"If you'd wiggled A, then B would've changed"

Causality and counterfactual conditionals

Schulz, K.

DOI
10.1007/s11229-010-9780-9

Publication date
2011

Document Version
Author accepted manuscript

Published in
Synthese

"If you'd wiggled A, then B would've changed" Causality and counterfactual conditionals

Katrin Schulz[†]

ILLC
University of Amsterdam
k.schulz@uva.nl

Abstract. This paper deals with the truth conditions of conditional sentences. It focuses on a particular class of problematic examples for semantic theories for these sentences. I will argue that the examples show the need to refer to dynamic, in particular causal laws in an approach to their truth conditions. More particularly, I will claim that we need a causal notion of consequence. The proposal subsequently made uses a representation of causal dependencies as proposed in Pearl (2000) to formalize a causal notion of consequence. This notion inserted in premise semantics for counterfactuals in the style of Veltman (1976) and Kratzer (1979) will provide a new interpretation rule for conditionals. I will illustrate how this approach overcomes problems of previous proposals and end with some remarks on remaining questions.

1. Introduction

(1) a. If you had practiced more, you would have won.
b. If you had been in Paris next week, we could have met.

It is surprising how often counterfactual conditionals like (1-a) and (1-b) occur in our daily conversations. They are regularly used when we evaluate previous actions and plan future behavior. But what do these sentences mean? In which circumstances do you agree that they are true?[1] These are the questions the present paper tries to answer. Of course, this is not the first article addressing this topic. In fact there exists an enormous literature on the meaning of counterfactual conditionals. Students from the most diverse disciplines have been interested in these sentences, among them philosophers, logicians, linguists, psychologists and computer scientists. A big advantage of studying a topic that lives at the intersection of different scientific areas is that one can learn from the authentic perspective and the methodology each of the sciences comes with. In the present paper we will systematically make use of this opportunity.

[†] The author is supported by the Netherlands Organization for Scientific Research.

[1] In this paper, we will simply assume that counterfactual conditionals have truth values.

To come back to the initial question: how would you describe the meaning of counterfactual conditionals (shortly: counterfactuals)? Lets simplify things a bit and assume as logical form of such conditionals $A \gg C$, where A is the antecedent, C the consequent and $\gg$ the conditional connective.[2] A first intuitive description of the meaning of $A \gg C$ is this: a counterfactual is true with respect to a world w_0 if the consequent follows from the antecedent: $[\![A \gg C]\!](w_0) = 1$ iff$_{\text{def}}$ $A \models C$. But what is the relevant notion of entailment in this definition? One could propose to use the classical notion of entailment: $A \models C$ holds if C is true in all models that make A true. But this is clearly too strong, as the following example from Lifschitz illustrates.

The circuit example. *Suppose there is a circuit such that the light is on (L) exactly when both switches are in the same position (up or not up). At the moment switch 1 is down ($\neg S1$), switch 2 is up ($S2$) and the lamp is out ($\neg L$).*

(2) If switch one had been up, the lamp would have been on.

Intuitively, the counterfactual (2) is true in the given context. But it does not hold that the lamp is on in all possible worlds where switch one is up. For instance, in a world where switch 1 is up but switch 2 is down, the lamp is off. The fact that switch 1 is up will only entail that the lamp is on in case switch 2 is still up as well. Hence, there is more information going into the derivation of the consequence from the antecedent. Certain singular facts of the evaluation world w_0 can be used as extra premises. Furthermore, also certain generalizations considered valid in the context of evaluation are available as additional premises. In our example this is the generalization that the light is on (L) exactly when both switches are in the same position. We conclude, the interpretation scheme of conditionals should be rewritten as follows.

A basic interpretation rule for conditional sentences

$[\![A > C]\!]^{D,w_0} = 1$ iff$_{\text{def}}$ $A + P_{w_0} \models_D C$,

where P_{w_0} is a set of singular facts true in the evaluation world w_0, D is a set of regularities considered valid in the evaluation context, and $\models$ the relevant notion of entailment.

