Part I

Counterfactuals

Chapter 1

Introduction

1.1 The Material Conditional

cnt:int:mat:sec

In its simplest form in English, a conditional is a sentence of the form "If ... then ...," where the ... are themselves sentences, such as "If the butler did it, then the gardener is innocent." In introductory logic courses, we earn to symbolize conditionals using the $\to$ connective: symbolize the parts indicated by ..., e.g., by formulas φ and ψ, and the entire conditional is symbolized by $\varphi \to \psi$.

The connective $\to$ is *truth-functional*, i.e., the truth value—$\mathbb{T}$ or $\mathbb{F}$—of $\varphi \to \psi$ is determined by the truth values of φ and ψ: $\varphi \to \psi$ is true iff φ is false or ψ is true, and false otherwise. Relative to a truth value assignment $\mathfrak{v}$, we define $\mathfrak{v} \vDash \varphi \to \psi$ iff $\mathfrak{v} \nvDash \varphi$ or $\mathfrak{v} \vDash \psi$. The connective $\to$ with this semantics is called the *material conditional*.

This definition results in a number of elementary logical facts. First of all, the deduction theorem holds for the material conditional:

$$\text{If } \Gamma, \varphi \vDash \psi \text{ then } \Gamma \vDash \varphi \to \psi \tag{1.1}$$

It is truth-functional: $\varphi \to \psi$ and $\neg\varphi \vee \psi$ are equivalent:

$$\varphi \to \psi \vDash \neg\varphi \vee \psi \tag{1.2}$$
$$\neg\varphi \vee \psi \vDash \varphi \to \psi \tag{1.3}$$

A material conditional is entailed by its consequent and by the negation of its antecedent:

$$\psi \vDash \varphi \to \psi \tag{1.4}$$
$$\neg\varphi \vDash \varphi \to \psi \tag{1.5}$$

A false material conditional is equivalent to the conjunction of its antecedent and the negation of its consequent: if $\varphi \to \psi$ is false, $\varphi \wedge \neg\psi$ is true, and vice versa:

$$\neg(\varphi \to \psi) \vDash \varphi \wedge \neg\psi \tag{1.6}$$
$$\varphi \wedge \neg\psi \vDash \neg(\varphi \to \psi) \tag{1.7}$$

The material conditional supports modus ponens:

$$\varphi, \varphi \to \psi \vDash \psi \tag{1.8}$$

The material conditional agglomerates:

$$\varphi \to \psi, \varphi \to \chi \vDash \varphi \to (\psi \wedge \chi) \tag{1.9}$$

We can always strengthen the antecedent, i.e., the conditional is *monotonic*:

$$\varphi \to \psi \vDash (\varphi \wedge \chi) \to \psi \tag{1.10}$$

The material conditional is transitive, i.e., the chain rule is valid:

$$\varphi \to \psi, \psi \to \chi \vDash \varphi \to \chi \tag{1.11}$$

The material conditional is equivalent to its contrapositive:

$$\varphi \to \psi \vDash \neg\psi \to \neg\varphi \tag{1.12}$$
$$\neg\psi \to \neg\varphi \vDash \varphi \to \psi \tag{1.13}$$

These are all useful and unproblematic inferences in mathematical reasoning. However, the philosophical and linguistic literature is replete with purported counterexamples to the equivalent inferences in non-mathematical contexts. These suggest that the material conditional $\to$ is not—or at least not always—the appropriate connective to use when symbolizing English "if ... then ..." statements.

1.2 Paradoxes of the Material Conditional

cnt:int:par:sec

One of the first to criticize the use of $\varphi \to \psi$ as a way to symbolize "if ... then ..." statements of English was C. I. Lewis. Lewis was criticizing the use of the material condition in Whitehead and Russell's *Principia Mathematica*, who pronounced $\to$ as "implies." Lewis rightly complained that if $\to$ meant "implies," then any false proposition p implies that p implies q, since $p \to (p \to q)$ is true if p is false, and that any true proposition q implies that p implies q, since $q \to (p \to q)$ is true if q is true.

