PHIL 210/510: Formal Methods — Yale, Spring 2025

Lecture 1: Sets, Relations, and Functions — Dr. Calum McNamara

1. Sets

Set theory is the foundation of mathematics—and also a useful field of study for philosophers. It's basic concept is that of a **set**. Roughly speaking, a set is a collection of objects.[1] For our purposes, it doesn't matter what these objects are. They could be numbers, people, countries, etc.

To illustrate, take three objects—say, Mishka (my cat), the number 3, and Iceland (the country). We can collect these things into a set. We can also denote this set by writing names for its objects, and enclosing them in curly braces:

$$\{\text{Mishka, 3, Iceland}\}.$$

We call the objects in this set its **elements** or **members**.

Alternatively, consider the set consisting of all the people currently living in Japan. We can write this set as follows:

$$\{x : x \text{ is a person currently living in Japan}\}.$$

Read this as: 'The set of all x such that x is a person currently living in Japan'. Here, we pick out the set by specifying some *property* shared by all its members. (This is called the **intensive** way specifying a set; the other way we looked at is called the **extensive** way.)

Challenge Question. For any property we can think of, can we form a set whose members have that property? Why or why not?

If A is a set, and x is an element of that set, then we write '$x \in A$' to say that x is an element of A. Here, '$\in$' is the symbol for **set membership**.

Sets are *defined* by their members. We call this the **axiom of extensionality**. To see what this means, consider this set:

$$\{\text{Mishka, Mishka, Mishka}\}.$$

How many members does this set have? The answer is 1 (not 3). Writing the *name* for an object in the set multiple times doesn't change how many *things* are in that set. (This idea turns out to be surprisingly important.)

2. Unions, Intersections, and Relative complements

Consider two sets. First: $\{1, 3, 5\}$. And second: $\{1, 2, 3, 4\}$. The **union** of these sets is the set: $\{1, 2, 3, 4, 5\}$. More generally, if A and B are sets, then their union is—written '$A \cup B$'—is the set:

$$A \cup B = \{x : x \text{ is an element of } A \textbf{ or } x \text{ is an element of } B\}.$$

(Note that, just as in logic, we take 'or' to be inclusive.) Conversely, the **intersection** of the two sets above is the set: $\{1, 3\}$. More generally, the intersection of two sets A and B—written '$A \cap B$'—is:

$$A \cap B = \{x : x \text{ is an element of } A \textbf{ and } x \text{ is an element of } B\}.$$

Finally, the **relative complement** of $\{1, 3, 5\}$ *in the set* $\{1, 2, 3, 4\}$ is the set $\{2, 4\}$. More generally, the relative complement of A in B—written '$B \setminus A$'—is the set:

$$B \setminus A = \{x \in B : x \notin A\}.$$

('$\notin$' is just our symbol for "non-membership".) This is also sometimes called the **set difference** of B and A. Note that it's not possible to have the complement of a set simpliciter—complementation is always "relative" to a given set. We'll see why that is in a couple of lectures' time.

3. Subsets and the Empty Set

[1] This definition is informal—and indeed, the notion of a set in mathematics is often taken as a kind of undefined primitive.

Consider the set consisting of all natural numbers, 0, 1, 2, 3,... This set is so important that we usually denote using a special symbol, '$\mathbb{N}$':

$$\mathbb{N} = \{0, 1, 2, 3, ...\}.$$

Consider also the set of even natural numbers:

$$\mathbb{E} = \{0, 2, 4, 6, ...\}.$$

Clearly, every element in $\mathbb{E}$ is also an element of $\mathbb{N}$. Thus, we say that $\mathbb{E}$ is a **subset** of $\mathbb{N}$.

More generally: a set A is a subset of another set B just in case every element in A is also an element in B. We write this as follows: '$A \subseteq B$'. (Two sets are equal just in case each is a subset of the other.)

Note that every set is a subset of itself. (After all, look at the definition for 'subset' we just gave!) When B contains some elements that A does not, however, then we say that A is a **proper subset** of B, and we write this '$A \subset B$'.

Note also that there's a special set, called the empty set, denoted $\varnothing$, that's a subset of every set. The empty set $\varnothing$ has no members at all.[2]

Challenge Question. Why is the empty set a subset of every set? (Hint: use the definition of 'subset'.)

4. Power sets

Consider the set $A = \{1, 2, 3\}$. How many subsets does it have? The answer is: 8: $\varnothing$, $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, $\{2, 3\}$, $\{1, 3\}$, $\{1, 2, 3\}$.

If we collect all the subsets of A into a single set, we get a set called the **power set of** A. We denote this set $Pow(A)$ (although different notations are sometimes used).

Challenge Question. If a set A has n members (where n is a natural number), can you say how many members the power set of A has?

5. Relations

Consider again the sets $\{1, 3, 5\}$ and $\{1, 2, 3, 4\}$. The **Cartesian product** of these sets is the set of all **ordered pairs**, $\langle x, y\rangle$, whose first member is an element of the first set, and whose second element is an element of the second.

More generally, if A and B are sets, then the Cartesian product of A and B, written '$A \times B$', is the set:

$$\{\langle x, y\rangle : x \in A \text{ and } y \in B\}.$$

(We can generalize this definition to include three sets, four sets, etc. But we'll focus mostly on just two sets, for present purposes.)

A **binary relation** is a subset of the Cartesian product of two sets.[3] (Likewise, a **ternary relation** is a Cartesian product of three sets, etc.)

To illustrate the notion of a binary relation, let A be the set of all living people. Then $A \times A$ is the set of all pairs of living people. One kind of binary relation might then be:

$$R = \{\langle x, y\rangle : x \text{ is a sibling of } y\}.$$

Clearly, R is a subset of $A \times A$. Note that we often use the notation 'xRy' or 'Rxy' to say that 'x stands in the relation R to y'. (Relations will come up *a lot* when we study modal logic.)

6. Functions

A **function** is a special kind of relation. It's a relation R such that, for every x, there's exactly one y such that x stands in the relation R to y.

To illustrate: consider the familiar function $f(x) = x^2$, where x is a natural number. This can be thought of as a set of pairs: $\{\langle 1, 1\rangle, \langle 2, 4\rangle, ...\}$.

We'll say a bit more about functions next time.

[2] It's permissible to write the empty set as '$\{\}$'.

[3] In this class, if I use the word 'relation' on its own, I'll almost always mean 'binary relation'.

PHIL 210/510: Formal Methods — Yale, Spring 2025

Lecture 2: Left-overs; Logic — Dr. Calum McNamara

1. More on Sets—and Venn Diagrams

Last time, we talked (a bit quickly!) about set theory. Among other things, we talked about the **algebra of sets**—things like unions, intersections, and relative complements.

Remember the definitions:

- The *union* of A and B is the set: $\{x : x \in A \text{ or } x \in B\}$.
- The *intersection* of A and B is the set: $\{x : x \in A \text{ and } x \in B\}$.
- The *relative complement* of A in B is the set: $\{x \in B : x \notin A\}$.

