Logic Primer*

Robbie Matyasi

PHL 232: Knowledge and Reality

This primer covers some basic concepts of logic that are necessary to engage with the topics of this class and that will be useful in your future as a student of philosophy and a critical thinker.

In sections 1-3 we will be concerned with *classical propositional logic* (CPL hereafter), namely the study of *reasoning* with *propositions* and *truth-functions*. In sections 4 & 5 we will explore topics outside of CPL and briefly discuss quantifiers and counterfactual conditionals.

1 Propositions

It will be useful for us to think about basic propositions as what simple declarative sentences express. Some examples:

(1) Leo is a lion.
(2) Cycling is fun.
(3) The moon is made of cheese.

For the purposes of this class we will define propositions as the primary bearers of truth-values: `true` and `false`. You may think that other things can be true or false, for example, sentences or beliefs can be true or false too. We will treat these as being true or false in a derived sense: sentences, beliefs, etc. are true if they express a proposition with the truth-value `true`, and false if they express proposition with the truth-value `false`.

*Special thanks to John Bunke, Adam Murray, Dominic Alford-Duguid, and Imogen Dickie for letting me implement and adapt their materials.

How do propositions relate to truth-values? A proposition says something is in a certain way, and they are true just in case that thing is in that way.

Now consider (2): it is true according to many people, but not everyone agrees. However, in CPL propositions are always either true or false, and never both or neither. CPL is a limited tool that only has the resources to deal with `true` and `false`. This doesn't mean that there is no uncertainty in everyday thinking and language, we just have to ignore this when we use CPL.

In CPL our primary focus is on the propositions themselves and not on sentences that express them. But why do we have to do that? We will not look at the detailed reasons to do so, but here is one basic consideration: sentences can be ambiguous, but propositions cannot. Consider (1): It is true just in case Leo is a lion. But we can mean other things by using the word 'Leo', for example the constellation Leo. In this case (1) is false because no constellation belongs to the biological species *Panthera leo*. Luckily, we can always disambiguate:

(1a) Leo, the biological organism, is a lion. (`true`)

(1b) Leo, the constellation, is a lion. (`false`)

So while we can use (1) to express different propositions, this is a not an issue if we focus on the intended propositions only.

Perhaps more importantly, in CPL we also distinguish between what a sentence *means* and how we *use* them. For example, consider (2) again: there are a number of situations in which we use this sentence for something else than just reporting the fact that cycling is fun. Someone may hope to convince you to invest in a bicycle by uttering (2). Or maybe someone utters (2), but immediately after they say "… but I prefer to take the subway". In the highly abstract setting of CPL, what a sentence *means* is a proposition, and it is independent of our intentions. In contrast, meaning and use in our everyday thinking are not that far apart—think about someone making their intentions clear by saying "I didn't mean that".

Our examples so far were basic propositions, but CPL also deals with propositions that are certain combinations of basic propositions. Here are two possible combinations of (1) and (3):

(4) Leo is a lion and the moon is made of cheese.
(5) If the moon is made of cheese, then Leo is a lion.

To see how CPL deals with these more complex propositions, we have to take a look at truth-functions.

2 Truth-functions

You may recall from a math class that there are mathematical objects called relations and that some relations take inputs and produce a certain output. *Functions* are those relations which, given an input, produce a single, determinate output: if you give a function some input, then it always gives you a single, particular output. A relation which, when you give it the same input a bunch of times, sometimes gives you output ϕ and sometimes output ψ is not a function.

For example, the successor function always gives you the next natural number:

n	$n+1$
1	2
85	86
124	125

But the "is a sibling of" relation may give you different outputs for a particular input:

x	sibling of x
Ashley	Mary-Kate
Ashley	Elizabeth
Joel	Ethan
Ethan	Joel

A *truth-function* is a function that takes truth-values as inputs and produces a single truth-value as its output. Since truth-functions are a kind of function that have a finite number of possible inputs, we can write them out as a table showing the values produced by the different possible inputs; this is called a *truth-table*. Here is the truth-table for a random truth-function:

input 1	input 2	output
`true`	`true`	`false`
`true`	`false`	`true`
`false`	`true`	`true`
`false`	`false`	`true`

Some propositions are combinations of other propositions and their truth-values are determined by the truth-functions they contain. A central thesis of CPL, which we will call Frege's thesis, is as follows.

