Conditionals and truth values

Larix Kortbeek

December 23, 2010

Contents

Chapter 1

Introduction

> "Contrariwise, if it was so, it might be; and if it were so, it would be; but as it isn't, it ain't. That's logic."
>
> **Lewis Carroll**, *Through the Looking Glass*

This thesis focuses on the logical nature of the word 'if', or any non-English equivalent also denoting a conditional. Conditionals are widespread in daily conversation. They are for instance used to give descriptions of causality, or to discuss future or hypothetical situations.

The use of conditionals displays an ability to imagine abstract situations and reason about them. Without conditionals, discourse and reasoning would be limited to situations which are in fact the case. There would be no ability to reason about the consequences of facts or to discuss what would be true in situations which are not currently the case. The whole of science is built on the idea of using theories and models to describe reality, to allow predictions for the future. This requires the ability to reason in terms of facts and their consequences, which requires the use of conditionals. It can therefore be said that the ability to use conditionals is a vital part of human intelligence.

A formal description of sentences with a conditional nature can play an important role in the development of artificial intelligence (Thomason, 2009). It may be a crucial element in building computer systems capable of human reasoning. These systems could be used to greatly enhance human-computer interaction by allowing computers to more quickly understand human rationale or to learn causal relationships in the same way as humans. Being able to formalize conditionals in a reliable manner would be a crucial part of building a system capable of understanding and synthesizing natural language as spoken by humans. Also, in order to create an autonomous robot capable of making decisions based on expectations of the results of its actions, this robot must be capable of understanding and using conditionals. These are only a few examples of the many situations in which a formal theory of conditionals will be helpful or even vital for applications of artificial intelligence.

1.1 Conditionals

Conditionals connect two statements in an asymmetrical manner. For instance, the sentence 'If it rains, the streets will get wet', may be separated into two

statements: 'it rains', and 'the streets will get wet'. Here 'if' acts as a binary operator connecting the two. An important task when developing a formal system is determining a semantical theory. This theory should answer the question of the nature of the conditional.

A straightforward way to formalize a complex sentence, like a sentence built around a conditional, would be to define its meaning in terms of the truth conditions of its parts. It is to be expected that the truth conditions of a complex sentence, as a whole, depend on the truth conditions of its constituents. For example the usage of 'and' as a connector of two statements can be interpreted using truth tables, as is done in many introductory classes of logic. In this case the complex sentence containing 'and' is true, if and only if both of its arguments are true (Carnap, 1985, p. 7).

As will be shown in chapter 2, unwanted, yet derivable, conclusions quickly arise when trying to define 'if' in a classical way by using Boolean truth values naively. Different logical systems have been developed to avoid these so-called paradoxes of the material conditional. Some attempt to do so by defining the truth value of conditionals in a more complex form of logics, while others completely deny the idea of conditionals having any form of truth value, in stead turning to ideas of acceptability and belief systems to formalize the conditional.

The aim of this thesis is to determine what problems arise when trying to formalize 'if' using compositional semantics, and whether the word 'if' can be formalized in a satisfactory manner at all in this way. To determine the answer to this question a comparison will be made between two approaches on conditionals: The work by Robert Stalnaker, which relies on possible world semantics to define the truth conditions of 'if' (Stalnaker, 1968), and an epistemic theory of conditionals developed by Peter Gärdenfors, which defines conditionals in terms of acceptance rather than truth (Gärdenfors, 1979).

1.2 Intuition

A question which must be answered before we can compare logical systems is how we are to determine which formal interpretation of conditionals is preferable. On what grounds should we decide which system should be chosen? Answering this question is not as simple as it may seem.

One approach is to view logical systems as a simplified model for human reasoning. In this case, the system should be able to interpret statements containing conditionals in a way comparable to how humans would intuitively interpret the same statement (Veltman, 1985, pp. 7-23).

This immediately leads us to a rather difficult question: How can we determine how humans intuitively interpret complex statements? Although there are logical statements whose interpretation may seem obvious, it is clearly not the case that logical intuitions are equal among all humans. Trained mathematicians, for instance, have rather different intuitions on implications than four-year-olds.

Another problem which arises when striving towards human-like intuition, is the fact that humans do make mistakes. A rather innocent looking logic puzzle known as the 'Wason selection task' tends to show human failures in conditional logic in as much as ninety percent of all test subjects (Wason and Shapiro, 1971).

The Wason selection task is a simple puzzle concerning four cards. Each of these cards has a number on one side and a color on the other. The four cards are placed in such a way that the faces show 3, 8, red, and brown. A test subject is now asked which card(s) should be turned over in order to test if the statement 'If a card shows an even number on one face, then its opposite face is red' is true.

In Wason's original experiment less than ten percent of his test subjects gave the correct answer, i.e. that the cards showing 8 and brown should be turned, and no other cards. In other words, intuitively most humans failed to correctly interpret the simple conditional statement. It should be noted that, when explained the rationale for the only correct answer, most people quite easily understood and accepted this explanation. This may signify a discrepancy between what we have as intuitions and what we are prepared to accept as correct interpretations of conditionals when given a systematic explanation.

A possible answer to the somewhat strange results of Wason's experiment also lies in the ambiguity of natural language. It may be the case that in this particular use of the conditional people tend to understand its meaning as 'only if', rather than just 'if'. The mistakes subjects made tend to support this idea. Also the fact people accepted the explanation given afterward, might be interpreted as those people only then understanding the form of 'if' intended.

