A Logical and Empirical Study of Right-Nested Counterfactuals

Katrin Schulz[1], Sonja Smets[1,2], Fernando R. Velázquez-Quesada[1], and Kaibo Xie[1(✉)]

[1] Institute for Logic, Language and Computation, Universiteit van Amsterdam, Amsterdam, The Netherlands
{K.Schulz,S.J.L.Smets,F.R.VelazquezQuesada,K.Xie}@uva.nl
[2] Department of Information Science and Media Studies, University of Bergen, Bergen, Norway

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 in general 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, and we will extend the well-known logic of actual causality with a new operator expressing an alternative notion of intervention that does not suffer from the problem (and thus can account for some critical examples). The core idea of the alternative approach is a new notion of intervention, which operates on the evaluation of the variables in a causal model, and not on their functional dependencies. Our result provides new insights into the logical analysis of causal reasoning.

1 Introduction

The meaning of counterfactual conditionals, sentences of the form *"If A were/had been the case, then B would be/have been the case"*, 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: philosophy, logic, psychology and others. But despite a lot of effort, no consensus has been reached yet about how the meaning of these sentences needs to be approached.

Following the similarity approach of Stalnaker and Lewis [1,2], which still is the dominant approach in the philosophical literature, counterfactuals are evaluated as follows. Given the antecedent A and the context of evaluation, we select certain (hypothetical) situations in which the antecedent is true, then checking whether they make the consequent B true as well. The question is how to define the relevant selection function correctly. According to Lewis and Stalnaker, the selection is based on similarity: we select 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.

P. Blackburn et al. (Eds.): LORI 2019, LNCS 11813, pp. 259–272, 2019.
https://doi.org/10.1007/978-3-662-60292-8_19

In recent years the interventionist approach to counterfactuals became very popular ([3–7] 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 [8,9]. The approach got its name from the way it describes the selection function. The antecedent is made true by intervention on the given causal dependencies: it is cut off its causal parents and stipulated to be true by law.[1]

Recently, this approach has been criticized by Fisher [12]. 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. We will propose a variation of the approach that drops strict interventionism and 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 [6]. 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 extend, they both make the same counterfactuals true. So, our proposal, though formalizing a slightly different take on what intervention means, is in terms of logical properties a very conservative change of the original interventionist approach.

2 The Interventionist Approach to Counterfactuals

Our presentation of the interventionist approach to counterfactuals is based on the one proposed by Briggs in [13]; still, we will only introduce the parts that are relevant for the discussion at hand. The two central ingredients of the approach are ***(i)*** the causal model, which contains information about the relevant causal dependencies, and ***(ii)*** the operation of intervention involved in the definition of the selection function, which maps a given causal model onto a class of models that make the antecedent of a given counterfactual true.

Causal models represent the causal dependencies between a given finite set of variables. For each variable V we fix its range $\mathcal{R}(V)$, the set of possible values the variable can take. The variables are sorted into the set $\mathcal{U}$ of *exogenous* variables (those whose value is independent from the value of other variables in the system), and the set $\mathcal{V}$ of *endogenous* variables (those whose value causally depends on the value of other variables in the system). Given a set of variables $\mathcal{U} \cup \mathcal{V}$, a *causal model* over $\mathcal{U} \cup \mathcal{V}$ is a tuple $\langle \mathcal{S}, \mathcal{A} \rangle$. The first component, $\mathcal{S}$, fixes the causal dependencies between the variables by assigning to each $V \in \mathcal{V}$ a function F_V that maps the values of a set of variables $PA_V \subseteq \mathcal{U} \cup \mathcal{V}$ (the *parents* of V) to a value of the variable V. The second component, $\mathcal{A}$, is a valuation

[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 [10,11].

function, assigning a value $\mathcal{A}(V) \in \mathcal{R}(V)$ to every $V \in \mathcal{U} \cup \mathcal{V}$ in a way that *complies with the causal dependencies in* $\mathcal{S}$: for all variables V, if $V \in \mathcal{V}$, then $\mathcal{A}(V) = F_V(\mathcal{A}(PA_V))$. Thus, if the values of the exogenous variables (those in $\mathcal{U}$) are given, the values of the variables in $\mathcal{V}$ can be calculated from these values and $\mathcal{S}$.[2] Finally, for talking about causal models, we use a simple propositional language extended with an operator for counterfactual conditionals.[3]