Frege's thesis: The truth-value of a proposition with truth-functional structure is purely a function of the truth-values of its parts.

The relevant 'parts' are the smaller propositions that are joined together by the truth-functional expressions.

Below you will find the most common truth-functions.

2.1 Negation

For negation we will use the standard symbol '¬'. Here is the truth-table for negation:

p	$\neg p$
T	F
F	T

Negation is a monadic truth-function, which means that it takes a single truth-value as input.

The natural language counterparts of negation are the expressions 'not' and 'it is not the case that'. As we would expect, if you have a proposition that is `true`, and then you apply negation to it, you get `false`. If you have a proposition that is `false`, and then you apply negation, you get `true`. This accords with how the word 'not' operates in English:

(6) The moon is made of cheese. (`false`)
(7) The moon is not made of cheese. (`true`)

2.2 Conjunction

For conjunction we will use the standard symbol '∧':

p	q	$p \wedge q$
T	T	T
T	F	F
F	T	F
F	F	F

Conjunction—as well as all the remaining truth-functions we discuss—is a dyadic truth-function, which means that it takes two truth-values as its input.

The most common natural language counterpart of conjunction is the expression 'and':

(8) Cycling is fun and Leo is a lion.

If you have a pair of propositions that are both `true`, and then you apply conjunction to them, you get `true`. In the above example, assuming that cycling is indeed fun and Leo is truly a lion, (8) is guaranteed to be `true`. But if any one of the propositions in the input is `false`, you get the truth-value `false`. Here's an example:

(9) The moon is made of cheese and Leo is a lion. (`false`)

2.3 Disjunction

For disjunction we will use the standard symbol '$\vee$':

p	q	$p \vee q$
T	T	T
T	F	T
F	T	T
F	F	F

The natural language counterpart of disjunction is the expression 'or', but only understood in a certain sense. There are basically two ways that we use the expression 'or' in English (when it is used to connect two propositions): one means *A* or *B* or both and the other means *A* or *B* but *not* both. The former is called 'inclusive or' and the latter 'exclusive or'. For example, if I am at a restaurant and the waiter says that with my hamburger I can have "fries or salad" then I may want to know whether I can have both; if the answer is "no," then 'or' in "fries or salad" was an instance of

exclusive 'or'. Sometimes we emphasize that we mean exclusive 'or' by adding 'either' and emphasizing it with intonation as in: "You may have either fries or salad."

In logic and especially in CPL, we typically stipulate that 'or' is always inclusive, that is, that if both disjuncts (i.e. the parts before and after the word 'or') are `true`, then the whole proposition is `true`. We do this primarily for technical reasons: for example, using inclusive 'or' entails that the following two equivalences are valid. These are called De Morgan's Laws ('iff' means 'if and only if' and we will discuss it in section 2.5):

De Morgan's Laws:

(i) $\neg(p \wedge q)$ iff $\neg p \vee \neg q$
(ii) $\neg(p \vee q)$ iff $\neg p \wedge \neg q$

Exercise: Check the first pair with a truth-table. Here is a correct truth-table for the second pair:

p	q	$p \vee q$	$\neg(p \vee q)$	$\neg p$	$\neg q$	$\neg p \wedge \neg q$
T	T	T	**F**	F	F	**F**
T	F	T	**F**	F	T	**F**
F	T	T	**F**	T	F	**F**
F	F	F	**T**	T	T	**T**

2.4 Material conditional

The material conditional is perhaps the most important and also the trickiest truth-function. For this reason, we need to pay extra attention to the material conditional and its natural language counterparts.

For the material conditional we will use the standard symbol '$\rightarrow$':

p	q	$p \rightarrow q$
T	T	T
T	F	F
F	T	T
F	F	T

In an expression such as '$p \to q$', we call the first element—here 'p'—the 'antecedent' and the second element—here 'q'—the 'consequent'.

Conditional statements in English feature the word 'if'; sometimes 'if' is accompanied by 'then'. For example:

(10) If the supermarket doesn't have lemons, then I will buy limes instead.
(11) Lili will be sad if the show is sold out.
(12) I will go to the dinner only if I can get those noodles again.

