Counterfactual Dependencies

MSc Thesis (*Afstudeerscriptie*)

written by

Kasper Højbjerg Christensen
(born September 18th, 1984 in Ringkøbing, Denmark)

under the supervision of **Prof Dr Frank Veltman**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:
August 29th 2011

Members of the Thesis Committee:
Prof Dr Frank Veltman
Prof Dr Jeroen Groenendijk
Dr Katrin Schulz

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract

This thesis is concerned with developing an adequate semantics for counterfactual conditionals.

A counterfactual conditional is standardly taken to be an expression of the form 'if it had been the case that φ, it would have been the case that ψ', where φ and ψ are sentences and φ furthermore expresses something false. Now since expressions of this form are not truth-functional in the standard sense, the task of coming up with an adequate semantics for them has proven to be a somewhat difficult affair. We will present some theories of counterfactuals and discuss the problems that these have. Many of these theories agree that to evaluate a counterfactual we need prior knowledge of certain relationships in the world. We agree on this point, but we will redefine what these relations are; namely *generation relations* where we say that X and Y are in this relation when X will bring about Y, while remaining silent on whether Y will obtain when X does not.

We incorporate this into a dynamic setting where the meaning of a sentence is an operation on the cognitive state of an agent. We also draw a distinction between a counterfactual being true in an absolute sense and a counterfactual being acceptable by an agent. Throughout the thesis we will concern ourselves mostly with the latter notion and propose that if one believes in such a thing as *the truth-value* of a counterfactual, then this is to be defined as acceptability by a certain idealized agent.

Contents

Acknowledgements

I would like to thank my supervisor, Prof Dr Frank Veltman, for useful discussions and equally useful comments on earlier versions of this thesis. Thanks also to Dr Katrin Schulz and Prof Dr Jeroen Groenendijk for being on my thesis committee and providing useful comments and feedback during my defense.

A special thanks to The MoL Gang (you know who you are!) and my flatmates; Lau Møller Andersen and Benjamin Gregory Lang. - I miss you guys... truly, madly, deeply!

Chapter 1

Introduction

Counterfactuals are notoriously difficult. They are not truth-functional in any straightforward sense. If they were, the following two conditionals would have to have the same truth-value:
(1) If I were exactly 7 meters tall, I would be more than 8 meters tall.
(2) If I were exactly 7 meters tall, I would be more than 6 meters tall.
Since I am not more than two meters tall, it follows that both conditionals have false antecedents and false consequents. Hence, if these counterfactuals were truth-functional in the standard sense, they would have to have the same truth-value, but it is straightforward that (1) is false, whereas (2) is true. So, it follows that if these conditionals are indeed truth-functional the notion of truth involved cannot be that of truth at the actual world. What has often been admitted is that, in the case of true counterfactuals, a certain kind of relationship between antecedent and consequent holds that does not hold in the case of false counterfactuals. What this relationship consists in is no easy puzzle to solve however. Many theories have been proposed and almost as many theories have been shown by counterexamples to be inadequate in one way or another. The present thesis is concerned with the question of finding an adequate semantics for counterfactual conditionals such as the above.

In this chapter of the thesis we will try to get clear on what a counterfactual is. We will briefly touch upon the relation between the epistemic and ontic readings of counterfactuals. We then present some theories to illustrate different approaches to counterfactuals, and in the last section we will draw a distinction between acceptability of a counterfactual and truth of a counterfactual.

In chapter 2, 3, and 4 the theories of Kratzer (2010), Veltman (2005), and Schulz (2009) will be presented and discussed, respectively. As will be clear when we present our own semantics, many of the main ideas that we use are already present in these three theories. However, the three theories have difficulties of their own, which will also be presented and discussed.

In chapter 5 we present our own semantics and discuss some of the notions we have defined in order to do so. Most notably our idea of a generation relation, where something *brings about* something else.

In chapter 6 we will discuss some of the issues that have come up throughout the thesis, and in chapter 7 we will briefly conclude this thesis with an overview of what we have done.

At some points throughout the thesis we will discuss the relationship between causality and counterfactuals. Most notably in relation to the theories of Veltman (2005), Schulz (2009), and our own semantics.

1.1 What is a counterfactual?

A counterfactual is standardly taken to be a conditional where the antecedent is false, that is, where the antecedent is *counter to fact*. A counterfactual therefore speaks about what would have been the case had that which is mentioned in the antecedent obtained. It is often (but not always) given the formulation "if it had been the case that φ, it would have been the case that ψ", where φ and ψ are sentences. For instance, now that I have just finished my cup of coffee I may say "if I had spilled all of that coffee, I would not have drunk it", thereby expressing that in other circumstances where the coffee had been spilled, I would not have been able to drink it.

It has been held that a counterfactual is a conditional where the antecedent is a past tense subjunctive statement. Therefore it is often claimed that a counterfactual is a special case of a subjunctive conditional, that is, a conditional where the antecedent is a subjunctive statement. A counterfactual is thus taken to be a subjunctive conditional with a false antecedent.

Concerning the classification of conditionals into subjunctive, counterfactuals, etc. there are of course many grammatical and linguistic issues at stake, and things are by no means as simple as they appear from the above description. However, since the purpose of this thesis is another we will not go into a discussion concerning the proper classification of conditionals.[1] Instead we will follow standard practise and take a counterfactual conditional (or counterfactual) to be a conditional where the antecedent expresses something false about the world.

On a syntactic level we will use $A \rightsquigarrow B$ to denote the counterfactual with antecedent A and consequent B.

1.2 Two vs. Three Parameters

One might wonder whether it makes sense to say that counterfactuals have truth-values in any absolute sense. That is, do we wish to construct a theory which holds that for every counterfactual, $A \rightsquigarrow B$, there is a definite answer as to whether it is true or false? To illustrate consider the situation where you are about to go to the airport and catch a flight at seven p.m. The time is now five p.m. and you are headed out the door.

(1) If it were seven p.m. you would be boarding your flight.

The question is whether the counterfactual in (1) is true or false. It seems that it is both true and false in the sense that one can just as plausibly claim it to be true as one can claim it to be false. That it must have a false reading as well is also seen by the fact that the counterfactual in (2) has a true reading.

(2) If it were seven p.m. you would be missing your flight.

Assuming of course that the two consequents are mutually exclusive and exhaustive, that is, either you are boarding your flight or you are missing it, it follows that a true reading of (2) implies a false reading of (1).

So, it would seem that in the case of (1) and (2) it would make little sense to speak of an absolute truth-value of the two.

Tichý (1984) draws a distinction between what he calls a Two Parameter theory and a Three Parameter theory. A proponent of a Two Parameter theory will claim that the meaning of a counterfactual is a function of the meaning of the antecedent and consequent only, whereas a proponent of a Three Parameter theory will claim that more is needed to fix the meaning of a counterfactual. The name *Two Parameter* theory thus reflects the fact that the meaning of a counterfactual is a function of two parameters; the meaning of the antecedent and of the consequent. As such, a *Three Parameter* theory agrees that the meaning of antecedent and consequent is needed, but it further admits of a necessary third parameter. This third parameter can be different things and we will see how, for instance, Lewis

[1] For such a discussion see Bennett 2003.

(1973a) takes it to be a primitive similarity ordering of possible worlds, whereas Veltman (2005) takes it to be the set of laws and generalizations that the agent takes to hold true of the actual world.

While we have been casting the above in terms of what is needed to determine the meaning of a counterfactual, we will from now on focus on what is needed to determine the truth-value. Here, in a straightforward way, a Two Parameter theory will say that the truth-value is a function of the antecedent and consequent only, whereas a Three Parameter theory will claim that a third parameter has a role in determining the truth-value. However, when we speak about truth instead of meaning, it is not clear that there is a substantial difference between a Two and a Three Parameter theory, since truth on a Three Parameter theory will always be truth relative to a setting of the third parameter, and thus not *the* truth or falsity of a counterfactual as the Two Parameter theory would claim to be speaking about.

We will return to discuss the difference between a Two and Three Parameter theory on the level of truth once we have seen some examples of different theories.

1.3 Two Readings

It is often held that (at least some) counterfactuals have both *epistemic* and *ontic* readings. It is not entirely clear from the literature what exactly the difference is, but one way of trying to capture it would be the following:

> **Epistemic**: An epistemic counterfactual is about what a rational agent should conclude (or is justified in concluding) if he came to learn the truth of the antecedent.
>
> **Ontic**: An ontic counterfactual is about what would have been the case had a certain fact about the world been different.

Now what exactly is the difference here? The difference is perhaps easiest to explain drawing on an example due to Hansson (1989) about hamburgers.[2] Suppose you enter a town of which you already know that it has two snack bars, A and B. You see a man with a hamburger. Now you drive past snack bar A and you see that it is in fact open. You tell yourself:

> (S) If snack bar A had not been open, then snack bar B would be open.

[2] see also Rott (1999).

This is presumably true in the described situation and most people share this intuition. However, suppose after you see that snack bar A is open you drive on to find snack bar B closed. Is (S) still true? Why is it not the case that (S') describes the situation just as good?

> (S') If snack bar A had not been open, I would not have seen the man with the hamburger.

The difference it is claimed, is to be found in the difference between an epistemic and ontic reading of the conditionals in question. (S) is strictly speaking only true under an epistemic reading. In the above, if you were to learn that A was in fact *not open* (somehow your senses have been deceiving you), then since you have still seen the man with the hamburger, you would be justified in concluding that the other snack bar, B, must be open. So, on an epistemic reading, (S) is true. However, on an ontic reading, (S') is true and (S) false. Why? The facts of the actual world are as follows. A is open, B is closed, and you have seen the man with the hamburger. It is reasonable to assume that a necessary condition for you seeing the man with the hamburger is that one of the snack bars are open. Since B is closed, we may conclude that A being open brings about the man with the hamburger. But, then when we counterfactually change the fact that A is open, we also change the fact about the man with the hamburger. Hence, if A is closed, it means there will be no man with a hamburger. Clearly, no such dependencies between A being open and B being open are present, so the counterfactual expressed by (S) is false, while the counterfactual expressed by (S') is true.[3]

It is clear that the epistemic reading of a counterfactual is intimately linked with the concept of belief revision. When an agent accepts an epistemic reading of a counterfactual it is because he, when revising his belief in the falsehood of the antecedent, gets into a state where he believes the consequent.

[3] In a sense I am already getting ahead of myself here. I am using insights from Tichý (1976) and Veltman (2005) to evaluate ontic counterfactuals. However, the exact evaluation procedure is not important. What is important on the level of ontic counterfactuals is that changing one fact also changes the facts that that fact brought about. We are concerned with the state of the world, and not with what we can conclude given a piece of information. On the level of the world, A being open *brought about* the man with the hamburger, but A being open did not bring about B being closed. So, changing A's status will only change the fact about the man with the hamburger, not anything that has to do with B's being open or closed. For more details see Veltman (2005).

On the other hand, an ontic counterfactual does not speak of what one would believe, but simply about what would have been the case. However, even if this is the case, the acceptability/assertability of a counterfactual by an agent will still depend on the belief state of that agent. It therefore makes little sense to say that an epistemic reading of a counterfactual is concerned with belief revision whereas an ontic reading is not. The difference between the two readings is a consequence of different ways of revising ones belief. It is often claimed that an epistemic reading of a counterfactual has to do with a minimal revision. That is, when evaluating an epistemic counterfactual one simply adds the antecedent to ones stock of belief and makes minimal adjustments to maintain consistency. If ones stock of belief supports the consequent, one should accept the conditional. However, as we just saw, when dealing with an ontic counterfactual, minimal adjustments to maintain consistency are not enough. As we saw in the example with the hamburger, when revising our belief in the falsehood of the antecedent we also have to revise our belief in the things that depended on the antecedent being false. The revision here is thus broader than just minimal adjustments to maintain consistency.

The exact difference between ontic and epistemic readings of counterfactuals is surely an interesting issue, but we will not pursue it here.[4] In what follows we will focus on ontic readings of counterfactuals; that is, the readings of counterfactuals to be discussed in what follows are the readings concerned with what would have been the case, had a certain fact about the world been different.

1.4 Goodman

Goodman (1954) begins his investigation of counterfactuals by noting that they are not truth-functional. Instead a "counterfactual is true if a certain connection obtains between the antecedent and consequent." (Goodman 1954). This connection, however, is not as straightforward as one could have hoped for, and Goodman starts by noting that it is rarely the case that the consequent follows from the antecedent by logic alone, as when someone says "if that match had been scratched, it would have lighted". More; the structure of the match, the presence of oxygen, etc., is needed to make the consequent follow from the antecedent, and even when these conditions are added explicitly, the entailment relation is still not one of pure logic.

[4] For further discussion see Rott (1999) and Bennett (2003).

> "But even after the particular relevant conditions are specified, the connection obtaining will not ordinarily be a logical one. The principle that permits inference of
>
> > That match lights
>
> from
>
> > That match is scratched. That match is dry enough.
> > Enough oxygen is present. Etc.
>
> is not a law of logic but what we call a natural or physical or causal law." (Goodman 1954).

That is, according to Goodman the connection obtaining between antecedent and consequent is one that depends partly on facts obtaining at the actual world and the true natural laws. We can thus state Goodman's truth-conditions for counterfactuals as follows:

> A counterfactual $A \rightsquigarrow B$ is true iff there exists some set C of true sentences, such that B follows from A in conjunction with the sentences of C by natural, logical or mathematical laws.[5]

Goodman observes that C can not be *any* set of true sentences whatsoever. If we are dealing with a genuine counterfactual then the negation of the antecedent will be true, and thus any consequent whatsoever will follow. This is not the intended meaning of counterfactuals, and so Goodman takes it on himself to come up with an answer to the first major problem of counterfactuals, that is, the problem of giving an adequate description of the set C.[6]

Goodman eventually settles for a circular definition; the sentences in C are allowed to describe conditions which would still obtain, had the antecedent been true. But this last formulation is of course itself counterfactual, and the truth-conditions ultimately circular. As interesting as Goodman's work and this latter problem is, we will not discuss these issues further in the present thesis.

[5] If you believe that logical and mathematical laws are just special cases of natural laws, you may simply ignore the mentioning of *logical* and *mathematical laws* in the truth-conditions.

[6] Goodman (1954) sets up two major problems that a theory of counterfactuals must provide an answer to. The second major problem of counterfactuals is how to define the laws by which the consequent follows from the antecedent and the true sentences of C.

Now, it should be clear why Goodman's theory is, on the level of truth, essentially a Two Parameter theory of counterfactuals: even though there are conditions on what can count as true sentences in the set C, whether or not a set meeting those conditions exists is a matter of brute fact. All that is needed to assess the truth of a counterfactual is thus the meaning of the clauses A and B.

However, as will be clear from future discussions the divide between Two and Three Parameter theories is perhaps, at least when it comes to truth, not a substantial one, but simply a matter of perspective. We may thus already note that in a certain sense Goodman's theory *is* a Three Parameter theory. The third parameter is of course the set C. What makes Goodman's theory a Two Parameter theory when it comes to truth is the contention that either there is such a set C meeting the conditions, or there is not. We will return to these issues in due course.

1.5 Lewis

One of the most well-known semantic theories of counterfactuals is due to Lewis 1973b. This approach builds heavily on possible worlds semantics and admits of a third parameter playing a role in giving the truth-conditions of a counterfactual. This parameter is a similarity relation on the set of possible worlds. Intuitively the idea is that some worlds are more similar to the actual world than others, and what determines the truth or falsity of a counterfactual is dependent not only on the content of the antecedent and consequent but also on this similarity relation. Take for instance a conditional such as 'if I had taken the bus an hour earlier, I would have arrived an hour earlier'. When we judge this counterfactual to be true it is simply because in the most similar world to the actual world where I take the bus an hour earlier, I will also arrive an hour earlier. There are of course worlds where the bus has a flat tire, and worlds where it simply vanishes into thin air, but the counterfactual is true because non of these worlds are as close to the actual world as the world in which I take the bus an hour earlier and arrive an hour earlier. Given such a similarity relation, the truth-conditions of a counterfactual on Lewis' approach are as follows:

> A counterfactual $A \rightsquigarrow B$ is true at a world w iff the consequent B is true at the closest world(s) to w, according to the ordering R, at which A is true *or* if A is true at no world at all.

Here R is the aforementioned relation of comparative overall similarity; that is, the closest worlds to w according to R are those worlds that are overall most similar to w. The third parameter here is thus the relation R, which is given by context and the content of the antecedent. That is, the truth of a counterfactual $A \rightsquigarrow B$ is a function of three things; the meaning of A, of B, and the relation R.[7] However, it is important to notice that there is not *one* relation which does the job for all counterfactuals. Instead the relation R changes from context to context. This also means that a counterfactual only has a truth-value relative to a choice of R, but since truth-relative-to-something-else is to be expected from Three Parameter theories, this should not worry us. What might worry us, however, is what constraints we can put on this relation R if we are to commit to Lewis' claim that this is a relation of comparative overall similarity. Kit Fine (1975) puts forward an example to show that the relation R might, in some cases, not be as straightforwardly related to similarity as Lewis would originally have liked it to be. The example is the famous Nixon example.

> Imagine that in the seventies there was indeed a button to which only The President of The United States of America had access, and which, when pressed, would launch the American nuclear missiles against Russia.

If the counterfactual 'if Nixon had pressed the button, then there would have been a nuclear holocaust' is true in the described scenario — as all intuitions seem to say it is — then it follows that R cannot be overall similarity. Clearly, overall similarity would deem the world that is exactly like our world, except a small malfunction occurs when the button is pressed, more similar than a world in which a nuclear holocaust occurs. So it would seem that the relation in question does not coincide with what we would normally intuitively judge as similarity.