The low scores would then be a result of human pragmatism in language, rather than the inability to have logically sound intuitions. This pragmatism could well be explained with Grice's theory of conversational implicatures (Grandy and Warner, 2009).

Whatever the cause may be, we should note that the explanation given for an interpretation of conditionals may be used to determine the correctness of this interpretation when human intuition is unclear.

In the field of artificial intelligence the connection between human reasoning and logics plays an important role. Modeling parts of natural intelligence using computers is still a very active field of research. For instances the fields of automated reasoning or natural language understanding would profit greatly from a formal theory of conditionals founded upon their human interpretation. A theory of conditionals which is to be used in these fields of artificial intelligence should therefore be able to produce results similar to human thinking. A logical system will however offer an idealization of human thought at best.

The discussion on the relation between human reasoning and logic is far from decided. Giving a complete description of this topic would go beyond the scope of this thesis. It is however important to note the connection between one's stance in this discussion and what one would expect from a logic of conditionals, and to also note the role this discussion plays in determining what makes a logical theory acceptable in general.

For the comparison made in this thesis it will suffice to hold the following criterion: The preferred logical theory will be the theory which defines conditionals in a way that is most comparable to human intuition, or, in cases where human intuition is unclear, the theory whose explanations of these cases is most acceptable. It may, after all, not only be useful to recreate and apply human intuition using artificial intelligence, but also to be able to use a logical theory to help interpret the way human reasoning works.

Chapter 2

Material implication

In this chapter we consider the most naive interpretation of what conditionals are. First we will give the common definition of this so-called *material* conditional. Later we will show in detail what problems arise when using this definition.

2.1 Definition

Classical logic has treated 'if' as a two-place logical connective, similar to 'and' and 'or'. Like the other Boolean connectives, this material implication (for which I will use the symbol '$\supset$') is defined depending on its arguments: The logical compound $p \supset q$ is true if and only if either the antecedent p is false or the consequent q is true, or both are true. This also means $p \supset q$ is false if and only if both p is true and q is false. The material implication $p \supset q$ is therefore equivalent to $\neg p \vee q$:

$$p \supset q \equiv \neg p \vee q \tag{2.1}$$

By definition the material implication is *truth-functional*. This means it is used as a function from the truth values of the arguments to a resulting truth value. The material implication is therefore either true or false, and nothing but the truth values of the input determines the resulting value. The validity of a material implication thus by definition entirely and only depends on the truth value of its arguments, and all its possible valuations can be captured in the four-row truth table 2.1.

p	q	$p \supset q$
0	0	1
0	1	1
1	0	0
1	1	1

Table 2.1: Truth table of the material implication, 0 stands for *false* and 1 stands for *true*

If limited to a truth-functional interpretation of conditionals, the definition given above is the most correct of the 16 possible definitions of a binary operator. It for instance correctly classifies a sentence of the form 'If it rains, the ground will get wet' as true in cases where it rains and the ground does get wet, and as false in cases where it rains but the ground stays dry. As mentioned above, it in fact interprets the sentence equivalently to 'Either it does not rain, or it does and the ground will get wet', an intuitively acceptable rewriting of the former sentence.

2.2 Paradoxes

The truth-functional definition of the conditional introduced above does quickly lead to unwanted, yet valid logical formula's (Wikipedia, 2010b). These so-called *paradoxes of material implication* are shown by applying regular rewriting rules for classical logic to simple formula's. Starting with perfectly acceptable logical formula's containing a material conditional, one can easily reach generally unacceptable formula's while cohering to the material definition of the conditional.

Many of the problems with the material conditional can be summed up in a single paradox: *the paradox of entailment*. This paradox may become more obvious when keeping in mind the equivalence between $p \supset q$ and $\neg p \vee q$. As can be seen from the last formula, or from just looking at the truth table 2.1 directly, any conditional containing an invalid antecedent (p in this case) will be valid.

This means for every false p and every q whether true or false, $p \supset q$ holds. As this means for instance $(p \wedge \neg p) \supset q$ will always hold, this is a very unintuitive interpretation of the conditional. Translated back into natural language, this means a sentence like 'If it is summer and it is not summer, then it is winter' would be valid. Undoubtedly, nobody with a decent grasp of the English language would find that sentence a certain truth.

One problem caused by this paradox of entailment is the material conditional's inability to correctly handle *counterfactual* conditionals. This category of conditionals consists of implicative statements where the antecedent is thought to be false, but assumed to be true.

Counterfactual conditionals are used to discuss what *would be* the case when something which is now false *were* true. They are therefore used when describing a hypothetical situation. This is in contrast with so-called *indicative* conditionals, which are used to discuss what *is* actually the case if an antecedent *is* actually true. An indicative conditionals is for instance 'If they improve their offense, Holland will win the cup next year', while a counterfactual conditional could be 'If they would have had a stronger offense, Holland would have won the cup this year'.

In English, counterfactuals are often recognized by the use of words like 'could', 'would', 'should', and 'might'. These statements are often heard in daily conversation in sentences like 'If I were not a physicist, I would be a musician', 'If you would have touched that wire, you would have hurt your hand', or 'Even if John would be here in 10 minutes, he would still miss the game' (where it is not likely that John will be able to arrive that soon). All of these statements can be categorized as counterfactual conditionals because their antecedent is false or highly unlikely.