Logicians of course know that *implication*, i.e., logical entailment, is not a connective but a relation between formulas or statements. So we should just

not read $\to$ as "implies" to avoid confusion.[1] As long as we don't, the particular worry that Lewis had simply does not arise: p does not "imply" q even if we think of p as standing for a false English sentence. To determine if $p \vDash q$ we must consider *all* valuations, and $p \nvDash q$ even when we use p to symbolize a sentence which happens to be false.

But there is still something odd about "if ... then..." statements such as Lewis's

> If the moon is made of green cheese, then $2 + 2 = 4$.

and about the inferences

> The moon is not made of green cheese. Therefore, if the moon is made of green cheese, then $2 + 2 = 4$.
>
> $2 + 2 = 4$. Therefore, if the moon is made of green cheese, then $2 + 2 = 4$.

Yet, if "if ... then ..." were just $\to$, the sentence would be unproblematically true, and the inferences unproblematically valid.

Another example of concerns the tautology $(\varphi \to \psi) \vee (\psi \to \varphi)$. This would suggest that if you take two indicative sentences S and T from the newspaper at random, the sentence "If S then T, or if T then S" should be true.

1.3 The Strict Conditional

cnt:int:str:sec

Lewis introduced the *strict conditional* $\varsubsetneq$ and argued that it, not the material conditional, corresponds to implication. In alethic modal logic, $\varphi \varsubsetneq \psi$ can be defined as $\Box(\varphi \to \psi)$. A strict conditional is thus true (at a world) iff the corresponding material conditional is necessary.

How does the strict conditional fare vis-a-vis the paradoxes of the material conditional? A strict conditional with a false antecedent and one with a true consequent, may be true, or it may be false. Moreover, $(\varphi \varsubsetneq \psi) \vee (\psi \varsubsetneq \varphi)$ is not valid. The strict conditional $\varphi \varsubsetneq \psi$ is also not equivalent to $\neg\varphi \vee \psi$, so it is not truth functional.

We have:

$$\varphi \varsubsetneq \psi \vDash \neg\varphi \vee \psi \text{ but:} \tag{1.14}$$
$$\neg\varphi \vee \psi \nvDash \varphi \varsubsetneq \psi \tag{1.15}$$
$$\psi \nvDash \varphi \varsubsetneq \psi \tag{1.16}$$
$$\neg\varphi \nvDash \varphi \varsubsetneq \psi \tag{1.17}$$
$$\neg(\varphi \to \psi) \nvDash \varphi \wedge \neg\psi \text{ but:} \tag{1.18}$$
$$\varphi \wedge \neg\psi \vDash \neg(\varphi \varsubsetneq \psi) \tag{1.19}$$

[1] Reading "$\to$" as "implies" is still widely practised by mathematicians and computer scientists, although philosophers try to avoid the confusions Lewis highlighted by pronouncing it as "only if."

However, the strict conditional still supports modus ponens:

$$\varphi, \varphi \mathrel{\neg\!\!\!\Rightarrow} \psi \vDash \psi \tag{1.20}$$

The strict conditional agglomerates:

$$\varphi \mathrel{\neg\!\!\!\Rightarrow} \psi, \varphi \mathrel{\neg\!\!\!\Rightarrow} \chi \vDash \varphi \mathrel{\neg\!\!\!\Rightarrow} (\psi \wedge \chi) \tag{1.21}$$

Antecedent strengthening holds for the strict conditional:

$$\varphi \mathrel{\neg\!\!\!\Rightarrow} \psi \vDash (\varphi \wedge \chi) \mathrel{\neg\!\!\!\Rightarrow} \psi \tag{1.22}$$

The strict conditional is also transitive:

$$\varphi \mathrel{\neg\!\!\!\Rightarrow} \psi, \psi \mathrel{\neg\!\!\!\Rightarrow} \chi \vDash \varphi \mathrel{\neg\!\!\!\Rightarrow} \chi \tag{1.23}$$

Finally, the strict conditional is equivalent to its contrapositive:

$$\varphi \mathrel{\neg\!\!\!\Rightarrow} \psi \vDash \neg\psi \mathrel{\neg\!\!\!\Rightarrow} \neg\varphi \tag{1.24}$$

$$\neg\psi \mathrel{\neg\!\!\!\Rightarrow} \neg\varphi \vDash \varphi \mathrel{\neg\!\!\!\Rightarrow} \psi \tag{1.25}$$

