Making Counterfactual Assumptions

Frank Veltman
Institute for Logic, Language and Computation
University of Amsterdam

Abstract

This paper provides an update semantics for counterfactual conditionals. It does so by giving a dynamic twist to the 'Premise Semantics' for counterfactuals developed in Veltman (1976) and Kratzer (1981). It also offers an alternative solution to the problems with naive Premise Semantics discussed by Angelika Kratzer in 'Lumps of Thought' (Kratzer, 1989). Such an alternative is called for given the triviality results presented in Kanazawa et al. (2005, this issue).

1 Introduction

Syntactically, counterfactual conditionals are quite complex. First, there is the antecedent starting with 'if' and consisting of a sentence in which a past perfect is used, and then there is the consequent with a verb phrase built from 'would', 'have', and a past participle, a so-called modal perfect, presumably[1] formed by taking the past tense of a future perfect.

In semantics this complexity has been neglected. The usual practice is to put all these modal, temporal, and aspectual modifications together in one single special arrow and to represent a counterfactual by a formula of the form $\ulcorner \varphi \rightsquigarrow \psi \urcorner$, where φ and ψ stand for arbitrary sentences. The meaning of $\ulcorner \varphi \rightsquigarrow \psi \urcorner$ is then defined in one go from the meanings of φ and ψ.

It would be nice to have a more stepwise analysis. This paper makes a start at this by decomposing counterfactuals in two pieces: the antecedent $\ulcorner$*If it had been the case that* $\varphi \urcorner$, and the consequent $\ulcorner$*it would have been the case that* $\psi \urcorner$.[2] Such a decomposition is called for because the modal perfect is not only used in counterfactual conditionals. Consider for instance the second sentence in the following text.

(i) *John did not drink any wine. He would have become sick.*

Sentences with a verb phrase consisting of 'would have' + past participle make no sense if they are presented without context. The first sentence in (i) provides a proper context for the second. Together they convey roughly[3] the same information as the one sentence

(ii) *If John had drunk any wine, he would have become sick.*

[1] Things are changing. Joyce Tang Boyland convincingly argues in Boyland (1996) that in present day English 'have' is becoming affixed to the preceding 'would', which makes 'would have' a single syntactic unit which is combined with a past participle.

[2] Condoravdi (2002) starts at the other end, giving a decompositional analysis of phrases like $\ulcorner$*it might have been the case that*$\ldots\urcorner$ and $\ulcorner$*it would have been the case that* $\ldots\urcorner$

[3] I do not want to get into the question whether it is a pragmatic or a semantic consequence of sentence (ii) that John did not drink any wine. If you believe that counterfactuals presuppose the falsity of their antecedent and that presupposition is a semantic notion, you can omit the qualification "roughly".

It would be interesting to know in exactly which contexts 'would have' + past participle can be used. This is by no means a trivial question, as is illustrated by (iii).

(iii)* *John drank too much wine. He would not have become sick.*

Why does (iii) make no sense? In particular, why do we not understand (iii) as (iv) unless something like 'otherwise' is inserted in front of the second sentence?[4]

(iv) *If John had not drunk too much wine, he would not have become sick.*

The dynamic outlook on semantics[5] offers a way to come to grips with questions like this, as it is designed to study meaning 'in context'. On the dynamic view, knowing the meaning of a sentence is knowing the change it brings about in the cognitive state of anyone who wants to incorporate the information conveyed by it. Formally, this amounts to this.

- The meaning $[\varphi]$ of a sentence φ is an operation on cognitive states.

In the following '$S[\varphi]$' denotes the result of applying the operation $[\varphi]$ to state S; it is the result of *updating* S with φ.

An important notion in this framework is the notion of *support*. A cognitive state S supports a sentence φ when updating S with φ adds no information over and above what is already in S. Instead of 'S supports φ', I will often say 'φ is accepted in S'.

- S supports a sentence φ, $S \models \varphi$, iff $S[\varphi] = S$.

Logical validity is defined in terms of this notion:

- $\varphi_1, \ldots, \varphi_n \models \psi$ iff for any state S, $S[\varphi_1] \ldots [\varphi_n] \models \psi$.

In other words, a sequence of premises $\varphi_1, \ldots, \varphi_n$ entails a conclusion ψ if updating a state with that sequence invariably leads to a state that supports the conclusion.

It is now possible to outline the general setup. Consider a sentence of the form

$\ulcorner$*If it had been the case that* φ*, it would have been the case that* $\psi\urcorner$

By interpreting the antecedent in state S one gets to state

$$S' = S[\textit{If it had been the case that } \varphi].$$

This state S' is stored in memory as a state subordinate to S, so that the consequent can be interpreted in the right context. The modal perfect *'it would have been the case that'* indicates that the message given in ψ pertains to this subordinate state S' rather than to the state S.

This can be generalised. Apparently, after processing the first sentence of (i) the stage is set for the interpretation of the subsequent sentence, whereas in (iii) this is not the case. When you have to interpret a negative sentence, such as the first sentence in (i), the interpretation process starts with an update with the positive subsentence, and then continues with some operation on this intermediate result. This intermediate result is kept in memory as an auxiliary state subordinate to the

[4]For readers who think that it is crucial that there be a negation in the first sentence, it will be worthwhile to look at

(v) *We asked Mary to taste the wine. John would have become sick.*

If 'Mary' is stressed in the first sentence and 'John' in the second, the second sentence makes perfect sense.

[5]This paper utilizes the framework presented in Veltman (1996).

main state, and differing from it mainly in that it supports a statement that is rejected in the main state. This subordinate state is the state the modal perfect in the second sentence of (i) is looking for. In interpreting the first sentence of (iii) no such subordinate state is created. Therefore the second sentence of (iii) finds nothing to pertain to. But if the phrase 'otherwise' is inserted in front of the second sentence, a subordinate state will be created and the interpretation of the second sentence runs smoothly.[6][7]

In this paper my main concern is not the interpretation of the consequent of a counterfactual, but the interpretation of the antecedent. What is it to make a counterfactual assumption? Given a state S and a sentence φ, what does

$$S[\textit{If it had been the case that } \varphi]$$

look like? In the next section I will discuss some problems with the standard answer to this question. In section 3 and 4 I will develop an alternative and show that it solves the problems discussed in section 2. In section 5 I will discuss the repercussions of the resulting theory of counterfactual conditionals for the treatment of indicative conditionals. Finally, section 6 is devoted to a comparison of the theory proposed here with its nearest neighbour, the theory proposed in Kratzer (1989).