Keep in mind that not all of these are translated to CPL in the same way. (10) is of the form '$p \to q$', where p is "the supermarket doesn't have lemons" and q is "I will buy limes instead"; but in (11) we don't have the same order: rather it says "If the show is sold out, then Lili will be sad". So everything of the form 'p if q' will be translated as '$q \to p$'. However, sentences of the form 'p only if q', such as (12), should be translated like (10) and not as (11).

- if p, then q = $p \to q$
- p if q = $q \to p$
- p only if q = $p \to q$

Exercise: If the translation of 'p only if q' seems weird, try to think it through with the following sentence "The match is burning only if there is oxygen in the room".

The natural language use of conditionals is far more complex than the material conditional, and there are many attempts to give them a good theory in linguistics and the philosophy of language, but this would take us very far afield. In any case, the material conditional does an *okay* job in capturing conditionals understood in a very limited sense. For this reason it is useful to go through the truth-table line by line.

The first line of the truth-table seems to align well with common sense. Consider:

(13) If the solar system has 8 planets, then at least 8 planets exist.

It is true that the solar system has 8 planets; it is true that at least 8 planets exist and the whole proposition is `true`. But this example is slightly misleading: the antecedent and the consequent in (13) are closely connected, since the solar system cannot have 8 planets while e.g. only 5 planets exist. In our ordinary understanding of conditionals we often take them to indicate some sort of connection between the antecedent and the consequent. But since the material conditional is just a truth-function, it is `true` if

both its antecedent and its consequent are `true`, even even if they are totally unrelated. For example:

(14) If the solar system has 8 planets, then dodos are extinct.

This is true when we treat the natural conditional as expressing the material conditional. The main point to remember is that it is sufficient that both the antecedent and the consequent are `true` for a material conditional to be `true`, whether or not they are related to one another.

The second line of the truth-table also makes good sense. Consider:

(15) If dodos are extinct, then you can spot dodos in High Park.

Given that dodos are extinct, of course you cannot spot dodos in High Park, so (15) is clearly false. Again, be aware that the material conditional is a truth-function so the antecedent and the consequent need not be related. So the following is `false` for the same reason as (15):

(16) If dodos are extinct, then the moon is made of cheese.

The two remaining lines of the truth-table are quite odd. We have to address this without getting too far into the details. The main reason why we find the last two lines strange is because typically we don't understand conditionals with false antecedents as clearly true or false; at best, they strike us as uncertain. Consider these:

(17) If the moon is made of cheese, then dodos are extinct.
(18) If the moon is made of cheese, then pigs fly.

One consideration to have is whether we would assign a different truth-value to propositions such as (17) and (18). Yet, we judge them to be odd for the same reason: if the antecedent is false we have no clear idea how to evaluate a conditional. Now if we decide both to be `false`, then '$\rightarrow$' has the same truth-table as '$\wedge$' and we are much better off with a separate truth-function for conjunction and the conditional. So we use the material conditional that assigns `true` to conditionals with a `false` antecedent, no matter the truth-value of the consequent.

2.5 Biconditional

The biconditional ('$\leftrightarrow$') is much simpler:

p	q	$p \leftrightarrow q$
T	T	T
T	F	F
F	T	F
F	F	T

Notice that $p \leftrightarrow q$ is equivalent to '$p \rightarrow q \wedge q \rightarrow p$'. (Try to show this to yourself with a truth-table.)

The natural language counterpart of the biconditional is "if and only if". (Remember what we said about 'only if'.)

3 Arguments

Much of what we will do in this course involves considering and evaluating arguments for various positions in epistemology and metaphysics. Here we will get through the basics of what makes an argument a good piece of reasoning as opposed to a bad piece of reasoning. Consider these examples:

- *Example 1:* If Amber, a student in PHL 232, completes the course, she either passes or fails. So: Amber completes and passes PHL 232 or Amber completes and fails PHL 232 (notice that this is an exclusive or). Now, let's suppose we get to the end of the semester and Amber tells you that she completed the course and didn't fail. On the basis of these two propositions—that either Amber completes the course and passes or she completes and fails and that Amber completed but didn't fail—you reason that Amber passed the course.
- *Example 2:* You know that if Delilah is at the park, then her dog Brutus is there too; perhaps the only reason Delilah ever goes to the park is to let Brutus run. So, we have the proposition that if Delilah is at the park, then Brutus is at the park. Now, suppose you find out that Brutus is at the park; maybe a friend saw Brutus running about, but the friend didn't see whether Delilah was there too. On the basis of these two propositions—that if Delilah is at the park then Brutus is at the park, and that Brutus is at the park—you reason that Delilah is at the park too.