Problem 1.1. Give **S5**-counterexamples to the entailment relations which do not hold for the strict conditional, i.e., for:

1. $\neg p \nvDash \Box(p \to q)$
2. $q \nvDash \Box(p \to q)$
3. $\neg\Box(p \to q) \nvDash p \wedge \neg q$
4. $\nvDash \Box(p \to q) \vee \Box(q \to p)$

Problem 1.2. Show that the valid entailment relations hold for the strict conditional by giving **S5**-proofs of:

1. $\Box(\varphi \to \psi) \vDash \neg\varphi \vee \psi$
2. $\varphi \wedge \neg\psi \vDash \neg\Box(\varphi \to \psi)$
3. $\varphi, \Box(\varphi \to \psi) \vDash \psi$
4. $\Box(\varphi \to \psi), \Box(\varphi \to \chi) \vDash \Box(\varphi \to (\psi \wedge \chi))$
5. $\Box(\varphi \to \psi) \vDash \Box((\varphi \wedge \chi) \to \psi)$
6. $\Box(\varphi \to \psi), \Box(\psi \to \chi) \vDash \Box(\varphi \to \chi)$
7. $\Box(\varphi \to \psi) \vDash \Box(\neg\psi \to \neg\varphi)$
8. $\Box(\neg\psi \to \neg\varphi) \vDash \Box(\varphi \to \psi)$

However, the strict conditional still has its own "paradoxes." Just as a material conditional with a false antecedent or a true consequent is true, a strict conditional with a *necessarily* false antecedent or a necessarily true consequent is true. Moreover, any true strict conditional is necessarily true, and any false strict conditional is necessarily false. In other words, we have

$$\Box\neg\varphi \vDash \varphi \mathrel{⥽} \psi \tag{1.26}$$

$$\Box\psi \vDash \varphi \mathrel{⥽} \psi \tag{1.27}$$

$$\varphi \mathrel{⥽} \psi \vDash \Box(\varphi \mathrel{⥽} \psi) \tag{1.28}$$

$$\neg(\varphi \mathrel{⥽} \psi) \vDash \Box\neg(\varphi \mathrel{⥽} \psi) \tag{1.29}$$

These are not problems if you think of $\mathrel{⥽}$ as "implies." Logical entailment relationships are, after all, mathematical facts and so can't be contingent. But they do raise issues if you want to use $\mathrel{⥽}$ as a logical connective that is supposed to capture "if ... then ...," especially the last two. For surely there are "if ... then ..." statements that are contingently true or contingently false—in fact, they generally are neither necessary nor impossible.

Problem 1.3. Give proofs in **S5** of:

1. $\Box\neg\psi \vDash \varphi \mathrel{⥽} \psi$
2. $\varphi \mathrel{⥽} \psi \vDash \Box(\varphi \mathrel{⥽} \psi)$
3. $\neg(\varphi \mathrel{⥽} \psi) \vDash \Box\neg(\varphi \mathrel{⥽} \psi)$

Use the definition of $\mathrel{⥽}$ to do so.

1.4 Counterfactuals

cnt:int:cnt:sec

A very common and important form of "if ... then ..." constructions in English are built using the past subjunctive form of *to be*: "if it were the case that ... then it would be the case that ..." Because usually the antecedent of such a conditional is false, i.e., counter to fact, they are called *counterfactual conditionals* (and because they use the subjunctive form of *to be*, also *subjunctive conditionals*. They are distinguished from *indicative* conditionals which take the form of "if it is the case that ... then it is the case that ..." Counterfactual and indicative conditionals differ in truth conditions. Consider Adams's famous example:

> If Oswald didn't kill Kennedy, someone else did.
>
> If Oswald hadn't killed Kennedy, someone else would have.

The first is indicative, the second counterfactual. The first is clearly true: we know President John F. Kennedy was killed by *someone*, and if that someone wasn't (contrary to the Warren Report) Lee Harvey Oswald, then someone else killed Kennedy. The second one says something different. It claims that

if Oswald hadn't killed Kennedy, i.e., if the Dallas shooting had been avoided or had been unsuccessful, history would have subsequently unfolded in such a way that another assassination would have been successful. In order for it to be true, it would have to be the case that powerful forces had conspired to ensure JFK's death (as many JFK conspiracy theorists believe).