Definition 2.1 (Language $\mathcal{L}_{\Box\!\!\rightarrow}$). Formulas ϕ of the language $\mathcal{L}_{\Box\!\!\rightarrow}$ over $\mathcal{U} \cup \mathcal{V}$ are given by

$$\phi ::= V = v \mid \neg\phi \mid \phi \wedge \phi \mid (\overrightarrow{V} = \overrightarrow{v}) \Box\!\!\rightarrow \phi \quad \text{for } V \in \mathcal{V},\ v \in \mathcal{R}(V),\ \overrightarrow{V} = (V_1, \ldots, V_n) \in \mathcal{V}^n,\ n \in \mathbb{N},\ V_i \neq V_j \text{ for } i \neq j,\ \overrightarrow{v} = (v_1, \ldots, v_n) \text{ with } v_i \in \mathcal{R}(V_i)$$

Sentences of the form $(\overrightarrow{V} = \overrightarrow{v}) \Box\!\!\rightarrow \phi$ should be read as *"if the variables in* $\overrightarrow{V}$ *were to be set to* $\overrightarrow{v}$*, then* ϕ *would hold"*.

The second important ingredient of the interventionist approach is the notion of intervention involved in the interpretation rule for counterfactual sentences $(\overrightarrow{V} = \overrightarrow{v}) \Box\!\!\rightarrow \alpha$. Given a causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{A} \rangle$ and an antecedent $\overrightarrow{V} = \overrightarrow{v}$, we need to define a model that makes the antecedent true;[4] in order to evaluate the consequent there. In the interventionist approach, this model is built by cutting the variables $\overrightarrow{V}$ off their causal parents $PA_{\overrightarrow{V}}$, forcing their value to be the one given by the antecedent $\overrightarrow{V} = \overrightarrow{v}$, as Definition 2.2 below details.[5]

Definition 2.2 (Intervention). Let $\langle \mathcal{S}, \mathcal{A} \rangle$ be a causal model. The semantic interpretation of the Boolean operators in $\mathcal{L}_{\Box\!\!\rightarrow}$-formula is as usual; for the rest,

$$\begin{array}{lll} \langle \mathcal{S}, \mathcal{A} \rangle \models V = v & \text{iff}_{def} & \mathcal{A}(V) = v \\ \langle \mathcal{S}, \mathcal{A} \rangle \models (\overrightarrow{V} = \overrightarrow{v}) \Box\!\!\rightarrow \phi & \text{iff}_{def} & \langle \mathcal{S}_{\overrightarrow{V} = \overrightarrow{v}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{V} = \overrightarrow{v}}} \rangle \models \phi \end{array}$$

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

[2] This is true for *recursive* causal models, the only ones that this paper will discuss. For their definition: from $\mathcal{S}$, define a relation $\rightarrowtail$ on the set of variables $\mathcal{U} \cup \mathcal{V}$ by writing $X \rightarrowtail Y$ if and only if X is among the parents of Y (the structure $\langle \mathcal{U} \cup \mathcal{V}, \rightarrowtail \rangle$ is called $\mathcal{S}$'s induced *causal graph*). Let $\rightarrowtail^+$ be the transitive closure of $\rightarrowtail$ (so $X \rightarrowtail^+ Y$ indicates that Y is *causally dependent* on X). A causal model is said to be *recursive* when $\rightarrowtail^+$ is a strict partial order (there are no circular dependencies between the variables).

[3] This language $\mathcal{L}_{\Box\!\!\rightarrow}$ extends the basic causal language (e.g., [6]) by allowing right-nested counterfactuals. Still, it is only a fragment of the language used in [13], as it does not allow Boolean combinations of atoms in the antecedent (which are not relevant to the discussion here).

[4] In [13]'s general setting, the selection function returns a set of models. However, for the possible antecedents of counterfactuals considered in our fragment of her language, the selected model is uniquely defined.

[5] Our language is a fragment of that in [13]. Thus, here we only recall the tools from [13] that are needed for our formulas' semantic interpretation.

(i) $\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}$ is as $\mathcal{S}$ except that, for each variable $V_i \in \overrightarrow{V}$, the function F_{V_i} is replaced by a constant function F'_{V_i} assigning the value v_i (i.e., $F'_{V_i} := v_i$).
(ii) $\mathcal{A}^{\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}}$ is the assignment to causal variables that is identical to $\mathcal{A}$ with respect to exogenous variables, and it complies with the causal dependencies in $\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}$ for the endogenous ones.