2 Ramsey's test and Tichy's puzzle

The starting point for our analysis is the informal recipe for evaluating counterfactual conditionals named after Frank Ramsey. It plays a fundamental role in many theories of counterfactuals, notably in the theories presenting some variant of Premise Semantics (Rescher (1964), Veltman (1976), Kratzer (1981)).

> *Ramsey test:* 'This is how to evaluate a conditional: first, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true.'

The quotation is taken from Stalnaker (1968, p.106). Ramsey's original suggestion only covered the case in which the antecedent is consistent with 'your' stock of beliefs. In that case no adjustments are required. In the above, Stalnaker generalizes this to the case in which the antecedent cannot be added to 'your' stock of beliefs without introducing a contradiction. In this case, which is typical of counterfactuals, adjustments are required.

The Ramsey test is in need of amendments. Making a counterfactual assumption does not boil down to a minimal belief revision, as is illustrated by the counterexample devised by Pavel Tichy:

> 'Consider a man, call him Jones, who is possessed of the following dispositions as regards wearing his hat. Bad weather invariably induces him to wear a hat. Fine weather, on the other hand, affects him neither

[6] To deal with the example of footnote 4 one needs to invoke a theory of focus like the one presented in Rooth (1985). According to this theory the general function of focus is evoking alternatives. (In this case, alternatives to Mary — which other people could we have asked to taste the wine?). These alternatives will give rise to subordinate states and are available for sentences in which the modal perfect is used to be interpreted in. (Apparently, we are ready to accommodate the idea that John is one of the alternatives.)

[7] Stefan Kaufmann (Kaufmann, 2000) uses stacks of states in a formalism to describe discourse phenomena like the one discussed here. See also van Rooy (2005, this issue) for a discussion of modal subordination.

> way: on fine days he puts his hat on or leaves it on the peg, completely at random. Suppose moreover that actually the weather is bad, so Jones *is* wearing his hat.' Tichy (1976, p. 271)

The question is: would you accept the sentence *'If the weather had been fine, Jones would have been wearing his hat'*?[8]

Presumably, your answer is 'no', but Ramsey's recipe yields 'yes'. We know (i) that Jones is wearing his hat, we know (ii) that it is raining. Now we must add the proposition (iii) that the weather is fine to this, thereby making the adjustments *minimally* required to maintain consistency. Clearly, this can be done without Jones having to take his hat off.

Tichy's criticism was not directed directly against the Ramsey test but against the analysis of counterfactuals developed by Robert Stalnaker and David Lewis. (Stalnaker (1968), Lewis (1973)). They proposed the following truth condition for counterfactual conditionals.

> A sentence of the form $\ulcorner$*If it had been the case that* φ*, it would have been the case that* $\psi\urcorner$ is true in the actual world w iff the consequent ψ is true in every world[9] in which the antecedent φ is true, and which in other respects differs minimally from w.

Tichy claims that the counterfactual *'If the weather had been fine, Jones would have been wearing his hat'*, asserted in the context described above, meets this condition. In the actual world, it is raining and Jones is wearing is hat. Given that it is a matter of chance whether or not Jones wears his hat when the weather is fine, it would seem that for any sunny world in which Jones is not wearing his hat there is an equally sunny world in which he does, and which – because of this – is less different from the actual world.

Lewis and Stalnaker are ready to admit that Tichy's example shows that the relevant conception of minimal difference 'needs to be spelled out with care' (Stalnaker (1984, p. 129)), but they do not think the example shows that the idea of minimal difference is wrong. Perhaps such contingencies like whether or not Jones is wearing his hat, do not matter when the differences and similarities of possible worlds have to be assessed. This is at least what Lewis suggests in Lewis (1979), where he formulates a system of weights that governs the notion of similarity involved. After some remarks on the important role of 'general' laws in this matter,[10] he says the following about the role of 'particular' fact.

> 'It is of little or no importance to secure approximate similarity of particular fact.' Lewis (1979, p. 472)

Here is a variant[11] of Tichy's puzzle which shows that this is not quite right.

> Suppose that Jones always flips a coin before he opens the curtains to see what the weather is like. Heads means he is going to wear his hat in case the weather is fine, whereas tails means he is not going to wear

[8] If you like the sentence better if there is a '*still*' between 'would' and 'have' in the consequent, then please read it that way.

[9] According to Stalnaker there is at most one such world, according to Lewis there may be more than one.

[10] As the first and the third criterion he mentions the following: *It is of the first importance to avoid big, widespread, diverse violations of law.... It is of the third importance to avoid even small, localized, simple violations of law.*

[11] The example was suggested to me years ago by my former student Frank Mulkens.

his hat in that case. Like above, bad weather invariably makes him wear his hat. Now suppose that today heads came up when he flipped the coin, and that it is raining. So, again, Jones is wearing his hat.

And again, the question is whether you would accept the sentence *'If the weather had been fine, Jones would have been wearing his hat'*. This time, your answer will be 'yes'. Lewis, too, would want to say 'yes', I guess. But can he? If similarity of particular fact did not matter in the first version of the puzzle, why would it now?

What really matters is this: In both cases Jones is wearing his hat *because* the weather is bad. In both cases we have to give up the proposition that the weather is bad — the very *reason* why Jones is wearing his hat. So, why should we want to keep assuming that he has his hat on? In the first case there is no special reason to do so; hence, we do not. In the second case there is a special reason. We will keep assuming that Jones is wearing his hat because we do not want to give up the independent information that the coin came down heads. And this, together with the counterfactual assumption that the weather is fine, brings in its train that Jones would have been wearing his hat.

In other words, similarity of particular fact is important, but only for facts that do not depend on other facts. Facts stand and fall together.[12] In making a counterfactual assumption, we are prepared to give up everything that depends on something that we must give up to maintain consistency. But we want to keep in as many independent facts as we can. In the next section I will develop this idea more precisely.

3 States, and what assumptions do to them.

In what follows I will assume that the reader is acquainted with the basic apparatus of possible worlds semantics.

