sect. 5 we will develop an alternative interventionist approach to the meaning of counterfactual conditionals that is not strictly interventionist. Finally, in Sect. 6 we will extend the discussion with some additional examples and investigate whether giving up strict interventionalism is sufficient to account for right-nested counterfactuals in general.

4 An Empirical Study on Fisher's Counterexamples

A possible objection against Fisher's observations and the conclusions he derives from them is that he confuses judging a sentence false with rejecting it as not well-formed. Maybe we are inclined to say "No" to the counterfactuals in (1) and (2), because they are very strange counterfactual sentences. To exclude this interpretation of the observations, we conducted a small empirical study in which we asked the participants to judge not only the counterfactuals (1) and (2), but also their counterparts (3-a) and (3-b) in which the final consequent has been negated. If participants judge the sentences (1) and (2) false because they consider them defective, they should judge (3-a) and (3-b) to be defective (and hence false) as well.

(3) a. If the match had lit, then if it had not been struck, it would not have lit.
b. If the headlamp were emitting light, then if it had had no batteries, the headlamp would not have been emitting light.

4.1 Method and Participants

We used the scenarios **Match** and **Headlamp** from page 6 and a third scenario containing a counterfactual $[\phi][\psi]\xi$ with ξ talking about a causal effect of ϕ. For each scenario we asked the participants to judge 3 counterfactuals: the target right-nested counterfactual, the counterfactual with the opposite final consequent and a filler item to check whether the participants where paying attention and understood the presented scenario correctly. This resulted in 9 questions that the participants had to answer. The

[9] Fisher discusses in Fisher (2017) an alternative definition of intervention, dubbed "side-constrained intervention", but admits that this variation is not really targeting the root of the problem.

Sentence	True	False	Unclear
(1)	4%	80%	16%
(3-a)	64%	13%	23%
(2)	7%	84%	9%
(3-b)	77%	9%	14%
(1)+(2)	6%	82%	12%
(3-a)+(3-b)	71%	11%	18%

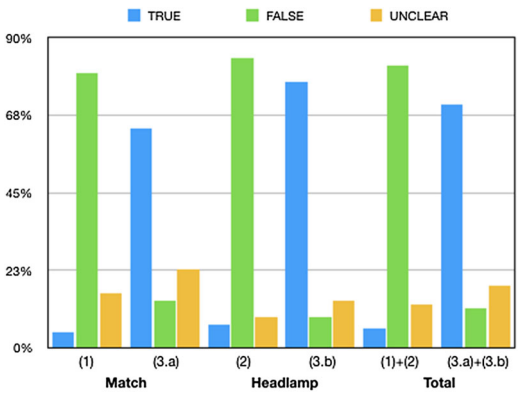

Fig. 2 Results of the 1st study

order of question was randomized. The participants had to judge the truth value of the counterfactual using a slider bar with five values, from 0 to 4. They were told that 0 means the sentence is false, 4 it is true and 2 that the truth value is unclear. The values 1 and 3 allowed them to indicate that they find a sentence weakly false or true.

The study was implemented in Qualtrics, a web-based survey tool. Participants were recruited via Prolific.ac, an online platform aimed at connecting researchers and participants willing to fill in surveys and questionnaires in exchange for compensation for their time Palan and Schitter (2018). We recruited native English speakers (British and American English). Fifty-two participants completed the task. Eight participants were excluded. Two participants did not answer the filler question for the match scenario correctly, seven participants did not answer the filler question for the headlamp scenario correctly, one also failed the match scenario. Thus, forty-four responses were included in the analyses reported below. Thirteen participants failed the control question for the third scenario we used. Because of the high number we concluded that there was a problem with the material used and excluded this scenario from the evaluations.

4.2 Results and Discussion

The table in Fig. 2 states the results of the study. We counted both values 3 and 4 on the scale as judging the sentence true and 0 an 1 as judging the sentence false. The graph in Fig. 2 plots the percentages of the different answers, first for both scenario's separately and then combined. The results show, first of all, that a majority of the participants agree with the intuitions reported by Fisher (2017). Furthermore, the results for the opposite counterfactuals (3-a) and (3-b) support the conclusion that the judgements are for the most part judgements about truth values and not about the well-formedness of the sentences under consideration.

Hence, we conclude with Fisher that these nested counterfactuals present a problem for the interventionist approach to their meaning. Fisher discussed the possibility to defend the approach by arguing that the conditionals under discussion are interpreted according to a different (epistemic) reading of counterfactuals and eventually

dismisses it. We agree with Fisher. Notice the particularity of the situation. Normally, the possibility of an epistemic reading is considered in case a counterfactual intuitively appears to be true, but the account under discussion cannot predict this.[10] Here we would have to explain why certain counterfactuals are intuitively false, while the approach predicts them to be true. In order to make this work, we would first have to argue that an intervention-based reading of these particular counterfactuals is not possible.

To sum up, the results of this study support Fisher's argumentation against the interventionist approach. But does that mean that we need to give up the interventionist approach to counterfactuals? We do not think so. We can give up the property of strict interventionism responsible for the problematic predictions, but still keep the general idea and all the strong predictions of the interventionist approach. The big conceptual step that needs to be taken is to apply intervention to the valuation $\mathcal{A}$ instead of the representation of the causal dependencies $\mathcal{S}$. In the next section we develop this idea in detail.

5 Non-strict Interventions

The goal is, then, to find a notion of intervention that coincides with Pearl (2000)'s proposal for non-nested cases (thus 'inheriting' the good behaviour of the strict interventionism approach in those situations), but also accounts for the results of the empirical study. The idea on which we will build our alternative proposal is that, although counterfactual assumptions might modify the value of some causal variables, they will not affect causal relationships. The underlying intuition is that counterfactuals reason about situations in which certain laws are violated, not situations ruled by a different set of laws. This way, the framework can satisfy Fisher's adequacy condition for approaches to the meaning of counterfactuals (see page 7): even after intervention on a particular variable, it remains connected to its causal parents. However, we are still capturing here the core idea of intervention: intervening in a variable X will only affect the value of variables that causally depend on X.[11]

An important consequence of the assumption that intervention can change valuations without modifying causal relationships is that we might end up with models in which the values of variables (as specified by the valuation) do not comply with the laws (as defined by the causal dependencies).[12] The notion of a causal model introduced in Sect. 2 does not allow for this. We need a more general notion of causal model, one that does not require the valuation $\mathcal{A}$ to comply with the structural functions in $\mathcal{S}$.

[10] A good example are backtracking counterfactuals: counterfactuals that reason backward in time. The interventionist approach predicts all backtracking to be impossible. However, sometimes backtracking seems to be possible. This is occasionally explained by distinguishing a possible epistemic reading that allows for backtracking.

[11] Thus, in particular, our notion of intervention will not allow for abductive reasoning from effect to cause.

[12] In other words, there might be variables whose values diverge from what is predicted by their structural function when applied to the values of their causal parents.

Definition 6 (General causal model) A *general causal model* for $\langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ is a tuple $\langle \mathcal{S}, \mathcal{A} \rangle$ in which $\mathcal{S}$ is defined as in Definition 1 and $\mathcal{A}$ is defined as a function assigning to every $X \in (\mathcal{U} \cup \mathcal{V})$ a value $\mathcal{A}(X) \in \mathcal{R}(X)$.

With this generalized notion of a causal model at hand, we can now introduce a new notion of intervention capturing the idea described above: it modifies the value of causal variables, but leaves causal relationships unaffected. This new form of intervention will be expressed by a different type of sentence in the formal language.