Thus, the proposal is that the selection function f discussed in the introduction should be defined as

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

It is worthwhile to emphasise that, in the model $\langle \mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}}\rangle$, the valuation $\mathcal{A}^{\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}}$ *complies* with the model's causal dependencies, $\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}$: for every $V \in \mathcal{V}$ we have $\mathcal{A}^{\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}}(V) = F'_V(\mathcal{A}^{\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}}(PA'_V))$. So, intervention happens at the level of $\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}$, and this change affects the valuation $\mathcal{A}^{\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}}$. In Sect. 5 we will introduce a notion of intervention that changes $\mathcal{A}$ directly and leaves $\mathcal{S}$ unaffected.

3 Fisher's Criticism

Fisher [12] criticizes the approach described above. More concretely, he claims that it makes incorrect predictions for right-nested counterfactuals. Concretely, he discusses the examples (1) and (2) below.[6]

- **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 in Fig. 1, where A_1 stands for the variable the first antecedent talks about and A_2 for the variable of the second antecedent.[7]

[6] Fisher also considers another example, involving 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.

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

Fig. 1. A causal model for **Match** and **Headlamp**, before and after interpreting the counterfactuals.

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. 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." ([12]: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 affect V itself. So, as long as ψ does not assign a new value to V, the counterfactual $(V = v) \boxdot\!\!\rightarrow (\psi \boxdot\!\!\rightarrow V = v)$ will always come out as true.

To solve this problem Fisher proposes that we have 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." ([12]:4942). However, he does not propose an alternative approach that has this property.[8] In the rest of the paper we want to do two things. First of all, we need to confirm Fishers judgments concerning the target examples (1) and (2) with an actual survey. After that, we will develop an alternative interventionist approach to the meaning of counterfactual conditionals that is not strictly interventionist.

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 we

[8] Fisher discusses in [12] an alternative definition of intervention, dubbed "side-constrained intervention", but admits that this variation is not really targeting the root of the problem.

conducted a small empirical study in which we did not only ask the participants to judge the counterfactuals (1) and (2), but also their counterparts (3-a) and (3-b). If participants judge the sentences (1) and (2) false because they consider the sentences defective, they should judge (3-a) and (3-b) to be 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** in Sect. 3 and a third scenario containing a counterfactual $\varphi \boxdot\!\!\rightarrow (\psi \boxdot\!\!\rightarrow \xi)$ 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 order of question was randomized. The participants had to judge the truth value of the counterfactual using a slider bar with 5 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 2 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 [14]. 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 that first of all a majority of the participants agree with the intuitions reported by Fisher [12]. 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 well-formedness of the counterfactuals under consideration.

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%

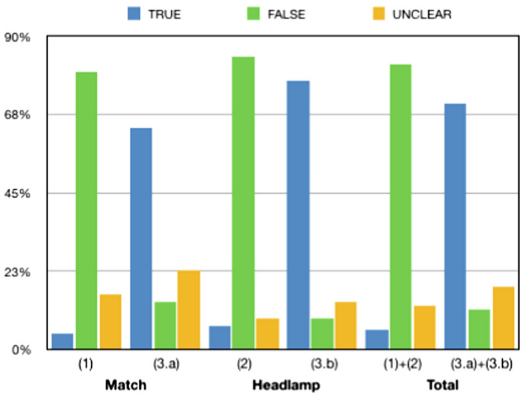

Fig. 2. Results of the 1^{st} study.

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 and only want to add that such a move does not make sense as long as there is no explanation for why the interventionist reading isn't available for the counterfactuals in question.

But does that mean that we need to give up the interventionist approach to counterfactuals? We don't 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 The Non-strict-intervention

The goal is, then, to find a notion of intervention that coincides with [8]'s account for non-nested cases (so it 'inherits' the good behaviour of the strict interventionism approach in those situations), but also satisfies Fisher's adequacy condition (thus agreeing with the results from our study). The definition below meets all these requirements. Its crucial idea is, again, that counterfactual assumptions might modify the value of causal variables, but preserve causal relationships.

Definition 5.1. Let $M = \langle \mathcal{S}, \mathcal{A} \rangle$ be a recursive causal model, and $\overrightarrow{V} = \overrightarrow{v}$ an intervention; let $\overrightarrow{V_d}$ be the variables in $\overrightarrow{V}$ whose current value (as given by $\mathcal{A}$) is different from their intended new value (as indicated by $\overrightarrow{V} = \overrightarrow{v}$). The selection function f is defined as $f(\langle \mathcal{S}, \mathcal{A} \rangle, \overrightarrow{V} = \overrightarrow{v}) := \langle \mathcal{S}, \mathcal{A}^{\overrightarrow{V}=\overrightarrow{v}} \rangle$, with the new assignment $\mathcal{A}^{\overrightarrow{V}=\overrightarrow{v}}$ calculated in the following way.