Definition 1 (Worlds and states) Fix a finite set $\mathcal{A}$ of atomic sentences.

(i) A *world* is a function with domain $\mathcal{A}$ and range $\{0,1\}$; a *situation* is a partial such function; a *proposition* is a set of worlds.

(ii) Let W be the set of possible worlds. A *cognitive state* S is a pair $\langle U_S, F_S\rangle$, where either (a) $\emptyset \neq F_S \subseteq U_S \subseteq W$; or (b) $F_S = U_S = \emptyset$.

In this definition a possible world is identified with the valuation that assigns the value 1 to the atomic sentences true in it, and 0 to the atomic sentences false in it. I am using 'p', 'q', and 'r' to refer to atomic sentences. I will often write '$\langle p,1\rangle \in w$' rather than '$w(p) = 1$', and use a similar notation when situations are at stake. So, '$\langle q,0\rangle \in s$' means that the atom q is false in the situation s. Pairs like $\langle p,1\rangle$ and $\langle q,0\rangle$ will sometimes be referred to as (positive and negative) *facts* constituting the situations they are elements of.

I will write '$[\![\varphi]\!]$' for the proposition expressed by φ, and assume the reader is acquainted with the fact that for formulas of propositional logic by definition the following holds:

$$\begin{array}{rcl} [\![p]\!] & = & \{w \in W \mid w(p) = 1\}, \\ [\![\neg\varphi]\!] & = & W \sim [\![\varphi]\!], \\ [\![\varphi \wedge \psi]\!] & = & [\![\varphi]\!] \cap [\![\psi]\!], \\ [\![\varphi \vee \psi]\!] & = & [\![\varphi]\!] \cup [\![\psi]\!], \\ [\![\varphi \rightarrow \psi]\!] & = & (W \sim [\![\varphi]\!]) \cup [\![\psi]\!]. \end{array}$$

[12] This is also the idea behind Kratzer's lumping semantics in Kratzer (1989).

An agent's cognitive state S is given with two sets of possible worlds, F_S and U_S, the former a subset of the latter. A world w is supposed to be an element of F_S if, for all the agent in state S knows, w might be the actual world. The set U_S is called the *universe* of the state S. A possible world belongs to U_S if all the propositions that an agent in state S considers to be *general laws* hold in it.

It has often[13] been noted that general laws play a special role in the interpretation of counterfactuals. Consider:

If John's boat had been made of wood, it would not have sunk.

Imagine that John's boat, an iron rowing boat, has sunk. It has been raining a lot lately, and John forgot to bail the water out. Probably, in this context you would prefer the above counterfactual to the next one.

If John's boat had been made of wood, it would (still) have sunk.

In making a counterfactual assumption we are not prepared to give up propositions we consider to be general laws. We will stick to a law of nature like *Wood floats on water*, at the cost of a contingent fact like *John's boat sank*.

It's not just *natural* laws that are at stake here. Take for instance the proposition that bad weather invariably induces Jones to wear a hat, and think about the role this proposition plays in the scenarios sketched above. It is not a law of nature, of course, but it's a law. We will not give it up when making counterfactual assumptions. When the weather is fine, we will assume that if the weather had been bad, Jones would have been wearing his hat, even if we have just seen him without it. Or take conventional laws like the rules of chess. In evaluating a statement like

If White had played 14.Nd5, Black would have lost.

we are not prepared to consider worlds where chess is not played by the rules, or played by different rules.

In any state S, $F_S \subseteq U_S$: the general laws set a limit to the factual information one can have. If the agent has no specific information about the actual world, then $F_S = U_S$: any world in which the general laws hold might be the actual world. In the *minimal state*, given by $\mathbf{1} = \langle W, W\rangle$, the agent neither has any factual information, nor is he acquainted with any law.

For a state to be *coherent* it is required that $F_S \neq \emptyset$. A state S in which $F_S = \emptyset$ is *absurd*: given the available information, there is no possible world left that might be the real one. In the mathematical setup we identify all absurd states and allow only one: $\langle\emptyset, \emptyset\rangle$, also known as $\mathbf{0}$, and as *the* absurd state. Agents will avoid getting into this state.

In our formal language $\Box$ will be used to express *'it is a law that...'*. Here, the dots have to be filled by a formula of propositional logic, $\Box$ can only occur as the outermost operator of a formula. Part (ii) of the next definition explains what an update with $\Box\varphi$ amounts to.

Definition 2 (Interpretation)
Let φ be a formula of propositional logic.

(i) (a) $S[\varphi] = \langle U_S, F_S \cap [\![\varphi]\!]\rangle$ if $F_S \cap [\![\varphi]\!] \neq \emptyset$;
(b) $S[\varphi] = \mathbf{0}$, otherwise.

(ii) (a) $S[\Box\varphi] = \langle U_S \cap [\![\varphi]\!], F_S \cap [\![\varphi]\!]\rangle$ if $F_S \cap [\![\varphi]\!] \neq \emptyset$;
(b) $S[\Box\varphi] = \mathbf{0}$, otherwise.

[13] A prominent example is John Pollock, who has stressed the point in all his writings on counterfactuals since Pollock (1976).

Updating with a propositional formula φ eliminates from F_S all possible worlds in which φ is false. Hence, only worlds in which φ is true are left as worlds that might be the actual world. If there are no such worlds left, one gets into the absurd state. Similarly, an update with $\Box\varphi$ eliminates from U_S all worlds in which φ is false. So, only worlds in which φ is true are left as worlds that might have been the actual world. The other ones are so outlandish, you do not have to reckon with them, not even in making the wildest counterfactual assumption.[14]

Below on the left a table is drawn representing the minimal state for a language with three atoms. Every row represents a world. The table in the middle represents the state that results when the minimal state is updated with $\neg q$, and on the right you see $\mathbf{1}[\neg q][\Box(r \to p)]$. For a given state S, worlds belonging to F_S are printed in boldface, and worlds that do not belong to U_S are struck through.