Another thing that we might want to say is this: if two worlds differ only with respect to one fact, then the world which agrees with the actual world on this fact is the more similar world, and thus the closer world according to R. As a famous example put forward by Tichý (1976) shows, this cannot be the case either.

> "Consider a man, call him Jones, who is possessed with 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 way: on fine days he puts his hat on or

[7] Because R is in turn a function of A and the context, we might want to say that the third parameter is in fact the context and not the relation R. However, for present purposes, nothing important turns on this choice.

> leaves it on the peg, completely at random. Suppose moreover that actually the weather is bad, so Jones is wearing his hat." (Tichý 1976).

Now we ask, is the counterfactual expressed by "if the weather had been fine, Jones would have been wearing his hat" true? Supposedly it is not. The story tells us that in such a situation whether or not Jones is wearing his hat is completely randomly decided. So, there is no reason to suppose that he will be wearing his hat (or that he won't be wearing his hat for that matter), and so the counterfactual is false. However, take two worlds, one where Jones is wearing his hat and one where he is not, but which otherwise agree completely with each other. According to the intuitive similarity constraints just mentioned this would mean that the former is closer to the actual world, because the former agree on the fact that Jones is wearing his hat. However, if this is the relation R we are working with it follows that the counterfactual is true, and hence, we cannot impose this constraint on the relation R. As for the relation R we must have $w_{hat} \geq_R w_{no\ hat}$ and $w_{no\ hat} \geq_R w_{hat}$, at least insofar as we do not want the any of the counterfactuals 'if the weather had been fine, Jones would have been wearing his hat' or 'if the weather had been fine, Jones would *not* have been wearing his hat' to be true. So, it seems the relation R is not a relation of similarity of worlds, at least not on our intuitive understanding of similarity.

But if we cannot spell out this relation between possible worlds, R, according to which some worlds are closer than others, then there is a sense in which the theory does not even get started. For every true (or false) counterfactual there will of course be some relation R such that the truth-conditions give the right prediction, but if we have no knowledge of this relation prior to judging the counterfactual true (or false), then the theory as such tells us nothing.

Lewis himself admits that the relation R between worlds is vague, but he maintains that this does not mean that it is ill-understood and thus cannot be used in explaining the meaning of counterfactuals.

> "It may be said that [...] the notion of comparative overall similarity of worlds is hopelessly unclear, and so no fit foundation for the clarification of counterfactuals or anything else. I think the objection is wrong. 'Unclear' is unclear: does it mean 'ill-understood' or does it mean 'vague'? Ill-understood notions are bad primitives because an analysis by means of them will be an ill-understood analysis. (It may yet be better than no analysis at all.) But comparative similarity is not ill-understood. It

> is vague—very vague—in a well-understood way. Therefore it is just the sort of primitive that we must use to give a correct analysis of something that is itself undeniably vague." (Lewis 1973b).

Lewis reasoning here is very clear. When determining the meaning of a counterfactual we make use of a (contextually given) primitive notion of similarity between worlds. This relation is *not* ill-understood, but it is very vague. However, since many counterfactuals are vague, it need not come as a surprise to us that in an analysis of their meaning we make use of a concept that is itself very vague. Hence, the vagueness from the similarity relation shines through on the level of counterfactuals. As such this seems to be an appropriate response from Lewis. Many counterfactuals are inherently vague, and so this vagueness has to come from somewhere.

However, the problem is that there are counterfactuals with very clear intuitions concerning their truth-value, but with very unclear intuitions as to how to spell out the underlying similarity relation, the relation we called R above. Take for instance the above example about Jones and his hat. Here intuitions are very clear that the counterfactual 'if the weather had been fine, Jones would be wearing his hat' is not true in the described scenario. But then, as we saw, it follows that for two worlds that are exactly like each other, except that in one Jones is wearing his hat and in the other he is not, the first is not more similar to the actual world than the other according to R. A story needs to be told here as to why this is the case. If the intuitions are very clear, then the underlying similarity relation should *not* be very unclear. However, it does not seem that this is the case.

It does not follow, however, that the theory as such is wrong. As Kratzer notes:

> "Notice that it is not that the similarity theory says anything false about examples like the [Jones] example. It doesn't say enough. It stays vague where our intuitions are relatively sharp. We should aim for a theory of counterfactuals that predicts vagueness for precisely the cases where our intuitions are vague, and makes sharp predictions for precisely the cases where our intuitions are sharp." (Kratzer 2010)

The point here is that the similarity relation involved in the truth-definitions should be such that when we encounter a counterfactual that intuitively has a very definite truth-value, we should be able to predict this, thus have a clear idea of the relation R. However, *we should not be allowed* to refer to the truth of the counterfactual in question in doing so, since then we will end up in a vicious circle.

There is also another worry, that may perhaps be a more serious worry for the proponent of a Lewisian approach to counterfactuals. The whole idea is that given an antecedent and a context some worlds are closer (or more similar) to the actual world, and once this is settled the truth-value of the counterfactual follows straightforwardly. In other words, the truth of a counterfactual is dependent on which worlds are closer than others. If this were true it would have some bearing on how we use and discuss counterfactuals. But as Tichý observes, this is not the case.

> "The [theory] of Lewis [is] completely divorced from the way subjunctive conditionals are argued over in practice. If world-similarity [is] what the truth-value of a subjunctive conditional turns upon, how is it that disputes about conditional statements are never settled by reference to such matters? Suppose that a dispute arises as to whether some nuclear missiles would have been launched had someone pushed a certain button yesterday. Are those who think the answer is 'Yes' ever likely to support their view by arguing that a situation in which the button was pushed and the rockets went off is more similar, overall, to the way things in fact are than is any situation in which the button was also pushed but nothing happened? [...] And will those who think the answer is 'No' try to refute these world-similarity or world-gerrymandering claims? I have yet to hear someone argue that way off the premisses of a philosophy department." (Tichý 1984).

The odd thing is of course that no one ever seems to mention what the relation is or argue by showing that one world is in fact closer than another. As such this is not a knock down argument against the theory, and it is not meant to be. However, if we take serious the claim of Lewis that the truth of a counterfactual is dependent on this relation R, then it does seem very odd that when discussing whether or not a counterfactual is true no one ever seems to mention this relation.

1.6 Mill, Ramsey, Chisholm

Another account of the Three Parameter approach to counterfactuals is due to Mill, Ramsey, and Chisholm. There are of course differences in the approaches of these three authors, respectively.[8] Here I will follow the presentation of the theory given in Tichý (1984), and as such I will ignore questions as to whether this is actually what Mill, Ramsey, and Chisholm, respectively, had in mind. This latter question is important for a historical overview of the debate, but for now it suffices to look at whether the theory as such is a good way to approach counterfactuals.

Tichý motivates his presentation of the theory in relation to the quote given above. He continues:

> "Among people for whom the correct truth-value of a conditional is a matter of genuine concern, such a dispute is likely to turn very soon into a dispute over some matters of fact and of ordinary logic. Those who think the conditional is true will typically invoke some facts (like the nature and state of the electrical circuits involved) and physical laws (or what they believe to be such) and then appeal to ordinary logic to show that these, together with the imaginary pushing of the button are related to the imaginary launching as the premisses of a valid argument to its conclusion. Their opponents, on their part, are likely to try and cast doubt on the alleged facts, or on the alleged laws, or on their adversaries' logic. The two parties will normally agree on which particular matters of fact are relevant to the problem at issue: yesterday's condition of the circuits, for example, will undoubtedly be deemed relevant. Today's condition of the missiles undoubtedly won't: neither party would dream of invoking this in favour of or against the conditional. No one would take seriously a clever logic-chopper who argued that, since the rockets are in fact still in their silos, then had the button been pushed yesterday, something would have been the matter with the circuits. Not because his conditional is unacceptable in some absolute sense. But because he appeals to a fact which does not belong to the class of facts which are relevant in the present context. In other contexts, where the class of relevant facts is circumscribed differently, the conditional may be quite to the point." (Tichý 1984).

The point here is worth discussing in detail. In the conversation some facts are relevant and some are not. As Tichý points out, in the above setting the conditional 'if someone had pushed the button yesterday, the rockets

[8] Mill (1868), Ramsey (1931), Chisholm (1955).

would not have launched' is not acceptable (not true), simply because the fact that the rockets did in fact not launch is not a relevant fact. There might however, be a context in which this fact would count as relevant, and the conditional therefore acceptable (true).

This observation is exactly the point of the Mill, Ramsey, Chisholm account according to Tichý. On their account the third parameter which plays a role in evaluating the counterfactual, is simply a set of auxiliary premises that are not stated, but simply tacitly assumed by the speaker. From these tacitly added premises and the antecedent of the conditional the consequent follows by pure logic: "On this theory, the logical relation involved in subjunctive conditionals is the familiar one of implication or entailment: subjunctive conditionals are explained simply as elliptical statements of logical consequence." (Tichý 1984). With this in mind we can give truth-conditions for counterfactuals as follows:

> A counterfactual $A \rightsquigarrow B$ is true relative to the set C of auxiliary premisses iff the members of the set C are true and $\{A\} \cup C$ logically implies B.

Now one might wonder how this set of auxiliary premises is given? How one continues this story will of course give rise to different specific theories of counterfactuals. Some, among them Mill, Ramsey, and Chisholm, would say that the set C is simply given by the intentions of the speaker (we might want to say that the speaker assumes the members of C).

There seems however to be some restrictions that one needs to put on this set of auxiliary premises. Whether or not one takes this set to be something which is given by the intentions (or assumptions) of the speaker, it seems that the negation of the antecedent is not allowed to be in there. If we are dealing with a pseudo counterfactual, in the sense that the antecedent is in fact not counter to fact, then this is taken care of by the truth-conditions; the set of auxiliary premises C need to contain only true statements, and if one of them is the negation of the true antecedent, then this will not be the case. However, if the counterfactual is genuine and C includes the negation of the antecedent, then this will be true, and as such any consequent whatsoever will follow logically, at least if the consequence relation is classical. If all true statements were allowed to be in the set C, then it follows that all genuine counterfactuals can be used truly — at least if C is given by the speakers intention. This, however, does not seem to be the case.

So it would seem that we need to put some restrictions on the set C in order to get a theory that is not simply "anything goes". As will be clear later, many of the theories to be discussed in this thesis try to do exactly that, and our own proposal for a semantics is no exception.

1.7 Acceptability vs. Truth

So far we have been speaking loosely about *truth* and *truth relative to something else*. It might be good to take a moment to think about what these different notions actually mean. For instance, when a speaker utters or accepts a counterfactual to be true, it must be because he takes the counterfactual to be true relative to his own belief state (whatever this belief state may reflect; i.e. a set of true laws and generalizations, a set of facts true about the world, or a similarity ordering among worlds). Similarly, when we speak about *the truth-value* of a counterfactual — if we in fact believe in such a notion, which of course is a different issue — we do not seem to make reference to the belief state of any speaker or hearer of the counterfactual. If there is such a thing as *the* truth-value of a counterfactual, it has this truth-value independent of what a particular agent may think about the world.[9] One could therefore claim that the notions of acceptability and assertability naturally relates to a Three Parameter theory of counterfactuals, whereas the notion of truth *as such* relates to a Two Parameter theory. Let me briefly explain how this could be put to work.

The truth-conditions for a Three Parameter theory is stated relative to a third parameter, call it σ. We have:

> A counterfactual $A \rightsquigarrow B$ is true relative to σ iff ...

where the dots have to be filled out by the appropriate conditions. What we propose this to reflect is the following.

> A counterfactual $A \rightsquigarrow B$ is acceptable/assertable by an agent iff $A \rightsquigarrow B$ is true relative to a setting δ of the third parameter σ, and the agent believes δ to reflect the truth.

What this means is simply that an agent will utter (or accept) a counterfactual when it is true *relative to something*, δ, which the agent also takes to be true. We might therefore replace the phrase '$A \rightsquigarrow B$ is true relative to

[9] This is of course an overly simplistic view of the matter. If you subscribe to an absolute notion of truth, it does not automatically follow that what determines what is true is independent of how people do (or are able to) think about the matter, as it is presented here. However, these issues are too delicate to discuss here.

a setting δ of the third parameter σ' with '$A \rightsquigarrow B$ is acceptable/assertable by an agent who believes δ to reflect the truth'.

Similarly we might believe that there is *one setting* on the third parameter σ which is in fact the true one. If we believe this we are able to give absolute truth-conditions for a counterfactual. Call the one true setting of the third parameter ϵ, then we can say the following:

> A counterfactual $A \rightsquigarrow B$ is true *in an absolute sense* iff $A \rightsquigarrow B$ is true relative to the setting ϵ of the third parameter σ.

Here we do not make reference to the belief state of any agents, but only to the true setting of the third parameter.

To illustrate we may present a small example. Smith did not get the promotion. Instead the promotion went to Jones. Smith believes he was the next best candidate, even though in fact, the next best candidate was Johnson. In this setting Smith will accept the counterfactual 'if Jones had not gotten the promotion, Smith would have gotten it' even though it is in fact false, because the true state of the matter is such that 'if Jones had not gotten the promotion, Johnson would have gotten it' is a true counterfactual.

We may note that the third parameter here is Smith's belief that he is the next best candidate for the job, whereas the true setting of the third parameter is the objective fact that Johnson is the next best candidate for the job. This, even though we express it as beliefs, may equally well be reflected in a similarity ordering of worlds. What this amounts to is just that Smith takes the similarity ordering (relative to the antecedent given) to be such that a world where he gets the promotion is closer to the actual world than is any world where neither him nor Jones gets it. But, in fact the true similarity ordering is such that a world where Johnson gets the promotion is closer than any world where Johnson or Jones does not get the promotion.

So, if we take this stance towards acceptability/assertability and truth, it would seem that the divide between Two Parameter theories and Three Parameter theories is not a substantial one; it is not two different approaches after all, because they speak about different things. Truth *relative to something else* which the latter kind speaks about is not to be confused with truth *in itself* which the former kind speaks about.

For instance, we saw that Goodman's theory was essentially a Two Parameter theory because it speaks of *the* truth or falsity of a counterfactual. A counterfactual is true if there is a set of true statements about the world such that the antecedent and these statements entail the consequent, and whether or not such a set exists is a matter of brute fact. However, it is not difficult to see how we would apply Goodman's theory to come up with a notion of acceptability and assertability. We just need to say that an agent will accept/assert a counterfactual when there is a set of statements such that the consequent follows from these in combination with the antecedent by natural, logical, or mathematical law, *and* that the agent in questions believes the statements in this set to be true.

It is also not hard to see how a Three Parameter theory can give rise to a Two Parameter theory. As we have seen, if there is one setting of the third parameter which can reasonably be said to be the correct one, then fixing the parameter on this will give rise to a Two Parameter theory; one that speaks about *the* truth or falsity of a counterfactual.

With this said it is not clear that there is in fact a substantial difference between a Two and a Three Parameter approach to counterfactuals. It seems the difference is instead one of the aspects of counterfactuals one wish to talk about; that is, truth in and off itself, *or* when and where an agent will accept or assert a counterfactual.

We therefore better keep in mind the difference between *the* truth-value of a counterfactual, which is objectively decided, and which we may or may not believe in, and the truth-value of a counterfactual *relative* to a third parameter, which, at least on our conception of the matter, is not related to truth as such, but to the notions of acceptability and assertability.

While at some points we will continue to speak loosely of Two vs. Three Parameter theories, the reader is asked to keep the contents of this section in mind.

Chapter 2

Lumps of Thought

Before moving on to present the framework of Veltman (2005) we will take a brief moment to explain the theory of Kratzer (2010) and her idea of *lumping semantics*. We do this because the theory of Kratzer in many ways can be seen as a forerunner of the theory of Veltman (2005). The approach of Kratzer is within the premise semantics approach, an approach that also Veltman (2005) falls under. The idea is simple.

> *Premise semantics approach*
> A counterfactual $A \rightsquigarrow B$ is true iff there is some set, $Prem$, such that $Prem \models\!\!\equiv B$.

Here $\models\!\!\equiv$ is some consequence relation which could be classical, causal etc. $Prem$ is a set containing the premises that are relevant for the given counterfactual, and a premise semantics approach is nothing more than an approach that takes it upon itself to explain and define how this set is given.

Kratzer says the following about the truth-conditions of counterfactuals:

> "There is an intuitive and appealing way of thinking about the truth-conditions of counterfactuals. It is an analysis, that in my heart of hearts, I have always believed to be correct [...]."

The analysis Kratzer is referring to is the following.

> "A 'would'-counterfactual is true in a world w if and only if every way of adding propositions that are true in w to the antecedent while preserving consistency reaches a point where the resulting set of propositions logically implies the consequent."[1] (Kratzer

[1] We may note that these truth-definitions are absolute in the sense that *either* every way of adding true propositions to the antecedent while preserving consistency will result in a set logically implying the consequent, *or* this will not be the case.

2010).

This, however, is highly controversial, and as Kratzer points out, it will only work if we take into account her notion of lumping.[2] Now, we will not go into the theory in detail, but since the notion of lumping is so central to the theory (and the theories building on this theory, even though they may not use the word *lumping*) we will present the main ideas.