It is a live debate whether the *indicative* conditional is correctly captured by the material conditional, in particular, whether the paradoxes of the material conditional can be "explained" in a way that is compatible with it giving the truth conditions for English indicative conditionals. By contrast, it is uncontroversial that counterfactual conditionals cannot be symbolized correctly by the material conditionals. That is clear because, even though generally the antecedents of counterfactuals are false, not all counterfactuals with false antecedents are true—for instance, if you believe the Warren Report, and there was no conspiracy to assassinate JFK, then Adams's counterfactual conditional is an example.

Counterfactual conditionals play an important role in causal reasoning: a prime example of the use of counterfactuals is to express causal relationships. E.g., striking a match causes it to light, and you can express this by saying "if this match were struck, it would light." Material, and generally indicative conditionals, cannot be used to express this: "the match is struck $\rightarrow$ the match lights" is true if the match is never struck, regardless of what would happen if it were. Even worse, "the match is struck $\rightarrow$ the match turns into a bouquet of flowers" is also true if it is never struck, but the match would certainly not turn into a bouquet of flowers if it were struck.

It is still debated What exactly the correct logic of counterfactuals is. An influential analysis of counterfactuals was given by Stalnaker and Lewis. According to them, a counterfactual "if it were the case that S then it would be the case that T" is true iff T is true in the counterfactual situation ("possible world") that is closest to the way the actual world is and where S is true. This is called an "ontic" analysis, since it makes reference to an ontology of possible worlds. Other analyses make use of conditional probabilities or theories of belief revision. There is a proliferation of different proposed logics of counterfactuals. There isn't even a single Lewis–Stalnaker logic of counterfactuals: even though Stalnaker and Lewis proposed accounts along similar lines with reference to closest possible worlds, the assumptions they made result in different valid inferences.

Chapter 2

Minimal Change Semantics

2.1 Introduction

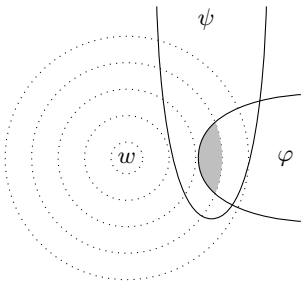
cnt:min:int:sec

Stalnaker and Lewis proposed accounts of counterfactual conditionals such as "If the match were struck, it would light." Their accounts were proposals for how to properly understand the truth conditions for such sentences. The idea behind both proposals is this: to evaluate whether a counterfactual conditional is true, we have to consider those possible worlds which are minimally different from the way the world actually is to make the antecedent true. If the consequent is true in these possible worlds, then the counterfactual is true. For instance, suppose I hold a match and a matchbook in my hand. In the actual world I only look at them and ponder what would happen if I were to strike the match. The minimal change from the actual world where I strike the match is that where I decide to act and strike the match. It is minimal in that nothing else changes: I don't also jump in the air, striking the match doesn't also light my hair on fire, I don't suddenly lose all strength in my fingers, I am not simultaneously doused with water in a SuperSoaker ambush, etc. In that alternative possibility, the match lights. Hence, it's true that if I were to strike the match, it would light.

This intuitive account can be paired with formal semantics for logics of counterfactuals. Lewis introduced the symbol "$\boxright$" for the counterfactual while Stalnaker used the symbol ">". We'll use $\boxright$, and add it as a binary connective to propositional logic. So, we have, in addition to formulas of the form $\varphi \to \psi$ also formulas of the form $\varphi \boxright \psi$. The formal semantics, like the relational semantics for modal logic, is based on models in which formulas are evaluated at worlds, and the satisfaction condition defining $\mathfrak{M}, w \Vdash \varphi \boxright \psi$ is given in terms of $\mathfrak{M}, w' \Vdash \varphi$ and $\mathfrak{M}, w' \Vdash \psi$ for some (other) worlds w'. Which w'? Intuitively, the one(s) closest to w for which it holds that $\mathfrak{M}, w' \Vdash \varphi$. This requires that a relation of "closeness" has to be included in the model as well.