This interpretation rule formulates the basic idea behind many approach towards the semantics of counterfactuals. A large part of the literature, particularly in the tradition of cotenability theory (Goodman, 1955) and premise semantics (Veltman, 1976; Kratzer, 1979) addresses

[2] This is the syntactic level to which we will analyze the logical form of conditionals within this paper. This is admittedly still a very coarse-grained view on the compositional structure of natural language conditionals, but sufficient for the goals pursued here.

the question of how to develop an adequate description of the set P_{w_0}, the singular facts of the evaluation world that can be used as additional premisses. The variable D has got less attention. The central claim of this paper is that in fact already the notion of entailment $\models$ is problematic and needs serious attention. I will argue that in order to obtain an adequate description of the truth conditions of counterfactuals we need a *causal* notion of entailment.

Central Claim
The semantics of (the dominant reading of) conditionals relies on a causal notion of entailment.

Section 2 contains arguments supporting this claim. In section 3 a formalization of causal reasoning is developed. This formalization is then put to use in section 4, where a semantic theory for counterfactuals will be presented. In section 5 we will discuss some philosophical implications of the proposal.

2. Motivation

Assuming a causal notion of entailment the basic receipt of how to interpret conditionals reads as follows: *A counterfactual conditional with antecedent A and consequent C is true if A will* **bring about** *C*. The necessity of such a causal notion of entailment can best be brought out in contrast with its most dominant competitor: an epistemic notion of entailment. In this case the basic receipt of how to interpret conditionals gets a different reading: *A counterfactual conditional with antecedent A and consequent C is true if on* **learning** *A you can* **conclude** *C*. In order to decide between the two approaches we have to look for examples they make different predictions for and compare these predictions with our intuitions. There is only space to discuss one type of example here[3]: the assessment example of (Harper, 1981). This example has been used as counterexample to epistemic approaches based on the Ramsey receipt.

The assessment example. *Jones is one of several rising young executives competing for a very important promotion. The company brass have found the candidates so evenly matched that they have employed a psychologist to break the tie by testing for personality qualities correlated with long run success in the corporative world. The test was administered to Jones and the other candidates on Thursday morning. The promotion was decided Thursday afternoon on the basis of the test*

[3] For more see (Schulz, 2007) and the homepage of the author.

scores, but will not be announced until Monday. On Friday morning Jones learnt, through a reliable company grapevine, that the promotion went to the candidate who scored highest on a factor called ruthlessness; but he is unable to discover which of them this is.

It is now Friday afternoon and Jones is faced with a decision. Smith has failed to meet his minimum output quota for the third straight assessment period, and a long standing company policy rules that he should be fired on the spot. Jones believes that his behavior in the decision will provide evidence about how well he scored on the ruthlessness factor.

This example is formulated in the context of probabilistic accounts to the meaning of conditionals. In this context conditionals are not assigned truth values, but probabilities expressing their acceptability. In this case, the epistemic interpretation rule for conditionals becomes: *A counterfactual conditional with antecedent A and consequent C is acceptable if on* **learning** *A the probability you assign to C gets high.* Given that Jones considers his decision to be relevant to his score on the ruthlessness factor, the probability he assigns to getting the promotion (C) conditional on firing Smith (A) is high. According to the epistemic receipt of when to accept counterfactuals, this should mean that his belief in the conditional (3) should be high as well. Intuitively, however, this conditional is neither true nor acceptable in the described scenario, because Jones' decision will in no way **affect** the decision of who will get the promotion.[4]

(3) If I fire Smith, I will get promotion.

Intuitively, the wrong predictions result because the semantics (or acceptability) of the conditional rather relies on a causal dependency than on an evidential relation: *"... what is relevant to deliberation is a comparison of what will happen if I perform some action with what would have happened if I instead did something else. A difference between* [the conditional probability of C given A, the author] *and* [the probability of C, the author] *represents a belief that A is evidentially*

[4] Some readers might object that the particular causal reading of counterfactual conditionals that we try to capture in this paper is not the only possible reading of conditional sentences like (3). There is, in particular, also an epistemic reading of conditionals available, at least to some speakers (see the debate about the famous Hamburger example from Hansson). For counterfactual conditionals the epistemic reading is rather marginal, clearly outperformed by the reading we are interested in here. However, the fact that there are other readings for conditionals available complicates the empirical assessment of semantic theories for counterfactuals.

relevant to the truth of C, but not necessarily a belief that the action has any causal influence on the outcome." (Stalnaker, 1981), p. 151.

To sum up, the assessment example shows that there is a relation between the interpretation of conditional sentences and causality. Still, there are different options left for how to explain this observation. One could argue that the sensibility of conditionals to causal dependencies is only an epiphenomenon (see (Lewis, 1973)). But contrary to Lewis, I claim that the truth conditions of conditional sentences build on the contextually salient causal dependencies. To work out the details of the proposal we first have to formalize a causal notion of entailment. This will be done in the next section.