	$p\ q\ r$
w_0	0 0 0
w_1	0 0 1
w_2	0 1 0
w_3	0 1 1
w_4	1 0 0
w_5	1 0 1
w_6	1 1 0
w_7	1 1 1

$\xrightarrow{\neg q}$

	$p\ q\ r$
$\boldsymbol{w_0}$	**0 0 0**
$\boldsymbol{w_1}$	**0 0 1**
w_2	0 1 0
w_3	0 1 1
$\boldsymbol{w_4}$	**1 0 0**
$\boldsymbol{w_5}$	**1 0 1**
w_6	1 1 0
w_7	1 1 1

$\xrightarrow{\Box(r \to p)}$

	$p\ q\ r$
$\boldsymbol{w_0}$	**0 0 0**
~~w_1~~	~~0 0 1~~
w_2	0 1 0
~~w_3~~	~~0 1 1~~
$\boldsymbol{w_4}$	**1 0 0**
$\boldsymbol{w_5}$	**1 0 1**
w_6	1 1 0
w_7	1 1 1

FIGURE 1

Definition 3 (Basis) Let $S = \langle U_S, F_S\rangle$ be a state.

(i) The situation s *forces* the proposition P *within* U_S iff for every $w \in U_S$ such that $s \subseteq w$ it holds that $w \in P$.[15]

(ii) The situation s *determines* the world w iff s forces $\{w\}$ within U_S.

(iii) The situation s *is a basis for* the world w iff s is a *minimal* situation determining w within U_S.

A basis for a world $w \in U_S$ is a part of w consisting of mutually independent facts which, given the general laws, bring the other facts constituting w in their train. It is easy to check that every world in the universe of the state pictured on the right above has exactly one basis. (For instance, the one basis for w_0 is $\{\langle p, 0\rangle, \langle q, 0\rangle\}$). However, generally speaking, it may very well be that a world has more than one basis.

Making a counterfactual assumption $\ulcorner$*If it had been the case that* $\varphi\urcorner$ in state S takes two steps. In the first step any information to the effect that φ is in fact false is withdrawn from S, and in the second step the result is updated with the assumption that the antecedent φ is true.

Definition 4 (ii) describes the first step, and definition 4 (iii) the second.

Definition 4 (Retraction) Let $S = \langle U_S, F_S\rangle$ be a state.

(i) Suppose $w \in U_S$, and $P \subseteq W$. The set $w{\downarrow}P$ is determined as follows:

$s \in w{\downarrow}P$ iff $s \subseteq w$ and there is a basis s' for w such that s is a *maximal* subset of s' not forcing P.

[14] Note that this definition would not work if we had allowed stacking of $\Box$'s etc.

[15] If there is no world $w \in U_S$ such that $s \subseteq w$, then, according to this definition, the situation s forces every proposition.

(ii) $S \downarrow P$, the *retraction* of P *from* S, is the state $\langle U_{S\downarrow P}, F_{S\downarrow P}\rangle$ determined as follows:

(a) $w \in U_{S\downarrow P}$ iff $w \in U_S$;

(b) $w \in F_{S\downarrow P}$ iff $w \in U_S$ and there are $w' \in F_S$ and $s \in w'\downarrow P$ such that $s \subseteq w$.

(iii) The state $S[\textit{if it had been the case that } \varphi]$ is given by $(S\downarrow[\![\neg\varphi]\!])[\varphi]$

A counterfactual $\ulcorner$*If it had been the case that* $\varphi, \ldots\urcorner$ is usually asserted in a context in which φ is known to be false. Let's concentrate on such contexts, and see what $S[\textit{if it had been the case that } \varphi]$ amounts to.[16] According to definition 4 (iii) we have to retract $[\![\neg\varphi]\!]$ from S first and then update the result $S \downarrow [\![\neg\varphi]\!]$ with φ. Definition 4 (ii) and 4 (i) add that to retract $[\![\neg\varphi]\!]$ from S the following has to be done for every world w in F_S, and every basis s' for w: Given that the basis s' forces the proposition $[\![\neg\varphi]\!]$, make *minimal* adjustments to s' to the effect that $[\![\neg\varphi]\!]$ is no longer forced. It is very well possible that there are various ways to do so. Let s be one of the results. The worlds in U_S extending s all belong to $S\downarrow[\![\neg\varphi]\!]$.

Readers acquainted with Premise Semantics will have recognized the melody. Just like in other versions of Premise Semantics we are interested in the maximal subsets of the 'premise set' that are consistent with the antecedent of the counterfactual. However, in this version the premise set for a world w is not given by the set of propositions that hold in w, as naive Premise Semantics would have it. In this version a premise set is given by a set of facts constituting a basis of the world w. And now, by 'consistent' we don't just mean 'logically consistent', but 'compatible with the general laws'.

The crucial trick is that actual retraction takes place at the level of the bases of the worlds. This is because we want to keep in as many independent facts as we can, but don't bother about facts that depend on other facts. This way we ensure that when a particular fact is retracted all the facts it takes in its train are retracted with it.

To give a formal analysis of the Tichy cases, we do not need an exact definition of what an update with a counterfactual amounts to. All we need to agree upon is that this definition must satisfy the following constraint:

$$S \models \textit{if had been } \varphi, \textit{would have been } \psi \text{ iff } S[\textit{if had been } \varphi] \models \psi$$

In other words, a state S supports a counterfactual conditional $\ulcorner$*if had been* φ, *would have been* $\psi\urcorner$ iff the subordinate state $S[\textit{if had been } \varphi]$ supports the consequent ψ.

(The reader will have noticed that I changed notation. For reasons of economy, I will henceforth write '*if had been*', and '*would have been*' rather than *'If it had been the case that'*, and *'it would have been the case that'*).

Tichy 1

Let p be short for *'The weather is bad'*, and q for *'Jones is wearing his hat'*. We are interested in the state $S = \mathbf{1}[\Box(p \rightarrow q)][p][q]$, which is pictured in the left table below.

[16] Again, in this paper I do not want to get into a discussion of the question whether counterfactuals presuppose the falsity of their antecedent. I am perfectly happy if this definition only works for cases in which both the speaker and the hearer believe that the antecedent is false. Maybe amendments are in order for the other cases.