(One thing I forgot to say: we often use the notation '$A \cup B$' for the union of A and B; '$A \cap B$' for the intersection of A and B; and '$B \setminus A$' (or sometimes '$B - A$' or even '$A^c \in B$') for the relative complement of A in B. I corrected this on the previous handout.)

It's surprisingly easy to visualize these concepts using **Venn diagrams**. For example, here's one way you can visualize sets A, B, and their intersection:

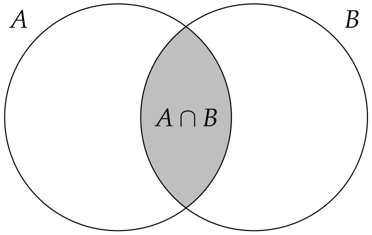


Figure1: Set intersection

Challenge Question. How would you alter this Venn diagram, so as to visualize the union of A and B, and the relative complement of A in B?

2. Relations and Functions

Consider the sets $\{1, 3, 5\}$ and $\{1, 2, 3, 4\}$. The **Cartesian product** of these sets is the set of all **ordered pairs**, $\langle x, y\rangle$, whose first member is an element of the first set, and whose second element is an element of the second—e.g., $\langle 1, 1\rangle$, $\langle 1, 2\rangle$, etc.

More generally, if A and B are sets, then the Cartesian product of A and B, written '$A \times B$', is the set:

$$\{\langle x, y\rangle : x \in A \text{ and } y \in B\}.$$

(We can generalize this definition to include three sets, four sets, etc. But we'll focus mostly on just two sets, for present purposes.)

A **binary relation** is a subset of the Cartesian product of two sets.[1] (Likewise, a **ternary relation** is a subset of the Cartesian product of three sets, etc.)

To illustrate the notion of a binary relation, let A be the set of all living people. Then $A \times A$ is the set of all *pairs* of living people. One kind of binary relation might then be:

$$R = \{\langle x, y\rangle : x \text{ is a sibling of } y\}.$$

Clearly, R is a subset of $A \times A$. Note that we often use the notation 'xRy' or 'Rxy' to say that 'x stands in the relation R to y'. (Relations will come up *a lot* when we study modal logic.)

A **function** is a special kind of relation. It's a relation R such that, for every x, there's exactly one y such that x stands in the relation R to y.

To illustrate: consider the familiar function $f(x) = x^2$, where x is a natural number. This can be thought of as a set of pairs: $\{\langle 1, 1\rangle, \langle 2, 4\rangle, ...\}$.

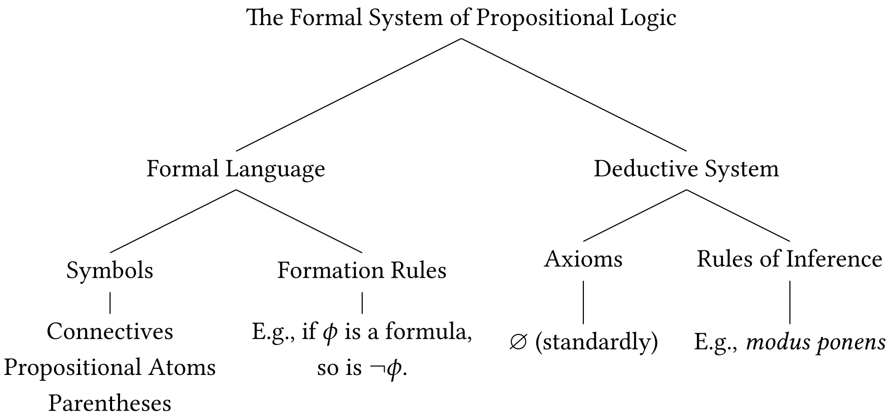
Let R be a relation (or a function—it doesn't matter). The **domain** of R, written '$dom(R)$' is the set: $\{x : \text{ there exists a } y \text{ such that } xRy\}$. Meanwhile, the **range** of R, '$ran(R)$', is the set: $\{y : \text{ there exists a } x \text{ such that } xRy\}$.

[1] In this class, if I use the word 'relation' on its own, I'll almost always mean 'binary relation'.

3. Propositional Logic—from 50k Feet

If you took Phil 115 with me, you probably got sick of me saying: 'Logic is the study of *arguments*'.[2] That's still true; but we're now going to think of logic a bit more abstractly. In particular, we're going to think about propositional logic as a **formal system**.

A formal system has two components: (i) a **formal language** and (ii) a **deductive system**. The tree below is a graphical representation of that situation for propositional logic:



In propositional logic, the formal language is (often) comprised of the following symbols:

- Parentheses: $($, $)$,
- Negation symbol: $\neg$,
- Conjunction symbol: $\wedge$,
- Disjunction symbol: $\vee$,
- Material conditional symbol: $\supset$,[3]
- Proposition atoms: $p_1, p_2, \ldots$; sometimes $p, q, r, \ldots$.

Then, we have **formation rules** for putting these symbols together, to form legitimate **formulas** of the language:

(i) Every propositional atom, $p_1, p_2, \ldots$ (or p, q,...) is a formula.

(ii) If ϕ is formula, then so is $\neg\phi$.[4]

(iii) If ϕ and ψ are formulas, then so is $(\phi \wedge \psi)$, $(\phi \vee \psi)$, and $(\phi \supset \psi)$.

(iv) Nothing else is a formula.

Meanwhile, the **deductive system** for propostional logic tells how, if we start with some formulas, we can **derive** other formulas, using some rules—so-called **rules of inference**. One way we can do this is to start with a very large stock of rules of inference. If you've learned **natural deduction** before, this is how you would've done things.

A different way we can specify the deductive system, however, is to start with a small stock of formulas, whose truth we take for granted. These are called **axioms**. For example, in propositional logic, one famous set of axioms is the following (due to Jan Łukasiewicz):[5]

- $(\phi \supset (\psi \supset \phi))$,
- $(\phi \supset (\psi \supset \chi)) \supset ((\phi \supset \psi) \supset (\phi \supset \chi))$,
- $((\neg\phi \supset \neg\psi) \supset (\psi \supset \phi))$.

We then take the following rule as our only rule of inference:

- **Modus Ponens**. Given φ and $(\varphi \supset \psi)$, infer ψ.

[2] I'm really hoping everyone in this class has some familiarity with logic, at least to the level of a course like Yale's Phil 115. If you don't, please speak to me—I can help you fill in the background.

[3] In previous courses, you may have used the symbol '$\rightarrow$' for the material conditional. In these notes, however, we reserve '$\rightarrow$' for the indicative conditional symbol (see chapter 4), and stick with '$\supset$' for the material conditional.

[4] Hang on! What the heck is 'ϕ' here? Think of it as a placeholder, which could stand for any formula. It's sort of analogous to how a variable, x, can stand for a number.