2.1 Lumping

This is perhaps easiest to explain using Kratzer's own example. Yesterday Paula painted a still life with apples and bananas. So, did Paula paint apples? Yes. Bananas? Yes. A still life? Yes. It is clear that Paula did in fact do all these things, but once I tell you that Paula painted a still life, you cannot tell me that that is true, but that she in fact *also* painted apples and bananas. Viewed in this way the fact that she painted apples and bananas is simply a part of the fact that she painted a still life. Paula painting a still life thus *lumps* her painting apples and her painting bananas in the actual world. Formally, Kratzer defines the lumping relation as follows:

> *Lumping*
> A proposition p lumps the proposition q in world w iff
>
> 1. p is true in w.
> 2. Whenever a situation s is part of w and p is true in s, then q is true in s as well.

Kratzer spends a great deal of her paper coming up with an adequate semantics based on situations. For present purposes it suffices to know that situations are partial functions from the atoms of the language into $\{0, 1\}$ (worlds are total such functions), and a proposition is a set of situations (intuitively, the situations where the proposition is true). A situation is part of a world exactly when $s \subseteq w$ in the standard way.[3] Now the definition says that a proposition p lumps q at a world w, if all the situations of w in which p is true, are also situations where q is true. To illustrate, the proposition that I own 6 place sets lumps the proposition that I own 6 forks in the actual world. In fact it lumps that proposition in every world where a fork is a

[2] See Kratzer (2010) for counterexamples when we do not take her concept of *lumping* into account.

[3] For further definitions, such as consistency, logical consequence, and the like, see Kratzer (2010). These are, however, just as one would expect.

necessary ingredient of a place set and where I own 6 of the latter. This is so because in every situation in which I have 6 place set, I also have 6 forks.

With this in mind the claim of Kratzer is that the truth-conditions given before are actually correct. We just need to pay attention to the lumping properties of propositions. That is, whenever we add a proposition, we also add the propositions that are lumped by it. This is of course given an exact formal definition in Kratzer (2010) using the aforementioned situation semantics. However, the exact formal details are not important for the present purposes, and we will not give them here.

With these truth-conditions Kratzer can give the right prediction in the Three Sisters example. This example is from Veltman (2005) and concerns itself with the following scenario.

> “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...” (Veltman 2005).

We want to ask if the counterfactual ‘if Ann had been in bed, Carol would be on the floor’ is true. According to our intuitions it should not be, because why would Billie not be on the floor?
The relevant propositions true of the actual world are thus the following:
(1) Whenever two of the sisters sleep in bed, the last one sleeps on the floor.
(2) Ann is on the floor.
(3) Billie is in bed.
(4) Carol is in bed.
Now the antecedent of the counterfactual in question is
(A) Ann is in bed.
To this we can add (1) and (3) which logically implies that Carol is on the floor. But, we can also add (1) and (4) which logically implies that Billie, and not Carol, is on the floor. Hence, the counterfactual ‘if Ann were in bed, Carol would be on the floor’ is false because not every way of adding true propositions to the antecedent result in a set that logically implies the consequent.

Now, as the reader might have noted (1) expresses something more general than the fact that whenever Ann and Billie are in bed, Carol is on the floor, even though this latter is the formulation given in the example from Veltman (2005). This will become important when we discuss our own proposal for a semantics, and we will return to this point later.

Before leaving Kratzer's theory we want to point to some of the shortcomings of the theory. Without getting technical we can describe the problem using an example from Kratzer, which Kratzer claims to provide a solution to in her paper. However, things are — at least not if we take Kratzer's own argumentation serious — not so simple. The example is about a zebra escaping from the Hamburg zoo.

> "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 2010).

Now the question is what the truth-value of the counterfactual 'if a different animal had escaped instead, it would have been a zebra' is. On a naive similarity approach, that is, an approach where the similarity relation is given as overall straightforward similarity, this is presumably true. This is so because a world where a (different) zebra escapes is intuitively more overall similar to the actual world than a world where a different animal escapes. However, it is not true according to Kratzer and strong intuitions count in her favor. She explains why using the lumping semantics. ("John" is the name for the zebra who in actual fact did escape).

> "On the present approach, we have an explanation: if the actual properties of the zebra mattered, it would be because of the following propositions:
> (13)
>
> a. A zebra escaped.
>
> b. A striped animal escaped.
>
> c. A black and white animal escaped.
>
> d. A male animal escaped.
>
> ....
>
> Given lumping, none of the propositions in (13) can be consistently added to the antecedent of a conditional of the form: 'If the animal that escaped had not been John ...'. In our world (at the time considered), every situation where a zebra escaped is a

> situation where John escaped. Every situation where a striped animal escaped is again a situation where John escaped and so forth, for all the properties of John. Hence in our world, the proposition that John escaped is lumped by (13a) to (13d)." (Kratzer 2010).

This is all good and well. Not every way of consistently adding true propositions to the antecedent results in a set that logically entails the consequent, which would have been the case if we could add (13a). This means that the counterfactual expressed by "if a different animal had escaped, it would have been a zebra" is false.

However, it also seems we can never add the proposition expressed by (13a) to the antecedent without reaching an inconsistency. This is so because (13a) lumps the proposition that John escaped, that is, the negation of the antecedent. But, if we have a look at Kratzer's truth-definitions for might counterfactuals, they state that a might-counterfactual is true if and only if not every way of adding true propositions to the antecedent results in a set such that adding the consequent to this set makes it inconsistent. What this means is the following. If Kratzer maintains that (13a) lumps the proposition that John escapes, it seems to follow that (13a) can never be added to the antecedent of 'if the animal escaping had not been John, it might have been a zebra' without making the resulting set inconsistent. Hence, the counterfactual expressed by "if a different animal had escaped, it might have been a zebra" is false on this approach. However, it is intuitively true. If a different animal had escaped, it might have been another zebra, just as well as it might have been a gazelle or a giraffe.[4]

There is another worry concerning the informal definitions of Kratzer (2010). The two definitions of would- and might-counterfactuals do not seem to be duals in the standard way, that is, that a might counterfactual 'if A, might B' is true iff 'if A, would $\neg B$' is false. As we saw before, there is evidence that under the truth-conditions presented 'if a different animal had escaped, it might have been a zebra' is false, so if the duality holds this means that 'if a different animal had escaped, it would not have been a zebra' is true. This, according to the definition means that every way of adding true propositions to the antecedent result in a set logically implying the consequent. This last

[4] I leave it open whether Kratzer would give a different formal treatment of this example. A treatment that does not render the counterfactual 'if a different animal had escaped, it might have been a zebra' false. However, the fact of the matter is that her intuitive explanation as to why 'if a different animal had escaped, it would have been a zebra' is false undermines itself.

thing does not seem to be the case though. If we add the true proposition that the animal that escaped was not an X, for every animal X different from a zebra, and the true proposition that these are all the animals, we seem to get a set that logically implies that the animal escaping was a zebra. Thus, it does not seem to be the case that the two definitions are duals, at least on an informal understanding of the truth-conditions.

However, to see whether this conjecture is in fact true of the theory, one would of course need to do a thorough investigation of the technical and formal details. We will not embark upon such an endeavor here, but suffice it to say that a lot of clarifying needs to be done before the truth-conditions, as they are stated without reference to formal and technical detail, make intuitive sense.

Chapter 3

How Facts Depend on Other Facts

Recall that in the Jones-example we were looking for a reason as to why the fact that Jones *is in fact wearing his hat* should not count when we evaluate the counterfactual ‘if the weather had been fine, Jones would have been wearing his hat’. And such a reason is exactly what Veltman (2005) presents us with. This is perhaps easiest to explain by using Veltman’s new version of Tichý’s Jones-example. Recall that in the original story, Jones invariably wears his hat if the weather is bad, but if the weather is fine, it is completely randomly decided. We now consider a variant of this story:

> “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 his hat in that case. [...] Now suppose that today heads came up when he flipped the coin, and that it is raining. So, again, Jones is wearing his hat.” (Veltman 2005).

The difference between the two scenarios is of course that now the counterfactual ‘if the weather had been fine, Jones would have been wearing his hat’ is intuitively true. But how can this be? If the fact that Jones is wearing his hat played no role in evaluating the counterfactual in the former scenario, how come it seems to matter in this new scenario? What exactly is the difference between the two? Veltman answers:

> “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. 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 wish to keep in as many independent facts as we can." (Veltman 2005).

The "similarity of particular fact" mentioned here is in relation to Lewis' theory, but as such this is not important. What is important is the difference between facts. Some facts depend on other facts, and when we give up the latter, we automatically give up the former. We see already that laws or generalizations play a special role in this framework. It is the "law" that Jones *invariably wears his hat when the weather is bad*, that makes us give up the assumption that he is wearing his hat, when we give up the assumption that the weather is bad.

Veltman's theory of counterfactuals is an attempt to formalize this insight. The framework uses a dynamic setting to capture the meaning of counterfactuals. In such a setting "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." (Veltman 2005). A cognitive state is simply a set of worlds — intuitively the worlds that the agent whose cognitive state we are talking about considers as candidates for the actual world — but because of the special role played by laws and generalizations, we need to define two sets of possible worlds. First, there is the set containing those worlds in which the laws and generalizations that the agent takes to be true hold. Second there is the set of worlds where not only all the laws and generalization hold, but where also the agents factual information is true. Veltman follows usual practise and models a possible world as a total function from the set of atoms into $\{0, 1\}$. A situation is just, in the standard way, defined as a partial such function.

> *Cognitive state*
> Let W be the set of possible worlds. A *cognitive state* S is a

pair $\langle U_S, F_S\rangle$, such that either (i) $\emptyset \neq F_S \subseteq U_S \subseteq W$ or (ii) $F_S = U_S = \emptyset$.

Here, the set U_S is the set of worlds in which all the general laws hold. It is therefore called *the universe* of the state S. Further, F_S contains the worlds that, for all the agent knows, could be the actual world, because it also contains the factual information of the agent. If the set F_S is empty, it means that there are no worlds left that can be the actual world. This state we call *the absurd state*, and denote it $\mathbf{0}$.[1] The state of complete ignorance, that is, the state $\langle W, W\rangle$, we denote by $\mathbf{1}$.

Now as mentioned above, the meaning of a sentence A is an operation on cognitive states. We denote the meaning of the sentence A as $[A]$, and the result of updating S with A as $S[A]$.
In a standard way we take $[\![A]\!]$ to be the proposition expressed by A, which in turn is just a set of possible worlds; the worlds where A is true. The proposition expressed by a propositional formula is thus given the standard recursive definition.

$$\begin{aligned}
[\![p]\!] &= \{w \in W | w(p) = 1\}, \text{ for atomic } p\\
[\![\neg A]\!] &= W \backslash [\![A]\!]\\
[\![A \wedge B]\!] &= [\![A]\!] \cap [\![B]\!]\\
[\![A \vee B]\!] &= [\![A]\!] \cup [\![B]\!]\\
[\![A \to B]\!] &= [\![\neg A]\!] \cup [\![B]\!]
\end{aligned}$$

Instead of always writing $w \in [\![A]\!]$ we will sometimes just say that A is true at w. The definition of updating a cognitive state with a sentence is as follows. Here, if A is a sentence, we will write $\Box A$ to mean "it is a general law that A".

Interpretation

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

(b) (i) $S[\Box A] = \langle U_S \cap [\![A]\!], F_S \cap [\![A]\!]\rangle$ if $F_S \cap [\![A]\!] \neq \emptyset$,
(ii) $S[\Box A] = \mathbf{0}$, otherwise.[2]

[1] It is clear from the definition of cognitive state that there can be only *one such* absurd state. We can therefore talk about *the absurd state*.

[2] Note that this does not allow stacking of boxes; that is, formulas of the kind $\Box\Box A$, which would mean "it is a general law that it is a general law that A", are not permitted in the language.

If we update with a fact it only affects F_S, whereas if we update with what is taken to be a general law it will affect both F_S and U_S.

A cognitive state supports a sentence when the information expressed by the sentence adds nothing to the state.

> *Support*
> S supports a sentence A, and we write $S \models A$, iff $S[A] = S$.

To capture what goes on when we make counterfactual assumptions Veltman first defines a basis for a world. A basis for a world w is just a minimal set of facts true at w such that this world is the only world with these facts true within U_S.

> *Basis*
> Let $S = \langle U_S, F_S \rangle$ be a cognitive state.
>
> (i) The situation s *forces* the proposition $[\![A]\!]$ within U_S iff for every $w \in U_S$ such that $s \subseteq w$ it is the case that $w \in [\![A]\!]$.
>
> (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.

> *Retraction*
> Again, let $S = \langle U_S, F_S \rangle$ be a cognitive state.
>
> (i) Let $w \in U_S$ and $[\![A]\!] \subseteq W$. The set $w \downarrow [\![A]\!]$ is the following: $s \in w \downarrow [\![A]\!]$ iff $s \subseteq w$ and there is a basis s' for w such that s is a *maximal* subset of s' not forcing $[\![A]\!]$.
>
> (ii) The *retraction* of $[\![A]\!]$ from S, $S \downarrow [\![A]\!]$, is the state $\langle U_{S\downarrow[\![A]\!]}, F_{S\downarrow[\![A]\!]} \rangle$ determined as follows:
>
> (A) $w \in U_{S\downarrow[\![A]\!]}$ iff $w \in U_S$,
>
> (B) $w \in F_{S\downarrow[\![A]\!]}$ iff $w \in U_S$ and there exist $w' \in F_S$ and $s \in w' \downarrow [\![A]\!]$ such that $s \subseteq w$.
>
> (iii) The state S[if had been A] is given by $(S \downarrow [\![\neg A]\!])[A]$.

> *Support of counterfactual*
> $S \models A \rightsquigarrow B$ iff S[if had been A] $\models B$.

The above formally captures the slogan that "facts stand and fall together". A counterfactual with antecedent A and consequent B is supported if the hypothetical state S[if had been A] supports B. And in turn we note that the hypothetical state S[if had been A] is determined in three steps. First, we give up the assumption that $\neg A$ is the case. Second, we give up anything that follows from this assumption. Third, we add that A is the case and everything that follows from this in accordance with the generalizations and laws. Also, we may note that while doing this we have not let go of any independent assumptions. The reason is that such independent assumptions would have to figure in the basis of the actual world (or world of evaluation), since if they did not, they would in turn not be *independent*.

We are now in a position to define when an agent will accept or assert a counterfactual. He will do so exactly when his cognitive state supports the counterfactual.

> *Acceptability/Assertability*
> An agent with cognitive state S will accept or assert a counterfactual $A \rightsquigarrow B$ exactly when $S \models A \rightsquigarrow B$.

Here we see that the third parameter S plays a role in determining when a counterfactual is acceptable/assertable. We will sometimes say that a counterfactual is true (or false) relative to a state S instead of saying that S supports (does not support) the counterfactual in question.

Now, if we believe that there is such a thing as a cognitive state reflecting all true information about the world, we can use this state to define *the* truth-value of a counterfactual. Let S' be this state containing all true information about the world, then we can define *the* truth-value of a counterfactual as follows:[3]

> *Truth*
> The counterfactual $A \rightsquigarrow B$ is true iff the state S' supports $A \rightsquigarrow B$.

This latter definition is not taken from Veltman (2005). Veltman does not speak of the truth of a counterfactual, but only about when a cognitive state supports a counterfactual.

[3] This is of course not as straightforward as it seems. The state S' is arguably a highly idealized state, and even if we believe that such a state as *the true state of the world* exists, it will most probably be more complex than what a cognitive state, as defined here, is able to reflect. However, if one believes in such a thing as *the* truth-value of a counterfactual, it seems that one needs to define it relative to exactly this state.

3.1 Examples

To see how these definitions work we look at some examples.

In the Jones-example we are only interested in the law that bad weather makes Jones wear his hat, and we know that the weather *is* bad and Jones *therefore is* wearing his hat. Let p and q be 'the weather is bad' and 'Jones is wearing his hat', respectively. Then the state we are interested in is $S = 1[\Box(p \to q)][p][q]$. We want to ask if this state supports 'if the weather had been fine, Jones would have been wearing his hat', that is, $\neg p \rightsquigarrow q$. The state S is depicted below.

	p	q
w_0	0	0
w_1	0	1
~~w_2~~	~~1~~	~~0~~
$\boldsymbol{w_3}$	**1**	**1**

The worlds that are not in U_S are struck through (in this case only w_2), while the worlds in F_S are marked in boldface (in this case only w_3). Now, a basis for w_3 is $\{\langle p, 1\rangle\}$, because this is a minimal set determining w_3 within U_S according to the above definition. We now want to know what the state $S \downarrow [\![p]\!]$ looks like. According to the definition the universe of this state is the same. Before calculating $F_{S\downarrow[\![p]\!]}$ we note that there is only one maximal subset of the basis $\{\langle p, 1\rangle\}$ *not forcing* $[\![p]\!]$ within U_S, namely $\emptyset$. This means that $w \in F_{S\downarrow[\![p]\!]}$ iff $w \in U_S$ and $\emptyset \subseteq w$. This in turn means that $S \downarrow [\![p]\!] = \langle U_S, U_S\rangle$. Hence, we update with $\neg p$ to get the state $(S \downarrow [\![p]\!])[\neg p]$ which is depicted below.

	p	q
$\boldsymbol{w_0}$	**0**	**0**
$\boldsymbol{w_1}$	**0**	**1**
~~w_2~~	~~1~~	~~0~~
w_3	1	1

Now since $F_{(S\downarrow[\![p]\!])[\neg p]} = \{w_0, w_1\}$ and q is false at w_0 it follows that $(S \downarrow [\![p]\!])[\neg p][q] \neq (S \downarrow [\![p]\!])[\neg p]$, which in turn just means that S[if had been $\neg p$] $\not\models q$. So by definition $S \not\models \neg p \rightsquigarrow q$, that is, the counterfactual 'if the weather had been fine, Jones would have been wearing his hat' is not supported by the state S, and hence, it is not acceptable by an agent with cognitive state S.