If you find the first example *good* and the second example *problematic*, you probably already have a good intuitive understanding of what makes a good piece of reasoning. Our goal from now is to get a systematic account of this distinction in CPL.

An argument is any collection of propositions in which one is supported by the others. The proposition supported by the others is called *conclusion*, while the others are called *premises*. You will often find arguments written in a standardized form like this:

1. If Delilah is at the park, then Brutus is at the park.
2. Brutus is at the park.

c. Therefore, Delilah is at the park.

(Don't forget that *Example 2* was bad reasoning, but this doesn't disqualify it from being an argument.)

We distinguish good arguments from bad arguments in two steps, both having to do with the truth-values of their premises and conclusions.

3.1 Validity

The first thing to check is whether the argument is *valid*. Validity is a property of arguments that requires a specific relationship between the propositions included in an argument. To understand validity we need to define this relation first.

Entailment: A proposition p entails proposition q (which we will write as '$p \vDash q$') just in case it is impossible that p is true while q is false.

Consider the following pair:

(19) Moon is made of cheese.
(20) Pigs fly.

We already have an intuitive idea that (19) and (20) have nothing to do with each other. Entailment is a concrete criterion aiming to capture this idea: (19) doesn't entail (20) because it is possible that the moon is made of cheese and pigs don't fly; and (20) doesn't entail (19) because it is possible that pigs fly and the moon is not made of cheese.

Now consider:

(21) The solar system has 8 planets.
(22) At least 8 planets exist.

These two propositions are clearly connected: (21) entails (22), since it is impossible that the solar system has 8 planets while less than 8 planets exist. However, (22)

doesn't entail (21): it is possible that at least 8 planets exist but there are less than 8 planets in the solar system.

Finally, consider:

(23) Leo is a lion.
(24) Leo is a member of the species *Panthera leo*.

These two propositions entail each other: being a lion is simply the same thing as being a member of the species *Panthera leo*, so it is impossible to be one while not being the other. This means that (23) and (24) entail each other.

Entailment and the material conditional are easy to confuse. The first thing to remember is that the material conditional is a truth-function while entailment is a more complex relationship between propositions. Consider these two definitions:

> $p \rightarrow q$ is true just in case if it is *not the case* that p is true while q is false.

> $p \vDash q$ is true just in case if it is *impossible* that p is true while q is false.

With this in hand, we can define validity as follows:

Validity: An argument is valid just in case the premises *entail* the conclusion.

So an argument is valid if it is impossible that all the premises of the argument are true together, while the conclusion is false.

With validity we can also systematically explain why we judge *Example 1* as good reasoning; and why we are not convinced by *Example 2*. Consider the latter again:

1. If Delilah is at the park, then Brutus is at the park.
2. Brutus is at the park.

c. Therefore, Delilah is at the park.

The problem with this reasoning is that it is possible that premises 1 and 2 are true together, while the conclusion is false. Too see this, consider all possible scenarios:

1. Delilah is at the park; Brutus is at the park.
2. Delilah is at the park; Brutus is not at the park.
3. Delilah is not at the park; Brutus is at the park.
4. Delilah is not at the park; Brutus is not at the park.

We have to rule out scenario 2 because to test the validity of the argument we have to assume that the material conditional in premise 1 is true. (If you have trouble seeing this, it is time to go back to the explanation of the truth-table for the material conditional.) Similarly, we have to rule out scenario 4 because we also assume that

premise 2 is true. So we are left with scenarios 1 and 3 that the premises do not rule out. But since the truth of the premises together does not rule out scenario 3, we can conclude that it is possible that the premises are true together while the conclusion is false.

1. Delilah is at the park; Brutus is at the park.
2. ~~Delilah is at the park; Brutus is not at the park.~~
3. Delilah is not at the park; Brutus is at the park.
4. ~~Delilah is not at the park; Brutus is not at the park.~~

We can also show the same thing with a simple truth-table:

Table 10: p = 'Delilah is at the park'; q = 'Brutus is at the park'

$p \rightarrow q$	q	p
T	**T**	**T**
~~F~~	~~F~~	~~T~~
T	**T**	**F**
~~T~~	~~F~~	~~F~~

Similarly, we can show that *Example 1* is valid:

1. Either Amber completes and passes PHL 232 or Amber completes and fails PHL 232.
2. Amber completed the class and didn't fail.

c. Therefore, Amber passed PHL 232.