3. Causal reasoning

3.1. Technical preliminaries

The semantics developed in this paper will interpret a simple propositional language to which a conditional connective $\gg$ has been added. Given a finite set of proposition letters $\mathcal{P}$ the language $\mathcal{L}^0$ is the closure of $\mathcal{P}$ under the connectives $\neg$, $\wedge$ and $\vee$. $\mathcal{L}^{\gg}$ is the union of $\mathcal{L}^0$ with the set of expressions $\phi \gg \psi$ where ϕ and ψ are elements of $\mathcal{L}^0$. We will deviate from classical two-valued logic and use Kleene's strong three valued logic to interpret the language $\mathcal{L}^{\gg}$. This logic distinguishes truth values $\{u, 0, 1\}$ with the partial order $u \leq 0$ and $u \leq 1$. The value u is not to be interpreted as a degree of truth, but rather expresses that the truth value is, so far, undecided. The chosen ordering reflects the intuition that u can 'evolve' towards one of the values 0 and 1. The three-valued truth tables for the connectives $\neg$, $\wedge$ and $\vee$ are given in figure 1.

p	0	0	0	1	1	1	u	u	u
q	0	1	u	0	1	u	0	1	u
$p \wedge q$	0	0	0	0	1	u	0	u	u

p	0	1	u
$\neg p$	1	0	u

p	0	0	0	1	1	1	u	u	u
q	0	1	u	0	1	u	0	1	u
$p \vee q$	0	1	u	1	1	1	u	1	u

Figure 1. Three-valued truth tables for $\neg$, $\wedge$ and $\vee$

An assignment of the truth values u, 1 and 0 to the set of proposition letters $\mathcal{P}$ will be called a situation for $\mathcal{L}^{\gg}$. If the assignment does not use the value u, the situation is also called a possible world for $\mathcal{L}^{\gg}$.

W denotes the set of all possible worlds. We will write $[\![\cdot]\!]^{D,s}$ for the function mapping formulas on truth values given a set of regularities D and a situation s. $[\![\phi]\!]^D$ denotes the set of possible worlds where ϕ is true. The meaning of sentences in $\mathcal{L}^0 \subseteq \mathcal{L}^{\gg}$ is determined based on the truth tables given in figure 1. The parameter D is only relevant for the meaning of conditionals $\phi \gg \psi$. The interpretation rule for conditionals is the central definition of the paper and will only be given at the end of section 4.

3.2. Representing causal dependencies

The goal of the present section is to develop a causal notion of entailment. More in particular, we want to define a notion of entailment that relates a set of literals Σ to a formula ϕ if (the truth of) ϕ causally depends on (the truth of) Σ. This relation can only be defined relative to some representation D of the relevant causal dependencies. If ϕ is a causal consequence of Σ given D we will write: $\Sigma \models_D \phi$.[5]
Before we can give a concrete definition of $\models_D$ we first have to clarify what it means for D to be a representation of the contextual relevant causal dependencies. The definition of a *dynamics* D given below is based on (Pearl, 2000)'s definition of a causal model. However, in some important aspects the notion of a dynamics differs from a causal model in order to overcome some shortcomings of the later notion.[6]
A dynamics is a structure that distinguishes between two groups of proposition letters. On the one hand there is the set B of background variables. These are proposition letters representing facts that are taken to be causally independent of any other proposition. On the other hand there are the inner variables $I = \mathcal{P} - B$. These are propositions letters representing facts that causally depend on others. The character of the dependency is described by the function F that associates every inner variable X with a set of proposition letters Z_X and a two-valued truth function f_X.The proposition letters in Z_X represent the facts that X directly causally depends on. The function f_X describes the character of the dependency: it describes how one can calculate the truth value of X given the values of the members of Z_X. Definition 1 makes use of the notion of rootedness which will be defined below.

[5] In the linguistic literature causal dependence is often analyzed as a primitive relation between events or events and states etc (see, for instance (van Lambalgen and Hamm;, 2005)). In the present article, however, direct causal dependence is a relation that holds between proposition letters. The reason for taking the propositional perspective within this paper is that it simplifies formal matters considerably and it is sufficient for the issues we want to pursue here.