[5] Strictly speaking, Łukasiewicz's axioms are all **axiom schemas**. We get legitimate axioms, when we replace the ϕ's and ψ's with formulas of propositional logic.

We call any formula in a deductive system, which we can derive from the axioms, using the rules, a **theorem** of the system.

4. A Toy Formal System

One of the best ways to get a feel for how formal systems—like the one we just constructed—work is by *playing* with them.

With that in mind, then, let's play a game with a toy formal system, from Douglas Hofstadter's famous book *Gödel, Escher, Bach* (1979): the so-called MIU system.

The formal language of the MIU system consists of "strings" of the following symbols: 'M', 'I', 'U'. We also have a single axiom: Ml. The theorems of the system are then strings we can "build" using these rules:

(1) If you have a string whose last letter is 'I', then you can add a 'U' at the end.

(2) Suppose you have a string Mϕ (where ϕ is any string). Then you can write M$\phi\phi$ (where ϕ again is a string).

(3) If you have three 'I's in a row in your string, i.e., your string contains 'III', then you can replace 'III' with 'U'.

(4) If 'UU' occurs inside your string, you can delete it.

Challenge Questions.

- Given the axiom MI, show that you can write 'MIU'.
- Now that we have MIU, show that we can write 'MIUIU'.
- Imagine that we have a string UMIIIMU. Show that we can write 'UMUMU'.
- Suppose we have the string MUU. Show that we can simply write 'M'.

Like I said, it's fun to play around with this system, and see what well-formed strings you can legitimately arrive at. (These are the theorems.) (For example, in illustrating rules 3 and 4, I assumed that we had the strings UMIIIMU and MUU. But are those theorems that we can arrive at given the axiom MI and the rules 1-4 in the first place? The answer isn't obvious.)

If you're really keen, try the following exercise: try "to make [the string] MU. Don't worry if you don't get it. Just try it out a bit—the main thing is for you to get the flavor of this MU-puzzle. Have fun" (Hofstadter, 2000, p. 35).[6]

5. Back to Propositional Logic

Now what we want to do is some of the same thing in propositional logic. In this case (again), we have the following three axiom schemas:

- $(\phi \supset (\psi \supset \phi))$,
- $(\phi \supset (\psi \supset \chi)) \supset ((\phi \supset \psi) \supset (\phi \supset \chi))$,
- $((\neg\phi \supset \neg\psi) \supset (\psi \supset \phi))$.

And we have the following rule of inference:

- **Modus Ponens**. Given φ and $(\varphi \supset \psi)$, infer ψ.

Challenge Question. Can you derive the formula $((p \supset q) \supset (p \supset p))$ from the axioms? How about $(p \supset p)$?

6. The Epistemology of Logic

Since this is a philosophy course, let's do some philosophy. Recall that our one rule of inference in propositional logic is *modus ponens*. Here's a natural language example—which Phil 115 students are probably sick of.[7]

P1 It's raining.

P2 If it's raining, then it's wet.

∴ It's wet.

[6] Hint: if you try this, you may be trying for a very long time…

[7] Recall that the symbol ∴ in the following means 'therefore'.

This argument certainly seems valid—and thankfully, it is according to the system of propositional logic we sketched earlier. In fact, every argument that exhibits this basic structure is valid in propositional logic. That seems like a good thing, prima facie. It's difficult to even think of natural language arguments that (intuitively) invalidate *modus ponens*.

It's difficult, but (arguably) it's not impossible. Vann McGee (1985), for example, believes the following argument is a counterexample to *modus ponens*:[8]

P1 If a Republican wins the election, then if it's not Reagan who wins it will be Anderson.

P2 A Republican will win the election.

∴ If it's not Reagan who wins it will be Anderson.

The first premise seems plausible. (Suppose a Republican wins. Then, if there are only two such Republicans in the running—Reagan and Anderson—if it's not Reagan, it has to be Anderson.) Similarly, the second premise is plausible. (Suppose you're back in 1985, just prior to the votes being counted. The polls heavily favor a Republican win.) But the conclusion seems implausible: Reagan was a popular politician at the time. But Anderson was a laughing stock.

More generally, McGee thinks that arguments that exhibit the following structure are counterexamples to *modus ponens*:

P1 $(\phi \supset (\neg\psi \supset \chi))$

P2 ϕ

∴ $(\neg\psi \supset \chi)$

McGee justifies his claim by saying that it's easy to find instances of the above schema where we believe P1 and P2, and yet do not believe the conclusion of the argument. This is strange since valid arguments—of which McGee's schema alleges to be one—are supposed to be "truth-preserving".

What do you think?

7. Predicate Logic

So far, we've focused on propositional logic. But we can extend our formal system to encompass predicate logic, too. In that case we extend our language with additional symbols:

- Names: $a, b, c, \ldots$
- Variables: $x, y, z, \ldots$
- Predicate symbols: $F, G, R, \ldots$
- Quantifiers: ∀, ∃

We also add the following formation rules:

- Fa, Fb, Ga, Gb, Rab, Rba, etc., are formulas:
- If ϕ is a formula in which the name a appears, then so is $(\forall x)\phi(a := x)$ and $(\exists x)\phi(a := x)$. (Here '$(a := x)$' means 'where each instance of the name a is replaced with the variable x'.)

The details of this extension to the language of propositional logic isn't so important. All I want to note is that it *can* be done.

Likewise, we can extend our deductive system with new axioms, and new rules. The additional axioms are often taken to be these:

- $(\forall x)\phi \supset \phi(x := a)$
- $(\forall x)(\phi \supset \psi) \supset (\phi \supset (\forall x)\psi)$.

And the additional rule is:

- **Universal Generalization**. From ϕ, infer $(\forall x)\phi$.

We'll talk more about predicate logic in subsequent weeks.

[8] This isn't quite true. Vann McGee believes that modus ponens isn't a valid rule of inference for *natural language conditionals*, like those below. But he also doesn't believe that natural language conditionals are the material conditional.

PHIL 210/510: Formal Methods Yale, Spring 2025
Lecture 3: Infinity Dr. Calum McNamara

1. Propositional Logic as a Formal System—Again

Last time, we talked a little bit about the **formal system** of propositional logic. Recall that, in this formal system, the **formal language** consists of formulas like $p \wedge q$ (‘p and q’), $p \supset q$ (‘If p, then q’), $\neg p$ (‘not p’), and so on. Meanwhile, our **deductive system**—or at least, one *version* of the deductive system—consists of the following three axiom schemas:

- $(\phi \supset (\psi \supset \phi))$,
- $(\phi \supset (\psi \supset \chi)) \supset ((\phi \supset \psi) \supset (\phi \supset \chi))$,
- $((\neg\phi \supset \neg\psi) \supset (\psi \supset \phi))$.

as well as the following rule of inference:

- **Modus Ponens**. Given φ and $(\varphi \supset \psi)$, infer ψ.

Challenge Question. Can you derive the formula $((p \supset q) \supset (p \supset p))$ from the axioms? How about $(p \supset p)$?