Because of the required falsity of the antecedent in counterfactual statements and the paradox of entailment described above, if the counterfactual conditional behaved as the material conditional, every counterfactual conditional would by definition be true. The material implication is therefore not a satisfactory formalization of a counterfactual conditional.

Another paradox of the material implication can be summed up in the formula $q \supset (p \supset q)$. The validity of this formula can be read from the truth table 2.1. For every row where the consequent of a material implication (q in this case) is true, this implication is automatically valid. Therefore, for any currently true q, and for any p whether true or false, $p \supset q$ holds. This means a sentence like 'If the sun shines, the streets will be wet' is valid whenever the streets are wet, only because the streets are then wet, regardless of the antecedent of the implication or any relation between antecedent and consequent. A sentence like 'If the world explodes, the streets will be wet' is equally true whenever the streets are wet. This behavior is obviously counterintuitive.

Because of the paradoxes that the equivalence between $p \supset q$ and $\neg p \vee q$ causes, the truth-functional material implication cannot be used as a correct formal description of the human interpretation of conditionals. The truth-functional interpretation leads to counterintuitive validities, and its explanations for these validities are unacceptable. Alternative theories of conditionals have been given. In the next chapter we will look at an interpretation based on possible world semantics.

Chapter 3

Possible worlds

This chapter will introduce the concept of possible worlds in respect to conditionals. Because of the use for conditionals in discussing hypothetical situations, using a semantic theory for conditionals based on possible worlds seems like an attractive idea. In this chapter we discuss two very different interpretations of the conditional both based on possible worlds.

3.1 Strict conditional

This section introduces the strict conditional, a fairly naive theory on conditionals, based on the material conditional, but pertaining to possible worlds. We will first give a definition of this interpretation of conditionals, before discussing its flaws and shortcomings, which show its close relation to the material implication.

3.1.1 Definition

A first possible solution to the paradoxes of the material conditional may be found in the so-called *strict conditional*. This interpretation of the conditional (for which I will use the symbol '$\dashv\!\exists$') builds on the idea of possible worlds in modal logic. This interpretation defines the conditional 'if p then q' as $\Box(p \supset q)$.

The placement of the modal operator ($\Box$) over the material implication gives this formula a meaning roughly equivalent to 'in every possible world where p holds, q also holds'. The modal operator serves to make the implication valid only when it is a necessity. This means the implication is invalid whenever there is a possible world in which the material interpretation of the conditional does not hold. This leads to the following equivalence:

$$p \dashv\!\exists\, q \equiv \Box(p \supset q) \equiv \neg\Diamond\neg(p \supset q) \equiv \neg\Diamond\neg(\neg p \vee q) \equiv \neg\Diamond(p \wedge \neg q) \tag{3.1}$$

The strict conditional was developed by Clarence Irving Lewis as a solution for the paradoxes of material implication described above (Hunter, 2008). It does avoid the problems with counterfactual conditionals described in the previous chapter: While $p \supset q$ is by definition valid whenever p is currently false, $p \dashv\!\exists\, q$, i.e. $\Box(p \supset q)$, is not automatically valid in this case. Using modal logics

allows us to imagine a possible world where the antecedent, which is false in the current or actual world, may be true.

Let for instance p stand for 'Einstein was not a physicist', a statement which may be assumed to be false in the current world, since the opposite is true, and let q stand for 'Einstein was an artist', of which the validity in the current world is not relevant. Let us now interpret the compound statement 'If Einstein wasn't a physicist, he would have been an artist'. This is a paraphrased version of an actual quote by Einstein. We intuitively do not interpret this statement as a necessary truth, as we can imagine a situation where the antecedent is true, yet the consequent is false.

Using the material implication, because of the falsity of the antecedent and the paradox of entailment, this implication would be immediately valid. Using modal logics however, we can imagine possible worlds, different from the actual world, where the antecedent is true, i.e. worlds where Einstein was not a physicist.

The question the strict implication asks now is whether in every possible world where the antecedent is true, the consequent is true as well. Or: Is there no world possible where Einstein was not a physicist, but neither was he an artist?

Of course, we can easily imagine such a world. We could imagine an awful world where Einstein would have been killed by lightning as a child. This would lead him to never become a physicist. Even though this world is quite distant from the current world, this is a possible world we may assume to exist. This will make the antecedent of our compound statement (p) valid in this world. He would have never been able to become an artist in this world either. Therefore the consequent of the implication (q) would be invalid. In the sad and completely improbable Einstein killing world imagined here, the statement ($p \wedge \neg q$) would thus be valid.

Because of the existence of at least one of such possible worlds, the formula $\Diamond(p \wedge \neg q)$ is true. By the equivalence given in 3.1, this means $\Box(p \supset q)$ is false. Therefore in this case $p \not\Rightarrow q$ does not hold. This corresponds to an intuitive understanding of the statement 'If Einstein wasn't a physicist, he would have been an artist'. Although a rather poetic thing to say, it is by no means an absolute truth. Therefore we can agree with the strict implication dismissing it as a false statement and its explanation, which refers to the possibility of a situation where the implication is not fulfilled.