[6] For discussion see (Schulz, 2007).

DEFINITION 1. *(Dynamics)*
A dynamics for $\mathcal{L}^{\gg}$ is a tuple $D = \langle B, F\rangle$, where

i. $B \subseteq \mathcal{P}$ is the set of background variables;

ii. F is a function mapping elements X of $I = \mathcal{P} - B$ to tuples $\langle Z_X, f_X\rangle$, where Z_X is an n-tuple of elements of $\mathcal{P}$ and f_X a two-valued truth function $f_X : \{0,1\}^n \longrightarrow \{0,1\}$. F is rooted in B.

The particular character of causal dependencies makes certain restrictions on the function F necessary. Firstly, causal dependencies cannot be circular; i.e. the fact A can not at the same time be an effect of a fact B and causally responsible for B. Furthermore, we demand that the background variables are those and only those variables that everything else depends upon. That means that if you walk backward in the history of dependencies for every variable you should always end up with background variables. The next definition summarizes these conditions under the notion of rootedness.

DEFINITION 2. *(Rootedness)*
Let $B \subseteq \mathcal{P}$ be a set of proposition letters and F be a function mapping proposition letters $I = \mathcal{P} - B$ to tuples $\langle Z_X, f_X\rangle$, where Z_X is an n-tuple of elements of $\mathcal{P}$ and f_X a two-valued truth function. Let R_F be the relation that holds between two proposition letters $X, Y \in \mathcal{P}$ if Y occurs in Z_X. Let R_F^T be the transitive closure of R_F. We say that F is rooted in B if $\langle \mathcal{P}, R_F^T\rangle$ is a poset and B equals the set of minimal elements of $\langle \mathcal{P}, R_F^T\rangle$.

An application. Let us illustrate how this approach models Lifschitz' circuit example, repeated here from section 1.

> **The circuit example.** *Suppose there is a circuit such that the light is on (L) exactly when both switches are in the same position (up or not up). At the moment switch 1 is down ($\neg S1$) , switch 2 is up ($S2$) and the lamp is out ($\neg L$).*

We need to distinguish three proposition letters for this example: $S1$ stands for *switch 1 is up*, $S2$ for *switch 2 is up* and L for *the lamp is on*. The state of the lamp causally depends on the position of the switches, hence $S1$ and $S2$ are background variables, while L is an inner variable. Thus in the dynamics D of the circuit example the function F maps L on the tuple $\langle\{S1, S2\}, f_L\rangle$ where f_L maps L on 1 if and only if $S1$ and $S2$ have the same truth-value (see figure 2).

D = <B, F>,
B = {S1, S2},
F(L) = <{S1, S2}, f_L>.

f_L:	S1	S2	L
	0	0	1
	0	1	0
	1	0	0
	1	1	1

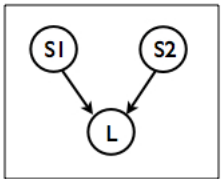


Figure 2. A dynamics for the circuit example

3.3. Causal reasoning formalized

How can we now define the causal notion of entailment $\Sigma \models_D \phi$, i.e. how to define the causal consequences of a set of literals Σ given a dynamics D? We will use to this purpose an idea from logic programming. We will define the causal consequences of Σ as the sentences true in a certain minimal model of Σ. This model will be defined as the least fixed point of an operator $\mathcal{T}_D$. The operator $\mathcal{T}_D$ maps situations s on new situations $\mathcal{T}_D(s)$, calculating the direct causal effects of the settings in s.

Let us illustrate this idea with an example. Figure 3 sketches a new dynamics. In this figure an arrow points from X to Y in case X is a direct cause of Y. For convenience we assume that the causal dependencies holding in this picture follow the formulas: $X_1 \wedge X_2 \leftrightarrow Y_1$, $X_3 \wedge X_4 \leftrightarrow Y_2$, and $Y_1 \wedge Y_2 \leftrightarrow Z_1$. Let us assume, furthermore, that Σ is the set $\{X_1, X_2, Y_2\}$. Let s_Σ be the situation making all formulas in Σ true and mapping all proposition letters not occurring in Σ to the value u. A first application of the operator $\mathcal{T}_D$ to the situation s_Σ calculates the value of Y_1 from the values of X_1 and X_2 and redefines the value of Y_1 from u to 1. In $s_3 = \mathcal{T}_D(\mathcal{T}_D(s_\Sigma))$ also the value of Z_1 is set to 1. This process continues as long as there are causal effects predicted by the regularities in D. But given the way the operator $\mathcal{T}_D$ is defined, the value of some dependent variable Y never can change the value of some variable X that Y depends on (see figure 3).