	$p\ q\ r$
w_0	0 0 0
w_1	0 0 1
w_2	0 1 0
w_3	0 1 1
~~w_4~~	~~1/0/0~~
~~w_5~~	~~1/0/1~~
$\boldsymbol{w_6}$	**1 1 0**
$\boldsymbol{w_7}$	**1 1 1**

$\xrightarrow{\textit{if had been } \neg p}$

	$p\ q\ r$
$\boldsymbol{w_0}$	**0 0 0**
$\boldsymbol{w_1}$	**0 0 1**
$\boldsymbol{w_2}$	**0 1 0**
$\boldsymbol{w_3}$	**0 1 1**
~~w_4~~	~~1/0/0~~
~~w_5~~	~~1/0/1~~
w_6	1 1 0
w_7	1 1 1

FIGURE 2

World w_6 has one basis, $\{\langle p,1\rangle,\langle r,0\rangle\}$. The one basis for w_7 is $\{\langle p,1\rangle,\langle r,1\rangle\}$. Applying definition 4, we see that $w_6{\downarrow}[\![p]\!] = \{\{\langle r,0\rangle\}\}$, and $w_7{\downarrow}[\![p]\!] = \{\{\langle r,1\rangle\}\}$. This means that $S\downarrow[\![p]\!] = \langle U_S, U_S\rangle$. Hence, the state $S[\textit{if had been } \neg p]$ is the state given by the right table above.

Clearly, $S[\textit{if had been } \neg p] \not\models q$. Therefore, $S \not\models \textit{if had been } \neg p, \textit{would have been } q$;. in other words, the theory says that an agent in state S should not accept the sentence *'If the weather had been fine, Jones would have been wearing his hat'*.

Tichy 2

Turning to the variant to Tichy's example, again, let p be short for *'The weather is bad'*, and q for *'Jones is wearing his hat'*. The atom r stands for *'The coin comes up heads'*. This time we are interested in the state $S = \mathbf{1}[\Box((p \vee r) \leftrightarrow q)][r][p]$, which is given by the left table below.

	$p\ q\ r$
w_0	0 0 0
~~w_1~~	~~0/0/1~~
~~w_2~~	~~0/1/0~~
w_3	0 1 1
~~w_4~~	~~1/0/0~~
~~w_5~~	~~1/0/1~~
w_6	1 1 0
$\boldsymbol{w_7}$	**1 1 1**

$\xrightarrow{\textit{if had been } \neg p}$

	$p\ q\ r$
w_0	0 0 0
~~w_1~~	~~0/0/1~~
~~w_2~~	~~0/1/0~~
$\boldsymbol{w_3}$	**0 1 1**
~~w_4~~	~~1/0/0~~
~~w_5~~	~~1/0/1~~
w_6	1 1 0
w_7	1 1 1

FIGURE 3

The (only) basis for $w_7 = \{\langle p,1\rangle,\langle r,1\rangle\}$. Furthermore, $w_7{\downarrow}[\![p]\!] = \{\{\langle r,1\rangle\}\}$, which means that $S{\downarrow}[\![p]\!] = \langle U_S, \{w_3, w_7\}\rangle$. Hence, the subordinate state $S[\textit{if had been } \neg p]$ is the state pictured by the table on the right above. Notice that $S[\textit{if had been} \neg p] \models q$. Therefore, $S \models \textit{if had been } \neg p, \textit{would have been } q$. In other words, the theory says that we were right when we accepted *'If the weather had been fine, Jones would have been wearing his hat'*.

4 Counterfactuals as tests

The next update condition for counterfactuals is the simplest condition in line with the constraint we formulated above:

Definition 5 (Counterfactuals as tests)

$S[\textit{if had been } \varphi, \textit{would have been } \psi] = S$, if $S[\textit{if had been } \varphi] \models \psi$

$S[\textit{if had been } \varphi, \textit{would have been } \psi] = \mathbf{0}$, otherwise.

Given this definition, sentences of the form ⌜*if had been* φ, *would have been* ψ⌝ do not convey new information — not directly at least. They provide an invitation to perform a test. By asserting ⌜*if had been* φ, *would have been* ψ⌝, a speaker makes a kind of comment: 'Given the general laws and the facts I am acquainted with, the sentence ψ is supported by the state I get in when I assume that φ had been the case'. The addressee is supposed to determine whether the same holds on account of his or her own information. If not, a discussion will arise, and in the course of this discussion both the speaker and the hearer may learn some new laws and facts, which could affect the outcome of the test.

Such things sometimes really happen. Consider the Tichy case once more. Imagine that someone with the information of the variant of Tichy's example says *'If the weather had been fine, Jones would have been wearing his hat'* to someone who only has the information available in the original example. The addressee will not accept the statement. But then, when he or she hears about the coin etc., this will change. So, ultimately the addressee gets some new information about the actual world, but in a very indirect way.

The reason why we cannot give a direct update rule for counterfactuals that works in all cases, is that it is not always clear which part of the new information is due to some hitherto unknown laws and which part to some hitherto unknown facts. More formally, in many cases it is not uniquely determined which worlds should be removed from the universe U_S and which worlds from F_S. It could be that you should accept ⌜*if had been* φ, *would have been* ψ⌝ because it is a general law that whenever φ is the case, ψ is the case as well, or it could be that you should accept it because it happens to be the case that χ, and it is a general law that you cannot have φ and χ without having ψ. And these are just two possibilities.

If you don't know on beforehand which laws are involved, there are various ways to decompose the new information. However, in a context where the laws are fixed – like when we are discussing a chess game, or when we are solving problems in classical mechanics, we can give a direct update rule. The key to this update rule is supplied by the next proposition.

Proposition
Let S be a state.

$F_{S[\textit{if had been}\,\varphi]} = \{w \in U_S \mid w \in F_{\langle U_S,\{v\}\rangle[\textit{if had been}\,\varphi]}$ for some $v \in F_S\}$.

This proposition says that the operation of making a counterfactual assumption is *distributive*: we can think of $F_{S[\textit{if had been}\,\varphi]}$ as the result of taking the union, for all $v \in F_S$, of all the sets $F_{\langle U_S,\{v\}\rangle[\textit{if had been}\,\varphi]}$.

Call $w \in F_{\langle U_S,\{v\}\rangle[\textit{if had been}\,\varphi]}$ a φ-*alternative* to v. Using this terminology, we can reformulate the proposition and say that $F_{S[\textit{if had been}\,\varphi]}$ consists of the φ-alternatives of all the worlds in F_S.