It is also easy to see that this approach also makes the right prediction in the new version of the Jones example. We now need to add an atom r representing 'the coin came up heads', and the law stating that if the coin comes up heads, then Jones is wearing his hat, that is, $\Box(r \to q)$. The state we are interested in is $S = 1[\Box(p \rightarrow q)][\Box(r \rightarrow q)][p][q][r]$ and it is given 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**

We see that w_7 has only one basis; $\{\langle p, 1\rangle, \langle r, 1\rangle\}$, and a maximal subset of this not forcing $[\![p]\!]$ within U_S is $\{\langle r, 1\rangle\}$. This means that $S \downarrow [\![p]\!] = \langle U_S, \{w_3, w_7\}\rangle$, and since of $w_3(p) = 0$ and $w_7(p) = 1$, it straightforwardly follows, in combination with $w_3(q) = 1$, that $(S \downarrow [\![p]\!])[\neg p] = (S \downarrow [\![p]\!])[\neg p][q]$, hence, $S \models \neg p \rightsquigarrow q$, that is, the counterfactual 'if the weather had been fine, Jones would have been wearing his hat' is supported by our state S, and it is supported exactly because we never give up the information that the coin came down heads, which can be seen from the fact that $\langle r, 1\rangle$ is in the basis of w_7.

One of the merits of this theory, according to Veltman, is that it is able to make predictions. That is, we can, given information about the laws and generalizations under consideration, and the facts true of the actual world, calculate whether a given counterfactual is supported — and if we believe that a state can contain all relevant true information, whether a counterfactual is true or false.
However, the theory does not always predict the intuitively right thing, as we will see in the next section.

3.2 The Problems

Veltman himself points to a case where his theory seems to give the wrong prediction. The story is from Kratzer (2010).

> "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 2010).

The theory of Veltman will predict that the counterfactual 'if the flag were up, the King would be in the castle' is false relative to the relevant state, and that the counterfactual 'if the flag were up, the lights might be out' is true relative to a state having the facts as described above along with the law $Flag \wedge Light \rightarrow King$.[4] This goes against our intuition, and in fact the problem is more general. As Schulz observes:

> "Veltman (2005) also discusses another type of example his approach has troubles with. These are *would have* conditionals that are based on a law that concludes from the truth of two premises to the truth of the consequent: $prem1 \wedge prem2 \rightarrow cons$. The critical predictions turn up when in the evaluation world the first premise is true, the second false, and the consequent false as well. In such a context a *would have* conditional *If the second premise had been true as well, the consequent would have been true* is sometimes intuitively true. Veltman's approach, however, in general predicts that in such a situation the conditional is false. The reason is that the basis of the described evaluation world consists of the true premise and the false consequent." (Schulz 2007).

As can be seen this fits the King of Bavaria example. However, this example has problems of its own. The theory fails when it takes the law in question to be that "whenever the flag is up and the lights are on, the King is in the castle", that is, on a formal level $Flag \wedge Lights \rightarrow King$. However, it is by no means clear that this is the law we are dealing with here. In fact, the law seems to be that whenever the King is in, then the flag will be up and the lights will be on, or perhaps even both laws combined. Even though the approach of Veltman does not give the right prediction with these laws either, we will not discuss the King of Bavaria example, simply because it is not totally clear what goes on there.

As Schulz mentions in the above quote, the approach of Veltman have problems giving the right predictions when we want to evaluate counterfactuals such as $prem2 \rightsquigarrow cons$ in the described scenario. The reason is that in such a case a basis for the world would be $\{\langle prem1, 1\rangle, \langle cons, 0\rangle\}$, and since this forces the second premise to be false, we have two maximal

[4] Here, we just define a state S to support a counterfactual 'if it had been the case that A, it *might* have been the case that B' iff $S[\text{if had been } A][B] \neq \mathbf{0}$. This definition is taken from Veltman (2005).

subsets of this such that this is not the case; $\{\langle prem1, 1\rangle\}$ and $\{\langle cons, 0\rangle\}$. But that the latter is a maximal subset in turn means that some world w in $F_{S[\text{if had been } prem2]}$ (for the appropriate S) is such that $\{\langle cons, 0\rangle\} \subseteq w$, and hence $prem2 \rightsquigarrow cons$ false.
Perhaps this is easier explained with a version of Lifschitz' Circuit-example.

> Suppose there are two switches and a light. The light is on when both switches are up. Right now the first switch is up, the second down, and the light out.

Now we ask, is the counterfactual expressed by "if switch two were up, the light would be on" true? Undoubtedly your intuitions say *yes*. But, unfortunately the theory of Veltman predicts it to be false relative to the relevant state. Let $s1$, $s2$, and l represent 'switch one is up', 'switch two is up', and 'light is on', respectively. Then we are interested in the state $S = 1[\Box(s1 \wedge s2 \rightarrow l)][s1][\neg s2][\neg l]$. This is depicted below.

	$s1$	$s2$	l
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**
w_5	1	0	1
~~w_6~~	~~1~~	~~1~~	~~0~~
w_7	1	1	1

Now a basis for w_4 is $\{\langle s1, 1\rangle, \langle l, 0\rangle\}$, which forces $s2$ to be false. Hence, $w_4 \downarrow [\![\neg s2]\!] = \{\{\langle s1, 1\rangle\}, \{\langle l, 0\rangle\}\}$, which in turn means that $S \downarrow [\![\neg s2]\!] = \langle U_S, \{w_0, w_2, w_4, w_5, w_7\}\rangle$ and $(S \downarrow [\![\neg s2]\!])[s2] = \langle U_S, \{w_2, w_7\}\rangle$. Since $\{\langle l, 0\rangle\} \subseteq w_2$ it follows that $(S \downarrow [\![\neg s2]\!])[s2] \neq (S \downarrow [\![\neg s2]\!])[s2][l]$, and therefore that the counterfactual 'if switch two were up, the light would be on' is false relative to S. The prediction of this theory seems to go wrong because it does not distinguish between $s2$ and l in any sense. These two facts are on a par, and were it the case that switch two is up, then it would be the case that *either* switch one is down *or* the light is on. The problem seems to be that while we understand an asymmetry in the example — the light being on depends on both switches being up in a way that the switches being up does *not* depend on the light being on — the formal set up does not seem to take this asymmetry into account. As we will see later, Schulz proposes that this asymmetry is due to the fact that the two switches being up *causes* the light to be on. The asymmetry is thus one of *cause and effect*. Later we will explain the asymmetry present here in a different way. One that does not necessarily ground itself in *cause and effect*, but which can be interpreted that way.

While it was not clear in the King of Bavaria example what exactly was going on, and intuitions about the truth or falsity of counterfactuals were not that strong and perhaps not entirely trustworthy, it seems that in this case it is clear what is going on and that the intuitions saying that 'if the second switch were up, the light would be on' should be true are very strong. Veltman admits that his theory gives false predictions in cases such as this, and that this goes against some very strong intuitions concerning the acceptability of these counterfactuals. However, he does draw attention to a case, seemingly with the same logical form, in which we do not want the counterfactual $prem2 \rightsquigarrow cons$ to be true. This is the aforementioned Three Sisters example in which three sisters have to share one bed only big enough for two of them, so that whenever Ann and Billie are in bed, Carol is on the floor.

As Veltman points out, nobody seems to be willing to accept the counterfactual 'if Ann were in bed, Carol would be on the floor' in the described scenario. It might as well have been Billie on the floor. However, if one takes the law to be $Ann \wedge Billie \rightarrow \neg Carol$, where Ann is just 'Ann is in bed' etc., it seems that this example has the same logical form as that above. It is straightforward to see that we have that 'if Ann were in bed, Carol would be on the floor' is false relative to the state $S = 1[\Box(Ann \wedge Billie \rightarrow \neg Carol)][\neg Ann][Billie][Carol]$. (The evaluation procedure is the same as in the example given above). What the theory predicts to be true (relative to the appropriate state S) is the — in this context — much more acceptable 'if Ann were in bed, then *either* Billie *or* Carol would be on the floor'. So, for this example it seems that Veltman's theory has the correct answer, and as we will later see, the framework of Schulz (2009) has some problems with this example.

We will later argue that these two examples do not have the same form. While it is true that whenever Ann and Billie are in bed, Carol is on the floor, the Three Sisters-example provides us with much more information than that. Once we take this information into account we can predict that the counterfactual 'if Ann were in bed, Carol would be on the floor' is unacceptable, while the counterfactual 'if switch two were up, the light would be on' is acceptable.

3.3 Is this Causality?

As we just saw, in the theory of Veltman a counterfactual is true relative to a state where some laws and generalizations are assumed about the world. These laws and generalizations are what the agent takes to be true about the world. As such there is nothing saying that these laws and generalizations have to be related to the concept of causality, nor is there anything saying that they cannot.

Lewis (1973b) proposes a definition of causality in terms of counterfactuals. On a naive view this goes as follows.

> C causes E, or E is an effect of C iff the counterfactual 'if C had not happened, then E would not have happened' is true.[5]

It is immediately clear that for this definition to be non-circular, the truth-definitions of counterfactuals cannot make reference to causal connections; at least not causal connections in any objective sense. The question now is whether the truth-conditions of ones theory make reference to such causal connections in a way that make these ultimately circular; and furthermore, circular in a *vicious* way. There are of course various replies to this question according to how one conceives of the relation between causality and counterfactuals. One can simply deny that causality can be defined in terms of counterfactuals, thus one need not worry as to whether notions of causality play a role in determining the meaning of counterfactuals. This is the approach taken by Pearl (2000). Pearl takes causality to be the more primitive notion of the two and tries to define counterfactuals in terms of causal networks. If one, however, believes that causality should be properly reduced to counterfactuals one can take one of two broad roads. One can either try to show that ones truth-conditions do not make any reference to causality. Or, one can show, that in case ones truth-conditions cannot be said to make no reference at all to causality, the reference does not amount to any vicious circularity.

[5] This is a naive view because it is overly simplified and therefore has obvious counterexamples. Take for instance the following: John and Susy both shoot their rifles at a bottle. Both are accurate aimers, but because John pulls the trigger before Susy, his bullet hits and shatters the bottle. So, we would like to say that Johns shooting caused the bottle to shatter. But, the counterfactual 'if John had not shot, the bottle would not have shattered' is intuitively false, because in the case of John not shooting, Susy would have hit the bottle with her shot. Lewis is of course aware of this, and this is why I call the above a naive view. (Lewis 1973a). However, how to exactly define causality in terms of counterfactuals is a delicate matter, and as such we will not discuss it here. What is important here is the question of whether one can in fact give the meaning of counterfactuals without a prior commitment to causality or causal laws.

The thing to note is that a counterfactual is always true *relative* to a cognitive state of some agent, and the content of a cognitive state is what the agent takes himself to know about the world. If one wishes to define causality in an objective way, one will need to have absolute truth-conditions for counterfactuals. We have proposed to have absolute truth-conditions given relative to an idealized state S' comprising all relevant information to the evaluation of the counterfactual. The question is therefore whether the information in the state S' can be given in non-causal terms in order to avoid a circular definition. And while there is nothing in principle saying that this cannot be done, we will later see that some relations which we need to evaluate counterfactuals *do* seem to be intimately linked to causality in a way that makes it hard to argue that they are not.

As we will later see, Schulz (2009), who proposes to define the meaning of counterfactuals in terms of causal networks, takes a somewhat different approach to the relationship between causality and counterfactuals. Even though she proposes to define the meaning of counterfactuals in terms of prior knowledge about causal relations, she does not wish to say that causality is not dependent on counterfactuals at all. The above definition of causality is in fact not a genuine definition. Instead we might understand it as a way of testing whether a causal relationship obtains. Thus the claim is nothing more than that one should separate the epistemic and ontological aspects of the relationship between causality and counterfactuals. On an epistemic level we may use counterfactuals as test cases to see whether a causal relationship obtains in the world. But before we get into a discussion of these questions, we will present the framework of Schulz (2009) in the next section.

Chapter 4

Causal Entailment

The theory of Schulz comes in two different variants: Schulz (2007) and Schulz (2009). Here we will present the theory of Schulz (2009), since this is the newer and simpler of the two. The main idea behind Schulz' theory is a redefinition of the notion of *causal model* found in Pearl (2000). Schulz calls her *causal model* counterpart a *dynamics*, which will be defined later. First, we need to say something about the language.

> *Language*
> Given a finite set of propositional atoms $\mathcal{P}$, we define the language $\mathcal{L}^0$ to be the closure under the standard logical connectives. The language of counterfactuals $\mathcal{L}^{\leadsto}$ is then defined as the union of $\mathcal{L}^0$ and the set of expressions $\varphi \leadsto \psi$, where $\varphi, \psi \in \mathcal{L}^0$.
>
> *Interpretation*
> The language $\mathcal{L}^{\leadsto}$ is interpreted using strong Kleene three-valued logic. We have the truth values 0 (false), 1 (true), and u (undefined). Undefined can develop into true or false; hence we have the partial order $u \leq 0$ and $u \leq 1$. The truth tables for $\neg$ and $\wedge$ are given below.

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

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

Situations and Worlds
An assignment of u, 0, and 1 to the propositional atoms is called a *situation* for $\mathcal{L}^{\rightsquigarrow}$. If the assignment does not make use of u, we call the assignment a *possible world* for $\mathcal{L}^{\rightsquigarrow}$.

The set of all possible worlds is denoted W and $\llbracket\varphi\rrbracket^D$ denotes the set of possible worlds where φ is true (relative to D). For formulas $\varphi \in \mathcal{L}^0$ we have $\llbracket\varphi\rrbracket^D = \llbracket\varphi\rrbracket$, and the truth-value of φ is just given by recursive definitions of strong Kleene three-valued logic. However, if φ contains the connective $\rightsquigarrow$, D, which will be defined in due course, plays a role in determining the truth-value.
$\llbracket\varphi\rrbracket^{D,s}$ denotes the truth-value of φ in the situation s (relative to D).

The central claim of Schulz is that before we can start to evaluate a counterfactual, we must have prior knowledge of the causal dependencies that are relevant in the given context. Recall that the theory of Veltman was unable to give the right prediction for the counterfactual ‘if switch two were up, the light would be on’ in the scenario mentioned before. We were not able to give this a different truth-value than the counterfactual ‘if switch two were up, switch one would be down’, even though this latter is intuitively false and the former true. According to Schulz this is a result of neglecting to pay attention to the causal dependencies of the example: the switches being up *causes* the light to be on, and it is this direction in the flow of causality which makes us judge the first counterfactual true and the second false. To represent the causal dependencies Schulz introduces the notion of a *dynamics*. Here we distinguish two types of variables, B, the set of background variables, and I, the set of inner variables. As will be clear from the definition, a variable is just a propositional atom.

Dynamics
A dynamics is a tuple $D = \langle B, F \rangle$, such that

1. $B \subseteq \mathcal{P}$ is the set of *background variables*.
2. F is a function mapping elements X of $I = \mathcal{P} \setminus B$ to tuples $\langle Z_X, f_X \rangle$. Here Z_X is an n-tuple of elements of $\mathcal{P}$ and f_X is a two-valued truth-function mapping a truth-value on X in accordance with the elements of Z_X, that is, $f_X : \{0,1\}^n \rightarrow \{0,1\}$. Furthermore, F is *rooted* in B.

Intuitively, the members of Z_X are the variables that X depends causally on, and f_X specifies the nature of that causal dependency.

The definition looks much more complicated than it is. To illustrate we can use the variant of Lifschitz' Circuit-example that we have seen before. We are dealing with three propositional atoms, $s1$, $s2$, and l, representing the same as before. Now, a dynamics for the situation is the following: l depends on $s1$ and $s2$, so the latter are members of the background variables B, since these do not depend on other variables. That is $D = \langle \{s1, s2\}, F \rangle$. Now F will map l to the tuple $\langle \{s1, s2\}, f_l \rangle$, which again, is because the value of l is causally dependent on the values of $s1$ and $s2$. Now, f_l specifies this dependency. We know that the light is *caused* by the switches to be on exactly when both are up. So, it follows that f_l is the function such that l is mapped to 1 exactly when $s1$ and $s2$ both have the value 1, and mapped to 0 otherwise.

Now there are further constraints on the dynamics D. Because we are representing causal dependencies, it follows that the background variables cannot themselves depend on a variable that originally depended on that background variable, or, in other words, the causal dependencies cannot be circular. In the above example this is just saying that $s1$ and $s2$ cannot causally depend on the value of l, since l was causally dependent on $s1$ and $s2$ in the first place. The definition of rootedness takes care of this.

> *Rootedness*
> Let $D = \langle B, F \rangle$ be a dynamics. Define the relation $\leq$ such that for $X, Y \in \mathcal{P}$; $X \leq Y$ if X occurs in Z_Y. Now let $\leq^T$ be the transitive closure of $\leq$. F is rooted in B if $\langle \mathcal{P}, \leq^T \rangle$ is a poset and B equals the minimal elements of this set.

We now need to define the notion of *causal entailment* that according to Schulz is the entailment-relation with which the antecedent and some facts to be specified later, entail the consequent. To define this notion, however, we need first to define the operation $\mathcal{T}$.