Note that the system of **(first-order) predicate** logic builds on this formal system. It does so by enriching the formal language with formulas like $(\forall x)Fx$ (‘for all x, x has property F’), and $(\exists x)Fx$ (‘there exists an x with property F’), as well as a couple of new axioms and rules. We’ll talk more about that system later in the course, when we briefly touch on **higher-order logic**.

2. The Numbers as a Formal System

Why are we introducing propositional logic in this extremely abstract way? One reason is that it helps to introduce the notion of a formal system in general. It’s useful to know what a formal system is, and how to construct one, because they pop up all the time (albeit, sometimes in disguised ways).

For example, take the **natural numbers**, $\mathbb{N} = \{0, 1, 2, ...\}$. As it turns out, these numbers, together with the usual rules of arithmetic, can also be viewed as a formal system. Here’s a hint of how that works.

Our formal language consists of the following symbols: the number 0, parentheses, and the letter S.[1] Well-formed formulas of the system look like: 0, $S(0)$, $S(S(0))$, etc. We can read ‘$S(0)$’ as ‘the successor of 0’—and this, of course, is the number 1. Likewise, ‘$S(S(0))$’ says ‘the successor of the successor of 0’—namely, the number 2, and so on. Thus, we allow ourselves to use ‘1’, ‘2’, etc., as shorthands for the relevant successors.

Now here are our axioms (these are often called the **Peano axioms**):

- 0 is a number.
- If n is a number then so is its $S(n)$, the successor of n.
- 0 is not the successor of any number.
- If $S(n) = S(m)$, then $n = m$. (In other words, every number has a unique successor.)
- Let $P(n)$ be any statement describing a property pertaining to the number n. Suppose that $P(0)$ is true, and suppose that, whenever $P(n)$ is true, then so it $P(S(n))$. Then $P(n)$ is true for every number n. (This is sometimes called the **Principle of Mathematical Induction**—we’ll talk about it more on Wednesday.)

We can think, here, of the second axiom as also describing our one rule of inference: if we have a number of n, then can conclude that $S(n)$ is also a number.

As it turns out, these five axioms characterize almost everything we know about the natural numbers.[2]

Challenge Questions. Prove—from the axioms!—that 3 is a natural number. How would you define **addition** in our system? How about **multiplication**?

3. Constructing the Numbers from Sets

[1] If you did the Russell reading, he uses ‘succ’ instead of S.

[2] I say ‘almost’ because when it comes to defining addition, etc., we have to introduce some additional definitions. For example, we have to stipulate that, for any n, $n + 0 = n$.

At this point, we have *assumed* the existence of something that "plays the role" of zero—that's what our first axiom tells us. But what exactly *is* the number zero?

One way to think about the numbers is in terms of *sets*. In particular, the mathematician John von Neumann[3] showed that the following **model** satisfies all of the axioms we gave above:

- $0 = \{\} = \varnothing$,
- $1 = S(0) = \{\varnothing\}$,
- $2 = S(S(0)) = \{\varnothing, \{\varnothing\}\}$,
- $3 = S(S(S(0))) = \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}$.
- Etc.

So, roughly: von Neumann's definition says that the natural numbers are what we get by starting with the empty sets, and then forming sets, recursively, out of everything that came before.

4. The Infinite

It's clear that our operation S—or the set operations used in von Neaumann's hierarchy—can be used to "generate" numbers indefinitely. That is, even though every number n is finite, the set of all natural numbers is **infinite**.

That's still a bit vague, however. So how can we get a handle on it? Well, let's start again by thinking about the natural numbers, $0, 1, 2, 3, \ldots$. Now think about the even numbers, $0, 2, 4, 6, \ldots$. Notice something weird. We can write a **list** in which every even number is paired off, one to one, with a natural number:

0. 0
1. 2
2. 4

$\vdots$

So, the weird thing here is that, even though it *looks* like there should be half as many even numbers as natural numbers, nevertheless we can pair them off one-to-one.

Interestingly, the same thing goes when we consider a set that looks like it should have *more* numbers than the natural numbers. For instance, consider the set of all integers: $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. Once again, we can pair the integers off one-to-one with the natural numbers:

0. 0
1. 1
2. -1
3. 2
4. -2

$\vdots$

This is *weird*. After all, the even numbers are a *proper subset* of the natural numbers. And the natural numbers are a *proper subset* of the integers. Thus, the examples we gave above lend themselves to a definition of 'infinite set'

Definition (Infinite Set). A set A is **infinite** iff it can be put into a one-to-one correspondence with one of its proper subsets. It's **finite** otherwise.

Definition (Countable). A set A is **countable** iff either (a) it's finite, or (b) it can be put into one-to-one correspondence with the set of natural numbers. It's uncountable otherwise.

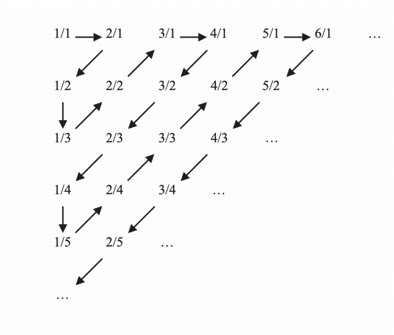
The above also lends itself to a definition of "size" for sets:

Definition (Cardinality). Two sets A and B have the same cardinality (viz., size) iff their members can be put into one-to-one correspondence.

5. The Uncountable

[3] You should Google von Neumann. He contributed to just about every field of human inquiry you can think of.

Let's consider one more example. Can the set of **rational numbers**—i.e., numbers of the form n/m—be put into a one-to-one correspondence with the set of natural numbers? In other words, can does the set of rationals, $\mathbb{Q}$, have the same cardinality as $\mathbb{N}$? Amazingly, the answer is 'Yes'. To see how, consider the following image (ripped from David Papineau's book *Philosophical Devices*):



The idea here is that the "fraction" $1/1$ goes in position 1; $1/2$ goes in position 2; $2/1$ goes in position 3; $1/3$ goes in position 4; and so on.

Thus, once again, even though it looks like there are "more" rational numbers than natural numbers, that turns out not to be true: $\mathbb{Q}$ can be put into one-to-one correspondence with $\mathbb{N}$. This might lead you to wonder whether *every* infinite set has the same size as $\mathbb{N}$. Bafflingly, the answer turns out to be 'No'. To illustrate, consider the set of **real numbers**, $\mathbb{R}$. This is the set of all numbers that can be expressed as an infinite decimal expansion. This includes the naturals, the integers, and the rationals, but also "irrational numbers" like π and e.

To see that there are more reals than naturals, consider the following (incidentally, the proof here also illustrates one of the proof techniques we'll talk about on Wednesday). Suppose all the real numbers between 0 and 1 can be put on a list, e.g.:

0. 0.1237263...

1. 0.43847485...

2. 0.4548457...

3. 0.3843758...