The strict implication also solves the paradox of $q \supset (p \supset q)$: The current validity of an implication's consequent is no longer sufficient for the validity of the implication as a whole:

Let p stand for 'the sun shines', let q stand for 'the streets are wet', and assume q is currently true, i.e. the streets are currently wet. Let us now consider the compound statement 'If the sun shines, the streets will be wet'. As mentioned above, using the material implication for 'if', this statement will be valid because of the validity of the consequent. But, although q is said to be currently valid, it is possible to imagine worlds where the streets are not wet, and therefore in those worlds $\neg q$. It is also possible that in some of those worlds the weather is sunny, and therefore p is valid. We now have that in some worlds $p \wedge \neg q$ holds, therefore, by definition of the modal operators, in the current or actual world, $\Diamond(p \wedge \neg q)$ holds. So $\neg(p \not\Rightarrow q)$ although $(p \supset q)$. Therefore, concerning this paradox, the strict implication better fits human intuitions on

implications as well.

3.1.2 Paradoxes of strict implication

There are also paradoxes of *strict* implication. These are similar to the paradoxes of material implication. The first paradox of strict implication is akin to the paradox of entailment for the material implication. It can be summed up in formula 3.2. This means an impossible proposition, i.e. a proposition which is not valid in any possible world, implies any consequent. This would mean a sentence like 'If 1 equals 2, you won't get any desert' is automatically valid when interpreted according to the strict conditional.

$$\neg\Diamond p \supset (p \dashv q) \tag{3.2}$$

$$\Box q \supset (p \dashv q) \tag{3.3}$$

The second paradox of strict implication is similar to the material paradox concerning $q \supset (p \supset q)$. This paradox is characterized by formula 3.3: If the consequent of a strict implication is a necessary truth, the implication is valid, regardless of the antecedent or how it is related to the consequent.

A statement is necessarily true when there is no possible world in which it does not hold. This is for example always the case for propositional tautologies. This means for instance the implication 'If you don't eat your vegetables, $2 + 2$ will equal 4' is valid when interpreted as a strict conditional.

This is certainly a strange thing to say. The correctness of this statement, however, may be actually not so straightforward. Using Grice's theory of conversational implicatures, it can be said that although, in most conversations, statements of this kind are inappropriate, they might actually be correctly seen as valid (Grandy and Warner, 2009).

The Gricean *Maxim of Quantity* can explain the problems with the statements. Their awkwardness is based on the fact that human conversation is bound by economical rules: Speakers will prefer saying only what is needed to share the information they are discussing. This also means that when speakers do add extra words to a sentence, these words are expected to add information.

Social awkwardness for example arises when adding an unnecessary conjunction to a proposition. The sentence 'It is raining and $2 + 2$ equals 4' is as informative as just the sentence 'It is raining', because '$2 + 2$ Equals 4' can be expected to be understood as always valid. It is strange and socially unwanted to needlessly complicate a sentence without adding information. This is exactly what would be predicted through Gricean Maxims.

Similarly, when hearing a necessary truth, people do not expect these truths to be embedded within an implication. When hearing an implication, people expect the connection between antecedent and consequent to offer extra information. Why else would the speaker have chosen to speak an implication where just speaking its consequent would have been enough?

It could therefore be said that $2 + 2$ does really equal 4 when you do not eat your vegetables, because $2 + 2$ always equals 4. This means, claiming the statement is valid, and using the idea of a necessary truth to explain this will probably be accepted by most people. Although people will generally find this interpretation somewhat annoying and forced. Similar to how responding to 'Do

you know what time it is?' with 'Yes', is seen as a technically correct but socially frustrating answer. This supports Grice's idea of conversational implicatures.

There has been much discussion about the connection between antecedent and consequent required to make a valid implication. A branch of conditional logics known as *relevance logics* is built around the idea that there should be a connection between antecedent and consequent (Mares, 2009). Implications would then base their validity on the existence of a derivable relevant relationship between antecedent and consequent.

The logics discussed in this thesis are however not based on such a relationship. This is based on the idea that even implications that lack such a connection are in general conversation sometimes accepted and are sometimes not (Stalnaker, 1968). Take for instance the sentence 'If an octopus had predicted Holland to win the cup, Holland would still have lost the final'. This sentence may be accepted by someone that thought Holland lost because their opponents simply had the better team. The predictions of an octopus would not change that, therefore the antecedent may be considered irrelevant to the consequent, yet the implication can be accepted or refuted.

The validity of the implication 'If $2 + 2$ equals 5, $2 + 2$ will equal 4' is actually quite difficult to convince people of. The correct interpretation of this implication might be better explained by the variation in the accessibility relation used in modal logics. This relation determines which worlds are considered accessible from the current world, and thus are considered possible worlds. In general conversation, the rules of arithmetics are accepted as always valid, necessary truths. However, using a contradicting arithmetic rule in the antecedent might cause the listener to think worlds possible where the arithmetic rules are different.

This would mean there could be, in the context of the conversation, a possible world where arithmetic is not as we know it. Therefore there might be a possible world where both $2 + 2$ equals 5 and $2 + 2$ does not equal 4. In this world the material conditional would not hold, therefore the strict conditional 'If $2 + 2$ equals 5, $2 + 2$ will equal 4' would actually be invalid. This seems to be an intuitive explanation.

The example shows how the variance of the accessibility relation helps determine the validity of a strict conditional. How the accessibility relation should be interpreted is a difficult question. The strict conditional relies heavily on its definition of possible worlds. Therefore, to be able to use the strict conditional in a formal system, it would require a clear formalization of its accessibility relation. The strict conditional by itself is therefore useless as an explanation of a conversational implication without a decent formal theory of what is considered possible by speaker and listeners.