> *The $\mathcal{T}$-operator*

> Let D be a dynamics and s a situation. The situation $\mathcal{T}_D(s)$ is defined as follows. For $p \in \mathcal{P}$
>
> 1. If $p \in B$, then $\mathcal{T}_D(s)(p) = s(p)$.
> 2. If $p \in I = \mathcal{P} \setminus B$ and $Z_p = \langle p_1, p_2, ..., p_n \rangle$, then
>   (a) If $s(p) = u$ and $f_p(s(p_1), ..., s(p_n))$ is defined, then $\mathcal{T}_D(s)(p) = f_p(s(p_1), ..., s(p_n))$.
>   (b) If $s(p) \neq u$ or $f_p(s(p_1), ..., s(p_n))$ is not defined, then $\mathcal{T}_D(s)(p) = s(p)$.

Intuitively the operator $\mathcal{T}$ simply calculates the causal effects of a situation s, and produces a new situation where the effects are realized. Now it can be proven that the operator $\mathcal{T}$ will reach a fixed point in finitely many steps and that this fixed point is unique.[1] Given a set of literals Γ we define s_Γ to be the situation making all the literals of Γ true, while the propositional atoms not mentioned in Γ are given the value u.[2] Further, given a dynamics D, we let $s*$ denote the fixed point of $\mathcal{T}_D$ when applied to s. We can now define causal entailment.

> Let D be a dynamics, Γ a set of literals, and φ a formula. We say that Γ causally entails φ (given D) and we write $\Gamma \models\!\!\!\models_D \varphi$ according to the following definition: $\Gamma \models\!\!\!\models_D \varphi$ iff $[\![\varphi]\!]^{D, s_\Gamma *} = 1$.

In words, Γ causally entails φ if φ is true in the situation obtaining when we calculate all the causal consequences of the literals in Γ.

We can now almost give the truth-conditions for counterfactuals as they are given by Schulz, however, we need to define what it is to be a basis of a world. Unsurprisingly, the notion of basis is also causal and it is therefore given relative to a dynamics D.

> *Basis*
> Let w be a possible world and D a dynamics. The basis for w (relative to D), $b_{w,D}$, is the minimal set of literals Δ such that $s_\Delta * = w$.

Intuitively the basis is a set of literals such that all other facts about the world follow via the causal dependencies expressed in the dynamics. It can be proven that a basis always exists, and that it will always be unique.

[1] A fixed point is a point where the operator $\mathcal{T}_D$ returns the same output as was the input. A fixed point is thus a situation, s, such that $\mathcal{T}_D(s) = s$.

[2] We leave details of consistency and the like untouched.

Therefore it makes sense to speak of *the* basis. We may also note that this is one of the points where the difference of this theory compared to that of Veltman (2005) is obvious. We are now dealing with a unique basis, whereas in Veltman (2005) we might have more than one basis. In Veltman's framework the agent, given his knowledge of the laws and generalizations, would be able to deduce which world was actual given a basis for the world. That is, the notion of basis relates to *excluding* other possible candidates for the actual world. With Schulz we see that this is not enough. The facts in the basis are the facts from which all other facts *follow*, so it is not enough to have information that permits us to exclude all other than the actual world, we further need exactly the facts that causally entail all other facts. We will reuse this notion of basis when we define our own semantics, and as such the differences between the two will be discussed later.

The last thing we need to define before the truth-conditions can be given is a function that revises a basis of a world with the antecedent of a counterfactual.

> *Revision*
> Let $A \in \mathcal{L}^0$ and $\Delta \subseteq \mathcal{L}^0$. The revision of Δ with A, $Rev(A, \Delta)$, is given as the set of sets $\Delta' \cup \{A\}$, where Δ' is a maximal subset of Δ logically consistent with A.

We may note that the revision function simply selects the maximal subsets of Δ that are logical consistent with A. Thus, for now all reference to causal dependencies have disappeared.

We can now give the truth-conditions for a counterfactual according to Schulz:

> Let w be the world of evaluation and D a dynamics describing the causal dependencies in w. A counterfactual $A \rightsquigarrow B$ is true (relative to D) iff $\forall \Gamma \in Rev(A, b_{w,D}) : \Gamma \models\!\!\!\equiv_D B$.[3]

The upshot of the definition is this: when we evaluate the counterfactual $A \rightsquigarrow B$, we break of the causal history leading to A and simply stipulate its

[3] The reader might have noticed that this definition presupposes that A is itself a literal and as such is inadequate. If A is not a literal then there is nothing guaranteing that the revision function will return a set of literals, and thus, since the causal notion of entailment only takes as input a set of literals, it follows that the definition, as given here, is inadequate. Since all the counterfactuals we will discuss in the present thesis have literals as antecedents, we will not discuss these matters further. Schulz of course takes these matters into account and for the solution see Schulz (2009).

truth. We then make minimal adjustments to secure consistency and check to see if B is a causal consequence of this "intervention".

It is clear from the definition of the operator $\mathcal{T}$ that even if we counterfactually assume the variable A to be true, we will never change the value of the causes for this variable. This is to be expected when we deal with ontic counterfactuals, since these talk about what would have been the case had some fact about the world been otherwise. Take for instance the counterfactual 'if he had left the interview smiling, then it would have to have gone well'. This is false on the ontic reading. Whether or not he leaves the interview smiling will have no bearing on whether the interview went well. Instead, the former is — under normal circumstances — an sign of the latter, and the theory therefore rightly predicts that the conditional is false. Changing how he left the interview does not change how the interview went on a causal interpretation. All of this is as it should be, since the theory of Schulz is only intended to cover what she calls the *dominant* reading of counterfactuals; the ontic reading. What can be said is that the conditional 'if he left the interview smiling, it would have to have gone well' is true on an epistemic reading. If we learn that he left the interview smiling, then we are certainly justified in concluding that it went well.

As we can see, the truth-definitions are relative to the relevant causal dependencies, that is, a dynamics D. This is of course nothing new, since we already had with Veltman's theory that counterfactuals are true or false relative to some generalizations and laws true of the world. However, this time we are being very definite with what in Veltman's theory was laws and generalizations. Here they are *causal dependencies* represented by a dynamics. We will later discuss what we should take this to mean and what import this has on the discussion of counterfactuals versus causality.

Because of the fact that the truth of a counterfactual is given relative to a dynamics D, we might again say that the truth of a counterfactual relative to a dynamics is to be equated with the acceptability/assertability of the counterfactual in question by an agent who takes D to express the true causal dependencies of the world.

The framework of Schulz is designed to solve cases such as the variant of Lifschitz' Circuit-example given above. So, what could be more appropriate than to illustrate how these definitions work using that example?

Recall that in this example we have two switches and a light. The light is on exactly when both switches are up. We thus have $s1$, $s2$, and l, representing 'switch one is up', 'switch two is up', and 'light is on'. We have already established that a dynamics for this situation is such that

$D = \langle\{s1, s2\}, F\rangle$, where $F(l) = \langle\{s1, s2\}, f_l\rangle$, and f_l is such that l gets mapped to true iff $s1$ and $s2$ are both true.
Right now the situation is such that $s1$, $\neg s2$, and $\neg l$ is the case. This means that a basis for the actual world, call it @, is $b_{@,D} = \{s1, \neg s2\}$. But, then the only set in $Rev(s2, b_{@,D})$ is $\{s1, s2\}$. And, it is easy to see that $\{s1, s2\} \mathrel{|}\!\!\!\equiv_D l$, since the first application of $\mathcal{T}_D$ to $s_{\{s1,s2\}} = \{\langle s1, 1\rangle, \langle s2, 1\rangle, \langle l, u\rangle\}$ yields $\{\langle s1, 1\rangle, \langle s2, 1\rangle, \langle l, 1\rangle\}$, which is also the fixed point for the operator $\mathcal{T}_D$. It follows that the counterfactual $s2 \rightsquigarrow l$ is true in the described scenario under this approach.

So, we now have a theory able to give the correct prediction in the troublesome cases that Veltman's theory was not able to handle. These were exactly the cases where two premises, *prem*1 and *prem*2, *causes* an effect, *cons*, along with the facts being such that *prem*1 is true, and *prem*2 and *cons* are false. In such a situation Veltman's theory will wrongly predict the falsity of $prem2 \rightsquigarrow cons$, whereas it is just as easy to see that — given that the described causal dependence is the only one — Schulz' framework will predict it to be true.
It therefore seems that in a direct comparison we should prefer the theory of Schulz above that of Veltman — at least as long as we do not have any preferences as to whether to define causality in terms of counterfactuals or not. However, Schulz' theory have some cases of its own where things do not go as smoothly as in the above example. We will discuss such cases later.

4.1 Manipulation and Control

At first sight it seems that the truth-conditions make use of the notion of causality, and so that we cannot on this approach stick to the contention that causality should be defined in terms of counterfactuals. Schulz notes that her semantics — along with other semantics for counterfactuals — make use of the notion of *dependencies*.

> "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 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." (Schulz 2009).

It is clear that the relation of manipulation and control mentioned here is just the relation which is formally expressed by A being one of the variables in Z_C. What Schulz is doubting in the above quote is whether a relationship of this sort can rightly be called a causal relationship. She presents an example to illustrate that it is not clear that it always can. This is the Math example.

> Suppose you have '3 + 4 = 7' written on a board somewhere. You now utter the two following conditionals:
> (1) If the first number had been even, the result would have been even.
> (2) If the result had been even, the first number would have been even.

The point is that the first conditional is intuitively true, while the second is intuitively false, since why wouldn't the second number be odd instead. We can explain this using the idea of manipulation and control. Changing the first and second number changes the result in a systematic way. However, changing the result does not change the first number in a *systematic* way; it only does so when the value of the second number is fixed. This means that in the dynamics describing this situation we would have the result depending on the two numbers. This is arguably not a relation of causality, at least not causality as we normally understand it. It is simply a relation of manipulation and control. By manipulating only the first number, we are *sure* what will be the outcome, whereas if we manipulate the result, we have different possibilities for the input; that is, the first and second number. The idea that when evaluating a counterfactual we take into account dependencies of this kind, that is, dependencies that gives us control over the outcome, will be central to the semantics we are to develop later, and as such it owes a great deal to Schulz for pointing this out.

Schulz points out that in an example such as the above the dependencies are not causal but rather dependencies of manipulation and control. What this means is also that we are not excluding causal dependencies, because we must assume that causal dependencies can be represented as relations of this former sort; relations of manipulation of control.

But if these relations of manipulation and control are sometimes causal, then if we wish to build our semantics upon these, we must give up the idea that causality is defined in terms of counterfactuals, unless we are content with a circular definition.

Schulz proposes to separate two sides of the counterfactual versus causality debate. First, there is the side pertaining to content; that is, the question of what can be defined in terms of what. And it is clear that if the meaning of a counterfactual is inherently given using causal relationships, then we *cannot* define causality in terms of counterfactuals.

> "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. 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 we impose on reality to make rational behavior possible." (Schulz 2009).

It is clear from the above quote that Schulz does not find it viable to try to define causality in terms of counterfactuals. As such, causality, as it is used in defining the meaning of counterfactuals, is instead a *heuristics*; a tool which we impose on reality to make rational behavior possible.

However, in addition to the content related side, there is also the epistemic side of the debate, where the question is how we establish the truth of a counterfactual or how we establish that a causal relationship obtains in the world. According to Schulz it might just be that counterfactuals are useful in determining when a causal relationship obtains in the world. As such they are not defining causality, they just make very good test cases for causal relationships.

With this said, there is of course also a second option when it comes to settling the content and epistemic sides of the debate. We can stick to the claim that causality is in fact a notion that is defined in terms of counterfactuals, but, that the truth-conditions of Schulz giving the meaning of a counterfactual do not *strictly speaking* provide the meaning of a counterfactual. The meaning of a counterfactual is not related to causal relationships in the way mentioned, but are in fact deprived of any relation to causality. However, using already established causal relationships provides a very useful tool to assess the truth-value of a counterfactual. The meaning of a counterfactual is therefore something different from the one presented above, but the truth of a counterfactual will coincide with the evaluation method prescribed by the conditions stated. We will briefly return to this discussion later, but for now we note and state the two different possibilities mentioned above:

> (1) The meaning of a counterfactual is determined in relation to causal relationships in the world. However, counterfactuals

are still a useful tool in establishing when a causal relationship obtains in the world. In this way, causality is a more primitive notion than counterfactuals.
(2) Causality is defined in terms of counterfactuals. However, causal relationships provides a useful tool in evaluating a counterfactual for truth, even though, strictly speaking, causality is not a factor in establishing the meaning of a counterfactual. This way counterfactuals are the more primitive notion.

Chapter 5

The Semantics

We are now ready to get to the part of this thesis which is new; the introduction of a new semantics for counterfactuals. The semantics is as such just an amendment to the theory of Veltman (2005), but this will all be clear in due course. The ideas used to develop this semantics owe a lot to the theories of Veltman and Schulz, which is also part of the reason why these two theories have been presented so thoroughly.

We set up this semantics in a dynamic framework as well. We therefore start by defining a state, just as in Veltman (2005).

> *State*
> Let W be the set of possible worlds. A *state* S is a triple $\langle U_S, F_S, G_S \rangle$, such that either (i) $\emptyset \neq F_S \subseteq U_S \subseteq W$ or (ii) $F_S = U_S = G_S = \emptyset$.

The idea behind this is the same as in Veltman (2005). U_S is the set of worlds where the relations between the facts (what we have called laws and generalizations before) that the agent takes to be true about the actual world hold. F_S is a subset of U_S that, besides the information about generalizations, also encodes the information that the agent takes to be facts about the actual world. G_S is a set of *generation relations* that will be defined later.

Worlds are again just total functions from the atoms of the language into $\{0, 1\}$ and situations are partial such functions.

Instead of having laws similar to the framework of Veltman we introduce a new notion, the notion of a *generation relation*. This is a relation that holds between two sets of literals. If the sets X and Y are in this relation we write $gen(X, Y)$ in the formal language. Intuitively we take $gen(X, Y)$

to mean that the literals mentioned in X *completely determines* the literals in Y. For instance, in the example with Jones and his hat we have that bad weather completely determines that Jones is wearing his hat. We thus have the generation relation obtaining between $\{bad\}$ and $\{hat\}$ such that $gen(\{bad\}, \{hat\})$. We note that the relation is not symmetric; $gen(X, Y)$ is a different piece of information than $gen(Y, X)$. It is easy to see why this should be so. In the Jones example it is clear that bad weather completely determines that Jones is wearing his hat, but that Jones is wearing his hat does not completely determine that the weather is bad; the weather may be perfectly fine. The analogy to Schulz' idea of *manipulation and control* should be obvious. When we have a generation relation between two sets of literals, it just intuitively means that by making the literals in the first set true, we have total control over the second set. However, where in Schulz framework, if Y is dependent on X we would know the value of the literals of Y for all possible valuations of the literals in X. In this framework, we can only be certain of the value of the literals in Y for some valuations of the literals in X. This is just to say that if $gen(X, Y)$ and the literals in X are not true, we have no information of whether the literals of Y will be true. This will later help us make the right prediction in examples such as that of Jones and his hat, where it is undetermined whether Jones is wearing his hat when the weather is fine.

The generation relations can thus be seen to carry some of the information that in Schulz' framework was encoded into the dynamics. However, we have no conditions of non-circularity on the generation relations. In fact, intuitively no such conditions should exist. That the dice is presently showing six on the upside completely determines that one is not showing, i.e. one is faced against the table, and vice versa. We therefore have two generation relations in the scenario such that $gen(\{\text{‘dice shows six on upside’}\}, \{\text{‘one is faced against the table’}\})$ and $gen(\{\text{‘one is faced against the table’}\}, \{\text{‘dice shows six on upside’}\})$.

The relation to Schulz' ideas will be discussed at greater length when we discuss “the philosophical” aspects of the generation relation. For now we will continue with the definition of the semantics to get the broader idea.

We now define the closure of a situation s with respect to a set of generation relations, G. If X is a set of literals, we take X^* to be the situation making all and only the literals of X true.[1]

Closure
The closure of a situation s with respect to the set G, $cl_G(s)$, is

[1] Technically, for all atoms p, $\langle p, 1\rangle \in X^*$ iff $p \in X$ and $\langle p, 0\rangle \in X^*$ iff $\neg p \in X$.

the *minimal* set such that:[2]

1. $s \subseteq cl_G(s)$.
2. $\forall X, Y$ if $X^* \subseteq cl_G(s)$ and $gen(X,Y) \in G$, then $Y^* \subseteq cl_G(s)$.

It is easy to prove that the closure is unique, and so it makes sense to speak of *the* closure.

Proof. Suppose not. Then there is a situation s, and two other situations s' and s'' such that $s' \neq s''$ and s' and s'' both fulfill the above conditions of being the closure of s. Let $A = s' \cap s''$. Now since $s' \neq s''$, $s' \not\subseteq s''$, and $s'' \not\subseteq s'$ it follows that $A \subset s'$ and $A \subset s''$. So, if we can prove that A fulfills the two conditions, we are done, since then s' and s'' are *not* minimal sets fulfilling the conditions, which is a contradiction.
First condition: Since $s \subseteq s'$ and $s \subseteq s''$ it immediately follows that $s \subseteq A$. ✓
Second condition: Suppose A does not fulfill it. Then there is X, Y such that $X^* \subseteq A$, $gen(X,Y)$, but $Y^* \not\subseteq A$. The former means that $X^* \subseteq s'$ and $X^* \subseteq s''$, so since s' and s'' both fulfill the second condition, we have $Y^* \subseteq s'$ and $Y^* \subseteq s''$. But then $Y^* \subseteq A = s' \cap s''$. Contradiction since we supposed $Y^* \not\subseteq A$. ✓ □

Again, we write $[\![A]\!]$ for the proposition expressed by A. We then have:

$$\begin{aligned}
[\![p]\!] &= \{w \in W | w(p) = 1\}, \text{ for atomic } p \\
[\![\neg A]\!] &= W \backslash [\![A]\!] \\
[\![A \wedge B]\!] &= [\![A]\!] \cap [\![B]\!] \\
[\![A \vee B]\!] &= [\![A]\!] \cup [\![B]\!] \\
[\![A \rightarrow B]\!] &= [\![\neg A]\!] \cup [\![B]\!]
\end{aligned}$$

We can now define an update rule for both sentences and generation relations.