1. The value of variables in $\overrightarrow{V}$ becomes $\overrightarrow{v}$ (as indicated by the intervention).
2. For each variable Y not in $\overrightarrow{V}$,
   (a) if Y is not causally affected[9] by any variable in $\overrightarrow{V_d}$, keep its value as in $\mathcal{A}$.
   (b) if Y is causally affected by some variables in $\overrightarrow{V_d}$, its value is calculated according to the causal laws in $\mathcal{S}$.[10]

The just defined model, $\langle \mathcal{S}, \mathcal{A}^{\overrightarrow{V}=\overrightarrow{v}} \rangle$, and the one that results from a strict intervention, $\langle \mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}} \rangle$ (Definition 2.2), differ in their causal laws. The latter (Briggs) changes the functions for the intervened variables (producing $\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}$); the former (ours) preserves the original causal information (the 'old' $\mathcal{S}$).[11]

There is a second difference, concerning the way the new assignment is defined. In the strict interventionist case, the values of *all* non-intervened variables are recalculated according to the (recall: new) causal rules. In our case, the only non-intervened variables for which the recalculation takes place (recall: with respect to the original causal laws) are those that are causally affected by variables whose *value* is directly affected by the intervention.[12]

Note that $\langle \mathcal{S}, \mathcal{A}^{\overrightarrow{V}=\overrightarrow{v}} \rangle$ can be equivalently defined as follows:

Proposition 5.1. *Let $\langle \mathcal{S}, \mathcal{A} \rangle$ be a causal model and $\overrightarrow{V} = \overrightarrow{v}$ an intervention. Let*

- *$\overrightarrow{V_d}$ be as before: the causal variables in $\overrightarrow{V}$ whose value (as given by $\mathcal{A}$) differs from their intended new value (as indicated by $\overrightarrow{V} = \overrightarrow{v}$);*
- *$\overrightarrow{Z}$ be the endogenous variables not causally affected by variables in $\overrightarrow{V_d}$, with $\overrightarrow{z}$ their values according to $\mathcal{A}$.*

Then, the assignment $\mathcal{A}^{\overrightarrow{V}=\overrightarrow{v}}$ (Definition 5.1) can be equivalently defined as the (unique) assignment that is identical with $\mathcal{A}$ with respect to exogenous variables, and complies with the causal dependencies in $\mathcal{S}_{(\overrightarrow{V}=\overrightarrow{v}, \overrightarrow{Z}=\overrightarrow{z})}$ (see Definition 2.2)[13].

[9] "Y is causally affected by Z" intuitively means changing the value of Z may change the value of Y under some setting of variables. Formally, it means there exists some variables $\overrightarrow{V}$, $\overrightarrow{v} \in \mathcal{R}(\overrightarrow{v})$, and some distinct value $y, y' \in \mathcal{R}(Y)$, such that the value of Z forced by setting $\overrightarrow{V}, Y$ to $\overrightarrow{v}, y$ is different from its value forced by setting $\overrightarrow{V}, \overrightarrow{Y}$ to $\overrightarrow{v}, y'$.

[10] Recall: the model is recursive. Hence, $\mathcal{S}$'s induced causal graph induces, in turn, a chain of sets of variables $S_0 \subseteq \cdots \subseteq S_n$ such that $S_0 = \mathcal{U} \cup \overrightarrow{V}$, $S_n = \mathcal{U} \cup \mathcal{V}$ and, for any S_i and S_{i+1}, the value of variables in $S_{i+1} \setminus S_i$ can be calculated from the causal dependencies and the value of variables in S_i.

[11] Note: $\mathcal{A}^{\overrightarrow{V}=\overrightarrow{v}}$ may not comply with the causal dependencies in $\mathcal{S}$.

[12] When the original assignment $\mathcal{A}$ complies with the causal dependencies in $\mathcal{S}$, both strategies produce the same result. This is the only case relevant for Briggs' purposes.

[13] Proofs were omitted due to space limitations, but are available online https://www.dropbox.com/s/0i0xy416rs5dmor/Lori_Proofs.pdf?dl=0.

We will redefine the logic of counterfactuals, using this new notion of intervention for the semantic interpretation of $\boxrightarrow$. However, the strict intervention operator will still be useful, in particular, for axiomatizing the non-strict intervention. Thus, it will appear in the language as well, albeit under a different symbol ([ ]).