Notice that a state S supports the sentence ⌜*if had been* φ, *would have been* ψ⌝, iff the consequent ψ holds in all φ-alternatives of all worlds in F_S. And if a state S does not support ⌜*if had been* φ, *would have been* ψ⌝, then we can turn it into one that does by removing from F_S all the worlds v that have some φ-alternative in which ψ does not hold.

Thus we arrive at the following update clause.

Definition 6
(a) If there is some $v \in F_S$ such that $\langle U, \{v\}\rangle[\textit{if had been}\,\varphi] \models \psi$, then
$S[\textit{if had been}\,\varphi,\ \textit{would have been}\,\psi] = \langle U_S, \{v \in F_S \mid \langle U, \{v\}\rangle[\textit{if had been}\,\varphi] \models \psi\}\rangle$.

(b) Otherwise, $S[\textit{if had been}\,\varphi,\ \textit{would have been}\,\psi] = \mathbf{0}$.

As I already noted, this clause only works in cases in which no new laws can arise. Only when the universe is fixed do counterfactuals express a fixed proposition:

$[\![\textit{if had been}\,\varphi,\ \textit{would have been}\,\psi]\!] = \{w \in W \mid \langle U, \{w\}\rangle[\textit{if had been}\,\varphi] \models \psi\}$

and only in those cases can we think of an update with a counterfactual as a propositional update:

$S[\textit{if had been}\,\varphi,\ \textit{would have been}\,\psi] = \langle U_S, F_S \cap [\![\textit{if had been}\,\varphi,\ \textit{would have been}\,\psi]\!]\rangle$

When the universe can change, counterfactuals get rather capricious. In that case they are not even persistent.

Definition 7 Let S and S' be states.

(i) *S is at least as strong as S'* iff $U_S \subseteq U_{S'}$ and $F_S \subseteq F_{S'}$.

(ii) A sentence φ *persistent* iff the following holds:
If S is at least as strong as S', and $S' \models \varphi$, then $S \models \varphi$.

If S is stronger than S', you know more laws and/or more facts in S than you know in S'. However, this does not necessarily mean that in S you will accept every sentence you accept in S'. Of course, intuitively this should hold for sentences that describe the facts, or that exemplify laws, but not all sentences do so. Well-known examples of sentences that are not persistent are sentences in which the epistemic modality *might* occurs. With $\ulcorner$*It might be the case that* $\varphi\urcorner$ a speaker expresses that φ is consistent with the information available. Obviously, as more information gets available this consistency might get lost. (See Veltman (1996) for details.)

The question is: are counterfactuals persistent? Here is a counterexample:

$\mathbf{1}[p][q] \models \textit{if had been}\,\neg p, \textit{would have been}\ q$, but
$\mathbf{1}[p][q][\Box(\neg p \rightarrow \neg q)] \not\models \textit{if had been}\,\neg p, \textit{would have been}\ q$.

It is crucial that a law is learnt here.

5 Are counterfactuals ambiguous?

The test condition for counterfactuals provided above fits in nicely with the theory of indicative conditionals proposed in Gillies (2004). Stated in our format, Gillies suggests the following:

$S[\textit{if}\,\varphi, \psi] = S$, if $S[\varphi] \models \psi$
$S[\textit{if}\,\varphi, \psi] = \mathbf{0}$, otherwise.

Some philosophers advocate a unified account of indicative and subjunctive conditionals. They believe that the only difference between indicatives and counterfactuals is that each is used in different circumstances. Indicatives are typically used in circumstances in which the speaker is ignorant about the truth value of the antecedent and counterfactuals in circumstances in which the agent thinks that the antecedent is false, but both express the same 'connection' between the antecedent and the consequent.

It would seem that anyone subscribing to this position is committed to the following.

> An agent who is ignorant about the truth value of φ, but entitled to entertain the indicative conditional $\ulcorner$*If* $\varphi, \psi\urcorner$, will later, after learning that φ is in fact false, be entitled to entertain the counterfactual $\ulcorner$*If it had been the case that* φ, *it would have been the case that* $\psi\urcorner$.

At first sight, this looks quite plausible, but it is false, as the next scenario shows.

The duchess has been murdered, and you are supposed to find the murderer. At some point only the butler and the gardener are left as suspects. At this point you believe

(i) *If the butler did not kill her, the gardener did.*

Still, somewhat later — after you found convincing evidence showing that the butler did it, and that the gardener had nothing to do with it — you get in a state in which you will *reject* the sentence

(ii) *If the butler had not killed her, the gardener would have.*

Actually, quite a few people believe that counterfactuals have two readings, an 'epistemic' reading and a 'ontic' one. They will maintain that on the epistemic reading sentence (ii) is true. In the epistemic case implicit reference is made to some previous epistemic state, in this example the state you were in when only two suspects were left. Thinking back, one can say that if it had not been the butler, it would have to have been the gardener.

Notice that only people who have gone through the same epistemic process as you did in your role of detective, will be able to appreciate this epistemic reading. People who have never been in a state in which the butler and the gardener were the only suspects left, and who just wonder which course history would have taken if the butler had not killed the duchess, will rightly think that that in that case the duchess might still have been alive. So, on this second, 'ontic' reading the sentence is plainly false.

I myself doubt that (ii), or any other counterfactual for that matter, has an epistemic reading. There are other means to express what the epistemic reading is supposed to express. In any case, the theory proposed here only covers the 'ontic' reading as the next formal picture shows.

Set p:= *'The butler killed the duchess'*, q:= *'The gardener killed the duchess'*, and r:= *'The duchess was killed'*. Consider first the state $S = \mathbf{1}[\Box((p \vee q) \to r)][r][p \vee q]$, which comprises what you know after you found out that it must have been the butler or the gardener.

	$p\ q\ r$
w_0	0 0 0
w_1	0 0 1
~~w_2~~	~~0 1 0~~
$\boldsymbol{w_3}$	**0 1 1**
~~w_4~~	~~1 0 0~~
$\boldsymbol{w_5}$	**1 0 1**
~~w_6~~	~~1 1 0~~
$\boldsymbol{w_7}$	**1 1 1**

FIGURE 5

Notice that $S[\neg p] \models q$. Gillies' theory says, as any decent theory of indicative conditionals would do, that under these circumstances the state S supports $\mathit{If}\ \neg p,\ q$.