Definition 7 Formulas ϕ of the language $\mathcal{L}_{[\,],\,\square\!\!\rightarrow}$ over $\langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ are given by

$$\phi ::= X{=}x \mid \neg\phi \mid \phi \wedge \phi \mid [\overrightarrow{X}{=}\overrightarrow{x}\,]\phi \mid (\overrightarrow{X}{=}\overrightarrow{x}\,) \,\square\!\!\rightarrow \phi$$

for $X \in \mathcal{U} \cup \mathcal{V}$, $x \in \mathcal{R}(X)$, $\overrightarrow{X} = (X_1, \ldots, X_k) \in (\mathcal{U} \cup \mathcal{V})^k$, $k \in \mathbb{N}$, $X_i \neq X_j$ for $i \neq j$ and $\overrightarrow{x} = (x_1, \ldots, x_k)$ with $x_i \in \mathcal{R}(X_i)$.

We have now two counterfactual formulas: $[\overrightarrow{X}{=}\overrightarrow{x}\,]\phi$ and $(\overrightarrow{X}{=}\overrightarrow{x}\,) \,\square\!\!\rightarrow \phi$. The former will be semantically interpreted using the well-known notion of intervention described in Definition 5. We will refer to this notion of intervention as *strict* intervention. The latter will be semantically interpreted using our new notion of *non-strict* intervention, which captures the following intuition: a non-strict intervention affects only the variables that are causally dependent on the intervened ones, with their values set according to the causal laws. Therefore, the rest of the variables (those causally independent of the intervened ones) should remain untouched. This idea is formally spelled out as follows.

Definition 8 Let $\langle \mathcal{S}, \mathcal{A} \rangle$ be a general causal model. Formulas in $\mathcal{L}_{[\,],\,\square\!\!\rightarrow}$ that are also in $\mathcal{L}_{[\,]}$ are interpreted as in Definition 5. For formulas of the form $(\overrightarrow{X}{=}\overrightarrow{x}\,) \,\square\!\!\rightarrow \phi$,

$$\langle \mathcal{S}, \mathcal{A} \rangle \models (\overrightarrow{X}{=}\overrightarrow{x}\,) \,\square\!\!\rightarrow \phi \qquad \text{iff}_{def} \qquad \langle \mathcal{S}, \mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}} \rangle \models \phi$$

with $\langle \mathcal{S}, \mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}} \rangle$ the general causal model whose valuation, $\mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}}$, is obtained in the following way. Let $\overrightarrow{X_d}$ be a vector containing the variables in $\overrightarrow{X}$ whose current value (given by $\mathcal{A}$) is different from their intended new value (indicated by $\overrightarrow{x}$).

(i) The value of variables in $\overrightarrow{X}$ becomes $\overrightarrow{x}$ (as indicated by the intervention).
(ii) For each variable Y not in $\overrightarrow{X}$,
   (a) if Y is not causally dependent on any variable in $\overrightarrow{X_d}$ (i.e., if there is no $X_d \in \overrightarrow{X_d}$ such that $X_d \rightarrowtail^{+}_{\mathcal{S}} Y$), keep its value as in $\mathcal{A}$.
   (b) if Y is causally dependent on some variables in $\overrightarrow{X_d}$ (i.e., if $X_d \rightarrowtail^{+}_{\mathcal{S}} Y$ for some $X_d \in \overrightarrow{X_d}$), its value is calculated according to the causal laws in $\mathcal{S}$ from the values already in $\mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}}$.[13]

[13] The model is recursive so, from $\mathcal{S}$'s induced *causal graph* $\langle \mathcal{U} \cup \mathcal{V}, \rightarrowtail \rangle$ and the antecedent $\overrightarrow{X}{=}\overrightarrow{x}$, one can create a chain of sets of variables $S_0 \subset \cdots \subset S_k$ where $S_0 = \mathcal{U} \cup \overrightarrow{X}$, $S_k = \mathcal{U} \cup \mathcal{V}$ and, for any S_i and

The notions of strict (Definition 5) and non-strict intervention (Definition 8) differ in two crucial points. The first is the structural functions of the models they produce. In the model resulting from a strict intervention, the intervened variables have been cut off from their causal parents; however, the model resulting from a non-strict intervention preserves the previous causal information. Because of this, after a non-strict intervention, the valuation ($\mathcal{A}^{\vec{X}=\vec{x}}$) may not comply with the structural functions ($\mathcal{S}$). The second difference concerns the way the new valuation is defined for endogenous variables. In the strict interventionist case, the values of *all* endogenous variables are recalculated according to the (recall: modified) structural functions. In the non-strict case, recalculation (recall: with respect to the original structural functions) takes place only for endogenous variables causally dependent on those intervened variables whose value actually changes.[14]

The following observation will be useful: in a model where the valuation complies with the structural functions, the valuation that results from a non-strict intervention can be equivalently defined in the following way.

Proposition 1 *Let $\langle \mathcal{S}, \mathcal{A} \rangle$ be a causal model (i.e., a general causal model where $\mathcal{A}$ complies with $\mathcal{S}$); let $\vec{X}=\vec{x}$ be the antecedent of a counterfactual formula. Let*

- *$\vec{X}_d$ be as in Definition 8: a vector containing the variables in $\vec{X}$ whose value (as given by $\mathcal{A}$) differs from their intended new value (as indicated by $\vec{x}$);*
- *$\vec{Z}$ be the endogenous variables not occurring in $\vec{X}$ that are not causally dependent on variables in $\vec{X}_d$, with $\vec{z}$ their values according to $\mathcal{A}$.*

Then, the valuation $\mathcal{A}^{\vec{X}=\vec{x}}$ (Definition 8) can be equivalently defined as the (unique) valuation that is identical with $\mathcal{A}$ with respect to exogenous variables not in $\vec{X}$, assigns to exogenous variables in $\vec{X}$ their respective value in $\vec{x}$, and complies with the causal dependencies in $\mathcal{S}_{(\vec{X}=\vec{x}, \vec{Z}=\vec{z})}$ (see Definition 5).

Thus note how, if $\langle \mathcal{S}, \mathcal{A} \rangle$ has no causal violations (i.e., $\mathcal{A}$ complies with $\mathcal{S}$), the valuation created by a non-strict intervention with $\vec{X}=\vec{x}$, given by $\mathcal{A}^{\vec{X}=\vec{x}}$, coincides with the one created by a strict intervention with ($\vec{X}=\vec{x}$, $\vec{Z}=\vec{z}$), given by $\mathcal{A}^{\mathcal{S}_{(\vec{X}=\vec{x}, \vec{Z}=\vec{z})}}$.

5.1 Fisher's Counter-Examples Revisited

The semantics for counterfactuals proposed here can deal with the examples **Match** and **Headlamp** discussed in Sects. 3 and 4. For reasons of space we will only discuss **Match** (**Headlamp** works analogously).

S_{i+1}, the value of variables in $S_{i+1} \setminus S_i$ can be calculated from $\mathcal{S}$ and the value of variables in S_i. Since the values $\mathcal{A}^{\vec{X}=\vec{x}}$ assigns to variables in S_0 are fixed (from the initial valuation $\mathcal{A}$ and the antecedent $\vec{X}=\vec{x}$), the values of the rest can be properly obtained.

[14] Note: when the original valuation $\mathcal{A}$ complies with the causal dependencies in $\mathcal{S}$, both strategies produce the same result.

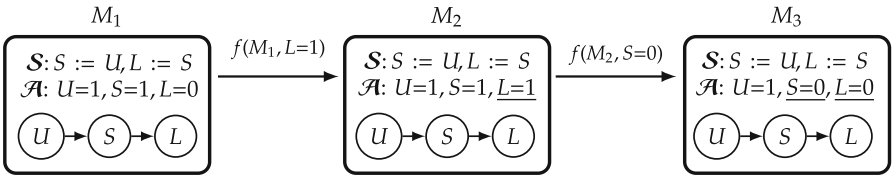

Fig. 3 The evaluation of the **Match** example with the selection function f

- **Match.** I hold up a match and strike it, but it does not light. I say
  (4) If the match had lit, then (even) if it had not been struck, it would have lit.