Interpretation
If A is a sentence, then:

1. $S[A] = \langle U_S, F_S \cap [\![A]\!], G_S \rangle$ if $F_S \cap [\![A]\!] \neq \emptyset$,
2. $S[A] = \mathbf{0}$, otherwise.

[2] When it is obvious what set of generation relations we are taking the closure relative to we will leave out the subscript and just write $cl(s)$.

However, for the update with generation relations we can not just reuse the definition of Veltman (2005).

> The state $S[gen(X,Y)]$ is defined as follows.
>
> 1. $w \in U_{S[gen(X,Y)]}$ iff $w \in U_S$ and $\forall s \subseteq w$: $cl_{\{gen(X,Y)\}}(s) \subseteq w$.
> 2. $w \in F_{S[gen(X,Y)]}$ iff $w \in U_{S[gen(X,Y)]}$ and $w \in F_S$.
> 3. $G_{S[gen(X,Y)]} = G_S \cup \{gen(X,Y)\}$

This has as a consequence that for any state $S = \langle U_S, F_S, G_S \rangle$, it will always be the case that: $\forall w \in U_S$, if $s \subseteq w$, then $cl_{G_S}(s) \subseteq w$. As can be seen G_S simply contains all the generation relations that we have updated with.[3] We need to store this information because our notion of a basis for w will differ from that of Veltman (2005).

> *Basis*
> A basis, b, for a the world w is a minimal situation such that $cl_{G_S}(b) = w$.

Since the notion of the closure of b only makes sense relative to the generation relations we are dealing with, it is clear why we need to store this information in G_S.

It is clear that a basis is not necessarily unique. Take the example from before, where p and q represents 'the dice roll was a six' and 'one is facing the table', respectively. Suppose now the dice is showing a six on the upside, so one faced against the table. A basis in this case could either be the set $a = \{\langle p, 1\rangle\}$ or the set $b = \{\langle q, 1\rangle\}$, because $cl(a) = cl(b) = \{\langle p, 1\rangle, \langle q, 1\rangle\}$.

A revised basis with respect to a proposition $[\![A]\!]$ is just a maximal subset of a basis not forcing A to be the case. The definition of forcing is the same as in Veltman 2005.

[3] Because of this it is clear that U_S is definable in terms of G_S and the set of possible worlds W. Let G_S and W be given, then U_S is the set such that: $w \in U_S$ iff (1) $w \in W$ and (2) if $s \subseteq w$, then $cl_{G_S}(s) \subseteq w$. We have chosen to define a state using all three notions — that is F_S, U_S, and G_S — but as such we could have defined a state as a tuple $S = \langle F_S, G_S \rangle$ and defined U_S as a derivative notion.
For now it suffices to note that it would have made no difference at all if we had left out U_S and defined it in terms of G_S. The important thing to note is that a state comprises two different kinds of information. Information about the generation-relations, which is encoded into U_S and G_S, and information about the facts of the world, which is encoded into F_S.

> *Forcing*
> The situation s *forces* the proposition $[\![A]\!]$ within U_S iff for every $w \in U_S$ such that $s \subseteq w$ it is the case that $w \in [\![A]\!]$.
>
> *Revised Basis*
> A revised basis for w (with respect to $[\![A]\!]$) is a maximal subset s' of a basis s for w such that s' does not force $[\![A]\!]$ within U_S.

We can now define the retraction of a proposition from a state S.

> *Retraction*
> The retraction of $[\![A]\!]$ from the state S is the state $S \downarrow [\![A]\!]$, given by:
>
> 1. $w \in U_{S\downarrow[\![A]\!]}$ iff $w \in U_S$.
> 2. $gen(X, Y) \in G_{S\downarrow[\![A]\!]}$ iff $gen(X, Y) \in G_S$.
> 3. $w \in F_{S\downarrow[\![A]\!]}$ iff $w \in U_S$ and there is a revised basis b for some world $w' \in F_S$ (with respect to $[\![A]\!]$) such that $b \subseteq w$.

As such these definitions are just those of Veltman 2005, with the obvious differences stemming from the fact that we work with a different notion of generalizations and of state.

We can now give the definition of support of a counterfactual. It is entirely as in Veltman (2005). First we need to recall what support is.

> *Support*
> S supports A, and we write $S \models A$, iff $S[A] = S$.
>
> *Support of counterfactual*
> S supports the counterfactual $A \rightsquigarrow B$ iff $(S \downarrow [\![\neg A]\!])[A] \models B$.

We will follow Veltman and write S[if had been A] for the state $(S \downarrow [\![\neg A]\!])[A]$. Now we may say that a counterfactual is true relative to a state S if and only if that state supports the counterfactual in question. But instead of referring to it as truth *relative* to S, we may also sometimes say, just as before, that an agent with cognitive state S will assert or accept $A \rightsquigarrow B$ exactly when S supports $A \rightsquigarrow B$.

If we believe in an idealized state being able to reflect all true and relevant information for a given counterfactual, $A \rightsquigarrow B$, we may define *the* truth-value of a counterfactual relative to this state, call it S'.

Truth of counterfactual
$A \rightsquigarrow B$ is true iff S' supports $A \rightsquigarrow B$.

Before discussing why, and if, this semantics will in general give the intuitively right predictions, it will be very instructive to look at a couple of examples.

5.1 Some Examples

Let us start with the variant of Lifschitz' Circuit example that has also been discussed earlier. We recall that in this example there is a light which is on exactly when both switches are up. Let, again, $s1$, $s2$, and l be the atoms. Now we have to figure out what we know in terms of generation relations. Well, first of all we know that the two switches being up determines the light to be on; that is, we have $gen(\{s1, s2\}, \{l\})$. But, we also know that if one of the switches is not up, the light is not on, so we have the two further generation relations; $gen(\{\neg s1\}, \{\neg l\})$ and $gen(\{\neg s2\}, \{\neg l\})$. In the scenario switch one is up, switch two is down, and the light is out. That is, the state we are interested in is $S = 1[gen(\{s1, s2\}, \{l\})][gen(\{\neg s1\}, \{\neg l\})]$ $[gen(\{\neg s2\}, \{\neg l\})][s1][\neg s2][\neg l]$, which is pictured below.

	$s1$	$s2$	l
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**
~~w_5~~	~~1~~	~~0~~	~~1~~
~~w_6~~	~~1~~	~~1~~	~~0~~
w_7	1	1	1

$G_S = \{gen(\{s1, s2\}, \{l\}), gen(\{\neg s1\}, \{\neg l\}), gen(\{\neg s2\}, \{\neg l\})\}$

The only basis for w_4 is $b = \{\langle s1, 1\rangle, \langle s2, 0\rangle\}$, since $cl(b) = w_4$. Now this set forces $\llbracket \neg s2 \rrbracket$ within U_S, so a revised basis is $b' = \{\langle s1, 1\rangle\}$. This means that $F_{S\downarrow\llbracket \neg s2 \rrbracket} = \{w_4, w_7\}$, so $F_{S[\text{if had been } s2]} = \{w_7\}$. But then S[if had been $s2$] = S[if had been $s2$][l], which just means that the counterfactual 'if switch two were up, the light would be on' is supported by this state.

The second example is the Three Sisters example of Veltman 2005. As we noted before, even though the only generalization mentioned in the example is that whenever Ann and Billie are in bed, Carol is on the floor, there seems to be more information available implicit. It is our contention that what we understand when we hear this example is that two sisters are in bed and the last one on the floor (at the time described). What this means is that we have three generation relations and their contrapositive. We know that if Ann and Billie are in bed, then Carol is indeed on the floor, but we also know that if Carol is on the floor, then Ann and Billie are in bed, and similar for other combinations of the sisters. So, we have all in all six generation relations, given by the set $G = \{gen(\{a, b\}, \{\neg c\}), gen(\{a, c\}, \{\neg b\}), gen(\{b, c\}, \{\neg a\}), gen(\{\neg a\}, \{b, c\}), gen(\{\neg b\}, \{a, c\}), gen(\{\neg c\}, \{a, b\})\}$, and the state we are interested in is $1[G][\neg a][b][c]$.[4] This is given below.

	a	b	c
~~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~~

$G_S = G$.

Now a basis for the actual world, w_3, is $\{\langle a, 0\rangle\}$ or $\{\langle b, 1\rangle, \langle c, 1\rangle\}$. From the former we get that a revised basis is $\emptyset$, which means that $F_{S\downarrow[\![\neg a]\!]} = \{w_3, w_5, w_6\}$. Hence, $F_{(S\downarrow[\![\neg a]\!])[a]} = \{w_5, w_6\}$, and it follows that 'if Ann had been in bed, Carol would be on the floor' is false relative to the relevant state S, because c is true at w_5.

5.2 The Generation Relation

What is expressed by the generation relation? As we have hinted at several times we take this to express relations of, what Schulz called, *manipulation and control*. Now, when this is admitted, there are of course different routes one can take in trying to explain what exaclty this is. One could deny any objective existence of these relations as such and hold that they are simply imposed upon reality by us — not entirely unlike the attitude of Schulz towards the phenomenon of causality — or we could take these relations to express something objective; something that actually exists out there in the world. We will return to discuss these matters shortly, but before we do so, there are some things which can be said about the generation relation regardless of ones stance on the ontological status of this.

[4] Here, and in general, the state $S[G]$, when G is a set of generation relations, will just be used to mean the state S updated consecutively with the members of G.

When we say manipulation and control, do we mean that by manipulating the "input" we have total control over the "output"? No, we cannot intend this meaning, because it would be wrong. Instead, when a generation relation holds between X and Y, $gen(X, Y)$, it only means that if X obtains, in the sense of the literals of X all being true, then it will bring Y along with it. X is thus a sufficient condition for Y. However, controlling the value of the literals in X only gives us total control over the literals in Y insofar as we specify it so that the literals of X are all true; when some of these are false we are not in control of the value of the literals of Y, and as such they may be true or false in an unsystematic way. So, when we say manipulation and control we mean relations of control in the sense that we *know how to bring about* the truth of the literals of Y; namely by making the literals of X true.
Some facts about this relation follow by its very characterization. For instance, if X brings about Y and Y we brings about Z, it follows that X will bring about Z. That is, the generation relation is transitive. In a similar fashion it is obvious that this relation is also reflexive, that is, for any X, X will bring itself about.

If one wishes to say that these relations have existence in an objective sense, then one must also say something about what they are. Are they causal relations? It seems some such relations may be accurately described as causal, but here we may be using the notion of causality like Schulz, that is, as a heuristics, a tool for better coping with reality. It seems possible for someone to claim that these relations are existing in the world, and as such are absolutely true or false, but that they are not causal relations. Hence, we may be able to define causality in terms of counterfactuals after all. However, to say that the generation relation has objective existence but is not at all related to causality seems dubious, at least in the broader meaning of *causality.* There are of course cases where we have an obvious example of a generation relation without it being causal. For instance, that the dice roll was a six is in a generation relation with one being faced against the table, simply because we know that when the dice roll is a six, then it will not show one, because that is how dices are made; one and six are on opposite sides. There are also examples where the relation seems to be causal, perhaps in the heuristic meaning of the word; Jones wears his hat *because* the weather is bad. Bad weather *causes* him (perhaps through some tiny biological processes) to wear his hat. And this notion of manipulation and control, whether or not it is in fact causality, is also captured by the generation relation.

One might wonder whether the notion of a generation relation is inherently a counterfactual notion, because if this is so, the truth-conditions would be circular. Let $\bigwedge X$ be the conjunction of the literals in X. We might want to ask whether the following holds: $S \models gen(X, Y)$ iff $S \models \bigwedge X \rightsquigarrow \bigwedge Y$. Let us break it down and ask the two directions separately. First, we might wonder if the following holds: if $S \models gen(X, Y)$, then $S \models \bigwedge X \rightsquigarrow \bigwedge Y$. This is generally so. Because of how we have defined $gen(X, Y)$ we can be sure of this, but that every generation relation gives rise to a true counterfactual should not worry or surprise us. We define the truth of a counterfactual to be relative to a state where some information is given. And when this information is a generation relation, which by definition means that one thing will bring about the other, then of course this gives rise to a true counterfactual. This counterfactual may be false relative to other states however.

On the other hand, it is not the case that if $S \models \bigwedge X \rightsquigarrow \bigwedge Y$, then $S \models gen(X, Y)$. Counterexample:

> Let p, q, and r be atoms. Assume $gen(\{p, q\}, \{r\})$. Let $\neg p$, q, and $\neg r$ be the case. Then it is easy to verify that $p \rightsquigarrow r$ (it is formally equivalent to the variant of the Lifschitz Circuit example presented above) is supported by the state $S = 1[gen(\{p, q\}, \{r\})][\neg p][q][\neg r]$, yet we do not have $S \models gen(\{p\}, \{r\})$. There is a world, $w \in U_S$, such that $w = \{\langle p, 1\rangle, \langle q, 0\rangle, \langle r, 0\rangle\}$, hence $s = \{\langle p, 1\rangle\} \subseteq w$, but $cl_{\{gen(\{p\},\{r\})\}}(s) \nsubseteq w$. That is, $S[gen(\{p\}, \{r\})] \neq S$.

So, we have that every generation relation gives rise to a true counterfactual (relative to a state that has been updated with this generation relation), which is to be expected. However, the generation relation is not equivalent to a counterfactual relation, since we can have true counterfactuals $\bigwedge X \rightsquigarrow \bigwedge Y$ without it being true in those states that $gen(X, Y)$.

Perhaps it is worthwhile to also say a little about what the generation relation *is not*. It is clear that the generation relation is *not* a strict conditional. This can be seen from the dice example given before. That the dice roll was a six is, as we argued above, in a generation relation with the fact that one is faced against the table. But, this is not a matter of necessity in the sense that it is true throughout all possible worlds that if the dice roll was a six, then one will be faced downwards. That one and six are on opposite sides of a standard dice is not a necessary fact. A dice might have had another design where six and four are on opposite sides.

It is also not a part-of relation between situations such as that presented in Kratzer's lumping semantics. That the weather is bad *brings about* that Jones is wearing his hat, but it is not true that any situation where the weather is bad is a situation where Jones is wearing his hat. After all, the weather is presumably bad even before Jones opens his curtains to check and in such a situation Jones is not wearing his hat while the weather is in fact bad.

5.3 We Need Circular Dependencies

While the framework of Schulz (2009) does not allow for circular dependencies it seems we need them to handle the case of the Three Sisters properly.

The "law" mentioned in the example is that whenever Ann and Billie are in bed, Carol is on the floor. Let a, b, and c represent 'Ann is in bed' etc. Now the obvious choice of background variables is thus a and b, which c then depends on. A dynamics would be $D = \langle\{a, b\}, F\rangle$, where F maps c to $\langle\{a, b\}, f_c\rangle$. Here f_c is specified as below:

a	b	f_c
0	0	0
0	1	1
1	0	1
1	1	0

This says that if Ann and not Billie is in bed, then Carol will be in bed, and similarly if Billie and not Ann is in bed. If both are in bed, Carol is on the floor.

A basis for the actual world according to this dynamics will be $\{\neg a, b\}$. There is only one revised basis which is $\{a, b\}$ and it is straightforward to see that the counterfactual 'if Ann had been in bed, Carol would be on the floor' is true relative to this choice of D.

This however, we do not want. We might say that this is only so because we give the wrong dynamics, instead, the background variables are a and c. If we do that however, the unacceptable 'if Ann had been in bed, Billie would be on the floor' becomes true. If, we choose b and c as background variables, we get the undesired result that 'if Ann had been in bed, Carol *might* be on the floor' and 'if Ann had been in bed, Billie *might* be on the floor' are both false, since in this case they would *both* still be in bed.

It thus seems that in this example there is no reasonable way to set up a dynamics which will make us predict the intuitively right thing. A way to fix this seems to be to give up the condition that the dependencies specified in the dynamics cannot be circular. This would of course have as a consequence that the notion of basis is no longer unique, and so we get different bases for the three sister example; $\{\neg a, b\}$, $\{\neg a, c\}$, and $\{b, c\}$. If we now define a counterfactual to be true iff the consequent follows causally from any revision of all bases, we are able to say that the counterfactuals 'if Ann had been in bed, Carol would be on the floor' and 'if Ann had been in bed, Billie would be on the floor' are both false, whereas the corresponding might-conditionals are both true.

This is of course only one way to go in trying to repair the semantics. However, for now it suffices to note that the dependencies of the Three Sisters example seem to such that a dynamics cannot capture it. This further seems to be because the dependencies are circular in such a way that the status of *any* two of the sisters determines the status of the third.

5.4 The Math Examples

We now have a way of explaining the asymmetry in the Math examples from Schulz. On the board is written '3 + 4 = 7', and you say:

> (1) If the first number had been even, the result would have been even.
> (2) If the result had been even, the first number would have been even.