It should be noted that possible worlds where the rules of arithmetics do not hold are generally not used in possible world semantics. These worlds could be said to be contradictory and therefore non-existent. There are however variants of modal logic which do allow the reasoning about such worlds, or even worlds where the rules of propositional logics are different. These worlds are then sometimes called *impossible* or *non-normal* worlds.(Zalta, 1997)

The strict implication also has some inherent characteristics which are different from the use of conditionals in daily conversation. These characteristics will be discussed in the following sections.

3.1.3 Transitivity

Let p stand for 'John lived in Japan', let q stand for 'John speaks Japanese fluently', and let r stand for 'John can come here every Wednesday to teach us Japanese'. Now $p \rightarrow q$ stands for 'If John lived in Japan, he would speak Japanese fluently', and $q \rightarrow r$ stands for 'If John would speak Japanese fluently, he could come here every Wednesday to teach us'.

It is quite intuitive to accept both $p \rightarrow q$ and $q \rightarrow r$, but to deny $p \rightarrow r$, which would mean 'If John lived in Japan, he could come here every Wednesday to teach us Japanese'. Therefore it seems intuitive for a preferred formalization of the implication to not be automatically transitive. The strict conditional, on the contrary, *is* transitive, because of the transitivity of the material conditional:

$$\dfrac{\dfrac{\Box(p \supset q) \qquad \Box(q \supset r)}{\Box(p \supset q \wedge q \supset r)}}{\Box(p \supset r)}$$

This means, regarding transitivity, the strict conditional does not behave as is intuitively expected of conditionals.

3.1.4 Negation

Another unwanted characteristic for the strict conditional has to do with its negation. For instance, what would it mean to deny a sentence like 'If Robben had been in better form, Holland would have won the cup'? A possible view on the negation of conditionals, based on work by Chisholm, is that while the antecedent does not change, the consequent will be negated when an implication is negated (Chisholm, 1946). It does indeed seem possible that someone would reply to the previous sentence with 'That's not true! If Robben had been fit, Holland would still not have won the cup'. This negation is written as a strict conditional in formula 3.4.

The naive negation of a strict conditional would however be interpreted differently. Consider again the sentence 'If Robben had been in better form, Holland would have won the cup'. The naive negation of the strict conditional interpretation of this sentence, as seen in formula 3.5, would be already valid when it is possible that Robben was in better form, and Holland still had not won. The more intuitive reading of the negation, as seen in formula 3.4, would be valid only when it is impossible that Robben was in better form, and Holland won the cup. This leads to the conclusion that the strict conditional does not naively handle negation the way it is intuitively understood.

It could be said that the problem lies not with the strict implication, but with the definition of the negation in this case. It would however require a formal definition of the negation of the strict conditional, before this can be used as a satisfactory interpretation of the implication.

$$p \dashv \neg q \equiv \Box(p \supset \neg q) \equiv \neg\Diamond(p \wedge q) \tag{3.4}$$

$$\neg(p \dashv q) \equiv \neg\Box(p \supset q) \equiv \Diamond(p \wedge \neg q) \tag{3.5}$$

3.1.5 Contraposition

A third unwanted characteristic of the strict conditional is that it allows the inference of contraposition. This concept can be seen in formula 3.6. It is a direct result of the characteristics of the material conditional described in 3.7.

$$(p \dashv q) \leftrightarrow (\neg q \dashv \neg p) \tag{3.6}$$

$$p \supset q \equiv \neg p \vee q \equiv \neg\neg q \vee \neg p \equiv \neg q \supset \neg p \tag{3.7}$$

There are instances of the implication where the inference of contraposition seems acceptable. By accepting 'If it rains, the streets will get wet' it seems intuitive to also accept 'If the streets don't get wet, it does not rain'.

But, this inference cannot always be made. Take for instance the sentence 'If Robben was in better form, Holland still would not have won the cup'. This could be accepted by someone who thought it was not Robben's physical condition but maybe the Dutch defense that caused Holland to not win the cup. The contraposition of the sentence would be 'If Holland would have won the cup, Robben was not in better form'. This would not be a necessarily acceptable implication, although it is the contraposition of an accepted implication. Therefore, the inference of contraposition does not generally hold for the implication in general conversation.

3.2 Stalnaker conditional

Robert Stalnaker proposed another definition of the implication, which is also based on modal logics. This conditional, for which he used the symbol $>$, is built around the idea of finding a possible world which differs minimally from the current one (Stalnaker, 1968).

3.2.1 Stalnaker's theory of conditionals

In his 1968 article 'A Theory of Conditionals' Stalnaker lays out the ground work for a system which tackles both what he refers to as the *logical problem of conditionals* and what he calls the *pragmatic problem of conditionals*. He defines these problems as follows:

The *logical* problem of conditionals is the search by logicians to find an acceptable formal definition of the truth conditions of the implication. This definition should provide the rules concerning the use of implication in logical formula's and proofs. Solving the problem of conditionals would mean having found a formal definition of conditionals which can be used to interpret implications and how they relate to the facts, in order to determine the validity of the implications.

Even with a solid formalization of the conditional, it may be possible that multiple valuations of a conditional are consistent with the facts. The task set by the *pragmatic* problem of conditionals is, according to Stalnaker, '[...] to find and defend criteria for choosing among these different valuations'.

The boundary between these two problems is somewhat unclear. A stricter semantic theory could leave less interpretation to pragmatics, while a more vague semantic formalization could rely more heavily on a pragmatic theory

to solve any ambiguities. A theory which does not give truth conditions, but in stead offers a formalization of conditions for justified belief, is considered purely pragmatic. Stalnaker intends to provide a theory which defines truth conditions for implications, but which also depends on the context of use, and thus pragmatics.