Next, consider the state $S' = \mathbf{1}[\Box((p \vee q) \to r)][r][p \vee q][p][\neg q]$, pictured below on the left hand side. The (only) basis for w_5 is given by $\{\langle p, 1\rangle\, \langle q, 0\rangle\}$. The state $S'{\downarrow}[\![p]\!] = \langle U_S, \{w_0, w_1, w_5\}\rangle$. The state $S'[\,\mathit{if had been}\ \neg p]$ is the state pictured below on the right hand side.

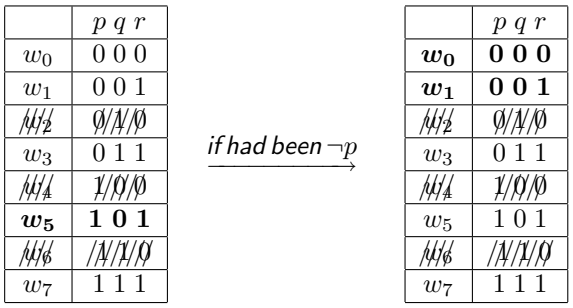

FIGURE 6

S'[*if it had been that* $\neg p$] $\not\models q$. In other words, $S \not\models$ *if had been* $\neg p$, *would have been* q.

The above illustrates that there is a huge difference between making a counterfactual assumption and revising one's beliefs. When you believe that φ is true and you imagine that φ had been false, you have to change your cognitive state, but it is it not the kind of change you would have to make if you were to discover that φ is *in fact* false. It is not a *correction*. Notice that $w_0 \in S'$[*if it had been that* $\neg p$], which means so much as that if the butler had not killed her, the duchess might have been still alive. However, if at some point you were to discover that your belief that the butler did it is in fact wrong, you would not automatically give up your belief that the duchess was killed. It is likely that you would reopen the investigations.[17]

6 A problematic case

The theory presented in this paper offers an alternative solution to the problems dealt with by Angelika Kratzer in Kratzer (1989). Such an alternative is called for given the defects in the formal set up of lumping semantics discussed in Kanazawa et al. (2005, this issue). I have tried to remedy these defects, and I have tried to do so keeping in as many informal ideas behind the lumping set up as I could. The result is a modification of naive Premise Semantics, just like Kratzer's theory, and just like hers it is a theory that makes concrete predictions in concrete cases.

The predictions are in many cases the same. For example, given the account above, there is no reason why the sentence

If a different animal had escaped from the zoo, it would have been a zebra.

should be accepted by an agent with the following information:

> Last year, a zebra escaped from the Hamburg zoo. The escape was made possible by a forgetful keeper who forgot to close the door of a compound containing zebras, giraffes, and gazelles. A zebra felt like escaping and took off. The other animals preferred to stay in captivity. Kratzer (1989, p. 625)

This example poses a problem for an account which just follows Ramsey's recipe. After all, it is possible to accommodate the counterfactual assumption that a different animal escaped from the zoo without giving up the idea that it was a zebra.

The reason why this example poses no problem for the theory presented here is because the information that a zebra escaped is not represented as an independent fact in the bases of the worlds that constitute the state S supporting this information. Every basis of every world in F_S will contain some object that is a zebra (fact

[17] See Rott (1999) for an insightful discussion of these points.

1) and that escaped (fact 2). This is all that is needed to enforce the proposition that a zebra escaped. In accommodating the assumption that a different animal escaped, in every basis of every world fact 2 will have to be replaced by a fact consisting of a *different* object that escaped. There is no reason why this object should be a zebra. Any other kind of animal will do.[18]

As far as I can see there is just one case where our theory does not give the outcome wanted by Kratzer. Here is the story.

> King Ludwig of Bavaria likes to spend his weekends at Leoni Castle. Whenever the Royal Bavarian flag is up and the lights are on, the King is in the Castle. At the moment the lights are on, the flag is down, and the King is away. Suppose now counterfactually that the flag were up. Well, then the King would be in the castle and the lights would still be on. But why wouldn't the lights be out and the King still be away? Kratzer (1989, p. 640)

Kratzer needs all the lumping machinery to exclude the latter possibility. The theory presented here cannot exclude it.

Here is a formal sketch of the situation: Let p be short for *'The flag is up'*, q for *'The lights are on'*, and r for *'The king is out'*. Consider the state $S = \mathbf{1}[\Box((p \wedge q) \rightarrow \neg r)][\neg p][q][r]$[19], given by the left table below.

	$p\ q\ r$
w_0	0 0 0
w_1	0 0 1
w_2	0 1 0
$\boldsymbol{w_3}$	**0 1 1**
w_4	1 0 0
w_5	1 0 1
w_6	1 1 0
~~w_7~~	~~1 1 1~~

$\xrightarrow{\textit{if had been } p}$

	$p\ q\ r$
w_0	0 0 0
w_1	0 0 1
w_2	0 1 0
w_3	0 1 1
w_4	1 0 0
$\boldsymbol{w_5}$	**1 0 1**
$\boldsymbol{w_6}$	**1 1 0**
~~w_7~~	~~1 1 1~~

FIGURE 7

The world w_3 has one basis, $\{\langle q, 1\rangle, \langle r, 1\rangle\}$, and to accommodate the counterfactual assumption p, one could give up either $\langle q, 1\rangle$ or $\langle r, 1\rangle$. Hence, $S[\textit{if had been } p]$ is the state pictured on the right. Notice that $S[\textit{if had been } p] \not\models \neg r$. Therefore, $S \not\models \textit{if had been } p, \textit{would have been } \neg r$. According to the theory presented here, it is not the case that if the flag were up, the King would be in. Indeed, the lights might be out and the King might still be away.

For those who share Kratzer's intuitions, this will be a drawback. However, there are examples with the same logical structure as the one above for which one wouldn't want that $S \models \textit{if had been } p, \textit{would have been } \neg r$.