⋮

I claim we can *construct* a number that's guaranteed not to be on this list. To do so, we make the first digit one more than the first digit of the first number in this list, the second digit one more than the second digit of the second number, the third digit one more than the third digit of the third number, and so on . . . (using 0 as 'one more than 9' whenever the nth digit in the nth number is 9). Thus, the number we can construct, given the list I wrote, is: 0.2454....

Notice, however, that given our supposed initial listing of the reals between 0 and 1, our new number *can't* be anywhere in the original list, since it differs from the first number in the first digit, from the second in the second digit, and so on.

Thus, what we've shown here is that *the real numbers cannot be put into one-to-one correspondence with the natural numbers*. More broadly, what we've shown is that *there are infinite sets of different sizes*. The set of natural numbers, despite being infinite, turns out to have strictly fewer elements than the set of real numbers. (The technique we used here is called **diagonalization**.)

6. The Continuum Hypothesis

As it turns out, this result is related to the *power set* operation, which we talked about in the first class. If we start with an infinite set A, then the power set of A, $pow(A)$, is also an infinite set, which is strictly "bigger" than A!

This turns out to be true of $\mathbb{R}$: its cardinality is equal to the cardinality of the power set of the natural numbers. But is there an infinite set whose cardinality is strictly between these two? This is known as the **continuum hypothesis**. The answer is: we don't—and *can't*—know. It can be shown the truth or falsity of this statement is **independent** of the formal system of set theory, which we considered in the first class.

PHIL 210/510: Formal Methods — Yale, Spring 2025
Lecture 4: Proof; Bits and Pieces — Dr. Calum McNamara

1. How to Prove Things

In this class, you'll sometimes be asked to *prove* a certain statement. For example, a typical problem set question might look like:

Problem. Show that conditionalization preserves conditional probabilities. That is, show: $p_A(B \mid A) = p(B \mid A)$.

(Don't worry if you don't understand what any of that means yet—you will soon.) In order to do this, it's worth having a few techniques in your back pocket. So that's what I'll introduce you to now.

1.1 Unpacking Definitions

One of the first things you'll want to do in a proof is unpack the definitions you've been given. For example, consider the problem above. It asks you to show that two conditional probabilities are equal: $p_A(B \mid A)$ and $p(B \mid A)$. Given this, it's often a good idea to start by *unpacking* the relevant definition—in this case, the definition of conditional probability: $p(B \mid A) = p(A \wedge B)/p(A)$. (Again, don't worry if you don't know what this means. You will soon!)

Let's try an example—one that you *will* know something about already:

Challenge Question. Show that, if $A \subseteq B$, then $A \in Pow(B)$.

1.2. Proving 'if' Statements

You will sometimes be asked to prove statements that make use of the word 'if'—statements like this:

Challenge Question. Show that, *if* $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$.

When you're asked to prove statements like this, the first thing you'll want to do is *suppose* the 'if' part. You can then prove the part after 'if', given this supposition.

1.3. Proving 'if and only if' Statements

You will sometimes be asked to prove statements that involve the words 'if and only if' (or 'iff' for short). Good news. This really just involves proving two 'if' statements. To illustrate, supposer we want to prove the following:

Challenge Question. Let A and B be sets. Then, prove that $A \subseteq B$ iff $A \cap B = A$.

Here, you start by supposing that $A \subseteq B$, and then show that $A \cap B = A$. You *then* suppose that $A \cap B = A$, and then show that $A \subseteq B$.

1.4. Proof by Contradiction

Suppose we want to show that the following is true:

Challenge Question. Show that: $A \cap (B \setminus A) = \varnothing$.

One way you can do this is to suppose the inequality does *not* hold, and then show that, given this supposition, we can reason our way to a contradiction.

1.5. Proof by Induction

Remember the Principle of Induction, from our discussion of the Peano axioms. It said that: if P is a property of numbers, $P(0)$ is true, and *if* $P(k)$ is true, then this *implies* that $P(k+1)$ is true (for some arbitrary k), then $P(n)$ is true for all n.

This technique—proof by induction—can be applied to more than just numbers. For example, it often works when we want to prove things about a logical language:

Challenge Question. Prove that every well-formed formula of propositional logic has an even number of parentheses.

Here, we start by showing that the "atomic formulas" have an even number of parentheses (namely, 0). We then *suppose* arbitrary formulas φ and ψ have an even number of parentheses. Next, we prove that every formula we can *build* from these formulas, using the formation rules, also has an even number of parentheses. Induction then lets us conclude that every well-formed formula has an even number of parentheses.

(Note: If you need to use induction to solve a homework problem, I'll usually mention this in a hint.)

2. Use and Mention

Let's now take a look at something completely different. Remember how, when I introduced propositional logic, I said that formulas like p, $p \wedge q$, etc., were formulas *in* our formal language. In contrast, when I wanted to speak in general, *about* statements of the formal language, I used Greek letters, like φ and ψ.

This distinction—between *using* words/sentences and merely speaking *about*, or "mentioning", words/sentences—is important. It's called the **use/mention** distinction. We can illustrate it, in English, with a simple example. Is the following sentence true or false? If it's false, how can we make it true?

(1) net is part of a clarinet.

A harder example. Suppose I supply the following instructions to a bakery:

(2) Bake me a cake, and write God bless everyone inside a heart.

It's hard to figure out exactly what these instructions are supposed to mean. One ambiguity results because the phrase 'God bless everyone' is being *used* here, when it really needs to be *mentioned*. How can we supply quotes to make the sentence less confusing?

Here's another example, involving the Pig and Whistle pub in Oxford. Suppose I write an email to a sign-writer, who's made the sign for the pub. I say:

(3) There needs to be more space between pig and and and and and whistle.

Again, this is extremely hard to parse. How can we make it clearer?

Thus, the general rule is: when you are *using* a word, you do not use quotes; but when you are (merely) *mentioning* that word, talking about the word itself, rather than the thing to which it refers, then you do use quotes.

If you think you've got your head around these ideas, here's a problem to ponder. Suppose that we used the word 'leg' to refer to a horse's tail. Then, how many legs does a horse have?

3. Quotes and Corner Quotes

Consider the following sentence from MacFarlane's notes (p. 1):

(4) Where ϕ and ψ are formulas, $(\phi \wedge \psi)$ is true in a model $\mathcal{M}$ iff ϕ is true in $\mathcal{M}$ and ψ is true in $\mathcal{M}$.

We could try to re-write this sentence to account for the distinction between use and mention as follows:

(5) Where 'ϕ' and 'ψ' are formulas, '$(\phi \wedge \psi)$' is true in a model $\mathcal{M}$ iff 'ϕ' is true in $\mathcal{M}$ and 'ψ' is true in $\mathcal{M}$.

But we run into problems here since ϕ and ψ are *meta-variables* whose *values* are true in the model $\mathcal{M}$. In other words, the expression 'ϕ' merely denotes the symbol ϕ, which is not itself a formula of propositional logic.