Central to the theory is the concept of a possible world which differs minimally from the actual world. The benefit of using possible worlds was shown when introducing the strict conditional and its ability to handle counterfactual conditionals in section 3.1.1. In order to discuss a situation where the antecedent of an implication is not currently the case, it would help to be able to determine a world where this antecedent *is* the case.

To be able to handle implications where the antecedent is impossible, Stalnaker introduces a world λ. This world is called the absurd world. In this world contradictions and all their consequences are true. This is an extension of usual modal semantics, where impossible statements would have no world to discuss them in.

Stalnaker defines a selection-function f which is used to find the nearest possible world in which the antecedent of an implication is true. For instance, $f(A, \alpha)$ would find the world, most similar to world α, where A is the case. For antecedents which are true in the actual world, the function f will return the actual world itself. This is because there would be no changes needed to fulfill the antecedent, and no world is more similar to the actual world, than the actual world itself.

The function f adheres to the following conditions: where $f(A, \alpha) = \beta$, A is the *antecedent*, α is the *base world*, and β is the *selected world*.

Condition 1. *For all antecedents A and base worlds α, A must be true in $f(A, \alpha)$.*

Condition 2. *For all antecedents A and base worlds α, $f(A, \alpha) = \lambda$ only if there is no world possible with respect to α in which A is true.*

Condition 3. *For all base worlds α and antecedents A, if A is true in α, then $f(A, \alpha) = \alpha$.*

Condition 4. *For all base worlds α and antecedents B and B', if B is true in $f(B', \alpha)$ and B' is true in $f(B, \alpha)$, then $f(B, \alpha) = f(B', \alpha)$.*

Stalnaker readily admits that these conditions do not define this function uniquely. Which selection-function is to be used differs between conversations. Therefore the further definition of a selection-function would be a task for pragmatics. The four conditions do allow the definition of a semantic theory of conditional validity. He does so by defining the formal system C2.

The system C2 uses the standard propositional connectives $\neg$, $\supset$, $\wedge$, and $\vee$, the modal operators $\Box$ and $\Diamond$, and a conditional connective $>$ (called the corner).

The rules of inference for C2 are *modus ponens* (if A and $A \supset B$ are theorems, then B is a theorem), and *necessity* (if A is a theorem, then $\Box A$ is a theorem). The system has seven axiom schemata.

Axiom 1. *Any tautologous formula is an axiom.*

Axiom 2. $\Box(A \supset B) \supset (\Box A \supset \Box B)$

Axiom 3. $\Box(A \supset B) \supset (A > B)$

Axiom 4. $\Diamond A \supset ((A > B) \supset \neg(A > \neg B))$

Axiom 5. $A > (B \vee C) \supset ((A > B) \vee (A > C))$

Axiom 6. $(A > B) \supset (A \supset B)$

Axiom 7. $((A > B) \wedge (B > A)) \supset ((A > C) \supset (B > C))$

The resulting conditional connective finds its place in between the strict conditional and the material conditional. This means, $A \dashv B$ entails $A > B$, and $A > B$ entails $A \supset B$, through respectively Axiom 3 and Axiom 6.

Unlike the material and strict conditional, the Stalnaker conditional is not transitive. Provided that its antecedent is not impossible, it handles the negation of a conditional as the negation of its consequent under the same antecedent. This corresponds with the intuitive idea as described in 3.1.4. The inference of contraposition is not possible for the Stalnaker conditional. Therefore the strange behavior of the strict and material conditional does not apply to the Stalnaker conditional.

3.2.2 Analysis

There is a great deal that Stalnaker's semantic theory does not cover. The selection of a nearest possible world, is only vaguely defined and therefore depends completely on the pragmatic context. This selection-function is at the core of the truth-conditions of the Stalnaker conditional. Therefore the question of validity of a conditional shifts from a semantic to a pragmatic problem, which Stalnaker does not solve.

In a later article, Stalnaker admits that his theory does not offer a reduction of the problem of conditionals (Stalnaker, 1975). He does however state that his theory offers insight into the *form* of the truth conditions for conditionals.

Can the Stalnaker conditional actually be seen as sufficiently specified, when the actual truth-conditions depend so much on the selection-function chosen by individual agents? As discussed in relation to the strict conditional, what is considered a possible world is subject to change. Stalnaker refers to the set of worlds considered possible as the *context set*.

According to Stalnaker, this context set is determined by the content of the statements uttered in a conversation. This happens because statements are only appropriate within a certain context. Therefore by uttering a statement, a context can be deduced.

It is however not certain that two people will deduce the same context from the same utterance. A more 'closed minded' speaker, for instance someone who would never doubt the existence of a Christian god, will hold a different context possible than an agnostic listener. It seems the context set is subject to the beliefs of the agents involved in the conversation.

If we were to look at this context set in a more objective way, there is no reason why we would not consider, for instance, a world in which the laws of physics do not hold. Why is the context set not the set of all possible worlds? The only reasons for excluding worlds from the context set can be personal beliefs. The selection function, which acts on the context set, seems therefore to be based more on personal beliefs than on an ontological theory of possible worlds.