Definition 5.2. Formulas ϕ of the language $\mathcal{L}_{\boxrightarrow,[\,]}$ over $\mathcal{U} \cup \mathcal{V}$ are given by

$$\phi ::= V = v \mid \neg\phi \mid \phi \wedge \phi \mid (\overrightarrow{V} = \overrightarrow{v}) \boxrightarrow \phi \mid [\overrightarrow{V} = \overrightarrow{v}]\phi$$

for $V \in \mathcal{V}$, $v \in \mathcal{R}(V)$, $\overrightarrow{V} = (V_1, \ldots, V_n) \in \mathcal{V}^n$, $n \in \mathbb{N}$, $V_i \neq V_j$ for $i \neq j$, $\overrightarrow{v} = (v_1, \ldots, v_n)$ with $v_i \in \mathcal{R}(V_i)$.

For $\mathcal{L}_{\boxrightarrow,[\,]}$'s semantics, atoms and Boolean operators are evaluated as before. The cases for the intervention operators [ ] and $\boxrightarrow$ are as follows.

Definition 5.3. (Intervention). Let $\langle \mathcal{S}, \mathcal{A} \rangle$ be a causal model. Then,

$$\begin{array}{lll} \langle \mathcal{S}, \mathcal{A} \rangle \models [\overrightarrow{V} = \overrightarrow{v}]\phi & \text{iff}_{def} & \langle \mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}} \rangle \models \phi \quad \text{(see Definition 2.2)} \\ \langle \mathcal{S}, \mathcal{A} \rangle \models (\overrightarrow{V} = \overrightarrow{v}) \boxrightarrow \phi & \text{iff}_{def} & \langle \mathcal{S}, \mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}} \rangle \models \phi \quad \text{(see Definition 5.1)} \end{array}$$

If $\langle \mathcal{S}, \mathcal{A} \rangle$ is a model without causal violations (i.e., $\mathcal{A}$ complies with $\mathcal{S}$), then the assignment created by our intervention ($\mathcal{A}^{\overrightarrow{V}=\overrightarrow{v}}$) coincides with the one created by a strict intervention ($\mathcal{A}^{\mathcal{S}_{\overrightarrow{V}=\overrightarrow{v}}}$). Thus, our proposal does extend the original causal modelling semantics [8], providing a non-strict-interventionist approach for nested counterfactuals.

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).

- **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 need to define the causal model $M_1 = \langle \mathcal{S}, \mathcal{A} \rangle$ with respect to which the counterfactual (4) is interpreted. We define $\mathcal{V} = \{S, L\}$ and $\mathcal{U} = \{U\}$, with S indicating whether the match has been struck (1: yes, 0: no), L indicating whether the match has lit (1: yes, 0: no). The exogenous variable U represents external factors causally responsible for S.[14] Furthermore, we define $\mathcal{S} = (S := U, L := S)$ and $\mathcal{A} = (U = 1, S = 1, L = 1)$ (model M_1 in Fig. 3). We need to account for the observation that the counterfactual $(L = 1) \boxrightarrow ((S = 0) \boxrightarrow L = 1)$ is intuitively false with respect to this model,

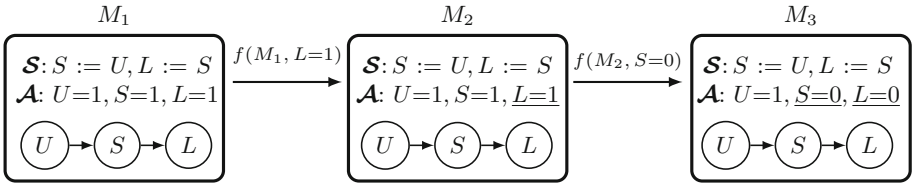

Fig. 3. The evaluation of the **Match** example with the selection function f

a prediction that a strict interventionist approach, as discussed in Sect. 3, is unable to make.

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 (Definition 5.1), affecting the original assignment 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 (4) is false in the given context.

5.2 The Axiomatization for the Logic

The modified notion of intervention can be axiomatized (Table 1) with the help of the axioms for the strict intervention operator $[\,]$ (see [15]) plus additional axioms for $\boxdot\!\!\rightarrow$. Axioms **A1** through **A9** characterise the behaviour of the strict intervention operator $[\,]$.[15]. From axioms **A10**–**A11**, every variable has exactly one value,[16] and axiom **A12** states that our modified version of intervention is still deterministic. Axioms **A13** and **A14** are the crucial ones, as they describe the relationship between the two forms of intervention. Axiom **A13** relies on Proposition 5.1 to describe the assignment after a non-strict intervention $\boxdot\!\!\rightarrow$ in terms of the assignment after a (different) strict intervention $[\,]$. It states that,

[14] If the model allows interventions on *exogenous* variables, the example can be modelled with only two variables: the exogenous one S and the endogenous one L. We use the additional U, as in the literature it is common to allow interventions only on *endogenous* variables.