> Consider the case of three sisters who own just one bed, large enough for two of them but too small for all three. Every night at least one of them has to sleep on the floor. Whenever Ann sleeps in the bed and Billie sleeps in the bed, Carol sleeps on the floor. At the moment Billie is sleeping in bed, Ann is sleeping on the floor, and Carol is sleeping in bed. Suppose now counterfactually that Ann had been in bed...

[18] This can only be made more precise in a predicate logic version of the theory presented here.
[19] Nothing much changes if one strengthens the law to $\Box((p \wedge q) \leftrightarrow \neg r)$.

I am pretty sure that this time you are not prepared to say: 'Well, in that case Carol would be sleeping on the floor'. Indeed, why wouldn't Billie be on the floor?

Still, this example has the same logical structure as the King Ludwig example. Let p stand for *'Ann sleeps in the bed'* , q for *'Billie sleeps in the bed'*, and r for *'Carol sleeps in the bed'*. The question we are interested in, is whether $\mathbf{1}[\Box((p \wedge q) \rightarrow \neg r)][\neg p][q][r] \models$ *if had been* p, *would have been* $\neg r$. We saw already that the answer is 'no'.

If Kratzer's intuitions are right, there must be some crucial factor that the theory presented here does not take into account. I would not know what factor that would be. Clearly, there are important diffferences between the two examples: the three atoms figuring in the second example refer to facts with an equal 'epistemic status', whereas in the first example there is an important difference between, on the one hand, the king's presence and, on the other hand, the light being on and the flag being up; the latter serve as external signs for the otherwise invisible occurrence of the former. I can imagine that an explanation of the difference between the two examples starts with this observation[20], but I have no idea how it would continue, let alone that I would know how to model it formally.

This is not the only issue I have to leave for another occasion. I have just taken a first step in getting a decompositional analysis of counterfactual conditionals. Further steps are called for. Most urgent: in the above I have neglected all matters having to do with the interplay of tense and mood in the *would+have+past participle* construction. This means there is a range of problems we have nothing sensible to say about.

Let me give one example. For an indicative conditional to make sense, it is not necessary that the event described in the antecedent precede the event in the consequent. There is nothing wrong with a sentence like:

If he left the interview smiling, it went well.

However, in the counterfactual mood, this cannot be done.

If the interview had gone well, he would have left smiling.

sounds perfect. But it is hard, if not impossible, to get a reading of

**If he had left the interview smiling, it would have gone well.*[21]

in which the event described in the consequent precedes the event described in the antecedent.

One wonders if this phenomenon is due to the peculiar way in which tense and mood are combined in the English modal perfect, or if there is a deeper, semantic or cognitive reason for it, which also affects the counterfactual mood in other languages. I hope it is possible to shed some light on this by combining some of the ideas put forward here with the event based semantics put forward in Condoravdi (2002).

Author's address

ILLC/ Department of Philosophy
University of Amsterdam

[20] I owe this observation to one of the referees, who illustrates the point by a variant on the Ann/Billie/Carol puzzle: 'Suppose Carol is invisible. Suppose further that you are a proud parent of Ann, Billie and Carol, and before you go to bed you go in and check on the kids. As described in the original version, Ann is on the floor, Billie is in bed and Carol (obviously) is also in bed. Now you turn to your spouse and comment: if Ann had been in bed, Carol would have been on the floor.'

[21] There should at least be a comprehensible *epistemic* reading of this sentence — if at least counterfactuals have such readings.

Nieuwe Doelenstraat 15
1012 CP Amsterdam
e-mail: veltman@hum.uva.nl

Acknowledgement
I thank the referees for their comments and suggestions, and the editors for their help and patience.

References

Boyland, J. T. (1996). *Morphosyntactic Change in Progress: A Psycholinguistic Treatment*. PhD thesis, Michigan State University.

Condoravdi, C. (2002). Temporal interpretation of modals. In Beaver, D., Martinez, L., Clark, B., and Kaufmann, S., editors, *The Construction of Meaning*, pages 59–88. CSLI Publications, Palo Alto.

Gillies, A. (2004). Epistemic conditionals and conditional epistemics. *Noûs*.

Kanazawa, M., Kaufmann, S., and Peters, S. (2005). Conditionals in possible worlds: Times three. *Journal of Semantics*. this issue.

Kaufmann, S. (2000). Dynamic discourse management. In Faller, M., Pauly, M., and Kaufmann, S., editors, *Formalizing the Dynamics of Information*, pages 171–188. CSLI Publications.

Kratzer, A. (1981). Partition and revision: the semantics of counterfactuals. *Journal of Philosophical Logic*, 10:242–258.

Kratzer, A. (1989). An investigation of the lumps of thought. *Linguistics and Philosophy*, 87(1):3–27.

Lewis, D. (1973). *Counterfactuals*. Basil Blackwell, Oxford.

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

Pollock, J. (1976). *Subjunctive Reasoning*. Reidel, Dordrecht.

Rescher, N. (1964). *Hypothetical Reasoning*. North Holland Publishing Company, Amsterdam.

Rooth, M. (1985). *Association with Focus*. PhD thesis, Amherst MA: University of Massachusetts.

Rott, H. (1999). Moody conditionals: Hamburgers, switches, and the tragic death of an american president. In Gerbrandy, J., Marx, M., de Rijke, M., and Venema, Y., editors, *JFAK. Essays dedicated to Johan van Benthem on the occasion of his 50th birthday*, pages 98–112. Amsterdam University Press, Amsterdam.

Stalnaker, R. (1968). A theory of conditionals. In Rescher, N., editor, *Studies in Logical Theory*, pages 98–112. Basil Blackwell, Oxford.

Stalnaker, R. (1984). *Inquiry*. A Bradford Book. MIT Press, Cambridge MA.

Tichy, P. (1976). A counterexample to the stalnaker-lewis analysis of counterfactuals. *Philosophical Studies*, 29:271–273.

van Rooy, R. (2005). A modal analysis of presupposition and modal subordination. *Journal of Semantics*. this issue.

Veltman, F. (1976). Prejudices, presuppositions, and the theory of counterfactuals. In Groenendijk, J. and Stokhof, M., editors, *Amsterdam Papers in Formal Grammar. Proceedings of the 1st Amsterdam Colloquium*, pages 248–281. University of Amsterdam.

Veltman, F. (1996). Defaults in update semantics. *Journal of Philosophical Logic*, 25:221–261.