Lewis introduced an instructive way of representing counterfactual situations graphically. Each possible world is at the center of a set of nested spheres containing other worlds—we draw these spheres as concentric circles. The

worlds between two spheres are equally close to the world at the center as each other, those contained in a nested sphere are closer, and those in a surrounding sphere further away.



The closest φ-worlds are those worlds w' where φ is satisfied which lie in the smallest sphere around the center world w (the gray area). Intuitively, $\varphi \boxdot\!\!\rightarrow \psi$ is satisfied at w if ψ is true at all closest φ-worlds.

2.2 Sphere Models

con:min:sph:sec

One way of providing a formal semantics for counterfactuals is to turn Lewis's informal account into a mathematical structure. The spheres around a world w then are sets of worlds. Since the spheres are nested, the sets of worlds around w have to be linearly ordered by the subset relation.

Definition 2.1. A *sphere model* is a triple $\mathfrak{M} = \langle W, O, V\rangle$ where W is a non-empty set of worlds, $V\colon \mathrm{At}_0 \to \wp(W)$ is a valuation, and $O\colon W \to \wp(\wp(W))$ assigns to each world w a *system of spheres* O_w. For each w, O_w is a set of sets of worlds, and must satisfy:

1. O_w is *centered* on w: $\{w\} \in O_w$.
2. O_w is *nested*: whenever S_1, $S_2 \in O_w$, $S_1 \subseteq S_2$ or $S_2 \subseteq S_1$, i.e., O_w is linearly ordered by $\subseteq$.
3. O_w is closed under non-empty unions.
4. O_w is closed under non-empty intersections.

The intuition behind O_w is that the worlds "around" w are stratified according to how far away they are from w. The innermost sphere is just w by itself, i.e., the set $\{w\}$: w is closer to w than the worlds in any other sphere. If $S \subsetneq S'$, then the worlds in $S' \setminus S$ are further way from w than the worlds in S: $S' \setminus S$ is the "layer" between the S and the worlds outside of S'. In particular, we have to think of the spheres as containing all the worlds within their outer surface; they are not just the individual layers.

The diagram in Figure 2.1 corresponds to the sphere model with $W = \{w, w_1, \ldots, w_7\}$, $V(p) = \{w_5, w_6, w_7\}$. The innermost sphere $S_1 = \{w\}$. The

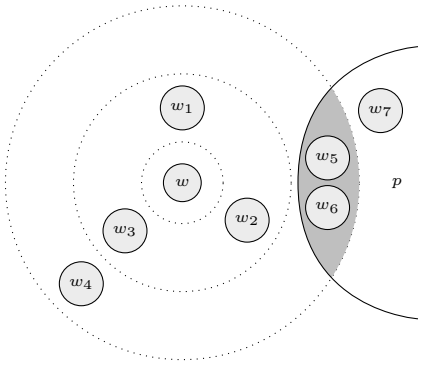


Figure 2.1: Diagram of a sphere model

con:min:sph: fig:sphere-model

closest worlds to w are w_1, w_2, w_3, so the next larger sphere is $S_2 = \{w, w_1, w_2, w_3\}$. The worlds further out are w_4, w_5, w_6, so the outermost sphere is $S_3 = \{w, w_1, \ldots, w_6\}$. The system of spheres around w is $O_w = \{S_1, S_2, S_3\}$. The world w_7 is not in any sphere around w. The closest worlds in which p is true are w_5 and w_6, and so the smallest p-admitting sphere is S_3.

To define satisfaction of a formula φ at world w in a sphere model $\mathfrak{M}$, $\mathfrak{M}, w \Vdash \varphi$, we expand the definition for modal formulas to include a clause for $\psi \boxdot\!\!\rightarrow \chi$:

Definition 2.2. $\mathfrak{M}, w \Vdash \psi \boxdot\!\!\rightarrow \chi$ iff either

1. For all $u \in \bigcup O_w$, $\mathfrak{M}, u \nVdash \psi$, or
2. For some $S \in O_w$,
   a) $\mathfrak{M}, u \Vdash \psi$ for some $u \in S$, and
   b) for all $v \in S$, either $\mathfrak{M}, v \nVdash \psi$ or $\mathfrak{M}, v \Vdash \chi$.

con:min:sph: sphere-vac
con:min:sph: sphere-nonvac

According to this definition, $\mathfrak{M}, w \Vdash \psi \boxdot\!\!\rightarrow \chi$ iff either the antecedent ψ is false everywhere in the spheres around w, or there is a sphere S where ψ is true, and the material conditional $\psi \rightarrow \chi$ is true at all worlds in that "ψ-admitting" sphere. Note that we didn't require in the definition that S is the *innermost* ψ-admitting sphere, contrary to what one might expect from the intuitive explanation. But if the condition in (2) is satisfied for some sphere S, then it is also satisfied for all spheres S contains, and hence in particular for the innermost sphere.