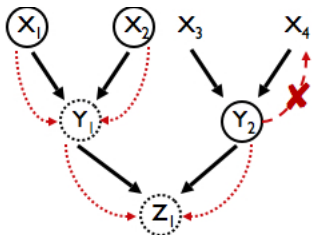


Figure 3. Constructing fixed points

A precise description of the operation $\mathcal{T}_D$ is given in definition 3. For an arbitrary proposition letter q, a situation s and a dynamics D the operation $\mathcal{T}$ determines the value of q in the new situation $\mathcal{T}_D(s)$ as follows: if q is among the background variables of D then causal dependencies cannot have any effect on its value, hence the value of q remains unchanged. If q is not part of the background variables and the laws predict an outcome for q given the value of the direct causes of q in s, then q is set to this predicted value – *provided q was still undetermined in s*. Otherwise the value of q is left unchanged.

DEFINITION 3. *The operation $\mathcal{T}$.*
Let D be a dynamics and s a situation for $\mathcal{L}^{\gg}$. We define the situation $\mathcal{T}_D(s)$ as follows. For all $q \in \mathcal{P}$,

(i) *If $q \in B$ then $\mathcal{T}_D(s)(q) = s(q)$.*

(ii) *If $q \in I = \mathcal{P} - B$ with $Z_q = \langle p_1, ..., p_n \rangle$, then*

a. *If $s(q) = u$ and $f_q(s(p_1), ..., s(p_n))$ is defined, then $\mathcal{T}_D(s)(q) = f_q(s(p_1), ..., s(p_n))$.*

b. *If $s(q) \neq u$ or $f_q(s(p_1), ..., s(p_n))$ is not defined, then $\mathcal{T}_D(s)(q) = s(q)$.*

The application of this operation $\mathcal{T}$ to a situation s will produce a new situation $\mathcal{T}(s)$ where the direct effects of the setting in s are calculated. This operation can be iterated to calculate direct effects again and again. But at some point this process will stagnate and the output situation will be identically to the input situation: a fixed point of $\mathcal{T}$ is reached. For the example in figure 3 this is already the case after two applications of the operation, see figure 4. It can be shown that such a fixed point always exist and that it can be reached in finitely many steps.[7] Based on this fact we can finally define our causal notion of entailment (see definition 4).

DEFINITION 4. *Causal entailment*
Let Σ be a set of literals and D a dynamics. We say that Σ causally entails ϕ given D if ϕ is true on the least fixed point s^*_Σ of $\mathcal{T}_D$ relative to s_Σ.

$$\Sigma \models_D \phi \;\; \textit{iff}_{def} \; [\![\phi]\!]^{D, s^*_\Sigma} = 1.$$

[7] Proofs of these claims can be found on the homepage of the author. Notice that the operation $\mathcal{T}$ is not in the sense monotone that from $s_1 \leq s_2$ it follows $\mathcal{T}_D(s_1) \leq \mathcal{T}_D(s_2)$. Instead, we have $s \leq \mathcal{T}_D(s)$. The reason is that $\mathcal{T}$ cannot change the truth value of propositional variable already set to 1 or 0, even if this contradicts the predictions made by causal regularities described in the dynamics D.

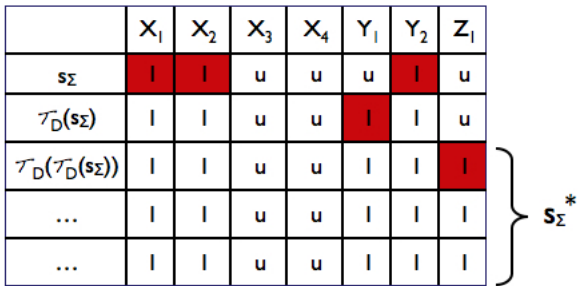


Figure 4. Constructing fixed points

4. Conditional semantics

Below we repeat the basic interpretation rule for conditional sentences developed in the first section of this paper (see page 2).

> A basic interpretation rule for conditional sentences
> $[\![A > C]\!]^{D,w_0} = 1$ iff$_{\text{def}}$ $A + P_{w_0} \models_D C$,
> where P_{w_0} is a set of singular facts true in the evaluation world w_0, D is a set of regularities considered valid in the evaluation context, and $\models$ the relevant notion of entailment.