First, we define the causal model $M_1 = \langle \mathcal{S}, \mathcal{A} \rangle$ for interpreting the counterfactual 5.1. Take $\mathcal{V} = \{S, L\}$ and $\mathcal{U} = \{U\}$, with S indicating whether the match has been struck (1:yes, 0:no) and L indicating whether the match has lit (1:yes, 0:no). The exogenous variable U represents external factors causally responsible for S.[15]

The sentence contains nested counterfactuals, so we need to intervene twice: first, with $L{=}1$ (the antecedent of the main counterfactual), and then, with $S{=}0$ (the antecedent of the embedded counterfactual). On the resulting model, we should check whether $L{=}1$ (the consequent of the embedded counterfactual) is true. The first intervention, $L{=}1$, produces model M_2 in Fig. 3 (see Definition 8), affecting the original valuation but preserving the original causal dependencies. For evaluating the embedded counterfactual $(S{=}0) \boxdot\!\!\rightarrow L = 1$, we apply the second intervention, $S{=}0$, to M_2. This results in the model M_3 in Fig. 3, with $S = 0$ as the intervention requires, and $L = 0$, as L's value is still causally sensitive to S. In this final model, the innermost consequent $L = 1$ fails; thus,

$$M_1 \not\models (L{=}1) \boxdot\!\!\rightarrow ((S{=}0) \boxdot\!\!\rightarrow L = 1).$$

We correctly predict that the counterfactual 5.1 is false in the given context.

5.2 The Import/Export Principle

The core examples discussed in this paper have an interesting connection with the famous Import-Export Principle (IEP), stating that if A and B share no variables, then $(A \wedge B) \boxdot\!\!\rightarrow C$ and $A \boxdot\!\!\rightarrow (B \boxdot\!\!\rightarrow C)$ are equivalent.[16] This principle got a lot of attention in the literature on the logic of conditional sentences, in particular in connection to indicative conditionals (Lewis 1973; Skyrms 1980 among many others). But while for indicative conditionals it is generally accepted that the principle should be valid, this is less clear for counterfactuals. The similarity approach predicts the principle to fail,[17]

[15] Note: our setting allows intervention on exogenous variables, so S can be taken to be exogenous, thus making U superfluous. Still, U is kept, in line with the common modelling strategy of representing external factors by means of exogenous variables.

[16] This principle is sometimes also called the Weak Import-Export Principle, while the principle without the restriction of non-common variables is called Import-Export Principle Fisher (2017).

[17] But see Starr (2014) for a dynamic semantic implementation of the similarity analysis that does validate the Import-Export Principle.

so people have looked for examples confirming this prediction (Etlin 2008; Kaufmann 2005; Starr 2019). Interestingly, these examples are often identical or very similar to the core examples of Fisher that this paper tries to account for. A rare exception is (5-a), brought forward in Skyrms (1980). Skyrms considers this sentence to be true and the related counterfactual (5-b) where both antecedents of (5-a) are combined in one antecedent to be false.

(5) a. If this sample were burning green (say it was barium) then it would still be true that had it been sodium it would have burned yellow.
b. If the sample were burning green and had been sodium, it would have burn yellow.

Another example can be found in Lange (1999).

> Suppose that you and I have just run a race, and I have won. I believe that I would always win if I really tried. Then I am willing to assert: "Suppose that you had won the race. Then I must not have been trying; had I tried, I would have won." This is $p > (q > r)$. I am not willing to assert the corresponding $(p \wedge q) > r$: Had you won and I really tried, I would have won. There is no logically possible world in which you and I both win the race. (Lange 1999, p. 259)

However, all the examples brought forward as violations of IEP share the causal structure of the core examples discussed in our paper: IEP is observed to fail in case the first antecedent in a right-nested counterfactual causally depends on the second, embedded antecedent. In this situation the second antecedent can overwrite the truth of the first antecedent. But if the counterfactual is reformulated in conjunctive form this overruling of the first antecedent is not possible anymore.

While the similarity approach does predict that IEP fails, it cannot explain the specific circumstances in which the principle breaks down and why it seems to hold in so many other cases. Now, one would expect an interventionist account to outperform the similarity approach here, because of the central role causality plays in the interventionist picture. But a strict interventionalist approach like the one introduced in Sect. 2 validates IEP (see Briggs (2012)). Recall that in strict interventionism the variables that are intervened on are cut off from their causal parents and forced to a particular value. Then all endogenous variables are calculated again from the values of the exogenous variables and the new causal dependencies. The order in which interventions are executed has no effect on the result: once a variable is intervened on, no later intervention on different variables can change its value. So, the particular observation that we are considering here, where a later intervention overrules an earlier one, cannot be modeled.

Our account, however, is made exactly to deal with these causal exceptions to the IEP (see Sect. 5.1). We allow for later interventions to overwrite the effect of earlier ones. This occurs exactly in the case the later intervention affects causes of the earlier intervention. If the antecedents of a right nested counterfactual are causally independent of each other, our approach will predict that IEP is valid.[18] But in general

[18] In this case the order in which the interventions are performed has no effect on the resulting model. This is because an intervention, as defined here, only affects the value of a variable if the variable is intervened or if it causally depends on an intervened variable.

IEP is not valid. Our approach improves on the similarity approach, because it puts the finger much more precisely on the point were IEP fails.

5.3 The Axiomatization

For an axiom system for $\mathcal{L}_{[\,],\,\Box\!\!\rightarrow}$ w.r.t. general causal models, the crucial observation is Proposition 1: the model that results from a non-strict intervention can be defined in terms of a strict intervention. Thus, the axiomatisation follows a small variation of the *reduction* strategy typically found in *dynamic epistemic logic* Baltag et al. (1998), van Ditmarsch et al. (2008), van Benthem (2011): provide axioms and rules that allow to translate any formula with a non-strict-intervention operator $\Box\!\!\rightarrow$ into a logically equivalent one where this operator does not occur. Then, rely on a sound and complete axiom system for the language without $\Box\!\!\rightarrow$.[19] Table 2 provides the axioms and rules that define the translation.

Theorem 2 *The axioms and rules in Table 2, together with the axioms and rules in Table 1, provide a sound and complete axiom system for* $\mathcal{L}_{[\,],\,\Box\!\!\rightarrow}$ *w.r.t. general causal models.*

Axioms in Table 2 indicate how to deal with a non-strict intervention operator depending on the form of its consequent. Axiom **A10** is the most important, relying on Proposition 1 to describe the valuation after a non-strict intervention $\Box\!\!\rightarrow$ in terms of the valuation after a (different) strict intervention $[\,]$. It makes use of the abbreviation $\rightsquigarrow^{+}$, which characterises syntactically the notion of causal dependency $\rightarrowtail^{+}$ (the transitive closure of direct dependency).[20] In words, it states that if there are vectors of variables $\overrightarrow{X_d}$ and $\overrightarrow{Z}$ such that $\overrightarrow{X_d}$ contains exactly the variables in $\overrightarrow{X}$ whose value would change (conjuncts 1 and 2 in the antecedent) and $\overrightarrow{Z}$ contains exactly the variables that are not causally dependent on those in $\overrightarrow{X_d}$ (conjuncts 3 and 4 in the antecedent), then a non-strict intervention with $\overrightarrow{X}=\overrightarrow{x}$ coincides with a strict intervention with $(\overrightarrow{X}=\overrightarrow{x},\ \overrightarrow{Z}=\overrightarrow{z})$. Observe how, given $\overrightarrow{X}=\overrightarrow{x}$, there are *always* such

[19] The 'small variation' detail refers to the fact that, although we do have an axiom system for the fragment of $\mathcal{L}_{[\,],\,\Box\!\!\rightarrow}$ without non-strict intervention (Sect. 2), the system characterises validities over a different class of models: *causal models*. In this particular case this is not a problem, the reason being that the setting of this section is a conservative extension of that in Sect. 2. Indeed, atoms, Boolean operators and $[\,]$ are interpreted as before. The class of models does change, but this affects neither atoms nor Boolean operators. Crucially, strict interventions force the valuation to agree with the structural functions, so formulas occurring under their scope behave just as in a causal model. The reduction translates non-strict interventions into strict ones, so this is enough.