Note also that the definition of sphere models does not require that there *is* an innermost ψ-admitting sphere: we may have an infinite sequence $S_1 \supsetneq S_2 \supsetneq \cdots \supsetneq \{w\}$ of ψ-admitting spheres, and hence no innermost ψ-admitting spheres. In that case, $\mathfrak{M}, w \Vdash \psi \boxdot\!\!\rightarrow \chi$ iff $\psi \rightarrow \chi$ holds throughout the spheres S_i, S_{i+1}, ..., for some i.

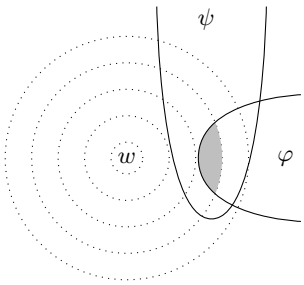


Figure 2.2: Non-vacuously true counterfactual

con:min:tf:
fig:true

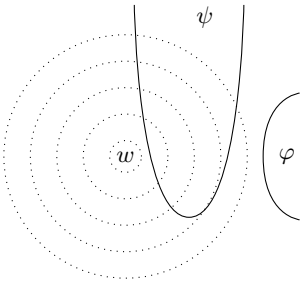


Figure 2.3: Vacuously true counterfactual

con:min:tf:
fig:vacuous

2.3 Truth and Falsity of Counterfactuals

con:min:tf:
sec

A counterfactual $\varphi \boxdot\!\!\rightarrow \psi$ is (non-vacuously) true if the closest φ-worlds are all ψ-worlds, as depicted in Figure 2.2. A counterfactual is also true at w if the system of spheres around w has no φ-admitting spheres at all. In that case it is *vacuously* true (see Figure 2.3).

It can be false in two ways. One way is if the closest φ-worlds are not all ψ-worlds, but some of them are. In this case, $\varphi \boxdot\!\!\rightarrow \neg\psi$ is also false (see Figure 2.4). If the closest φ-worlds do not overlap with the ψ-worlds at all, then $\varphi \boxdot\!\!\rightarrow \psi$. But, in this case all the closest φ-worlds are $\neg\psi$-worlds, and so $\varphi \boxdot\!\!\rightarrow \neg\psi$ is true (see Figure 2.5).

In contrast to the strict conditional, counterfactuals may be contingent. Consider the sphere model in Figure 2.6. The φ-worlds closest to u are all ψ-worlds, so $\mathfrak{M}, u \Vdash \varphi \boxdot\!\!\rightarrow \psi$. But there are φ-worlds closest to v which are not ψ-worlds, so $\mathfrak{M}, v \nVdash \varphi \boxdot\!\!\rightarrow \psi$.