The possible scenarios are the following:

1. Amber completes PHL 232 and passes.
2. ~~Amber completes PHL 232 and fails.~~
3. ~~Amber doesn't complete PHL 232 and passes.~~
4. ~~Amber doesn't complete PHL 232 and fails.~~

If premise 1 is true, Amber either completes PHL 232 and passes, or she completes PHL 232 and fails. This rules out scenarios 3 and 4. If premise 2 is true, scenario 2 is ruled out too. So the only possible scenario is 1, which means that the conclusion of this argument is guaranteed by the truth of the premises.

Exercise: Check this argument with a truth-table in which p = 'Amber completes PHL 232 and passes'; q = 'Amber completes PHL 232 and fails'.

3.2 Soundness

Our foremost aim in reasoning is to reach truths, but valid arguments do not always guarantee true conclusions: if an argument is valid then the conclusion is guaranteed to be true *if the premises are true*, but this does not rule out that the premises are in fact false. Consider the following valid argument:

1. If the moon is made of cheese, then pigs fly.
2. The moon is made of cheese.

c. Therefore, pigs fly.

The argument is perfectly valid, but premise 2 is clearly false so the argument does not guarantee the truth of the conclusion.

Exercise: Check this with a truth-table. Make sure you focus on the lines in which premise 2 is false and premise 1 is true.

Thus the second thing we have to look for is whether the argument's premises are in fact true. This is captured by the property of *soundness*:

Soundness: An argument is sound just in case (i) it is valid; and (ii) all of its premises are true.

Once an argument is valid, which means that it is impossible for its premises to be true while its conclusion is false, and we can also show that all the premises in the argument are in fact true, then we can be certain that the argument's conclusion is true too. We reached our goal!

3.3 Basic rules of inference

An inference rule characterizes a pattern or form of argument—usually one or two premises and one conclusion—that is valid. The idea is that the premises entail the conclusion so the move from the premises to the conclusion is valid.

There are many other rules of inference, but the following four are all you need for this class.

Modus ponens	*Modus tollens*
1. $p \to q$	1. $p \to q$
2. p	2. $\neg q$
c. q	c. $\neg p$

You can see modus tollens as an instance of modus ponens if you notice that the contrapositive of a material conditional is equivalent to the original conditional: that is, '$p \to q$' is equivalent to '$\neg q \to \neg p$'.

Conjunction elimination	
1. $p \wedge q$	1. $p \wedge q$
c. q	c. p

Disjunction introduction
1. p
c. $p \vee q$

Note that in an instance of disjunction introduction, q can be anything; you can add a disjunct to a proposition any time you like while still preserving validity.

It is important that you not get the behaviours of conjunction and disjunction mixed up: you can't just add conjuncts whenever you'd like, and you can't just eliminate disjuncts whenever you want.

Exercise: Demonstrate the validity of these rules of inference with truth tables.

Exercise: Demonstrate with truth-tables why the following bad inference rules are invalid.

BAD modus ponens	*BAD disjunction elimination*	*BAD conjunction introduction*
1. $p \to q$	1. $p \vee q$	1. p
2. q	c. q	c. $p \wedge q$
c. p		

4 Quantification

There is another layer of logical analysis outside of CPL that we can apply to words such as 'there is', 'there are', 'all', 'some', 'none', and related expressions. These are called *quantificational expressions* and the branch of logic that deals with them is called *predicate logic*. Predicate logic is more complex than CPL, and it would take is much too far afield to say even some very basic things about it. However, it is useful for you to know how to read some of the standard notation for predicate logic.

We define two more symbols:

- We will use the symbol '$\exists$' to mean 'There is some … that is …'.

  For example, '$\exists xFx$' means: 'There is *some* x that is F.'

  The symbol '$\exists$' is called the *existential quantifier*.

  F here is technically variable for a predicate, but we will often refer to it as a property.

  '$\exists xFx$' doesn't say that there's only one x that is F ; it says there is *at least one*, but there might be more than one.

- We will use '$\forall$' to mean 'For all…, …'.

  For example, '$\forall xFx$' means: 'For *all* x, Fx'. The symbol '$\forall$' is called the *universal quantifier*.