[20] The syntactic characterisation of causal dependency $\rightarrowtail^{+}$ relies in that of $\rightarrowtail$. Indeed, given that $|\mathcal{U}|=m$ and $|\mathcal{V}|=n$, the fact that V is causally dependent on X can expressed by the formula

$$(X \rightsquigarrow V) \vee \bigvee_{k=1}^{m+n-2} \bigvee_{\{\langle X_1,\ldots,X_k\rangle \mid X_i \in (\mathcal{U}\cup\mathcal{V}\setminus\{X,V\})\}} \Big((X \rightsquigarrow X_1) \wedge \bigwedge_{j=1}^{k-1} (X_j \rightsquigarrow X_{j+1}) \wedge (X_k \rightsquigarrow V)\Big)$$

which is abbreviated as $X \rightsquigarrow^{+} V$. Note how, while k runs over the number of needed intermediate variables (at least 1, and at most $m+n-2$), the intermediate disjunct runs over all possible tuples of k variables.

Table 2 Axiom system for $\mathcal{L}_{[\,],\,\square\!\!\rightarrow}$ w.r.t. general causal models over the given signature

A10	$\bigvee_{\overrightarrow{X_d} \subseteq \overrightarrow{X}} \bigvee_{\overrightarrow{Z} \subseteq \mathcal{U} \cup \mathcal{V}} \bigwedge \left\{ \begin{array}{l} \bigwedge_{X_i \in \overrightarrow{X} \cap \overrightarrow{X_d}} (X_i \neq x_i), \\ \bigwedge_{X_i \in \overrightarrow{X} \setminus \overrightarrow{X_d}} (X_i = x_i), \\ \bigwedge_{Z \in \overrightarrow{Z}} \neg \bigvee_{X_d \in \overrightarrow{X_d}} (X_d \rightsquigarrow^{+} Z), \\ \bigwedge_{Z' \in \mathcal{V} \setminus \overrightarrow{Z}} \bigvee_{X_d \in \overrightarrow{X_d}} (X_d \rightsquigarrow^{+} Z') \end{array} \right\} \rightarrow \left(((\overrightarrow{X} = \overrightarrow{x}) \,\square\!\!\rightarrow Y = y) \leftrightarrow [\overrightarrow{X} = \overrightarrow{x}, \overrightarrow{Z} = \overrightarrow{z}\,] Y = y \right)$
A11	$((\overrightarrow{X} = \overrightarrow{x}) \,\square\!\!\rightarrow \neg\phi) \leftrightarrow \neg((\overrightarrow{X} = \overrightarrow{x}) \,\square\!\!\rightarrow \phi)$
A12	$\left((\overrightarrow{X} = \overrightarrow{x}) \,\square\!\!\rightarrow (\phi \wedge \psi) \right) \leftrightarrow \left((\overrightarrow{X} = \overrightarrow{x}) \,\square\!\!\rightarrow \phi \;\wedge\; (\overrightarrow{X} = \overrightarrow{x}) \,\square\!\!\rightarrow \psi \right)$
A13	$((\overrightarrow{X} = \overrightarrow{x}) \,\square\!\!\rightarrow [\overrightarrow{Z} = \overrightarrow{z}\,]\phi) \leftrightarrow ([\overrightarrow{X'} = \overrightarrow{x'}][\overrightarrow{Z} = \overrightarrow{z}\,]\phi)$ with $\overrightarrow{X'} = \overrightarrow{x'}$ the subassignment of $\overrightarrow{X} = \overrightarrow{x}$ for $\overrightarrow{X'} = \overrightarrow{X} \setminus \mathcal{V}$
RE	From $\psi_1 \leftrightarrow \psi_2$ derive $\phi \leftrightarrow \phi[\psi_2/\psi_1]$, with $\phi[\psi_2/\psi_1]$ a formula obtained by replacing one or more occurrences of ψ_1 within ϕ by ψ_2.

vectors $\overrightarrow{X_d}$ and $\overrightarrow{Z}$; the role of the (always true) antecedent is simply to 'calculate' their contents.

Axioms **A11**-**A12** are the rules for Boolean operators. Then, Axiom **A13** states that a non-strict intervention on endogenous variables does not affect the truth of formulas within the scope of strict intervention; this is because causal relationships are invariant under non-strict interventions. Finally, **RE** makes the reduction work for nested non-strict interventions by means of an inside-first strategy: apply the reduction over the innermost intervention operator in the formula until the operator disappears, and then proceed with the next.[21]

6 Discussion and Conclusions

In this paper we proposed a new approach to the semantics of counterfactual conditionals. Our proposal builds on the well-known interventionist approach, but spells out the notion of intervention differently. There are two separate steps that we took in defining our proposal. First, we made a substantial conceptual shift in what we understand to be the target of intervention. We propose that intervention does not take place at the level of structural dependencies, but at the level of the (incidental) valuations of the variables. Conceptually, this means that we see intervention not as a hypothetical modification of the underlying laws of nature, but as the hypothetical assumption of exceptions to the laws (see Schulz 2011, 2014 for a similar move). As a consequence, after intervention, no information on causal dependencies in the actual world is lost. The second part of the proposal lies in how exactly we define the valuation resulting from intervention. We propose, on the one hand, that the value of variables not causally affected by the intervened variables remains unchanged and, on the other hand, that the value of the causally dependent variables is recalculated according to the laws, the new value of the intervened variables and the old values of the causally independent variables (see Definition 8). This approach allows us to satisfy our objectives: (i) the predictions made for the truth conditions of counterfactuals that are not right-nested are the same as made in Briggs (2012), and (ii) the approach correctly deals with the counterexamples brought forward in Fisher (2017).

The change we propose for the concept of intervention, though minor in terms of predictions, is conceptually quite substantial. This places our approach in the middle between central strongholds in the landscape of approaches to counterfactuals. The new approach is still in all rights an *interventionist* approach to counterfactuals: it still takes counterfactuals to reason about hypothetical changes to the world and not alternative beliefs about what the facts are. However, it makes different assumptions about what underlies these changes. In particular, intervention is not taken to make any changes to the laws of nature. Instead, the truth of the antecedent is modelled as a small and local exception to the course or nature. This seems to place our approach in the same boat as epistemic accounts of counterfactuals, which also assume that no

[21] Thus, the generalisation rule for non-strict intervention is not required. For details on this, see Wang and Cao (2013) (in particular, Theorem 11).

laws are given up.[22] But epistemic approaches take counterfactuals to reason about what an agent would have inferred had she believed the antecedent to be true. Thereby, they in general allow for abductive reasoning in counterfactuals, which goes against the core idea of interventionism and is excluded in our account. The law information is kept in our approach for the sole purpose of allowing a recovery of violated causal laws for the evaluation of embedded counterfactuals. This is what we take to be the main point of Fisher's examples: they show that embedded counterfactuals can make use of causal laws that the antecedent of the embedding counterfactual intervened on.