[15] More precisely, **A1**–**A8** are the axioms for non-nested intervention from [15], and **A9** deals with nested strict-intervention [13,16].

[16] In [15] there are no causal violations; thus, $V = v$ is equivalent to $[\,](V = v)$, and axioms **A1** and **A2** suffice. This is not the case in our setting, as causal violations might occur; hence the need of **A10**–**A11**.

Table 1. Axiom system for $\mathcal{L}_{\Box\!\!\rightarrow,[\,]}$ w.r.t. causal models.

A0	Propositional tautologies	**MP** from ϕ and $\phi \rightarrow \psi$ infer ψ
A1	$[\overrightarrow{V}{=}\overrightarrow{v}](Y = y) \rightarrow \neg[\overrightarrow{V}{=}\overrightarrow{v}](Y = y')$	for $y, y' \in \mathcal{R}(Y)$ with $y \neq y'$
A2	$\bigvee_{y \in \mathcal{R}(Y)} [\overrightarrow{V}{=}\overrightarrow{v}](Y = y)$	
A3	$\Big([\overrightarrow{V}{=}\overrightarrow{v}](Y = y) \wedge [\overrightarrow{V}{=}\overrightarrow{v}](Z = z)\Big) \rightarrow [\overrightarrow{V}{=}\overrightarrow{v}, Y{=}y](Z = z)$	
A4	$[\overrightarrow{V}{=}\overrightarrow{v}, Y{=}y](Y = y)$	
A5	$\Big([\overrightarrow{V}{=}\overrightarrow{v}, Y{=}y](Z = z) \wedge [\overrightarrow{V}{=}\overrightarrow{v}, Z{=}z](Y = y)\Big) \rightarrow [\overrightarrow{V}{=}\overrightarrow{v}](Z{=}z)$	for $Y \neq Z$
A6	$(V_0 \rightsquigarrow V_1 \wedge \cdots \wedge V_{k-1} \rightsquigarrow V_k) \rightarrow \neg(V_k \rightsquigarrow V_0)$ [a]	
A7	$[\overrightarrow{V}{=}\overrightarrow{v}](\phi \wedge \psi) \leftrightarrow ([\overrightarrow{V}{=}\overrightarrow{v}]\phi \wedge [\overrightarrow{V}{=}\overrightarrow{v}]\psi)$	**A8** $[\overrightarrow{V}{=}\overrightarrow{v}]\neg\phi \leftrightarrow \neg[\overrightarrow{V}{=}\overrightarrow{v}]\phi$
A9	$([\overrightarrow{V}{=}\overrightarrow{v}][\overrightarrow{Y}{=}\overrightarrow{y}]\psi) \leftrightarrow [\overrightarrow{V'}{=}\overrightarrow{v'}, \overrightarrow{Y}{=}\overrightarrow{y}]\psi$	for $\overrightarrow{V'} = \overrightarrow{V} \setminus \overrightarrow{Y}$
A10	$(V = v) \rightarrow \neg(V = v')$	for $v, v' \in \mathcal{R}(V)$ and $v \neq v'$
A11	$\bigvee_{v \in \mathcal{R}(V)} (V = v)$	
A12	$(\overrightarrow{V} = \overrightarrow{v}) \,\Box\!\!\rightarrow \bigvee_{y \in \mathcal{R}(Y)} (Y = y)$	
A13	$\bigwedge \left\{ \begin{array}{l} \bigwedge_{V_i \in \overrightarrow{V_d}} (V_i \neq v_i), \\ \bigwedge_{V_i \in \overrightarrow{V} \setminus \overrightarrow{V_d}} (V_i = v_i), \\ \bigwedge_{Z \in \overrightarrow{Z}} \neg \bigvee_{V \in \overrightarrow{V_d}} (V \rightsquigarrow Z), \\ \bigwedge_{Y \in \mathcal{V} \setminus \overrightarrow{Z}} \bigvee_{V \in \overrightarrow{V_d}} (V \rightsquigarrow Y) \end{array} \right\} \rightarrow \Big(((\overrightarrow{V} = \overrightarrow{v}) \,\Box\!\!\rightarrow X = x) \leftrightarrow [\overrightarrow{V}{=}\overrightarrow{v}, \overrightarrow{Z} = \overrightarrow{z}]X = x\Big)$	
A14	$((\overrightarrow{V} = \overrightarrow{v}) \,\Box\!\!\rightarrow [\overrightarrow{X} = \overrightarrow{x}]\phi) \leftrightarrow ([\overrightarrow{X} = \overrightarrow{x}]\phi)$	
A15	$\Big((\overrightarrow{V} = \overrightarrow{v}) \,\Box\!\!\rightarrow (\phi \wedge \psi)\Big) \leftrightarrow ((\overrightarrow{V} = \overrightarrow{v}) \,\Box\!\!\rightarrow \phi \;\wedge\; (\overrightarrow{V} = \overrightarrow{v}) \,\Box\!\!\rightarrow \psi)$	
A16	$((\overrightarrow{V} = \overrightarrow{v}) \,\Box\!\!\rightarrow \neg\phi) \leftrightarrow \neg((\overrightarrow{V} = \overrightarrow{v}) \,\Box\!\!\rightarrow \phi)$	