Exercise: Convince yourself using the definitions of '$\exists$' and '$\forall$' that the following are true:

- $\exists xFx$ if and only if $\neg\forall x\neg Fx$
- $\forall xFx$ if and only if $\neg\exists x\neg Fx$

One place you will encounter this notation later on this term is in Ney's statement of Leibniz's Law, which she writes:

$$\forall x \forall y \forall F(x = y \rightarrow [Fx \leftrightarrow Fy])$$

We read this as: 'For all x, for all y, for all F , if x is identical to y, then Fx if and only if Fy'. This means that, for all objects x and y, if x and y are identical, then x and y have all and only the same properties.

5 Counterfactual conditionals

So far we discussed the material conditional and its relation to some 'if … then …' statements in English. However, we neglected a kind of conditional that we use quite often and we can't plausibly write out as a truth-function in CPL: the *counterfactual conditional*. The counterfactual conditional plays a central role in Robert Nozick's truth-tracking theory of knowledge as well as David Lewis's counterfactual-dependence account of causation. To give you the tools necessary to understand these accounts, in this section we will discuss the most famous account of the logic of counterfactual conditionals.

Counterfactuals are conditional statements about what would be the case if something else were the case. Consider the following:

(25) If the moon had been made of cheese, it would still stay in orbit.

This is a counterfactual because it is a claim about what would have been the case if the moon had been made of cheese.

We represent counterfactual conditionals as follows:

$$p \mathrel{\square\!\!\rightarrow} q$$

(In English: if it had been the case that p, it would have been the case that q)

5.1 Counterfactual vs. material conditionals

To see why we have to interpret '$p \mathrel{\square\!\!\rightarrow} q$' differently than '$p \rightarrow q$', consider the following pair:

(26) If Oswald didn't kill Kennedy, then someone else did.
(27) If Oswald hadn't killed Kennedy, then someone would have.

The first statement is clearly true, since Kennedy was in fact assassinated. If it wasn't Oswald who did it, then it was *someone else*. But the second statement is intuitively false: assuming Oswald acted alone, then if Oswald had stayed put in the Soviet Union, presumably no one else would have killed Kennedy.

Moreover, the material conditional has different logical properties than counterfactuals. Let's look at one of these properties. According to our theory of the material conditional, the following is valid:

1. If it is raining, then the ground is wet.

c. Therefore, if the ground is not wet, then it is not raining.

This is called *contraposition* and it works because both of these propositions are false just in case the ground is not wet while it is raining (perhaps in a scenario in which someone is covering the ground with a giant tent).

Exercise: Show contraposition for any possible antecedent and consequent by showing that '$(p \to q) \vDash (\neg q \to \neg p)$' is true.

Contraposing counterfactuals can lead us to bad results very easily. Consider:

1. If Lee Harvey Oswald hadn't killed John F. Kennedy, then Lyndon Johnson wouldn't have been President in 1963.

c. Therefore, if Lyndon Johnson had been President in 1963, then Lee Harvey Oswald would have killed John F. Kennedy.

The first seems true, while the second is seems false. The most likely scenario in which Johnson was President in 1963 involves him defeating Kennedy for the Democratic nomination in 1960. In that scenario, Kennedy might have been Vice President (if chosen by Johnson for the ticket) or he would still have been a Senator from Massachusetts. In either case, it is hard to see why Oswald would be attempting to assassinate Kennedy rather than the sitting President, which is Johnson in this scenario. So, contraposition is invalid for counterfactual conditionals.

The second property of material conditionals we should look at is *transitivity*. Consider:

1. If Delilah is at the park, then Brutus is at the park.

2. If Brutus is at the park, then Brutus has a wonderful time.

c. Therefore, if Delilah is at the park, then Brutus has a wonderful time.

It is easy to see that this is a valid argument if we understand the propositions with the material conditional.

Exercise: Show '$(p \to q) \wedge (q \to r) \vDash p \to r$'. Make sure you consider all the 8 possible combinations of truth-values for p, q, and r.

p	q	r	$p \to q$	…
T	T	T	T	
T	T	F	T	
T	F	T	F	
T	F	F	F	
F	T	T	T	
F	T	F	T	
F	F	T	T	
F	F	F	T	

In contrast, consider these:

1. If Hoover had been born a Russian, then he would have been a communist.
2. If Hoover had been a communist, then he would have been a traitor.
c. Therefore, if Hoover had been born a Russian, then he would have been a traitor.