Chapter 4

Conditional belief

In this chapter we will discuss a theory of conditionals based on belief. The theory which will be examined is developed by Peter Gärdenfors and relies on changes of belief to allow for the analysis of counterfactual conditions. We will give a general description of Gärdenfors's theory, compare this with Stalnaker's theory of conditionals and provide a general analysis of the strengths and weaknesses of Gärdenfors's theory.

4.1 Definition

In his article 'Conditionals and Changes of Belief', Gärdenfors sets out a theory of conditionals based on epistemic notions, rather than ontological notions (Gärdenfors, 1979). The fundamental concepts of his theory are *states of belief* and *changes of belief*. The criteria of acceptability he defines are based on a suggestion made by Ramsey: 'Accept $A \rightarrow B$ in a state of belief P if and only if the minimal change of P needed to accept A also requires accepting B'.

Gärdenfors defines a *belief set* P as a set of formulas which satisfies the following conditions:

Condition 1. *P is non-empty*

Condition 2. *if $A \in P$ and $B \in P$, then $A \wedge B \in P$*

Condition 3. *if $A \in P$ and $A \supset B$ is a truth-functional tautology, then $B \in P$*

Here A and B are formulas in a truth-functional, propositional logic. The belief set contains all sentences which an agent accepts or knows to be true. This is of course an idealized view on beliefs as, according to Gärdenfors's definition, every agent would believe all truth-functional consequences of his beliefs. It is however also allowed for belief sets to contain contradictory beliefs. This means it is perfectly possible for a formula p, that a belief set P contains both p and $\neg p$. This could obviously never be the case for possible worlds.

Belief sets need not be complete: It is not the case that for every sentence A, either A or $\neg A$ is part of an agents belief set. If this were the case, the agent would be omniscient.

A special belief set is the set of all formula's. This set is called the *absurd* belief set, as for any formula A, both A and $\neg A$ will be an element of this set.

This absurd belief set is of use when discussing conditionals with an impossible antecedent. This is similar to Stalnaker's introduction of an absurd world λ which is used in the same way.

The concept of a belief set can now be used to facilitate a change in beliefs. This allows us to discuss conditionals with an antecedent which is not currently part of a persons belief set. This is for instance the case for many counterfactual conditionals.

For each belief set P and each sentence A, Gärdenfors assumes that there is a unique belief set P_A, which represents the state of belief which is the result of minimally changing P to include A. Using this idea, $A > B \in P$ iff $B \in P_A$. The conditional is thus accepted if and only if its consequent is accepted in the closest belief set which accepts the antecedent.

Gärdenfors gives some conditions to which the function which allows a change of belief to a minimally differing belief set will need to adhere:

Condition 4. *for all belief sets P, $A \in P_A$.*

Condition 5. *for all belief sets P, if $A \in P_B$, and $B \in P_A$, then $P_A = P_B$.*

Condition 6. *for all belief sets P, if $A \in P$, then $P = P_A$.*

This change of belief sets reminds us of the selection-function in Stalnaker's theory. For every change of belief there is a unique belief set the result of applying this change of belief, similar to how Stalnaker's selection-function returns one single minimally different possible world for an antecedent which is interpreted as true.

4.2 Analysis

In Gärdenfors's theory, belief sets take up a similar role as possible worlds do in Stalnaker's theory. They both allow for the analysis of situations which are not currently the case. Possible worlds however correspond with complete belief sets. As explained above this means in a possible world w, for every sentence A, either A or $\neg A$ is the case.

In general, a distinction can be made between two forms of belief change: A belief *update* concerns adapting beliefs after a change in the situation, facts which were previously believed to be true may have changed in the new situation. A belief *revision* concerns adapting previously held beliefs because of new information on a situation. Although the situation has not changed in this case, added information may require an agent to change some previously held belief, which turns out to be unreliable.

As was the case with Stalnaker's selection-function for the closest possible world, finding the minimal required change of beliefs to allow for an antecedent to be accepted is a somewhat vague task. If we view this problem as related to belief updates, this problem seems to be closely related to the epistemic frame problem.

This philosophical problem concerns the uncertainty for an agent performing an action, and thereby changing the world, in readjusting its beliefs to take in account the consequences of this action (Shanahan, 2009). For instance, after pushing a cart across a room, an agent needs to adjust its beliefs concerning the location of the cart. It is natural to assume that the agent does not need

to adjust its beliefs concerning the color of the walls of the room, or the atomic weight of carbon, after moving a cart.

There is however no way to tell what beliefs need to be adjusted without a substantial knowledge of relevance relations in the world of an agent. And even with a perfect knowledge of all relevance relations, the computational complexity of a change of belief would still span the total of all beliefs. For each belief an agent holds, it should decide whether, based on his knowledge of relevance, the belief still holds after performing an action. This makes the epistemic frame problem a large hurdle in, for instance, developing autonomous agents in open world environments.

Determining the change of beliefs required to make the antecedent of a conditional true, leads to the same need for a substantial knowledge of relevance relations and an unmaintainable computational complexity. There is again substantial knowledge of the world required for an agent to correctly revise its beliefs without missing any required changes.

If we take for instance the sentence "If I push this cart to the north of the room, the color of the northern wall of the room will become red". For an agent to correctly accept or reject this sentence, it is required to know the consequences of pushing the cart to the north of the room. The agent should perhaps have knowledge of the influence of colored lights on nearby surfaces, the behavior of things mounted on top of cars, and so on. Without this knowledge, the agent will be unable to reach the correct belief set when revising its beliefs based on the antecedent of the conditional.