[a] With $Y \rightsquigarrow Z$, indicating that *"Y has causal effect on Z"*, given (see, e.g., [15]) by

$$\bigvee_{\overrightarrow{v} \in \mathcal{R}(\overrightarrow{V}), \{y, y'\} \subseteq \mathcal{R}(Y), y \neq y', \{z, z'\} \in \mathcal{R}(Z), z \neq z'} [\overrightarrow{V}{=}\overrightarrow{v}, Y{=}y](Z = z) \wedge [\overrightarrow{V}{=}\overrightarrow{v}, Y{=}y'](Z = z').$$

if $\overrightarrow{V_d}$ contains exactly the variables in $\overrightarrow{V}$ whose value would change (conjuncts 1 and 2 in the antecedent), and $\overrightarrow{Z}$ contains exactly the variables that are not causally affected by those in $\overrightarrow{V_d}$ (conjuncts 3 and 4 in the antecedent), then a non-strict-intervention with $\overrightarrow{V} = \overrightarrow{v}$ coincides with a strict intervention with $\overrightarrow{V} = \overrightarrow{v}, \overrightarrow{Z} = \overrightarrow{z}$. Axiom **A14** then uses strict intervention to state that causal relationships are invariant under non-strict interventions. Finally, axioms **A15**–**A16** are the rules for Boolean operators.

Theorem 5.1. *This axiom system is sound and strongly complete with respect to recursive causal models (see Footnote 13).*

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 uses a different approach to intervention. 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 object 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 [4,17] for a similar move). As a consequence, 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 that the value of all variables not causally affected by those variables that we intervene on remain unchanged and that then the value of the remaining variables is calculated from this information (as the model is assumed to be recursive) and the unchanged causal dependencies (see Definition 5.1). 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 [13] and (ii) the approach correctly deals with the counterexamples brought forward in [12].

But does that mean that this way all problems with the interventionist approach to counterfactuals are solved? Certainly not. First of all, notice that we target here only the issue of right-nested counterfactuals. But even if we only focus on right-nested counterfactuals, there are still open questions. This approach was specifically designed to deal with the examples and intuitions reported on in [12] and confirmed in Sect. 4. Fisher suggest that the observations he makes generalize to arbitrary right-nested counterfactuals where variables in the first antecedent causally depend on variables in the second antecedent. But whether this is true has to be investigated first. We performed a second study to test whether Fishers expectations are confirmed when using slightly larger models containing a third variable C (see the two scenarios in Fig. 4). While we could confirm, using the same method as before, that still the majority of the participants consider counterfactual of the form (i) $B \boxdot\!\!\rightarrow (\neg A \boxdot\!\!\rightarrow B)$ false (left diagram in Fig. 4), this effect becomes weaker when the consequent is substituted with the third variable C (the counterfactual becomes (iii) $B \boxdot\!\!\rightarrow (\neg A \boxdot\!\!\rightarrow C)$) and basically disappears in combination with scenario 2 (right diagram in Fig. 4). In a third study focusing in particular on this scenario and counterfactuals of the form (iii) we could not find any difference between the number of participants that consider this sentence true and those that considered its counterpart (iv) true.

Based on the work of Fisher [12] and the empirical results presented here it seems clear that the first part of our proposal is on the right track: 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 then gave to intervention on the valuation is correct needs to be studied in future work. In some cases, like the examples discussed in [12], it seems to be exactly what is needed, in other cases it is still unclear what we should predict.