The first and the second both seems true. But the argument is invalid: even assuming the premises to be true, the Russian-born Hoover still could have been a patriotic communist.

Lastly, *antecedent strengthening*:

1. $p \to q$
c. Therefore, $(p \wedge r) \to q$

Antecedent strengthening is tricky because it is valid for the material conditional only, but it most likely fails for counterfactuals *as well as* ordinary English 'if … then …' statements. (Remember that the material conditional is already problematic as an interpretation of English.)

Exercise: Show '$(p \to q) \vDash (p \wedge r) \to q$'.

Here we will only look at a bad case of counterfactual reasoning:

1. If Kangaroos had no tails, they would topple over.

c. Therefore, if Kangaroos had no tails and they wore tiny jetpacks, they would topple over.

These examples show that the material conditional does not provide us a good theory of counterfactuals; they in fact motivated a quite different, non truth-functional account for them.

5.2 The Stalnaker-Lewis account

The most famous account of counterfactuals is commonly attributed to David Lewis and Robert Stalnaker.[1] To start, we will need the notion of a *possible world*. This is a tool that philosophers use to formalize talk about what *could have been* the case. For example, I am 180 centimetres tall, but I might have been shorter than that. For example, by using the notion of a possible world we can state "Robbie could have been 170 centimetres tall" like this: "There are some possible worlds in which it is true that Robbie is 170 centimetres tall".

A possible world is a completely specific way the universe might have been. There is no limit on ways the universe might have been, *except* that there are no possible worlds in which a contradiction is true—here we won't get into the reasons why. Otherwise, anything goes: there are possible worlds in which I am 2 kilometres tall, in which the laws of physics are so different that no planets ever formed, and in which everything is exactly as it is in this world except that the movie *The Matrix* is a documentary.

It is crucial for the Stalnaker-Lewis account that some possible worlds are more similar to one another than others are. For example, the world in which I am 170 centimetres tall but everything else is the same is more similar to the actual world than the one in which *The Matrix* is a documentary. The idea is that from the perspective of a particular world, we can rank other worlds in terms of their similarity to that world.

[1] As is often the case, there were many others working on similar accounts at the same time that are often uncredited: e.g. Donald Nute, Timothy Sprigge, and William Todd.

In such a cases, we often use spatial metaphors and say that a world is *closer* to one world than another. For example, from the perspective of the world in which *The Matrix* is a documentary, any world in which we all live in a computer simulation is closer—i.e. more similar—to that world than the ones in which we roam freely.

Also, it's important to note that there may be worlds that are *equally close* to a particular world. This idea gave rise to much controversy in the literature on counterfactuals that we won't do full justice here. But we need to briefly discuss this because both Nozick and Lewis relies on the assumption that there are in fact equally similar worlds—let's call them "ties". Consider the following:

(28) If Drake and Rihanna were compatriots, they would be Canadian.
(29) If Drake and Rihanna were compatriots, they would be Barbadian.

Neither of these are uncontroversially true: it's unclear whether we are even able to decide between (28) and (29). Perhaps even more confusingly, maybe both are true (i.e. they may have overlapping dual citizenship status). For all we know, it seems better to settle with something like this:

(30) If Drake and Rihanna were compatriots, they would be Barbadian or Canadian.

Intuitively, by allowing ties, we can capture the truth of (30) and the unclear status of (28) and (29). This means that we consider the worlds in which they are both Canadian and the worlds in which they are both Barbadian as equally similar to the actual world.[2]

With this in hand, we can state the truth-conditions for '$p \boxrightarrow q$' as follows:

Stalnaker-Lewis account of counterfactuals: '$p \boxrightarrow q$' is true in a world w just in case in all the closest worlds to w in which p is true, q is true.

In most cases, we will be only interested in counterfactuals evaluated in the actual world (for which we will use the symbol '@'):

Stalnaker-Lewis account of counterfactuals (actuality): '$p \boxrightarrow q$' is true in @ just in case in all the closest worlds to @ in which p is true, q is true.