In future work we hope to provide more evidence for the conceptual change in the notion of intervention that we propose. For instance, we should look for other counterfactuals for which both notions of intervention make different predictions, and then test which approach better matches the intuitions of speakers. Remember that Fisher only discusses examples of the form (i) $B \boxdot\!\!\rightarrow (\neg A \boxdot\!\!\rightarrow B)$ (see the top-right corner of Fig. 4), where A is a cause of B, while he claims that the observation extends to arbitrary right-nested counterfactuals. One way to test our approach would be to look at other types of right-nested counterfactuals, for instance examples of the form (iii) $B \boxdot\!\!\rightarrow (\neg A \boxdot\!\!\rightarrow C)$, where C is a direct cause of B. In a scenario where A causes B and B causes C, if in the actual context we have $A = 1, B = C = 0$, the strong interventionist approach would predict that both (i) and (iii) should be true, while according to our approach these counterfactuals should be false. We performed a preliminary study to test these predictions (see the two scenarios in Fig. 4). While we could confirm, using the same method as before[23], that still the majority of the participants consider counterfactual of the form (i) false (left diagram in Fig. 4), this effect becomes weaker for counterfactuals of type (iii) and basically disappears in combination with scenario 2 (right diagram in Fig. 4). Notice that additional variables do not mean a simple increase in uncertainty in the given answers. People still feel that they have intuitions about the truth values of these sentences; it is just that their opinions differ.

These results are problematic for strict interventionism as well as the alternative we proposed here. But one should be careful with over-interpreting the experiments reported here. Future work will have to show whether the results obtained are stable. And only when we have a clear picture of the phenomenon that we need to account for does it make sense to continue modifying the approach. Still, our proposal makes an important step in the right direction. Fisher's examples clearly show that sometimes we need to be able to recall causal dependencies after an intervention has violated them. This means that the structural information about these dependencies should not be the locus of the intervention. So, what we certainly want to defend here is the proposed step from intervention on the causal dependencies to intervention on the valuation of the variables. Whether the exact form we gave to intervention on the valuation is correct needs to be studied in future work.

[22] Except in the case that the antecedent explicitly denies a certain law.

[23] The questionnaire and data are available at https://osf.io/5t4uf/.

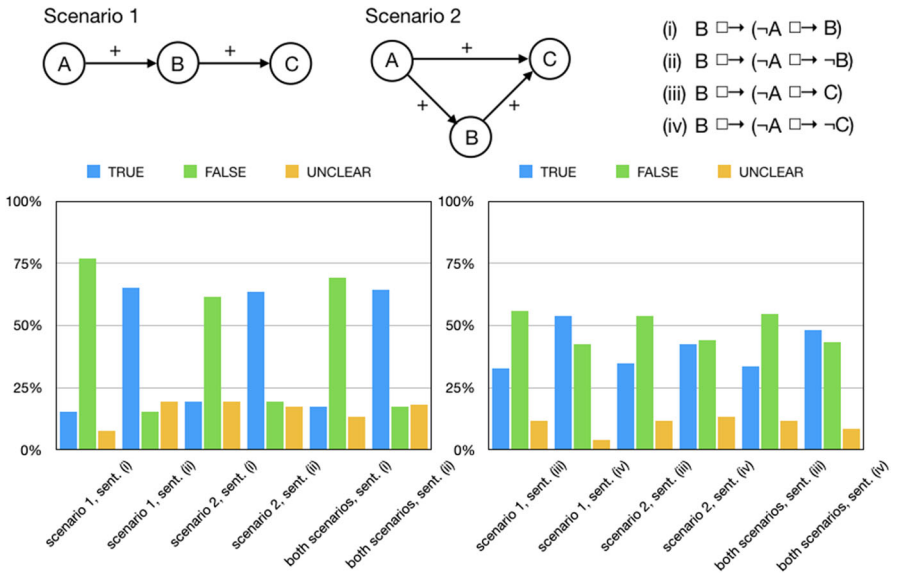

Fig. 4 Overview of the results of the second study; the sentences (i)-(iv) are those that we asked participants to judge in the two scenarios

A Proofs

Proof of Theorem 1. Soundness follows from the validity of axioms and rules. Completeness relies on Halpern (2000), which shows that **MP** and the axioms **D0**-**D11** (Table 3) are sound and complete with respect to recursive causal models *for a slightly different language* $\mathcal{L}^+$. Here is a sketch of the argument.

First, formulas in the language $\mathcal{L}^+$ are Boolean combinations of expressions of the form $[\overrightarrow{X}=\overrightarrow{x}]\gamma$, with $\overrightarrow{X} \subseteq \mathcal{V}$, $x_i \in \mathcal{R}(X_i)$ and γ a Boolean combination of atoms of the form $Y(\overrightarrow{u})=y$, with $Y \in \mathcal{V}$, $y \in \mathcal{R}(Y)$ and $\overrightarrow{u}$ a vector of values for all variables in $\mathcal{U}$. An atom $Y(\overrightarrow{u})=y$ is true in a model when setting variables in $\mathcal{U}$ to $\overrightarrow{u}$ gives Y the value y. As mentioned, **MP** and axioms **D0**-**D11** are sound and complete for recursive causal models (Theorem 3.3 in Halpern (2000)). [24] Then, note that the referred paper also mentions not only that **D10** implies **D2** and **D9**, but also that, within recursive models, **D2** and **D10** are equivalent. Thus, **MP** together with axioms **D0**-**D8** and **D11** are sound and complete for $\mathcal{L}^+$ within recursive models.

Now, for our purposes, observe first that axioms **D0**, **MP**, **D4**, **D6** and **D8** in the table above are identical to our axioms **A0**, **MP**, **A4**, **A6** and **A8**. Moreover, axioms **D1**

[24] Axioms **D0**-**D11** here are written in our style. Note that (i) $\langle\overrightarrow{X}=\overrightarrow{x}\rangle\phi$ abbreviates $\neg[\overrightarrow{X}=\overrightarrow{x}]\neg\phi$, and (ii) for $\overrightarrow{X}=(X_1,\ldots,X_k)$ and $\overrightarrow{x}=(x_1,\ldots,x_k)$, occurrences of $\overrightarrow{X}=\overrightarrow{x}$ *outside intervention operators* abbreviate the conjunction $X_1=x_1\wedge\cdots\wedge X_k=x_k$. The version of **D11** shown here looks drastically different from the original, mainly because it includes nested interventions, which are forbidden in $\mathcal{L}^+$. Yet, our version is correct, as atoms in $\mathcal{L}^+$, expressions of the form $Y(\overrightarrow{u})=y$, essentially allow for interventions on *exogenous variables* (changes in the values of variables in $\mathcal{U}$) to occur within the scope of interventions on endogenous variables $[\overrightarrow{X}=\overrightarrow{x}]$.

Table 3 Axiom system for $\mathcal{L}^{+}$ w.r.t. recursive causal models over the given signature (from Halpern (2000))