I have argued in section 2 that in order to specify the details of this interpretation rule (in particular P_{w_0} and $\models_D$) we need a causal notion of entailment. In section 3 such a notion of entailment has been introduced. Still, we are not yet in a position to apply the interpretation rule to concrete examples. Two questions need to be answered first. On the one hand, the set P_{w_0} of singular facts of the evaluation world w_0 has to be specified. This is the topic of section 4.1. On the other hand, we have to clarify what "+" stands for in the formula given above. This sign has to represent an operation of revising P_{w_0} with the new information A: $A + P_{w_0} = Rev(P_{w_0}, A)$. This revision function will be defined in section 4.2.

4.1. A causal notion of basis

The set P_{w_0} of singular facts of the evaluation world w_0 will be defined as a minimal set $B_D(w_0)$ of primitive facts (literals) of w_0 that determine everything else in w_0. Following (Veltman, 2005), we will call this set the *basis* of w_0.[8] The notion "determine" is in the spirit of

[8] In fact, the basic idea of how to define a basis is also directly adopted from (Veltman, 2005). The only difference is that he uses an epistemic interpretation of "determine".

the present paper interpreted causally: the basis specifies the "initial conditions" of w_0; everything else in w_0 is a direct or indirect causal effect of the basis facts. This can be formalized using the fixed point operator $\mathcal{T}$ of section 3: the basis has to be such that when you apply $\mathcal{T}$ the evaluation world w_0 emerges as fixed point.[9]

DEFINITION 5. *The basis*
Let w_0 be a possible world and D a dynamics. The basis $B_D(w_0)$ of w_0 with respect to D is a minimal set of literals B such that $s^*_B = w_0$.

FACT 1. *For a possible world w_0 and a dynamics D the basis $B_D(w_0)$ exists and is uniquely defined.*

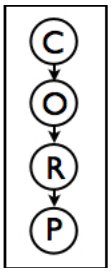
An application: bases in the shooting squad scenario.

> *There is a court, an officer, a rifleman and a prisoner. If the court orders the execution of the prisoner, the officer will give a signal to the rifleman, the rifleman will shoot and the prisoner will die.*

In this scenario 4 proposition letters has to be distinguished: C for *the court orders the execution*, O for *the officer gives the signal*, R for *the rifleman shoots*, and P for *the prisoner dies*. The description of the dynamics D is given in figure 5; f_O, f_R, and f_P are identity functions. The table on the right side of the figure describes a number of possible interpretations (possible worlds) for the proposition letters C, O, R, P. We want to calculate the bases of the possible worlds described in figure 5. For the world w_0 this is the set of literals $\{C\}$. It is easy to see that $s^*_{\{C\}} = w_0$. Starting with a situation that evaluates C as true and the three other proposition letters as undefined, the least fixed point of $\mathcal{T}$ is reached after 3 steps. The basis of world w_1 can be equally unproblematic calculated to be $\{\neg C\}$. The situation is somewhat more problematic for the worlds w_2, w_3 and w_4. These worlds are special because they violate some of the laws described in the dynamics. In w_2, for instance, the officer does not give the signal even though the court orders the execution. As basis for this world the fact C is not enough, because from this fact it does not causally follow that $\neg O$. To causally determine this world one needs additionally the fact that violates the law: $\neg O$. This set, $\{C, \neg O\}$, is sufficient as basis of w_2. In world w_4 the laws are violated twice: Even though the officer gives the signal, the rifleman doesn't die, and even though the rifleman does not shoot, the prisoner dies. Therefore, he basis of w_4 contains even three elements: $B_D(w_4) = \{C, \neg R, P\}$.

[9] For the proof of fact 1 see the homepage of the author.

D = <B, F>
B = {C}
F(O) = <C, f_O>,
F(R) = <O, f_R>,
F(P) = <R, f_P>.