Notice that we're requiring that all the closest p-worlds are q-worlds: this is so because sometimes we have to consider ties—i.e. worlds that are equally similar. If among the closest p-worlds there are some q-worlds and some $\neg q$-worlds, they are tied in

[2] You may think that the problem is that 'compatriots' is not specific enough for us to evaluate. Good thinking! Stalnaker thinks the same and ends up with a slightly different semantics than what we will use. However, for the purpose of this course we will have to stick to Lewis's solution that allows for ties.

the same way as the two options in the compatriots example above, and so we will assume that in these cases '$p \boxrightarrow q$' is false.

Let's see how our account works with our example from before:

(31) If Kangaroos had no tails, they would topple over.

Our account says that this statement is true (in the actual world) just in case in all of the closest possible worlds in which kangaroos don't have tails, they also aren't able to stand upright. That is, these worlds are more similar to the actual world than any possible world in which kangaroos don't have tails and they are able to stand upright.

To put it differently (t = 'Kangaroos have tails'; u = 'Kangaroos able to stand upright'):

'$\neg t \boxrightarrow \neg u$' is true in @ if and only if in all the closest worlds in which $\neg t$ is true, $\neg u$ is true. It follows that there is no world in which $\neg t \wedge u$ is true that is closer to @ than any of the closest worlds in which $\neg t \wedge \neg u$.

And this should seem right: suppose we kept everything the same as the actual world, including how kangaroos' bodies are set up, their muscles, etc., but changed just one thing: kangaroos now don't have tails. (In the actual world, kangaroos do use their tails to keep themselves upright.) In worlds like this, kangaroos flop over, at least for a little while until they got used to the new situation. These are worlds in which $\neg t \wedge \neg u$.

Now consider a possible world in which kangaroos don't have tails, but they are still able to stand upright because all kangaroos, it turns out, have tiny jetpacks strapped to them that tip them back up if they are about to fall over. That's a world in which $\neg t \wedge u$. But clearly the kangaroos-all-have-jetpacks world is less similar to the actual world than the worlds in which they don't have tails and they fall over (because we have kept everything else the same). A world in which kangaroos don't have tails and can't stand upright is going to be more similar to the actual world than any world in which kangaroos don't have tails but for some weird reason (e.g. jetpacks) they can still stand upright. This means that "If kangaroos didn't have tails, then they wouldn't be able to stand upright" is true; and that's the right result.

Exercise: Pick a counterfactual that you think is intuitively correct or intuitively incorrect, and then see if our account yields the right result.

One very important thing to notice is that Lewis and Nozick in your readings do not use the exact same method to evaluate counterfactual conditionals. They have

good reasons to do so that we won't address here. However, I ask you to keep in mind that our own account is intentionally simplified to make it easier for you to get used to the similarity-based accounts of counterfactuals. Our goal is not to have a fully developed and stable theory—perhaps that doesn't even exist yet. We will discuss the most important differences from the readings when they come up in class or in your assignments.

In the remainder we will explain with our account why contraposition and transitivity failed for counterfactuals in our examples.

Contraposition:

1. If Lee Harvey Oswald hadn't killed John F. Kennedy, then Lyndon Johnson wouldn't have been President in 1963.

c. Therefore, if Lyndon Johnson had been President in 1963, then Lee Harvey Oswald would have killed John F. Kennedy.

The premise is true in the actual world just in case in all the closest worlds in which Oswald did not kill Kennedy, Johnson is not a president in 1963. So, any of these worlds are closer to actuality than any world in which Oswald did not kill Kennedy and Johnson is president in 1963. This seems right: in the closest possible worlds to actuality in which Kennedy is not assassinated, he finishes his term. Yet the conclusion is false in the actual world because in the closest worlds in which Johnson is president in 1963 Oswald has no reason to kill Kennedy.

Transitivity:

1. If Hoover had been born a Russian, then he would have been a communist.

2. If Hoover had been a communist, then he would have been a traitor.

c. Therefore, Hoover had been born a Russian, then he would have been a traitor.

Evaluating from the actual world, the conclusion is false while the premises seem right. Plausibly a world in which Hoover is both Russian and a traitor is less close to the actual world than any world in which (while all else is fixed) Hoover is equally patriotic as his American counterpart but he is a Russian communist.

Exercise: Try to evaluate our example with antecedent strengthening.

Exercise: Try to show that our semantics gives the desired results for similarity ties. Remember to use the truth-table for disjunction. (You can use (28), (29), (30).)