D0	Propositional tautologies	**MP**	From ϕ and $\phi \rightarrow \psi$ infer ψ
D1	$[\overrightarrow{X}=\overrightarrow{x}]\big(Y{=}y \rightarrow \neg(Y{=}y')\big)$ for $y, y' \in \mathcal{R}(Y)$ with $y \neq y'$	**D2**	$[\overrightarrow{X}=\overrightarrow{x}]\Big(\bigvee_{y\in\mathcal{R}(Y)} Y{=}y\Big)$
D3	$\Big(\langle\overrightarrow{X}=\overrightarrow{x}\rangle(Y{=}y \wedge \overrightarrow{Z}=\overrightarrow{z})\Big) \rightarrow \langle\overrightarrow{X}=\overrightarrow{x}, Y{=}y\rangle(\overrightarrow{Z}=\overrightarrow{z})$	**D4**	$[\overrightarrow{X}=\overrightarrow{x}, Y{=}y](Y{=}y)$
D5	$\Big(\langle\overrightarrow{X}=\overrightarrow{x}, Y{=}y\rangle(Z{=}z \wedge \overrightarrow{W}=\overrightarrow{w}) \wedge \langle\overrightarrow{X}=\overrightarrow{x}, Z{=}z\rangle(Y{=}y \wedge \overrightarrow{W}=\overrightarrow{w})\Big) \rightarrow \langle\overrightarrow{X}=\overrightarrow{x}\rangle(Z{=}z \wedge Y{=}y \wedge \overrightarrow{W}=\overrightarrow{w})$ (for $\overrightarrow{W} = \mathcal{V} \setminus (\overrightarrow{X} \cup \{Z, Y\})$)		
D6	$(X_0 \rightsquigarrow X_1 \wedge \cdots \wedge X_{k-1} \rightsquigarrow X_k) \rightarrow \neg(X_k \rightsquigarrow X_0)$	**D7**	$[\overrightarrow{X}=\overrightarrow{x}](\phi \rightarrow \psi) \rightarrow ([\overrightarrow{X}=\overrightarrow{x}]\phi \rightarrow [\overrightarrow{X}=\overrightarrow{x}]\psi)$
D8	$[\overrightarrow{X}=\overrightarrow{x}]\phi$ for ϕ a propositional tautology		
D9	$\langle\overrightarrow{X}=\overrightarrow{x}\rangle\top \wedge \bigvee_{y\in\mathcal{R}(Y)}[\overrightarrow{X}=\overrightarrow{x}]Y{=}y$ if $\overrightarrow{X} = \mathcal{V} \setminus \{Y\}$	**D10**	$\langle\overrightarrow{X}=\overrightarrow{x}\rangle\top \wedge \bigvee_{y\in\mathcal{R}(Y)}[\overrightarrow{X}=\overrightarrow{x}]Y{=}y$
D11	$[\overrightarrow{V}=\overrightarrow{v}]([\overrightarrow{U}=\overrightarrow{u_1}]\phi_1 \wedge \cdots \wedge [\overrightarrow{U}=\overrightarrow{u_\ell}]\phi_\ell) \leftrightarrow \big([\overrightarrow{V}=\overrightarrow{v}][\overrightarrow{U}=\overrightarrow{u_1}]\phi_1 \wedge \cdots \wedge [\overrightarrow{V}=\overrightarrow{v}][\overrightarrow{U}=\overrightarrow{u_\ell}]\phi_\ell\big)$ (for $\overrightarrow{V} \subseteq \mathcal{V}$ and $\overrightarrow{U} = \mathcal{U}$)		

and **D2** are derivable in our system by using **A8** over, respectively, **A1** and **A2**. Then, axioms **D3** and **D5** are also derivable in our system, this time via propositional reasoning and $\langle \overrightarrow{X}=\overrightarrow{x}\rangle\phi \leftrightarrow [\overrightarrow{X}=\overrightarrow{x}]\phi$, which is derivable from **A7.1** and propositional reasoning.[25] Note also how **D7** can be derived from **A7.1**, **A7.2** and propositional reasoning.[26] Finally, axiom **D11** is derivable in our system too (an instance of repetitive applications of **A7.2**). Thus, **MP** and axioms **A0**-**A8** are sound and complete within recursive models *for the fragment of* $\mathcal{L}_{[\,]}$ *that corresponds to* $\mathcal{L}^{+}$.

To complete the argument note that, besides allowing explicit interventions on exogenous variables variables (with axioms **A4** and $\mathbf{A}_{\mathcal{U}}$ describing the effects), the only significant way in which $\mathcal{L}_{[\,]}$ extends $\mathcal{L}^{+}$ is by imposing no restrictions on a counterfactual's consequent. The extension means, effectively, that nested counterfactuals are allowed. In order to deal with them, our axiom system has axioms **A7.1**, **A7.2** and **A9**: the first and the second 'push' intervention operators inside the counterfactual's consequent (commuting with negations and distributing over conjunctions), and the third collapses two sequential interventions into a single one. Thus, with the help of axiom $\mathbf{A}_{[]}$ for the basic cases, every formula in $\mathcal{L}_{[\,]}$ can be translated into a semantically equivalent one in $\mathcal{L}^{+}$, for which **MP** and **A0**-**A8** are sound and complete.

Proof of Proposition 1. Let $\langle \mathcal{S}, \mathcal{A}\rangle$ be a causal model. Let $\overrightarrow{X}=\overrightarrow{x}$ be a counterfactual antecedent, with (i) $\overrightarrow{X_d}$ the variables in $\overrightarrow{X}$ whose $\mathcal{A}$-value differs from the intended $\overrightarrow{x}$, and (ii) $\overrightarrow{Z}$ the endogenous variables not occurring in $\overrightarrow{X}$ that are not causally dependent on variables in $\overrightarrow{X_d}$, with $\overrightarrow{z}$ their $\mathcal{A}$-values. By construction, $\overrightarrow{X}$ and $\overrightarrow{Z}$ are disjoint. Now, use $\mathcal{A}_1$ to abbreviate $\mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}}$ (Definition 8), so $\mathcal{A}_1$ assigns the respective x_i to each variable X_i in $\overrightarrow{X}$ and, for variables not in $\overrightarrow{X}$, follows $\mathcal{A}$ for those not causally dependent of any variable in $\overrightarrow{X_d}$, using $\mathcal{S}$ and the values already in $\mathcal{A}_1$ to calculate those causally dependent of some variable in $\overrightarrow{X_d}$. Finally, let $\mathcal{A}_2$ be the (unique) valuation that is identical with $\mathcal{A}$ with respect to exogenous variables not in $\overrightarrow{X}$, assigns to exogenous variables in $\overrightarrow{X}$ their respective value in $\overrightarrow{x}$, and complies with the structural functions in $\mathcal{S}_{(\overrightarrow{X}=\overrightarrow{x},\,\overrightarrow{Z}=\overrightarrow{z})}$. Take any $Y \in \mathcal{U}\cup\mathcal{V}$; it will be shown that $\mathcal{A}_1(Y) = \mathcal{A}_2(Y)$.

First, for *exogenous* variables. (i) If $Y \in \mathcal{U}$ is not in $\overrightarrow{X}$, then both $\mathcal{A}_1$ and $\mathcal{A}_2$ agree with $\mathcal{A}$, the first because Y is not causally dependent on any other variable, and the second by definition. (ii) If $Y \in \mathcal{U}$ is in $\overrightarrow{X}$, both $\mathcal{A}_1$ and $\mathcal{A}_2$ assign to it the value indicated by $\overrightarrow{x}$.

[25] By definition, $\langle \overrightarrow{X}=\overrightarrow{x}\rangle\phi \leftrightarrow \neg[\overrightarrow{X}=\overrightarrow{x}]\neg\phi$; by **A7.1**, $[\overrightarrow{X}=\overrightarrow{x}]\neg\phi \leftrightarrow \neg[\overrightarrow{X}=\overrightarrow{x}]\phi$. Thus, by propositional reasoning (from $A \leftrightarrow \neg B$ and $B \leftrightarrow C$ derive $A \leftrightarrow \neg C$), it follows that $\langle \overrightarrow{X}=\overrightarrow{x}\rangle\phi \leftrightarrow \neg\neg[\overrightarrow{X}=\overrightarrow{x}]\phi$ and hence, by propositional reasoning once again (from $A \leftrightarrow \neg\neg B$ infer $A \leftrightarrow B$), the required $\langle \overrightarrow{X}=\overrightarrow{x}\rangle\phi \leftrightarrow [\overrightarrow{X}=\overrightarrow{x}]\phi$ is obtained.

[26] From **A7.1** we get $[\overrightarrow{X}=\overrightarrow{x}]\neg(\phi\wedge\neg\psi) \leftrightarrow \neg[\overrightarrow{X}=\overrightarrow{x}](\phi\wedge\neg\psi)$; from **A7.2** we get $[\overrightarrow{X}=\overrightarrow{x}](\phi\wedge\neg\psi) \leftrightarrow ([\overrightarrow{X}=\overrightarrow{x}]\phi \wedge [\overrightarrow{X}=\overrightarrow{x}]\neg\psi)$. Then, by propositional reasoning, $[\overrightarrow{X}=\overrightarrow{x}]\neg(\phi \wedge \neg\psi) \leftrightarrow \neg([\overrightarrow{X}=\overrightarrow{x}]\phi \wedge [\overrightarrow{X}=\overrightarrow{x}]\neg\psi)$, that is (axiom **A7.1** and propositional reasoning), $[\overrightarrow{X}=\overrightarrow{x}]\neg(\phi \wedge \neg\psi) \leftrightarrow \neg([\overrightarrow{X}=\overrightarrow{x}]\phi \wedge \neg[\overrightarrow{X}=\overrightarrow{x}]\psi)$. Hence, by the definition of "$\rightarrow$", it follows that $[\overrightarrow{X}=\overrightarrow{x}](\phi \rightarrow \psi) \rightarrow ([\overrightarrow{X}=\overrightarrow{x}]\phi \rightarrow [\overrightarrow{X}=\overrightarrow{x}]\psi)$.