	C	O	R	P
w_0	1	1	1	1
w_1	0	0	0	0
w_2	1	0	0	0
w_3	1	1	0	0
w_4	1	1	0	1

Figure 5. A dynamics for the shooting squad example

4.2. Premise semantics for causal entailment

Now that we have specified the set P_{w_0}, the last thing that needs to be defined is the revision function, which calculates given P_{w_0} and the antecedent A the set Σ of singular facts that serves as input for the causal notion of entailment (see the rule on page 10). This revision function is defined using premise semantics ((Veltman, 1976), (Kratzer, 1979)). According to premise semantics $Rev_D(B_D(w_0), A)$ is the set of maximal subsets of $B_D(w_0)$ that are consistent with A and the relevant laws (encoded in the dynamics D), plus the antecedent. The question is what consistency means in the present frame work. We could interpret it in standard ways and just demand that there is some possible world where this subset of the basis together with the antecedent and all laws is true. But such an approach would not be able to account for the data we want to account for. There is, for instance, no way to predict exclusion of backtracking (neither weak, nor strong) within such a framework. Exclusion of backtracking can only be predicted, if one does allow for worlds where laws are violated, or, to use Lewis' words, miracles occur. The way standard premise semantics is defined the output can never contain words violating laws represented in D. Following the spirit of this paper, one might suggest to use a causal notion of consistency instead, relying on the causal notion of entailment. But what would be a causal notion of consistency? The semantic definition of consistence says that a set S of formulas is consistent if it has a model. Analogously, we define causal consistence as $S \not\models_D \bot$, what comes down to the claim that the minimal fixed point s^*_S has to exist. But when does this point not exist? Have we not shown in section 3 that it always exists? Well, there was one hedge: the initial conditions Σ have to be classically consistent. Otherwise, the situation s_Σ does not exist. We conclude that within the present framework causal consistency relative to a dynamics D comes down to logical consistency independent of D. Thus, an improved proposal for

the revision function would be to stick to revision as defined by premise semantics and just erase all reference to laws (causal dependencies).

DEFINITION 6. *Premise semantics for causal entailment*
Assume $A \in \mathcal{L}^0$ and $B \subseteq \mathcal{L}^0$. We define the revision of B with A relative to D, $Rev_D(B, A)$, as the set of sets $B' \cup \{A\}$ where B' is a maximal subset of B logically consistent with A.

There is one further complication. The causal notion of entailment takes as input a set of literals. Hence, we have to make sure that the revision function returns a set of literals. A closer look at definition 6 reveals that this is only a problem in case the antecedent is not a literal. However, if the antecedent can be rewritten as conjunction of literals, we can use the set of conjuncts as input of the revision function and will end up with a set of literals as result. That still leaves us with antecedents that can only be rewritten in disjunctive normal form with a non-trivial number of disjuncts. In this case all disjuncts have to be considered individually as input of the revision function.

DEFINITION 7. *For every formula $\phi \in \mathcal{L}^0$ $Lit(\phi)$ is the set of sets of literals with the following property: $\{\varphi_1, ..., \varphi_k\} \in Lit(\phi)$ iff $\varphi_1 \wedge ... \wedge \varphi_k$ is one of the disjuncts in the disjunctive normal form of ϕ.*

DEFINITION 8. *The truth conditions of conditionals*
Let $A \gg C$ be an element $\mathcal{L}^{\gg}$, D a dynamics an w_0 a possible world.

$$[\![A \gg C]\!]^{D,w_0} = 1 \text{ iff}_{def} \forall S \in Lit(A) \forall B \in Rev_D(B_{w_0}, S) : B \models_D C$$

An application: the circuit example (see page 2). Let's go back to the circuit example discussed in the beginning of the paper. Its dynamics has been described in section 3.2, page 8. We want to check whether the theory correctly predicts that the counterfactual (2) *If switch 1 had been up, the light would have been on.*, i.e $S1 \gg L$ is true in the world where switch one is down, switch two is up and the lamp is off. The first thing we have to do is to calculate $Lit(S1)$. Because $S1$ is itself a literal in the given model, this is trivial: $Lit(S1) = \{\{S1\}\}$. The next step is to calculate the basis B_{w_0} of the evaluation world w_0. Also this is easy given the simple scenario we are working in: $B_{w_0} = \{\neg S1, S2\}$. With these results at hand we can check the truth condition of the conditional sentence: $\forall S \in Lit(S1) \forall B \in Rev_D(B_{w_0}, S) : S1 \models_D L$. Because $Lit(S1)$ contains only one element we have to calculate the revision function only once. $Rev_D(\{\neg S1, S2\}, \{S1\}\} = \{S1, S2\}$. The last step is to calculate whether $\{S1, S2\} \models_D L$. Thus, we have to construct the smallest fixed point for $\mathcal{T}$ applied to $\{S1, S2\}$ and check

whether L is true on this model. The fixed point is, of course, the world w where $S1$, $S2$ and L are all true. We see that the conditional $S1 \gg L$ comes out as true in world w_0, as intended.