We needed a way to explain why the first is intuitively true and the second intuitively false. As already hinted at by Schulz, this is due to an asymmetry in the relation having to do with manipulation and control. By controlling the numbers on the left we have total control over the result. However, controlling the result leaves us with different possibilities for the numbers on the left. We can capture this using the generation relations. Let p, q, and r be 'the first number is even', 'the second number is even', and 'the result is even', respectively. Now we have the generation relations given by the set $G = \{gen(\{p, q\}, \{r\})$, $gen(\{\neg p, q\}, \{\neg r\})$, $gen(\{p, \neg q\}, \{\neg r\}), gen(\{\neg p, \neg q\}, \{r\})\}$. In total this just amounts to saying that the result is even if and only if both numbers are either odd or even. Right now the first numbers is odd, the second even, and the result therefore odd. We are interested in the state $1[G][\neg p][q][\neg r]$, which is pictured below:

	p	q	r
~~w_0~~	~~0~~	~~0~~	~~0~~
w_1	0	0	1
$\boldsymbol{w_2}$	**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

$G_S = G$.

A basis for the actual world, w_2, is just $\{\langle p, 0\rangle, \langle q, 1\rangle\}$. So a revised basis not forcing $[\![\neg p]\!]$ within U_S is $\{\langle q, 1\rangle\}$. So, it follows that $F_{S\downarrow[\![\neg p]\!]} = \{w_2, w_7\}$, and it is immediately clear that $(S \downarrow [\![\neg p]\!])[p][r] = (S \downarrow [\![\neg p]\!])[p]$, or in other words, that $p \rightsquigarrow r$ is true relative to this state.

On the other hand, a revised basis with respect to $\neg r$ is either $\{\langle p, 0\rangle\}$ or $\{\langle q, 1\rangle\}$, from which it follows that the counterfactual $r \rightsquigarrow p$ is unacceptable, since $w_1 \in F_{S\downarrow[\![\neg r]\!]}$, $w_1 \in [\![r]\!]$ and $w_1 \in [\![\neg p]\!]$.

Thus, using Schulz idea of manipulation and control to define the generation relations have left us with a way to explain away our intuitions regarding these math examples. And, it has let us do so by predicting the asymmetry in judgement of (1) and (2) above, exactly because the underlying relations, that which we have called generation relations, are also asymmetric.

Chapter 6

Discussion

In what follows we will discuss some of the issues that have come up throughout this thesis.

6.1 Causal and Analytic Relations

While one is not forced to hold that causality is ultimately defined in terms of counterfactuals, the question of which notion is "more primitive" still emerges.

One might therefore wonder how, and in what sense, causality is linked to the notion of the generation relations. There is an obvious part of this relation that resembles causality. For instance, when we say that bad weather always makes Jones wear his hat are we not saying that bad weather *causes* him to wear his hat? It would seem so, but the counterfactual 'no bad weather $\rightsquigarrow$ no hat' is not true relative to the relevant state. So, the notion of causality involved cannot be the simple notion of 'no cause $\rightsquigarrow$ no effect' being true. It is of course also very dubious whether anyone would ever accept such a reduction of causality to counterfactuals; it does not seem that the proposed condition is necessary for establishing a causal relationship.

Even so, what are we saying about the situation of evaluation when we propose that $gen(\{bad\}, \{hat\})$ holds true of it? It seems we are saying nothing more than the knowledge we presently have of the situation is such that bad weather *brings about* Jones wearing his hat. This is simply what we know about the world, it is given by the description of the example, and as such it is no wonder that we pay special attention to this when evaluating the counterfactual in question.

It is thus obvious to define the generation relations to be a piece of knowledge that the evaluator of the counterfactual possesses about the world. This is well in tune with how we have set up our semantics. We defined a state of an agent to reflect exactly the generation relations that the agent takes to hold true of the world, and we said that a counterfactual is acceptable by an agent *iff* the state of the agent supports the counterfactual in question. The knowledge represented by these generation relations can be causal or it can be of another kind. The good thing is that right now we do not need to say anything more about what it is, we need only observe that under such a treatment of counterfactuals the predictions of the theory coincide with our intuitions.

There does seem to be two different kinds of generation relations at work though. One is more accurately described by notions such as *causality*, while the second could plausibly be better described by using the term *analytic relation*. It does not seem that we can say that John being a bachelor *causes* him to be an unmarried man. It is true that whenever we have the former, we also have the latter, but it does not seem that the relation is accurately described by using the concept of causality. When John is a bachelor, the fact that he is an unmarried man follows purely by the meaning of the terms involved. We might therefore choose to call such relations analytic.

So, it is clear that there are two kinds of relations falling under the category of generation relation. One might be described as causal relations, while the other would more accurately be described by the term *analytic*. However, since their relevant properties as relations, that is, that by settling on the "input" *we know exactly* what the "output" is, are the same, it seems there is no reason to work with two different notions of relations.

If we do wish to make a distinction between the relations that are analytic and the relations that are not, we seem to have an obvious way of doing so. The state U_S is defined as a subset of the set of all possible worlds W. However, if we take serious the claim that in no possible world can someone be a bachelor and not an unmarried man, and vice versa — a claim that, as long as we are working relative to the meaning we *in this world* ascribe to the term *bachelor*, seems reasonable — it seems the analytic relations do not need to be presented as generation relations. This is so because these relations will hold throughout all members of the set W and worlds where these relations do not hold will therefore not be able to enter into the state U_S. This is a solution very similar to that of Schulz (2007). In that framework we work with a set of possible worlds where the relations that we have called analytic all hold true. On this set of worlds we then impose a *causal structure*, which is just another name for a dynamics. The approach of Schulz (2007) thus also has this two step procedure that we are proposing for our semantics. First we sort through all worlds to get rid of the worlds that are impossible.[1] Then, on the remaining worlds we impose the relations that we take ourselves to know about the world; in Schulz case these are the relations represented by a dynamics, whereas in our case these are the relations represented by the generation relations.

However, in the case of our semantics this represents a problem in case the agent does not know the analytic relations in question. We have defined the truth of a counterfactual relative to a state of an agent; which we have also called acceptability/assertability of a counterfactual. But, it seems clear that an agent to whom it is unknown that all bachelors are unmarried men should not straightforwardly accept the counterfactual ‘if John had been a bachelor, he would have been an unmarried man’. But, if we exclude all worlds where something is a bachelor yet not an unmarried man, and vice versa, from the set W we are unable to predict this. So we might want to make the set W relative to the agent in question, but then we need the following procedure; first we take the set W, then we impose the *analytic relations* that the agent takes to hold on these worlds, and then we impose the *generation relations* that the agent takes to hold on this set. We could of course do this by having two different kind of relations in our semantics; generation relations, which we have already presented and discussed, and then *analytic relations*, which we could abbreviate as $ana(X, Y)$ when the literals in X and Y are in such a relation.

There are thus many ways one can build a semantics which will make the same predictions as the one we have proposed, but where we are more clear on what is an analytic relation and what is a relation of *bringing about*, that is, a generation relation. For matters of simplicity and ease of exposition we have chosen not to make this distinction in the presentation of the semantics.

6.2 Picking a Basis

There is an obvious difference between the framework of Veltman (2005) and the framework developed here. The definition of a basis for a world. In Veltman’s framework a basis is a minimal situation determining the world

[1] It is a delicate matter what this ‘impossible world’ might mean. Here we take it to mean nothing more than worlds where the relations we have called analytic do not hold. It is of course another question whether such worlds are in fact *impossible* and thus deserve the name we have just given them. This debate however falls without the purpose of this thesis.

in question within U_S. What this means is the following; given the information in the basis, the agent is able to deduce which world is the actual given the laws and generalizations. This notion of basis thus has a more epistemic flavor than the one we have presented in our own semantics. What matters is ultimately that the agent, given the information in the basis and the generalizations and laws, is able to deduce which world is the actual and not whether all facts of the world actually *follow from* or *are brought about* by the facts in the basis. The idea of basis as it is presented in the present framework is exactly the latter. While it does make reference to the generation relations that the agent takes to hold, and so cannot escape the epistemic flavor entirely, we are not concerned with what the agent can deduce given the information in the basis, but what follows from this information in accordance with the generation relations. The difference is easily explained with an example. In the variant of the Lifschitz Circuit example that we have seen earlier, a basis for the actual world w is $\{\langle s1, 1\rangle\langle s2, 0\rangle\}$ according to our present framework because $cl(\{\langle s1, 1\rangle\langle s2, 0\rangle\}) = w$. However, according to Veltman's framework, the basis is $\{\langle s1, 1\rangle, \langle l, 0\rangle\}$ because this situation forces w within U_S. It is easy to see that on Veltman's framework we get the undesired revised basis $\{\langle l, 0\rangle\}$ which is exactly responsible for making the counterfactual 'if switch two were up, the light would be on' false.

The notion of basis we work with in the present semantics is thus a more ontic notion. A basis is a set of facts that bring about all other facts true in that world. It plays no role whether the agent can deduce which world is the actual given the basis, even though he will be able to do so. In other words, that the agent can deduce which world is actual is no longer a sufficient condition when defining a basis.[2]

There are thus two general ways of explaining why the choice made here is the better. On one explanation we refer to the pragmatic effect of this choice. Because we choose the basis as something from which all other facts follow, we get the desired predictive power; we rightly make predictions about counterfactuals that coincide with our intuitions. On this view we remain agnostic about whether the meaning of a counterfactual is in fact given in accordance with our definitions, and focus solely on the fact that this works; i.e. gives the right predictions when it comes to acceptability or unacceptability of a counterfactual.

However, we could explain the difference in picking a basis by saying that this *is in fact just* the correct meaning of counterfactuals. In specifying the

[2] See also Schulz 2008.

meaning and truth-conditions for a counterfactual we do not care what the agent can in fact deduce about the actual world, only what facts of the actual world need to obtain to bring all other facts along with it. Therefore we need to choose a basis as a set of facts, which along with the generation relations, brings every other fact of the world along with it. But now we ask, does not the fact that switch one is up and the light is out bring along the fact that switch two is down? In a certain sense it does, and in another sense it does not. The first sense is exactly an epistemic sense. I can imagine a world in which the light is out and switch one is up, and if I further take the generation relation $gen(\{s1, s2\}, \{l\})$ to be true, it follows that *I will have to* imagine this world as one where switch two is down. However, this is not what we are interested in for present purposes. The generation relation is one of bringing about something other, and we cannot, by changing the status of the light change the status of the switch, because there is no such connection present; there is indeed only the opposite connection; namely that changing the status of the switches will change the status of the lights. And it is this relation of bringing about that we pick a basis according to. It is admittedly very hard to see what this relation is in the present example if it is not a relation of causality. We do not want the basis $\{\langle s1, 1\rangle, \langle l, 0\rangle\}$ because these two facts do not *cause* the second switch to be down. Instead the fact that switch two is down *causes* the light to be out, and it is therefore in our basis along with the, in this setting, *independent* fact that switch one is up.

6.3 Causality

We must admit that we simply do not know how to answer the question of how causality and counterfactuals are related if they are related at all. However, with this said it does seem that, as already observed by Schulz, that "the semantics of counterfactuals exploits certain invariant relationships, certain dependencies". (Schulz 2009). The question is thus whether these relationships can be given in non-causal terms, which we need in order for a possible definition of causality not to be circular. In the present thesis we have chosen to call these relationships generation relations, and we have characterized such a relationship as something *bringing about* something other. With this said, however, it seems hard to defend a position defining causality in terms of counterfactuals. And this is so simply because this "something *bringing about* something other" seems to be inherently linked to the concept of "something *causing* something other".

But, as we briefly touched upon, it seems we can also hold that the truth-conditions, as we have stated them in our semantics, do not provide the meaning of a counterfactual, but is only a useful tool for evaluating these for truth and acceptability. On this view we can then further hold that the meaning of a counterfactual is something which is given independently of causal relations and thus propose to define causality in terms of counterfactuals. However, even though we are somewhat sympathetic to this view, we have to admit that a position such as this seems hard to argue for. As mentioned by Tichý (1984), it does seem that when discussing counterfactuals, people pay extra attention to laws and generalizations true of the actual world. If these laws and generalizations sometime express causal relations it seems hard to uphold that the meaning of counterfactuals is deprived of any relation to such causal relations.

With this said, however, if one wants to say something conclusive about the relationship between counterfactuals and causality one would need an in-depth analysis of what we take the concept of *causality* to mean in order to determine whether we can define *this concept* in terms of counterfactuals, or vice versa.

6.4 Counterexamples

As the astute reader might have noticed there are certain examples that the theory presented here does not seem to handle correctly. These are counterfactuals of the form $\neg p \rightsquigarrow \neg q$, when it is known that p will bring about q. To take a simple example. Suppose that Jones has just shot Smith and Smith therefore has died. Intuitively the counterfactual 'if Jones had not shot him, Smith would still be alive' should be acceptable in this scenario, but the problem seems to be that the theory does not predict this relative to the obvious choice of state S. Let p and q represent 'Jones shoots' and 'Smith dies' respectively. Then we have the generation relation $gen(\{p\}, \{q\})$, and as such we are interested in the state $S = 1[gen(\{p\}, \{q\})][p][q]$ which looks as pictured below.

	p	q
w_0	0	0
w_1	0	1
~~w_2~~	~~1~~	~~0~~
$\boldsymbol{w_3}$	**1**	**1**

$G_S = \{gen(\{p\}, \{q\})\}$

A basis for w_3 is $\{\langle p, 1\rangle\}$. But since we are interested in the counterfactual $\neg p \rightsquigarrow \neg q$ we have to revise this relative to p which is just $\emptyset$. Thus $S \downarrow [\![p]\!] =$

$\langle U_S, U_S, G_S \rangle$, which means that $F_{S[\text{if had been } \neg p]} = \{w_0, w_1\}$, so we do not have $S \models \neg p \rightsquigarrow \neg q$, since $\{\langle q, 0\rangle\} \nsubseteq w_1$.

What this means is that if we accept that the state S above reflects all relevant information to evaluating the counterfactual, then we will have to say that the counterfactual is in fact unacceptable, which seems to go against our intuition. However, it is our contention that strictly speaking, in lack of more information, the counterfactual should be false relative to the state given above. The reason is that information about Smith is missing. It might just be that one hundredth of a second before the shot was fired Smith suffered a heart attack that technically did not kill him, but surely would have, had the bullet not done the job an instant before.

This is perhaps clearest if we allow ourselves to speak about *the* truth-value of the counterfactual. In lack of information such as the above we cannot say that the counterfactual is in fact true; to say this we would need to establish the truth of the counterfactual relative to the *true state* of the world, and arguably this state would have to contain information such as Smith not being about to have a heart attack etc. And as such, if the counterfactual is true in any absolute sense, this would be because the true state of the world would have some generation relation $gen(X, \{q\})$ along with all the literals in X being true.[3] And it is straightforward that the counterfactual would be true relative to this state.

Before moving on to discuss why we feel that the counterfactual — even though, perhaps strictly speaking, it is not true in lack of more information — is acceptable in the described scenario, we might pause to highlight a theory that does give the desired prediction; the theory of Schulz (2009) which have been presented above. Now a dynamics representing the above scenario will be such that p is a background variable which q depends on. Now, since a dynamics must specify a truth-value for q for all values of p it is reasonable that q should be true exactly when p is. This means that changing the value of p to false will have as a causal effect that q becomes false as well, and hence the counterfactual $\neg p \rightsquigarrow \neg q$ will be true. One might see this as a strength of the theory of Schulz, but there are cases where the theory is brought into trouble. The definition of a dynamics states that a truth-value of the inner variables must be given for every truth-value of the background variables that it depends on. However, there are cases where the truth-value of the inner variables ought to remain undetermined even though the background variables are fully determined. The case of Jones

[3] Arguably, the set X might be infinite, but we will not concern ourselves with problems such as this here.

and his hat is such an example. To make the right predictions in this case we need it to be the case that the value of the inner variable, i.e. that Jones is wearing his hat, is fully determined when the background variable is true, i.e. when the weather is bad. However, we also need it to be the case that when the weather is fine, it is undetermined whether Jones is in fact wearing his hat, and as we have seen, the framework we have developed here is able to incorporate this information. One could of course go about this problem by redefining the notion of dynamics such that it is possible to map an inner variable to u for some values of the background variables that it depends on. This would straightforwardly allow us to predict the counterfactuals 'if the weather had been fine, Jones might have been wearing his hat' and 'if the weather had been fine, Jones might not have been wearing his hat' to both be true, which is intuitively what we want.

However, the possibility of having undetermined values of the literals in Y when $gen(X, Y)$ is given and the literals in X are not all true is an essential feature of our proposed semantics, and as such we need not look to other semantics to make the right predictions in the example with Jones and his hat.

As we saw before, it makes sense to say that the counterfactual 'if Jones had not shot, Smith would not have died' is not strictly speaking true in the scenario given — at least not in the absence of more information about Smith. However, we are very inclined to judge this as true, which in our present framework means that it should be acceptable relative to our cognitive state. We have seen that if the only generation relation that the agent in question takes himself to know is $gen(\{p\}, \{q\})$, then the counterfactual will not be true relative to this state.

At this point in the thesis we can of course only conjecture, but one way to explain this is that agents, in the absence of other information, sometimes assumes that a negative "input" will bring about a negative "output"; or in the formalism of our present framework, that agents, when given the information that $gen(\{p\}, \{q\})$ and nothing else, sometimes accepts the corresponding generation relation, $gen(\{\neg p\}, \{\neg q\})$, as true *and* incorporates this into his cognitive state. For instance, when presented with the example where Jones shoots Smith it is very natural to assume that in the absence of the shooting, Smith would still be alive; that is, it is very natural to assume that "no shooting brings about no death", or gen({'Jones does not shoot'}, {'Smith does not die'}), even though, strictly speaking, no such information is given in the example. The reason seems to be that people tend to stay alive if nothing interferes. In the example only one possible interference is mentioned, and it seems reasonable to conclude that if the mentioned interference does not obtain, then no interference will obtain, and Smith will stay alive. We thus conjecture that a certain form of default reasoning takes place here.