Now, for *endogenous* variables. (i) If $Y \in \mathcal{V}$ is in $\overrightarrow{X}$, then $\mathcal{A}_1$ gives it the value indicated by $\overrightarrow{x}$ (by definition), and so does $\mathcal{A}_2$ (by complying with $\mathcal{S}_{(\overrightarrow{X}=\overrightarrow{x},\overrightarrow{Z}=\overrightarrow{z})}$). (ii) If $Y \in \mathcal{V}$ is in $\overrightarrow{Z}$ then, $\mathcal{A}_1$ uses the value in $\mathcal{A}$ (by definition), and so does $\mathcal{A}_2$ (via the values $\overrightarrow{z}$, taken from $\mathcal{A}$). (iii)Finally, suppose Y is neither in $\overrightarrow{X}$ nor in $\overrightarrow{Z}$; then, the structural functions used by $\mathcal{A}_1$ and $\mathcal{A}_2$ to calculate Y's value are the same: from $\mathcal{S}$ for the first, and from $\mathcal{S}_{(\overrightarrow{X}=\overrightarrow{x},\overrightarrow{Z}=\overrightarrow{z})}$ for the second. On its own, this does not guarantee that Y's *value* under both $\langle \mathcal{S}, \mathcal{A}_1\rangle$ and $\langle \mathcal{S}_{(\overrightarrow{X}=\overrightarrow{x},\overrightarrow{Z}=\overrightarrow{z})}, \mathcal{A}_2\rangle$ is the same: there is a unique function F_Y, and yet the values of its parameters (all other variables) might be different. But, according to the previous items, the only variables in which $\mathcal{A}_1$ and $\mathcal{A}_2$ might differ are precisely the endogenous variables in neither $\overrightarrow{X}$ nor in $\overrightarrow{Z}$. Then, relying on the *recursiveness* of the model, one can use an inductive argument to show that, when the process that assigns values to variables calculates the value of such a Y, the values of all its *parents* will be the same in both $\mathcal{A}_1$ and $\mathcal{A}_2$. Indeed, the step $\#_0$ in the process assigns values to the variables in the set $S_0 := \{Y \in \mathcal{V} \setminus (\overrightarrow{X} \cup \overrightarrow{Z}) \mid$ the parents of Y are in $\mathcal{U} \cup \overrightarrow{X} \cup \overrightarrow{Z}\}$. The valuations $\mathcal{A}_1$ and $\mathcal{A}_2$ coincide in the values of the variables in $\mathcal{U} \cup \overrightarrow{X} \cup \overrightarrow{Z}$, so they will coincide in the values of variables in S_0. Crucially, the model is recursive, so $S_0 \neq \emptyset$. Then, each step $\#_{k+1}$ assigns values to the variables in the set $S_{k+1} := \{Y \in \mathcal{V} \setminus (\overrightarrow{X} \cup \overrightarrow{Z}) \mid$ the parents of Y are in $\mathcal{U} \cup \overrightarrow{X} \cup \overrightarrow{Z} \cup \bigcup_{0\leq i\leq k} S_i\}$. Now, $\mathcal{A}_1$ and $\mathcal{A}_2$ coincide in the values of the variables in $\mathcal{U} \cup \overrightarrow{X} \cup \overrightarrow{Z} \cup \bigcup_{0\leq i\leq k} S_i$, so they will coincide in the values of variables in S_{k+1}. But, again, the model is recursive, so $S_{k+1} \neq \emptyset$. Thus, eventually the values of all variables in $\mathcal{V} \setminus (\overrightarrow{X} \cup \overrightarrow{Z})$ will be calculated, and the values will be the same in both $\mathcal{A}_1$ and $\mathcal{A}_2$.

Proof of Theorem 2. The proof follows the *reduction* axioms strategy frequently used in *dynamic epistemic logic* Baltag et al. (1998), van Ditmarsch et al. (2008), van Benthem (2011). In our case, the strategy relies on a sound and complete axiom system for the $\Box\!\!\rightarrow$-less fragment of $\mathcal{L}_{[\,],\Box\!\!\rightarrow}$. For the remaining formulas, those involving $\Box\!\!\rightarrow$, the strategy uses 'reduction axioms': valid formulas and validity-preserving rules indicating how to translate a formula with occurrences of $\Box\!\!\rightarrow$ into a provably equivalent one without them. Soundness follows from the validity and validity-preserving properties of the new axioms and rules (so a formula and its translation are semantically equivalent); completeness follows from the completeness of the axiom system for the $\Box\!\!\rightarrow$-less fragment, as the recursion axioms define a recursive validity-preserving translation from the full $\mathcal{L}_{[\,],\Box\!\!\rightarrow}$ into the latter. The reader is referred to Wang and Cao (2013) and (van Ditmarsch et al. 2008, Chapter 7) for a detailed explanation of this technique.

For the underlying system, **MP** and axioms **A0**-**A9**, $\mathbf{A}_{\mathcal{U}}$ and $\mathbf{A}_{[]}$ constitute a sound and complete axiomatization for $\mathcal{L}_{[\,]}$ over causal models (Theorem 1). But, the fragment in $\mathcal{L}_{[\,],\Box\!\!\rightarrow}$ belonging to $\mathcal{L}_{[\,]}$ is not affected by the change of models (in particular, because strict interventions force the valuation to agree with the structural functions, thus always producing a causal model). Thus, as argued in Footnote 19, they also constitute a sound and complete axiom system for $\mathcal{L}_{[\,]}$ over general causal models.

For dealing with $\boxrightarrow$, axioms **A10**-**A13** define the recursive translation that takes any formula in $\mathcal{L}_{[\,],\,\boxrightarrow}$ and returns a logically equivalent one without $\boxrightarrow$. Here are arguments for their validity.

- Axiom **A10** is the basic case for the translation, as it eliminates $\boxrightarrow$ by showing how the valuation that results from a non-strict intervention is equivalent to a valuation that results from a strict intervention. Its validity follows from Proposition 1.
- Axioms **A11** and **A12** indicate how to deal with negations (commute $\boxrightarrow$ and $\neg$) and conjunctions (distribute $\boxrightarrow$ over $\wedge$). In particular, the validity of the former comes from the fact that a non-strict intervention is deterministic.
- Axiom **A13** eliminates a non-strict intervention that precedes a strict one. For its validity, take a causal model $\langle \mathcal{S}, \mathcal{A} \rangle$, and note the following.
  - $\langle \mathcal{S}, \mathcal{A} \rangle \models (\overrightarrow{X}=\overrightarrow{x}) \boxrightarrow [\overrightarrow{Z}=\overrightarrow{z}]\phi$ if and only if $\langle \mathcal{S}, \mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}} \rangle \models [\overrightarrow{Z}=\overrightarrow{z}]\phi$, that is, if and only if $\langle \mathcal{S}_{\overrightarrow{Z}=\overrightarrow{z}}, (\mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}})^{\mathcal{S}_{\overrightarrow{Z}=\overrightarrow{z}}} \rangle \models \phi$.
  - $\langle \mathcal{S}, \mathcal{A} \rangle \models [\overrightarrow{X'}=\overrightarrow{x'}][\overrightarrow{Z}=\overrightarrow{z}]\phi$ with $\overrightarrow{X'}=\overrightarrow{x'}$ the subassignment of $\overrightarrow{X}=\overrightarrow{x}$ for $\overrightarrow{X'} = \overrightarrow{X} \setminus \mathcal{V}$ if and only if $\langle \mathcal{S}, \mathcal{A} \rangle \models [\overrightarrow{X''}=\overrightarrow{x''}, \overrightarrow{Z}=\overrightarrow{z}]\phi$ with $\overrightarrow{X''}=\overrightarrow{x''}$ the subassignment of $\overrightarrow{X'}=\overrightarrow{x'}$ for $\overrightarrow{X''} = \overrightarrow{X'} \setminus \overrightarrow{Z}$ (Axiom **A9**), that is, if and only if $\langle \mathcal{S}_{\overrightarrow{X''}=\overrightarrow{x''}, \overrightarrow{Z}=\overrightarrow{z}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{X''}=\overrightarrow{x''}, \overrightarrow{Z}=\overrightarrow{z}}} \rangle \models \phi$.