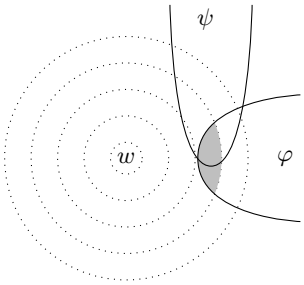


Figure 2.4: False counterfactual, false opposite

con:min:tf:
fig:false

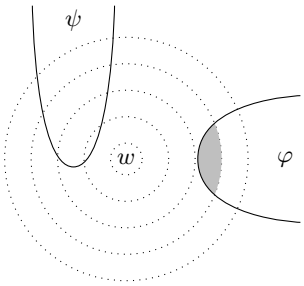


Figure 2.5: False counterfactual, true opposite

con:min:tf:
fig:false-opposite

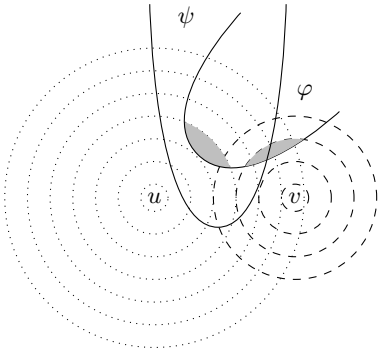


Figure 2.6: Contingent counterfactual

con:min:tf:
fig:contingent

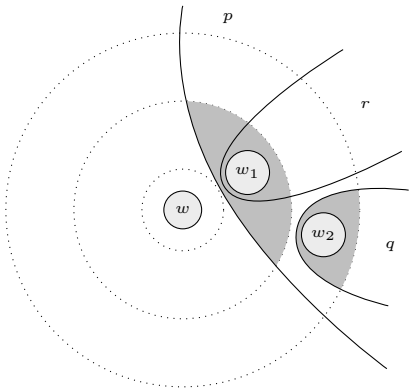


Figure 2.7: Counterexample to antecedent strengthening

cnt:min:agg:
fig:antecedent-strengthening

2.4 Antecedent Strengthenng

cnt:min:agg:
sec

"Strengthening the antecedent" refers to the inference $\varphi \to \chi \vDash (\varphi \wedge \psi) \to \chi$. It is valid for the material conditional, but invalid for counterfactuals. Suppose it is true that if I were to strike this match, it would light. (That means, there is nothing wrong with the match or the matchbook surface, I will not break the match, etc.) But it is not true that if I were to light this match in outer space, it would light. So the following inference is invalid:

> I the match were struck, it would light.
>
> Therefore, if the match were struck in outer space, it would light.

The Lewis–Stalnaker account of conditionals explains this: the closest world where I light the match and I do so in outer space is much further removed from the actual world than the closest world where I light the match is. So although it's true that the match lights in the latter, it is not in the former. And that is as it should be.

Example 2.3. The sphere semantics invalidates the inference, i.e., we have $p \boxrightarrow r \nvDash (p \wedge q) \boxrightarrow r$. Consider the model $\mathfrak{M} = \langle W, O, V \rangle$ where $W = \{w, w_1, w_2\}$, $O_w = \{\{w\}, \{w, w_1\}, \{w, w_1, w_2\}\}$, $V(p) = \{w_1, w_2\}$, $V(q) = \{w_2\}$, and $V(r) = \{w_1\}$. There is a p-admitting sphere $S = \{w, w_1\}$ and $p \to r$ is true at all worlds in it, so $\mathfrak{M}, w \Vdash p \boxrightarrow r$. There is also a $(p \wedge q)$-admitting sphere $S' = \{w, w_1, w_2\}$ but $\mathfrak{M}, w_2 \nVdash (p \wedge q) \to r$, so $\mathfrak{M}, w \nVdash (p \wedge q) \boxrightarrow r$ (see Figure 2.7).

2.5 Transitivity

cnt:min:tra:
sec

For the material conditional, the chain rule holds: $\varphi \to \psi, \psi \to \chi \vDash \varphi \to \chi$. In other words, the material conditional is transitive. Is the same true for counterfactuals? Consider the following example due to Stalnaker.

> If J. Edgar Hoover had been born a Russian, he would have been a Communist.
>
> If J. Edgar Hoover were a Communist, he would have been be a traitor.
>
> Therefore, If J. Edgar Hoover had been born a Russian, he would have been be a traitor.

If Hoover had been born (at the same time he actually did), not in the United States, but in Russia, he would have grown up in the Soviet Union and become a Communist (let's assume). So the first premise is true. Likewise, the second premise, considered in isolation is true. The conclusion, however, is false: in all likelihood, Hoover would have been a fervent Communist if he had been born in the USSR, and not been a traitor (to his country). The intuitive assignment of truth values is borne out by the Stalnaker–Lewis account. The closest possible world to ours with the only change being Hoover's place of birth is the one where Hoover grows up to be a good citizen of the USSR. This is the closest possible world where the antecedent of the first premise and of the conclusion is true, and in that world Hoover is a loyal member of the Communist party, and so not a traitor. To evaluate the second premise, we have to look at a different world, however: the closest world where Hoover is a Communist, which is one where he was born in the United States, turned, and thus became a traitor.[1]