To illustrate the problem with belief updates, take for example an agent α which holds the following beliefs:

1. I am in a room.
2. In this room there is a cart.
3. I can push the cart from the southern end of the room to the northern end or back.
4. The cart is at the southern end of the room.
5. On the cart is a bomb.
6. The bomb is at the southern end of the room.

How would agent α now respond to a sentence like 'If you move the cart to the northern end of the room, the bomb will be at the northern end of the room'? The agent will first have to revise its fourth belief, since in the antecedent of the conditional, the cart is no longer at the southern end of the room. The agent must change its beliefs to accommodate this.

The agent should however also revise any beliefs which are influenced by the fourth belief. One of these is the last belief of the list. How is agent α supposed to know to also revise this last belief, without knowledge on the behavior of bombs placed on top of carts?

To correctly accept or reject conditional statements using Gärdenfors's theory of conditionals seems to require knowledge of the consequences of the antecedents of these statements. This theory can therefore be said to offer no

reduction of the problem of conditionals. Being able to make correct judgements of conditional requires the ability to perform correct changes of belief. The ability to make correct changes of belief requires being able to know the consequences of any antecedent, which would already solve the problem of conditionals.

Gärdenfors's theory can therefore be said to offer no reduction of the problem. It can however provide a formal framework which could also be of use in multi agent systems. Being able to discuss the consequences of beliefs of individual agents, based on their individual beliefs, could be useful in determining the behavior of agents in complex models. This is something Stalnaker's theory of minimal change cannot account for.

Gärdenfors's theory also is at its current form of no direct practical use in the development of an artificially intelligent reasoning agent capable of judging conditionals. The mechanism for changing beliefs requires a significant knowledge of the world and would lead to an unmanageable computational complexity when implemented naively.

Chapter 5

Conclusion

The truth-functional material conditional quickly leads to unwanted behavior and is therefore no usable formalization of the conversational implication. The strict conditional does avoid some of the more obvious paradoxes of the material implication, and is still extremely straightforward and simple in terms of definition. There are however still certain characteristics to the truth-conditions of the strict conditional which make it unfit as a proper description.

Stalnaker's theory, based on the ontological notion of possible worlds, leads to many doubts: Although it is only meant as a basis for a more complete conditional theory, the use of a unique closest possible world leads to some difficulties. Determining the validity of a counterfactual conditional would even require omniscience. It does therefore seem that certainly concerning counterfactual conditionals the use of truth-values leads to an unsolvable problem. Concerning counterfactual conditionals, nothing more than acceptability based on belief could in practice be formalized.

Gärdenfors gives a usable formal description of belief states. It seems plausible to found a theory of conditionals on personal beliefs. This does however lead to the inability to assign truth values to conditional statements. In an AI environment where an artificial agent would partake in general conversation, a theory based on beliefs could provide enough certainty. In general conversation, humans are not required to be objectively certain of the validity of their statements, although their willingness to accept a statement *is* generally important.

The formalization in both cases still lacks a most important tool. The selection of a closest situation where a counterfactual antecedent is made true, is at the core of the problem of conditionals. The mechanism of belief revision and world selection seems to be closely related to the frame problem, which prevents much practical use to both of these theories. This selection function depends largely on context and knowledge of the world. It would therefore seem that the development of a strong pragmatic theory of conditionals could be helpful in the path to a usable formal system describing conditionals.

Bibliography

Adams, E. (1966). Probability and the Logic of Conditionals*. *Studies in Logic and the Foundations of Mathematics*, 43:265–316.

Adams, E. (1975). *The logic of conditionals: An application of probability to deductive logic*. Reidel.

Carnap, R. (1985). *Introduction to symbolic logic and its applications*, volume 453 of *Dover books on western philosophy*. Courier Dover Publications.

Chisholm, R. (1946). The Contrary-to-fact Conditional. *Mind*, 55(219):289.

Gärdenfors, P. (1979). Conditionals and Changes of Belief. In Niiniluoto, I. and Tuomela, R., editors, *The logic and epistemology of scientific change*, pages 381–404. North-Holland, Amsterdam.

Grandy, R. and Warner, R. (2009). Paul Grice. In Zalta, E., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2009 edition.

Hunter, B. (2008). Clarence irving lewis. In Zalta, E., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition.

Jackson, F., editor (1991). *Conditionals*. Oxford university press Oxford.

Mares, E. (2009). Relevance logic. In Zalta, E., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2009 edition.

Shanahan, M. (2009). The frame problem. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2009 edition.

Stalnaker, R. (1968). A theory of conditionals. *Studies in logical theory*, 2:98–112.

Stalnaker, R. (1975). Indicative conditionals. *Philosophia*, 5(3):269–286.

Thomason, R. (2009). Logic and artificial intelligence. In Zalta, E., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2009 edition.

Veltman, F. (1985). *Logics for conditionals*. Druk Jurriaans bv., Amsterdam.

Wason, P. and Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *The Quarterly Journal of Experimental Psychology*, 23(1):63–71.

Wikipedia (2010a). Conditional sentence — wikipedia, the free encyclopedia. [Online; accessed 10-August-2010].

Wikipedia (2010b). Paradoxes of material implication — wikipedia, the free encyclopedia. [Online; accessed 4-August-2010].

Zalta, E. (1997). A classically-based theory of impossible worlds. *Notre Dame Journal of Formal Logic*, 38(4):640–660.