  Thus, it is enough to show both

$$\mathcal{S}_{\overrightarrow{Z}=\overrightarrow{z}} = \mathcal{S}_{\overrightarrow{X''}=\overrightarrow{x''}, \overrightarrow{Z}=\overrightarrow{z}} \qquad \text{and} \qquad (\mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}})^{\mathcal{S}_{\overrightarrow{Z}=\overrightarrow{z}}} = \mathcal{A}^{\mathcal{S}_{\overrightarrow{X''}=\overrightarrow{x''}, \overrightarrow{Z}=\overrightarrow{z}}}.$$

  with $\overrightarrow{X''}=\overrightarrow{x''}$ the subassignment of $\overrightarrow{X}=\overrightarrow{x}$ for $\overrightarrow{X''} = \overrightarrow{X} \setminus (\mathcal{V} \cup \overrightarrow{Z})$ (thus, $\overrightarrow{X''}$ contains only exogenous variables). The first part is straightforward, as $\overrightarrow{X''}$ contains no endogenous variables. For the second part, it will be shown that both valuations agree in the values of every variable. Here are the cases for each *exogenous* variable $U \in \mathcal{U}$:
  - if $U \in \overrightarrow{Z}$, then $U \notin \overrightarrow{X''}$ and both valuations return the appropriate value in $\overrightarrow{z}$;
  - if $U \notin \overrightarrow{Z}$ but $U \in \overrightarrow{X}$, then $U \in \overrightarrow{X''}$ and both valuations return the appropriate value in $\overrightarrow{x}$;
  - if $U \notin \overrightarrow{Z}$ and $U \notin \overrightarrow{X}$, both valuations return the value $\mathcal{A}(U)$;

  Then, here are the cases for each *endogenous* variable $V \in \mathcal{V}$ (thus, $V \notin \overrightarrow{X''}$):
  - if $V \in \overrightarrow{Z}$, then both valuations return the appropriate value in $\overrightarrow{z}$.
  - if $V \notin \overrightarrow{Z}$, then both valuations return the unique solution for the respective (but identical) structural equations under the (identical) values for variables in $\mathcal{U} \cup \overrightarrow{Z}$.

References

Baltag, A., Moss, L. S., & Solecki, S. (1998). The logic of public announcements, common knowledge, and private suspicions. In I. Gilboa (Ed.), *TARK* (pp. 43–56). San Francisco, Morgan Kaufmann.

Barbero, F., Sandu, G. (2019). Interventionist counterfactuals on causal teams. In: B. Finkbeiner, S. Kleinberg (Eds.), *Proceedings 3rd workshop on formal reasoning about causation, responsibility, and explanations in science and technology, Thessaloniki, Greece, 21st April 2018. Volume 286 of Electronic Proceedings in theoretical computer science*, pp 16–30. Open Publishing Association

Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies, 160*(1), 139–166.

Ciardelli, I., Zhang, L., & Champollion, L. (2018). *Two switches in the theory of counterfactuals*. Linguistics and Philosophy: A study of truth conditionality and minimal change.

Etlin, D.J. (2008). Desire, belief, and conditional belief. PhD thesis, Massachusetts Institute of Technology

Fine, K. (2012). Counterfactuals without possible worlds. *The Journal of Philosophy, 109*(3), 221–246.

Fisher, T. (2017). Causal counterfactuals are not interventionist counterfactuals. *Synthese, 194*(12), 4935–4957.

Gabbay, D. M. (1984). Theoretical foundations for non-monotonic reasoning in expert systems. In: K. R. Apt (Ed.), *Logics and models of concurrent systems—conference proceedings, colle-sur-loup (near nice), France, 8-19 October 1984. Volume 13 of NATO ASI Series* (pp. 439–457). Springer.

Galles, D., & Pearl, J. (1997). Axioms of causal relevance. *Artificial Intelligence, 97*(1–2), 9–43.

Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science, 3*(1), 151–182.

Halpern, J. Y. (2016). *Actual causality*. MIT Press.

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. *The Review of Symbolic Logic, 6*(2), 305–322.

Kaufmann, S. (2005). Conditional predictions. *Linguistics and Philosophy, 28*(2), 181–231.

Kaufmann, S. (2013). Causal premise semantics. *Cognitive Science, 37,* 1136–1170.

Koons, R. (2017). Defeasible reasoning. In: E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Winter 2017 edn. Metaphysics Research Lab, Stanford University

Kraus, S., Lehmann, D., & Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence, 44*(1–2), 167–207.

Lange, M. (1999). Laws, counterfactuals, stability, and degrees of lawhood. *Philosophy of Science, 66*(2), 243–267.

Lewis, D. (1973). Counterfactuals and comparative possibility. In: *Ifs* (pp. 57–85). Springer.

Makinson, D. (1988). General theory of cumulative inference. In: M. Reinfrank, J. de Kleer, M.L. Ginsberg, E. Sandewall (Eds.), *Proceedings of 2nd international workshop, non-monotonic reasoning, Grassau, FRG, June 13-15, 1988. Volume 346 of Lecture Notes in Computer Science*, pp 1–18. Springer.

Marti, J., & Pinosio, R. (2014). Similarity orders from causal equations. In: *European workshop on logics in artificial intelligence*, pp. 500–513. Springer.

Palan, S., & Schitter, C. (2018). Prolific.ac: A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17,* 22–27.

Pearl, J. (2000). *Causality. Models, reasoning, and inference*. Cambridge University Press.

Pearl, J. (2013). Structural counterfactuals: A brief introduction. *Cognitive Science, 37,* 977–85.

Schulz, K. (2011). if you wiggle A, then B will change. Causality and counterfactual conditionals. *Synthese, 179*(2), 239–251.

Schulz, K. (2014). Minimal models vs. logic programming: The case of counterfactual conditionals. *Journal of Applied Non-Classical Logics, 24*(1–2), 153–168.

Skyrms, B. (1980). The prior propensity account of subjunctive conditionals. In: *Ifs* (pp. 259–265). Springer

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*. MIT Press.

Stalnaker, R. C. (1968). A theory of conditionals. In: *Ifs* (pp. 41–55). Springer.

Starr, W. (2019). Counterfactuals. In: Zalta, E. N. (Ed.), *The stanford encyclopedia of philosophy*. Fall 2019 edn. Metaphysics Research Lab, Stanford University

Starr, W. B. (2014). A uniform theory of conditionals. *Journal of Philosophical Logic, 43*(6), 1019–1064.

van Benthem, J. (2011). *Logical dynamics of information and interaction*. CUP

van Ditmarsch, H., van der Hoek, W., & Kooi, B. (2008). *Dynamic epistemic logic*. Springer.

Wang, Y., & Cao, Q. (2013). On axiomatizations of public announcement logic. *Synthese, 190*(1), 103–134.