Thus, we need to find a way around this problem. Quine—creative fellow that he was—invented the method of so-called *quasi-quotation*, or *corner-quotes*, for this purpose. Using corner quotes, we can rewrite the initial sentence like this:

(6) Where ϕ and ψ are formulas, $\ulcorner(\phi \wedge \psi)\urcorner$ is true in a model $\mathcal{M}$ iff ϕ is true in $\mathcal{M}$ and ψ is true in $\mathcal{M}$.

We can thus think of corner-quotes as a kind of notational shortcut. In particular, $\ulcorner(\phi \wedge \psi)\urcorner$ is a notational shortcut for: ϕ concatenated with '$\wedge$' concatenated with ψ.

Here's a harder example:

Challenge Question. Supply quotes and/or corner quotes to the following sentence, to make it true: For several definite descriptions D, Winston Churchill said We shall fight on D.

4. Types and Tokens

Consider this sentence:

(7) A rose is a rose is a rose is a rose.

How many words does this sentence contain? On the one hand, it seems sensible to say that it contains three words, namely ‘A’, ‘rose’, and ‘is’. But on the other hand, it seems equally sensible to say that it contains eleven words.

In fact, both answers are conceivably correct, because the question I asked was ambiguous. To disambiguate it, we can say: the sentence contains three word *types*, but eleven word *tokens*.

Challenge Question.

5. Analyticity, Necessity, A Prioricity

Statements (in English) can be true for different reasons. Moreover, there are different ways in which we can *discover* that a given statement is true, or false. For example, compare the following two statements:

(8) $2 + 2 = 4$

(9) It’s Sunny outside.

Both of these statements are true (at least at the time of writing). But the first seems to have a special property that the second lacks. Likewise, consider:

(10) I think, therefore I am.

(11) All bachelors are unmarried.

Arguably, these sentences also have special characters, which the second sentence, above, lacks. With this in mind, let us introduce some distinctions.

The Analytic/Synthetic Distinction. A sentence is said to be **analytically true** (or just **analytic**) if it’s true *purely in virtue of the meanings of the words.* For example:

(12) Vixens are female foxes.

It’s said to be **synthetic** otherwise.

The A Priori/A Posteriori Distinction. A sentence is said to be true *a priori* if it’s possible to discern it’s truth “prior to experience”. For example, you don’t have to go out into the world, conduct experiments, etc., to see that a certain sentence is a priori true. To illustrate, the first sentence here is often thought to be true *a priori*, while the second isn’t—as we say, it’s true *a posteriori.*

(13) $1 \neq 0$.

(14) Nothing travels faster than light.

The Necessary/Contingent Distinction. Finally, a sentence is said to be **necessarily true** if (roughly) it couldn’t possibly be false.[1] It’s merely **contingently true** otherwise. For example, the first sentence below is necessarily true, the second merely contingently true:

(15) $1 \neq 0$.

(16) It’s Sunny outside.

Challenge Question. Go through the four sentences that I started this section with. Which (if any) are true necessarily? A priori? Which are analytic?

[1] This definition is a bit circular. We’ll clarify things more, when we get to the next section—and, more importantly, when we get to modal logic.

Challenge Question. Historically, it was often thought that analytic = a priori = necessary. (That's one reason there's some overlap in the examples I gave above.) Nowadays, that view is widely rejected.[2] Can you think of any examples of a sentence which is, e.g., necessary, but a posteriori? How about contingent, but a priori? What about synthetic a priori?

Finally, it's probably worth knowing the following general facts about these distinctions:

- The analytic/synthetic distinction is usually taken to be a **semantic** distinction.
- The a priori/a posteriori distinction is usually taken to be an **epistemic** distinction.
- The necessary/contingent distinction is usually taken to be a **metaphysical** distinction.

6. Possible Worlds

Let's go back to the rough definition I gave of necessary truth. I said: a truth is necessary *if it couldn't possibly have been false*. One issue with this definition, however, is that appeals to the notion of possibility. And you might think a sentence's being *possibly* true is itself a notion that needs to be defined. How, then, are we to do this?

A common definition of 'necessary truth' in philosophy is *truth in all possible worlds*. (This is still rough; we'll make it more precise in our unit on modal logic.)

The notion of a possible world is arguably one of the most important notions to pop up in philosophy—especially in the last hundred-or-so years. We will use this notion in *all* of the units to come. But what *is* a possible world?

Often, in philosophical theorizing, we take the notion of a possible world as an unanalyzed primitive (the way that mathematicians take the notion of a set as an unanalyzed primitive). The best we can do is give it an informal gloss: a **possible world** is a *completely specific way the world could be*. It's something that "decides" the truth of every question you can ask. For example, you might wonder 'Is it raining?' Then, at any given possible world, the answer to that question will be either 'Yes' or 'No'. Likewise: 'Can things travel faster than light?' There may be possible worlds at which the answer is 'Yes'. But the important point is just that, *at* any given possible world, the question has an answer.

Later on, we'll see that we can analyze various things in terms of possible worlds—e.g., propositions. One interesting thing, however, is that we could (alternatively) think of possible worlds *themselves* as propositions (and leave 'proposition' as an unanalyzed primitive). On this view: a possible world is a proposition w such that, for any other proposition p, w either entails p or p's negation. This view is popular among higher-order logicians.

One last thing: Are possible worlds *real*? Almost everyone agrees, the answer is 'No'. They're usual fictions we invent for philosophical theorizing, the same way, e.g., the frictionless plain is a fiction useful for theorizing.

The great philosopher—the *greatest*, in my view—David Lewis, however, thought they are *real*. They are so useful in theorizing, he argued, that we should admit their existence. The argument is similar to the way we admit the existence of numbers—or better, *sets*—into our ontology, because numbers/sets are so useful in our theorizing. Mathematicians acknowledge the existence of sets, for example, because, in doing so, we can give a foundation for almost all other mathematical theorizing. The same thing goes, Lewis thought, for possible worlds.

This, however, strikes many as absurd. It's silly (they say) to think there's a *real* world where there's a talking donkey. Or two dragons fight for five minutes, and the world ceases to exist. And so on.

Lewis was well aware of these objections. But he thought this line of argument wasn't enough. As he put it, the most common reaction to his *arguments* for **modal realism**—the view that possible worlds are real—is the **incredulous stare**. But as he famously quipped: "I cannot refute an incredulous stare".

[2] Somewhat relatedly, many philosophers now reject the analytic/synthetic distinction altogether. Can you think of examples which seem to cast the legitimacy of that distinction into doubt?

PHIL 210/510: Formal Methods — Yale, Spring 2025

Lecture 5: Formal Semantics I — Dr. Calum McNamara

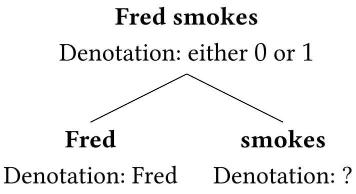
1. Frege on Compositionality

Thanks to our first two weeks spent on mathematical background, you're now familiar with the notion of a **function**. Indeed, you've been familiar with the notion of a function, as it applies to *numbers*, for some time. We all know, for example, how to compute the value of 7^2, given our knowledge of what 7 is, what the square function, $f(x) = x^2$, is, etc.