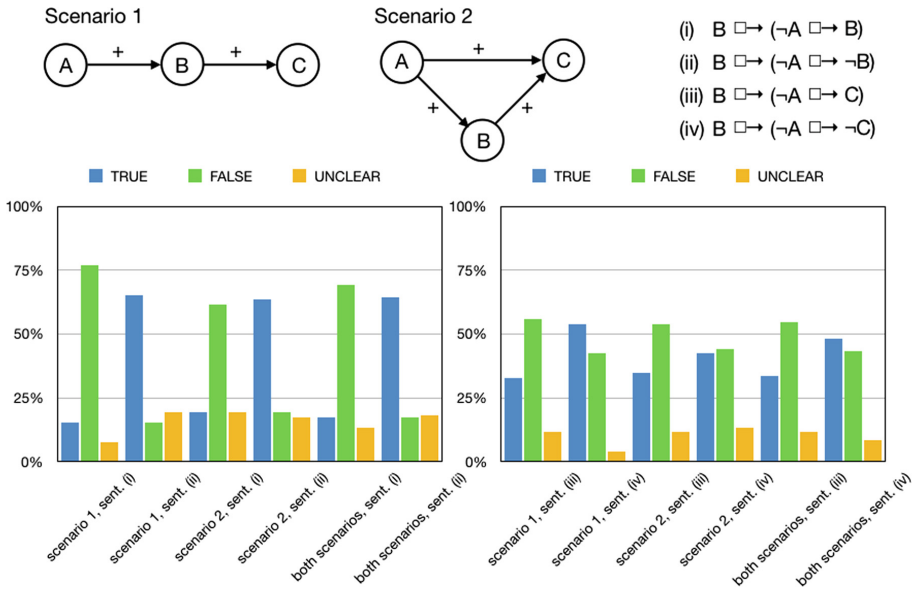

Fig. 4. Overview of the results of the second study; the sentences (i)–(iv) are those that we asked participants to judge in the two scenarios.

References

1. Lewis, D.: Counterfactuals and comparative possibility. In: Harper, W.L., Stalnaker, R., Pearce, G. (eds.) IFS. WONS, vol. 15, pp. 57–85. Springer, Dordrecht (1973). https://doi.org/10.1007/978-94-009-9117-0_3
2. Stalnaker, R.C.: A theory of conditionals. In: Harper, W.L., Stalnaker, R., Pearce, G. (eds.) IFS. WONS, vol. 15, pp. 41–55. Springer, Dordrecht (1968). https://doi.org/10.1007/978-94-009-9117-0_2
3. Pearl, J.: Structural counterfactuals: a brief introduction. Cogn. Sci. **37**, 977–985 (2013)
4. Schulz, K.: "If you wiggle A, then B will change". Causality and counterfactual conditionals. Synthese **179**(2), 239–251 (2011)
5. Kaufmann, S.: Causal premise semantics. Cogn. Sci. **37**, 1136–1170 (2013)
6. Halpern, J.Y.: Actual Causality. MIT Press, Cambridge (2016)
7. Ciardelli, I., Zhang, L., Champollion, L.: Two switches in the theory of counterfactuals. A study of truth conditionality and minimal change. Linguist. Philos. **41**(6), 577–621 (2018)
8. Pearl, J.: Causality. Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
9. Spirtes, P., et al.: Causation, Prediction, and Search. MIT Press, Cambridge (2000)
10. Halpern, J.Y.: From causal models to counterfactual structures. Rev. Symbolic Log. **6**(2), 305–322 (2013)
11. Marti, J., Pinosio, R.: Similarity orders from causal equations. In: Fermé, E., Leite, J. (eds.) JELIA 2014. LNCS (LNAI), vol. 8761, pp. 500–513. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11558-0_35

12. Fisher, T.: Causal counterfactuals are not interventionist counterfactuals. Synthese **194**(12), 4935–4957 (2017)
13. Briggs, R.: Interventionist counterfactuals. Philos. Stud. **160**(1), 139–166 (2012)
14. Palan, S., Schitter, C.: Prolific.ac — a subject pool for online experiments. J. Behav. Exp. Finan. **17**, 22–27 (2018)
15. Halpern, J.Y.: Axiomatizing causal reasoning. J. Artif. Intell. Res. **12**, 317–337 (2000)
16. Barbero, F., Sandu, G.: Interventionist counterfactuals on causal teams. arXiv preprint arXiv:1901.00593 (2019)
17. Schulz, K.: Minimal models vs. logic programming: the case of counterfactual conditionals. J. Appl. Non-Class. Log. **24**(1–2), 153–168 (2014)