It is also clear that if information about a possible interference is given, the agent in question will not assume $gen(\{\text{'Jones does not shoot'}\}, \{\text{'Smith does not die'}\})$. For instance, suppose Smith is a convicted murderer who is to be executed. The execution procedure is death by shooting, and to be sure that the execution will be successful the institution in charge works with two shooters, Jones and Johnson. They both shoot when the order is given and Smith, as a result, dies. Now, the crucial point is that no agent would accept the counterfactual 'if Jones had not shot, Smith would be alive' in this scenario. And this is so because it is unreasonable to assume that $gen(\{\text{'Jones does not shoot'}\}, \{\text{'Smith does not die'}\})$ is true in such a situation; in this scenario another reason for Smith to die is mentioned explicitly, that is, we are told that $gen(\{\text{'Johnson shoots'}\}, \{\text{'Smith dies'}\}$ and we therefore evaluate the counterfactual in accordance with this information.

As we have seen, the framework of Schulz incorporates the assumption of a fixed value of the inner variables for all values of the variables that it depends on. We have chosen not to incorporate this into our framework because there are examples such as that of Jones and his hat, where we do not want it to be determined whether or not Jones is wearing his hat when the weather is fine. However, we also acknowledge that in many cases, even though the information given strictly speaking only allows us to determine the value of the literals of Y when the literals of X are true, the value of the literals of Y gets determined in *more cases.* We conjecture that in cases such as the above, this is because when presented with the information that a will bring about b, that is, $gen(\{a\}, \{b\})$, and nothing else, agents sometimes conclude that $\neg a$ will bring about $\neg b$, that is, $gen(\{\neg a\}, \{\neg b\})$.

Admittedly, things are not entirely as simple as they are presented here, and to try to give a solution along the lines of this, one would need to do some thorough research into actual human reasoning.

The claim of the thesis is thus that an agent will accept a counterfactual when it is true relative to the set of generation relations that he takes to be true about the world. However, which generation relations these are can not always be read straight off from how the example is presented.

Chapter 7

Conclusion

In this thesis we have sought to develop a semantic theory of counterfactuals.

We started out by drawing a distinction between Two Parameter and Three Parameter theories and presented some older theories of counterfactuals. We then proposed that the difference between a Two and a Three Parameter theory makes the most sense when we view it as a distinction between *truth in itself*, that is, truth as an objective phenomena, and truth relative to a setting of a third parameter; which we have called acceptability/assertability by an agent who takes the world to be truly reflected by the setting on this third parameter.

When we evaluate counterfactuals some facts about the actual world seem to matter more than others. With Kratzer's (2010) theory of lumping we have the beginning of an explanation as to why this is. Some facts bring along other facts — which in Kratzer's theory was presented by her concept of lumping — and we need to pay attention to this when we evaluate counterfactuals. However, on the face of it, Kratzer's theory have some problems of its own and as such we did not go into any detail with the theory.

The theory of Veltman (2005) which we presented is a formalization of the slogan that facts stand and fall together. This just means that if one fact is responsible for another, then making a counterfactual assumption about the former will affect the latter.

With Veltman's theory we are also able to draw a clear distinction between acceptability/assertability which is just support of a counterfactual by a cognitive state, and *truth in itself* if one believes in such a thing. The latter will then be support of a counterfactual by a special idealized cognitive state, namely that reflecting *the true state of the world.*

With a formal semantics reflecting these ideas we are able to make apt predictions about counterfactuals in many cases. However, there are obvious counterexamples to the theory of Veltman. One such is the Lifschitz Circuit example where a light is on exactly when two switches are up. Assuming that the first switch is up, the second down, and the light out, we are not able to predict the acceptability of the counterfactual ‘if switch two were up, the light would be on’, which we would like to be able to do.

The theory of Schulz (2009) is able to give the right predictions in examples such as that of Lifschitz’ Circuit example. This is so because Schulz works with an asymmetry in the underlying generalizations of the example: the switches being up *causes* the light to be on, and not vice versa. Schulz develops this idea into the notion of a dynamics, which is just a specification of the dependencies true in the situation of evaluation. She further calls these causal relations, but admits that this might not be the most accurate description and that they are instead relations of *manipulation and control*. From this she defines the truth of a counterfactual in terms of causal entailment of the consequent from the antecedent and a set of base facts of the actual world. This is able to give correct predictions in most cases. However, Schulz theory suffers from two minor problems. In order to handle cases such as that of The Three Sisters example it seems we would need circular dependencies which her framework does not allow for. Secondly, the definition of a dynamics does not allow for undetermined effects. That is, for every value of the causes we will have an exact value of the effect. This means that we are unable to handle examples such as that of Jones and his hat, because when the weather is fine it is undetermined whether Jones is wearing his hat.

The semantics we have developed builds heavily on Schulz’ idea of manipulation and control. We have introduced a new concept, that of a generation relation, where certain facts bring about certain other facts. However, the difference between this and Schulz’ idea is that we allow for the possibility of the value of the “output” not being determined for some values of the “input”. This just means that when X and Y are in a generation relation and the literals of X not all true, then we cannot be sure about the truth-value of the literals of Y. This, along with the fact that we allow for circular dependencies, helps us predict the right thing in many cases. However, there are certain cases where we seem to make the wrong prediction. As for these cases there seems to be some reason in saying that they are strictly speaking not counterexamples because more information is needed to settle *the* truth-value.

Further, we use a dynamic setting for our semantics and as such we are more interested in the acceptability of a counterfactual by an agent. And when it comes to acceptability of these problematic counterfactuals we have conjectured that the acceptability-conditions we have given are still accurate; it is just that the agent in question will sometimes assume more information that is given explicitly in the example

We have further touched upon the question of the relationship between counterfactuals and causality, but we must admit that we do not know how to say something conclusive about this. In so far as we can conclude anything, it does seem that at least some of the relations expressed by the generation relations are inherently causal, and as such, a reduction of causality to counterfactuals seems problematic. At least so when we are talking about *the content* side of the relationship; about what can possibly be reduced to what. We grant that, as mentioned by Schulz, on the epistemic side of the relationship, counterfactuals might provide very good test cases to determine when a causal relationship obtains in the world.

Chapter 8

Postscript

In the light of some of the questions raised during my defense, there are a couple of points that I would like to say a little more about.

Backtracking

First, there is the question of "backtracking" counterfactuals. Suffice it to say that I am not entirely persuaded that all backtracking counterfactuals, that is, a counterfactual that goes against the flow of "causality" such that the event described in the consequent precedes the event described in the antecedent, should be clearly false on the ontic reading. I surely agree that most should, but there are also some where my intuitions are not clear.

One might object that the true ones are only the ones where the relation between antecedent and consequent is analytic, as in if I say: "If it was the case that tomorrow is my birthday, I would have been born on October 8th." Here the event that I was born clearly precedes the event that I have a birthday, so this counterfactual goes against the "flow of causality", which is just to say that on a time line the event in the consequent comes before the event in the antecedent. But, it is held, this counterfactual is true because of the analytic relation between having a birthday and ones birth. Because "birthday" means what it does, it follows that birthdays can only be had on the same date as that you were born on, and so there is no problem with the counterfactual, since the relation that binds the antecedent and consequent is an analytic relation.

All of this I grant. Backtracking counterfactuals where the relation between antecedent and consequent is analytic are often, if not always, true. However, it is also said that a backtracking counterfactual, where the relationship between antecedent and consequent is of a more causal nature, should be false (on the ontic reading of them).

This surely seems true in many cases, and it is true that my theory cannot predict this. Suppose we have the generation relation holding between $\{p\}$ and $\{q\}$; that is, $gen(\{p\}, \{q\})$, and that this expresses a genuine causal relation. Then, if we make the (counterfactual) asssumption that q is false, then it will follow that p is false, because the world $w = \{\langle p, 1\rangle, \langle q, 0\rangle\}$ is excluded from U_S by the generation relation. Hence we will get true counterfactuals of the form $\neg q \rightsquigarrow \neg p$. This means that in a great many cases (as many as there are false backtracking counterfactuals of the form $\neg\varphi \rightsquigarrow \neg\psi$ when ψ brings about φ) my theory will give the wrong prediction. This of course is a major drawback of the theory - unfortunately one I did not think about when constructing the theory.

However, I am not entirely persuaded that a theory like Schulz', which in all cases proclaims the causal backtracking counterfactuals to be false, is the right way to go either. Surely, it is in most cases, but there also seem to be some cases where (at least my own) intuitions are not entirely clear. Suppose Jones cuts Smith's head of with a sword. Smith surely dies from this. The question is what the truth-value should be of the counterfactual expressed by "if Smith had not died, his head would still have been cut off"? Schulz' theory will predict this to be true given that the underlying dynamics stipulates that Smith's dying is caused by Jones cutting his head off. This is so because changing the value of a variable in Schulz' theory can never effect the value of variables that come before that in the "causal hierarchy". In other words, if variable X depends on Y, then making a counterfactual assumption about X will never effect the value of Y.

My intuition with regards to the counterfactual is that there is a certain sense in which it is false. Surely, if Smith had not died, then it could not be the case that he was still without a head.

The way out of this for Schulz' theory would probably be to claim that my reading of the counterfactual 'If Smith had not died, his head would still have been cut off' as false (or equivalently, the reading of the backtracking conditional 'If Smith had not died, he would not have lost his head' as true) relies on an epistemic reading of the conditional in question. Actually, I am inclined to accept this. On an epistemic reading the backtracking counterfactual is surely true; "If Smith had not died, then he would *have to* have not lost his head" expresses something true without a doubt. But even granted that the epistemic reading of the conditional in question is true, the question remains, however, whether there is an ontic reading of the counterfactual 'If Smith had not died, he would not have lost his head' that expresses something true. My intuitions say that there might be. If an ontic counterfactual has to do with what would be the case had another fact been different, then surely one can claim that had it been a fact that Smith didn't die, it would also be a fact that Smith didn't lose his head.

You could of course claim that this is because the relation between antecedent and consequent is analytic. This, however, seems an inappropriate response. It doesn't seem to be the case that the meaning of "being alive" presupposes that one has a head; there are plenty of living things on earth which do not have heads.[1] Instead, it seems to be such, that in the light of the nature of human beings (in the actual world), dying is a causal consequence of getting ones head cut off.

I am not entirely sure what to make of this discussion. I am inclined to accept that the only true reading of 'If Smith had not died, he would not have lost his head' is an epistemic reading. This is made plausible when considering the explanation I just gave that "had it been a fact that Smith didn't die, it would also be a fact that Smith didn't lose his head" seems to be more accurately expressed with an extra "have to" inserted; that is, that "had it been a fact that Smith didn't die, it would *have to* also be a fact that Smith didn't lose his head." This extra "have to" seems to serve as a mark that upon reflecting on the information that Smith has not died, we come to conclude that then he would not have lost his head. And this, of course, looks a lot like the evaluation procedure for an *epistemic* counterfactual. However, with this said, I still can't escape the feeling that there is a true reading of 'If Smith had not died, he would not have lost his head' which expresses *a relationship between facts* (that is, an ontic reading), and not just a true reading that expresses what we can justifiably conclude were we to learn that Smith did not die (that is, an epistemic reading).

With this said though, I grant that Schulz' theory has the upper hand when it comes to handling backtracking counterfactuals. For reasons mentioned above it is clear that my theory will give lots of wrong predictions, whereas — as the preceding discussion aims to show — it is not clear that Schulz' theory gives *even one* wrong prediction in the matter of backtracking counterfactuals.

[1] And even though, for humans in this world, it is a necessary condition for being alive that you have a head, that *might have been* different.

Particular Facts

Second, there is the question of what the generation relation expresses. In particular it was questioned during the defense if the relations in the Three Sisters example can meaningfully be regarded as a generation relations.

In this example I say that two sisters being in bed is in a generation relation with the last sister being on the floor. The question is, however, whether it is actually meaningful to express it this way. It does not seem that $\{Ann, Billie\}$ generates $\{\neg Carol\}$ on any reading of the generation relation, that is, as causal or analytic. Instead, it seems that the fact that when two of them are in bed, the last must be on the floor is due to some particular fact of the world; namely that their bed is only big enough for two of them. So, it is asked, is it really appropriate to represent this as $gen(\{Ann, Billie\}, \{\neg Carol\})$ being true, when it seems this is not a relationship stemming from the fact that Ann and Billie being in bed *generates* Carol being on the floor (on the appropriate reading of *generates*)?

To this I will only say something very brief. It is indeed true that in and off itself Ann and Billie's position do not determine anything about Carol's position. However, the description of the example is such that we know that they are all sleeping, and we know that the bed is only big enough for two of them. And these facts make it so that $gen(\{Ann, Billie\}, \{\neg Carol\})$ holds true of the situation. There is nothing mysterious here. Sometimes particular facts give rise to relationships that can be expressed with the generation relation. The relationship between the Three Sisters' positions is not one of any kind of metaphysical necessity. Sure, they could have had a bigger bed, or one of them could have been the size of Thumbelina (who is no bigger than a thumb), so that they all would fit in the bed. However, the generation relation is not supposed to express relations that are necessary, and so the fact that the bed is the size it actually is gives rise to a true generation relation.

To give another example. The sun rises in the east, so if you stand on the western side of the Empire State Building at sunrise, you are sure to stand in shadow. We therefore have the true generation relation $gen(\{$'you stand on the western side of the Empire State Building at sunrise$'\}, \{$'you stand in shadow$'\}$. But that this relation is true is conditioned on the fact that the Empire State Building is a large building. So to speak, the fact that the Empire State Building is the size it actually is *gives rise* to the true generation relation. In the same manner with the Three Sisters example. Sure, it is a particular fact of the matter; that the bed is only big enough for two. However, this fact gives rise to a true generation relation between the positions of the three sisters; a relation that we use when we evaluate the counterfactual in question.

Even if this explanation is not satisfactory as to why we call it a generation relation in the example of the Three Sisters, it really doesn't matter much. The important thing is that in the scenario described we know that two sisters in bed will make the last sister be on the floor, and that one sister on the floor will make the other two be in bed. This seems to be exactly the information provided by the example, whether we call these relations generation relations or introduce a different relation that makes it clear that the relation is due to a particular fact of the actual world.

Bibliography

[1] Bennett, Jonathan (2003): *A Philosophical Guide to Conditionals*, Oxford University Press, (reprinted 2006).

[2] Chisholm, Roderick (1955): ‘Law Statements and Counterfactual Inference’, in *Analysis*, 15, pp. 97-105.

[3] Fine, Kit (1975): ‘Critical notice of *Counterfactuals*’, in *Mind* 84, pp. 151-158.

[4] Goodman, Nelson (1954): *Fact, Fiction, and Forecast*, Harvard University Press, reprinted 1979.

[5] Hansson, Sven O. (1989): ‘New Operators for Theory Change’, in *Theoria* 55, pp. 114-136.

[6] Henderson, Robert (2010): ‘“If not for” Counterfactuals: Negating Causality in Natural Language’, February 22, Draft.

[7] Kratzer, Angelika (2010): ‘An Investigation of the Lumps of Thought’, (revised version) in Kratzer, A: Papers on Modals and Conditionals, unpublished. Original version of the paper in *Linguistics and Philosophy* 12, pp. 607-653, 1989.

[8] Lewis, David (1973a): ‘Causation’, in *The Journal of Philosophy*, Vol. 70, No. 17, pp. 556-567.

[9] Lewis, David (1973b): *Counterfactuals*, Blackwell Publishing.

[10] Mares, Edwin & Fuhrmann, André (1995): ‘A Relevant Theory of Conditionals’, in *Journal of Philosophical Logic* 24, pp. 645-665.

[11] Mill, John Stuart (1868): *System of Logic*, Longmans, London.

[12] Pearl, Judea (2000): *Causality: Models, Reasoning, and Inference*, Cambridge University Press.

[13] Ramsey, Frank (1931): 'General Propositions and Causality', in Ramsey, F.: The Foundation of Mathematics and Other Logical Essays, (ed. Braithwaite), New York, pp. 237-255.

[14] Rott, Hans (1999): 'Moody Conditionals: Hamburgers, Switches, and the Tragic Death of an American President', http://www.illc.uva.nl/j50/contribs/rott/rott.pdf

[15] Schulz, Katrin (2007): *Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals*, ILLC Dissertation Series, DS-2007-04.

[16] Schulz, Katrin (2008): 'If you'd wiggled A, then B would've changed: Causality and Counterfactual Conditionals', long version, http://home.medewerker.uva.nl/k.schulz/bestanden/CausalityPaperPreRutgers.pdf.

[17] Schulz, Katrin (2009): 'If you'd wiggled A, then B would've changed: Causality and Counterfactual Conditionals', short version, http://home.medewerker.uva.nl/k.schulz/bestanden/CausalityPaper.pdf.

[18] Tichý, Pavel (1976): 'A counterexample to the Stalnaker-Lewis Analysis of Counterfactuals', in *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* vol. 29, No. 4, pp. 271-273.

[19] Tichý, Pavel (1984): 'Subjunctive Conditionals: Two Parameters vs. Three', in *Philosophical Studies* 45, pp. 147-179.

[20] Veltman, Frank (2005): 'Making Counterfactual Assumptions', in *Journal of Semantics* 22, pp. 159-180.