The great German mathematician, logician, and philosopher Gottlob Frege had the brilliant idea that **meaning** in natural language works in a similar way to the function $f(x) = x^2$. In particular, he thought that we can compute the meaning of whole sentences, from the meanings of their parts, in a fashion similar to the way we compute $7^2 = 49$.[1]

Frege's Conjecture. Semantic composition is functional application.

To illustrate this, consider a simple example: the sentence **Fred smokes**.[2] According to Frege, the denotation of a proper name, like **Fred** is just the person Fred. Similarly, the denotation of a declarative sentence, like **Fred smokes**, is just it's truth-value—i.e., either 0 and 1. But what about the predicate/intransitive verb **smokes**?



Frege's idea was that we should think of the denotation of a predicate like **smokes** as a *function*—namely, the function which takes in "entities" (like Fred), and maps them to truth values. Thus we have the following:[3]

- $[\![\textbf{Fred}]\!] = \text{Fred}$
- $[\![\textbf{Fred smokes}]\!] = 1$ iff Fred smokes
- $[\![\textbf{smokes}]\!] =$ a function that takes in entities, and maps them to 1 iff the entity in question smokes.

Applying Frege's idea, with **smokes** as the function and Fred as the entity, we thus get:

$$[\![\textbf{smokes}]\!]([\![\textbf{Fred}]\!]) = \ 1 \text{ if Fred smokes, 0 otherwise.}$$

In this example, the phrase '= 1 if Fred smokes, 0 otherwise' gives the **truth-conditions** for the sentence 'Fred smokes'. In turn, the truth-conditions tell us what the world would have to be like, in order for a given sentence to be true.

2. A Digression on Category Mistakes

Consider the following example, which looks a lot like the one we just encountered:

$$[\![\textbf{smokes}]\!]([\![\textbf{two}]\!]) = \ 1 \text{ if two smokes, 0 otherwise.}$$

Is this function defined or not? If the function *is* defined, then it should of course output 0 (false), since the number two can't smoke.

But then again, you might think that the function should simply "crash" here, because the number two isn't something to which the verb **smokes** can apply in the first place.

[1] In fact, Frege's idea leads to a brilliant, alternative foundation for mathematics known as *type theory*. Type theory is having a rivival at the moment. It plays a part in much recent work in philosophical logic—especially so-called *higher-order logic*—and metaphysics—especially so-called *higher-order metaphysics*. If you're curious, you should do some googling.

[2] Following Heim and Kratzer, I'll often use **boldface** text, when I'm mentioning a word/phrase/sentence, rather than using it, instead of using quotes. Note that this means by use of boldface does double duty!

[3] The double brackets here, '$[\![$', '$]\!]$', also denote a function—the **denotation function**. It maps words/phrases/sentences to their semantic values.

This issue divides semanticists (and philosophers!): some say that the function should output the value 0; others say that it shouldn't output a value at all. We're going to be simple-minded here, however, and assume that functions like the one above need not "crash" when they're given funny arguments.

3. Semantic Types

We have already referred (implicitly) to numerous **semantic types**. For example, we have spoken of the type of **entities** (like Fred); and we have discussed the fact that these entities are the inputs to functions (predicates). Furthermore, we have discussed the type of **truth-values**, which are just the numbers 0 and 1.

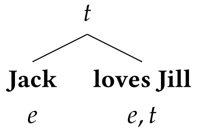
In this section, we are going to specify recursive rules for determining the semantic type of *any* linguistic object. First, let's denote the domains of the semantic types we already have at hand as follows:

(i) D_e is the domain of **entities**,

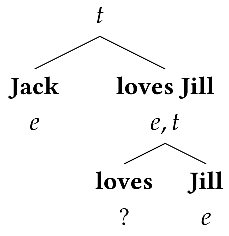
(ii) D_t is the domain of **truth-values**, i.e., the set $\{0, 1\}$,

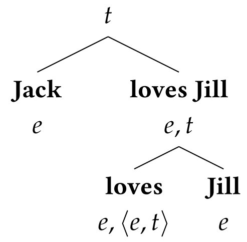
(iii) $D_{\langle e,t\rangle}$ is the domain of functions that have entities as their arguments and truth-values as their values.

This is a good start. But we need more semantic types than this. To illustrate why. Consider the following sentence: **Jack loves Jill**, alongside the corresponding syntactic tree: Now, it seems clear that **loves Jill** should be treated



as a function of type e, t. But it's also clear that we can break **loves Jill** down even further. Doing so results in the tree on the following page. From there, it's clear that **Jill** should have type e. But what about the transitive verb **loves**? Well, we want **loves** to be something that takes in an entity (in this case Jill), and outputs a *function* of type e, t. Thus, we can conclude that **loves** has semantic type $e, \langle e, t\rangle$, and that the tree may be completed as in the second tree on the next page.





Examples like this one motivate the idea that we're going to need many more semantic types than we currently have. Thus, we're going to introduce an infinite family of semantic types, according to a recursive procedure—namely, the following:

(1) e is a semantic type,

(2) t is a semantic type,

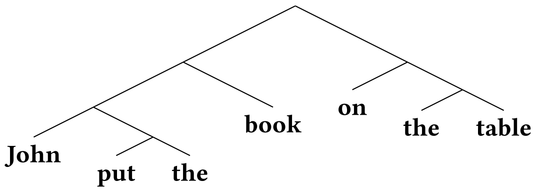
(3) If ϕ is a semantic type and ψ is a semantic type, then so is $\langle\phi, \psi\rangle$ (i.e., the function that takes in things of type ϕ and outputs things of type ψ).

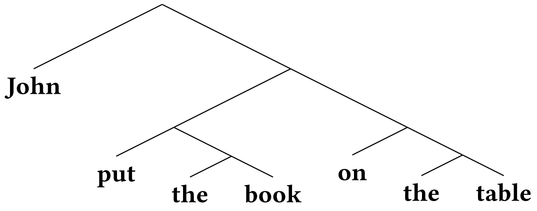
(Note that we'll often drop the angle brackets, writing things like 'e, t' instead of '$\langle e, t\rangle$', provided doing so doesn't introduce ambiguity.)

Of course, it's going to turn out that not every semantic type in this infinite hierarchy is one that we'll find commonly in natural language. Nevertheless, it's good to have them all at our disposal—we're going to need more of them than you'd think.

4. Syntactic Trees, Briefly

Above we've been "breaking down' sentences (like **Jack loves Jill**) into their constituent parts. There are better and worse ways to do this. For example, which of the following is the correct syntactic tree for the sentence **John put the book on the table**? Intuitively, it's the second. But we need to be able to say why, exactly, the second tree is the correct one.