5. Conclusions

The approach presented in this paper raises various questions. We can only touch on two of them here. It seems indisputable that the semantics of conditionals exploits certain invariant relationships, certain dependencies. According to the position defended here, the best way to characterize these dependencies is as relations of manipulation and control: a fact A stands in this relation to a fact C, if manipulating A will change C in a systematic way. I have called this type of dependency causal dependency. But one might wonder whether this is the right characterization. Consider the following example.

> **The math class example.** *It is a simple fact of basic math that if you add two natural numbers that are both even or uneven, the sum will be even. If one of the numbers is even and the other uneven, their sum is uneven. Suppose you are explaining this fact to some school kids and you have on the board* $3 + 4 = 7$*. You say ...*

(4) a. If the first number had been even, the result would have been even.
b. If the result had been even, the first number would have been even.

Intuitively, the first counterfactual is true, while the second is not. Thus, even in this case we see that in assessing the truth conditions of these conditionals we assume an asymmetry between the arguments of the operation + and its result. The present proposal would explain this asymmetry as one of manipulation and control: manipulating the arguments of an operation has effects on the result, while manipulating the result will not change the arguments. But in the context of this paper this is called causal dependency.

Another very interesting question for future work is what the present theory says about the relation between causality and counterfactuals. The approach seems to go right contra (Lewis, 1979), because it describes the meaning of conditional sentences based on causal dependencies. However, whether this is true depends on the perspective one takes. It is crucial to distinguish between the content of claims exploring causal relationships and the epistemological issues of how we test and establish causal relationships. The present paper is concerned with the

content-related side of this coin: the content of conditional sentences is determined with reference to causal regularities. The proposal made here is silent on the epistemological issue of how to establish causal relationships.

But what, then, is causality? The paper is silent on this point as well. But let me sketch a direction to go that fits very well with the proposal made here.[10] Causality, as presupposed by the meaning of conditionals, is a heuristics, something we use because it is enormously effective in dealing with reality. But as a heuristics, causality is nothing that can be reduced to something else. Causality is an *a priori* form that we impose on reality to make rational behavior possible.

References

Beth, E.: 1964, *Door wetenschap tot wijsheid. Verzamelde wijsgeerige studien.* Assen, NL: van Gorcum.

Goodman, N.: 1955, *Fact, fiction and forecast.* Indianapolis/New York/Kansas City: The Bobbs-Merrill Company, Inc.

Harper, W.: 1981, 'A sketch of some recent developments in the theory of conditionals'. In: W. Harper et al. (eds.): *IFS. Conditionals, belief, decision, chance, and time.* Dordrecht: Reidel, pp. 3–38.

Kratzer, A.: 1979, 'Conditional necessity and possibility'. In: R. Bäuerle, U. Egli, and A. von Stechow (eds.): *Semantics from different points of view.* Berlin/Heidelberg/New York: Springer, pp. 387–394.

Lewis, D.: 1973, 'Causation'. *Journal of Philosophy* **70**.

Lewis, D.: 1979, 'Counterfactual dependence and time's arrow'. *NOÛS* **13**, 455–476.

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

Schulz, K.: 2007, 'Minimal models in semantics and pragmatics: Free choice, exhaustivity, and conditionals'. Ph.D. thesis, University of Amsterdam, Amsterdam.

Stalnaker, R.: 1981, 'Letter to David Lewis'. In: W. Harper et al. (eds.): *IFS. Conditionals, belief, decision, chance, and time.* Dordrecht: Reidel, pp. 151–152.

van Lambalgen, M. and F. Hamm;: 2005, *The proper treatment of events.* Malden, USA: Blackwell Publishing.

Veltman, F.: 1976, 'Prejudices, presuppositions and the theory of counterfactuals'. In: J. Groenendijk et al. (eds.): *Amsterdam Papers in Formal Grammar*, Vol. 1. Centrale Interfaculteit, Universiteit van Amsterdam.

Veltman, F.: 2005, 'Making counterfactual assumption'. *Journal of Semantics* **22**, 159–180.

[10] This is not a new idea, see, for instance, (Beth, 1964).