Problem 2.1. Find a convincing, intuitive example for the failure of transitivity of counterfactuals.

cnt:min:tra: ex:trans-counterex

Example 2.4. The sphere semantics invalidates the inference, i.e., we have $p \boxright q, q \boxright r \not\vDash p \boxright r$. Consider the model $\mathfrak{M} = \langle W, O, V\rangle$ where $W = \{w, w_1, w_2\}$, $O_w = \{\{w\}, \{w, w_1\}, \{w, w_1, w_2\}\}$, $V(p) = \{w_2\}$, $V(q) = \{w_1, w_2\}$, and $V(r) = \{w_1\}$. There is a p-admitting sphere $S = \{w, w_1, w_2\}$ and $p \to q$ is true at all worlds in it, so $\mathfrak{M}, w \Vdash p \boxright q$. There is also a q-admitting sphere $S' = \{w, w_1\}$ and $\mathfrak{M} \nVdash q \to r$ is true at all worlds in it, so $\mathfrak{M}, w \Vdash q \boxright r$. However, the p-admitting sphere $\{w, w_1, w_2\}$ contains a world, namely w_2, where $\mathfrak{M}, w_2 \nVdash p \to r$.

Problem 2.2. Draw the sphere diagram corresponding to the counterexample in Example 2.4.

Problem 2.3. In Example 2.4, world w_2 is where Hoover is born in Russia, is a communist, and not a traitor, and w_1 is the world where Hoover is born

[1] Of course, to appreciate the force of the example we have to take on board some metaphysical and political assumptions, e.g., that it is possible that Hoover could have been born to Russian parents, or that Communists in the US of the 1950s were traitors to their country.

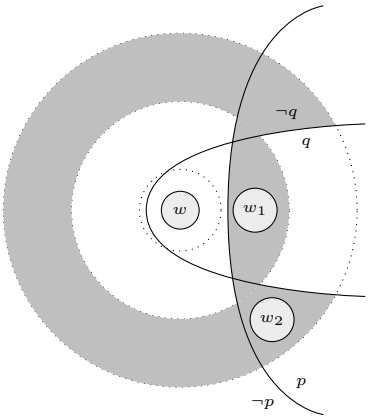


Figure 2.8: Counterexample to contraposition

cnt:min:cpo: fig:contraposition

in the US, is a communist, and a traitor. In this model, w_1 is closer to w than w_2 is. Is this necessary? Can you give a counterexample that does not assume that Hoover's being born in Russia is a more remote possibility than him being a Communist?

2.6 Contraposition

cnt:min:cpo: sec

Material and strict conditionals are equivalent to their contrapositives. Counterfactuals are not. Here is an example due to Kratzer:

> If Goethe hadn't died in 1832, he would (still) be dead now.
>
> If Goethe weren't dead now, he would have died in 1832.

The first sentence is true: humans don't live hundreds of years. The second is clearly false: if Goethe weren't dead now, he would be still alive, and so couldn't have died in 1832.

cnt:min:cpo: ex:contraposition-counterex

Example 2.5. The sphere semantics invalidates contraposition, i.e., we have $p \boxrightarrow q \not\models \neg q \boxrightarrow \neg p$. Think of p as "Goethe didn't die in 1832" and q as "Goethe is dead now." We can capture this in a model $\mathfrak{M}_1 = \langle W, O, V\rangle$ with $W = \{w, w_1, w_2\}$, $O = \{\{w\}, \{w, w_1\}, \{w, w_1, w_2\}\}$, $V(p) = \{w_1, w_2\}$ and $V(q) = \{w, w_1\}$. So w is the actual world where Goethe died in 1832 and is still dead; w_1 is the (close) world where Goethe died in, say, 1833, and is still dead; and w_2 is a (remote) world where Goethe is still alive. There is a p-admitting sphere $S = \{w, w_1\}$ and $p \to q$ is true at all worlds in it, so $\mathfrak{M}, w \Vdash p \boxrightarrow q$. However, the $\neg q$-admitting sphere $\{w, w_1, w_2\}$ contains a world, namely w_2, where q is false and p is true, so $\mathfrak{M}, w_2 \nVdash \neg q \to \neg p$.

Photo Credits

Bibliography