Linguists have developed a battery of tests (of which we'll look at three) for determining **syntactic constituents** of sentences. These tests are not, unfortunately, water-tight. But they do function as relatively good heuristics for determining syntactic constituents.

The first such test is called the **short answers test**. The idea is that, if some piece of a sentence can function as a short answer to a question, then it's likely to be a syntactic constituent. Here are a few examples.

- *Question*: What did John put on the table? *Answer*: The book.
- *Question*: Where did John put the book? *Answer*: On the table.

This suggests that **the book** and **on the table** are syntactic constituents of **John put the book on the table**.

The second test is called the **pro-form substitution test**. Pro-form is the general category of words including pronouns and proverbs. For example, alongside **He**, **she**, etc., we have words like **did**. Altogether, these words form a category called the pro-form category.

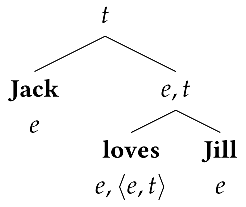
Thus, our next test is: if some piece of a sentence can be replaced with a pro-form, then it is likely to be a syntactic constituent of the sentence. To illustrate:

- **John put the book there** (with **there** replacing **on the table**).
- **Who put the book on the table? John did** (with **did** replacing **put the book on the table**).

Our final test is called the *movement test*. This test is best explained by means of example. So consider that we can transform **John put the book on the table** into (the rather sententious) **On the table John put the book**. Or alternatively, we can say **On the table is where John put the book**. The idea behind the movement test, then, is that certain units in the sentence naturally move together. These units are (likely to be) the syntactic constituents of the sentence.

More on Functional Application and λ-Notation

Having taken the necessary detour through the theory of syntactic structure, let's now take another look at a (slightly simplified version of the) tree we considered before:



Earlier, we said that it was possible to determine the semantic type of **loves** by looking at the semantic type of **Jill** and the node connecting **Jill** to **loves**. However, we also want a rule which tells us how put words like **loves** and **Jill** together to get the semantic value of **loves Jill**. The rule we need is this:

> **Functional Application (FA)**. If α is a branching node, $\{\beta, \gamma\}$ is the set of its daughters, and $[\![\beta]\!]$ is a function whose domain contains $[\![\gamma]\!]$, then:
>
> $$[\![\alpha]\!] = [\![\beta]\!]([\![\gamma]\!]).$$

Here, it's important to keep in mind the following distinction:

(i) The **semantic value** of a word is just the thing it denotes. For example, $[\![$**Jill**$]\!]$ = Jill (the person).

(ii) The **semantic type** of a word is the domain of things to which the thing denoted belongs. Thus, the semantic type of **Jill** is D_e, the domain of entities, since the person, Jill, is an entity (and not, e.g., a function or a truth-value).

Thus, to compute the semantic value of the node **loves Jill** in the tree above, we use functional application. That is, if **loves Jill** $= \alpha$, then we have:

$$[\![\textbf{loves Jill}]\!] = [\![\textbf{loves}]\!]([\![\textbf{Jill}]\!]).$$

And as we know, this is going to be a function f, which takes in entities, and returns more functions. A more traditional notation denoting this function would be extremely cumbersome, and is worse than useless when the function has infinitely many arguments and/or values. Thus, for this reason, semanticists often use a somewhat unusual notation—λ-**notation**.

Consider, for example, the way Heim and Kratzer (1998) define the successor function using the following notation:

$$f(n) = [\lambda n : n \in \mathbb{N} \,.\, n + 1].$$

This reads: "λn is the function that maps every natural number n to its successor, $n + 1$." More generally, when we see a function rendered in this so-called *λ-notation*, we read it as follows. Consider:

$$[\lambda \alpha : \phi \,.\, \psi].$$

In words: λ is the function that maps every α such that α is in the domain specified by ϕ to its value, ψ. Thus, α is the **argument variable**, ϕ is the **domain condition**, and ψ is the **value description**.

When it is obvious, we suppress the domain condition in the λ-notion. For example, if it is clear from context that we're talking about the natural numbers, then we can define the successor function as:

$$[\lambda n \,.\, n + 1].$$

This reads: "the function that maps every natural number n to its successor, $n + 1$."

With the Functional Application rule and λ-notation clearly in mind, then, here's a challenge question to end with:

Challenge Question. Which of the following is the correct denotation of **loves**?

(1) $[\![\textbf{loves}]\!] = [\lambda x \,.\, [\lambda y \,.\, y \text{ loves } x]]$,

(2) $[\![\textbf{loves}]\!] = [\lambda x \,.\, [\lambda y \,.\, x \text{ loves } y]]$.

PHIL 210/510: Formal Methods — Yale, Spring 2025
Lecture 6: Formal Semantics I — Dr. Calum McNamara

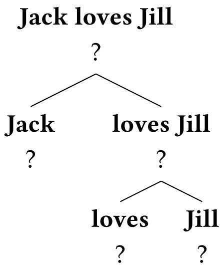
1. Unfinished Business

Last time, we talked about simple sentences like **Fred smokes**. Recall that:

- $\llbracket \textbf{Fred} \rrbracket = \text{Fred}$
- $\llbracket \textbf{smokes} \rrbracket$ = a function that takes in entities, and returns the value 1 iff the entity smokes.
- $\llbracket \textbf{smokes} \rrbracket(\llbracket \textbf{Fred} \rrbracket) = 1$ iff Fred smokes.

What we've done in the last line is compute the **truth-conditions** for the sentence **Fred smokes**. (Note that functions of this form—which output 1 if an entity has some property, and 0 if it doesn't have the property—are called **characteristic functions**.)

We also ended with some remarks on syntactic structure—we looked at a few tests, which help us to identify the "syntactic constituents" of sentences. Having done that, we then "broke down" the sentence **Jack loves Jill**:



Challenge Question. As a warm-up, let's begin today by replacing the question-marks in the tree structure above with **semantic types** for the various expressions.[1]

The process we just went through to answer this question illustrates the general rule we use for "combining" basic parts of sentences, to get meanings for more complex parts. (Our first example did so as well.) The rule I have in mind:

> **Functional Application (FA)**. If α is a branching node, $\{\beta, \gamma\}$ is the set of its daughters, and $\llbracket \beta \rrbracket$ is a function whose domain contains $\llbracket \gamma \rrbracket$, then:
>
> $$\llbracket \alpha \rrbracket = \llbracket \beta \rrbracket(\llbracket \gamma \rrbracket).$$

Remember that $\llbracket \cdot \rrbracket$ is the **interpretation function**.[2] It maps word tokens to their denotations. Here, it's important to keep in mind the following distinction:

(i) The **semantic value** of a word is just the thing it denotes. For example, $\llbracket \textbf{Jill} \rrbracket = \text{Jill}$ (the person).

(ii) The **semantic type** of a word, in contrast, is the domain of things to which the thing denoted belongs. Thus, the semantic type of **Jill** is e—viz., entities—since the person, Jill, is an entity (and not, e.g., a function or a truth-value).

To compute the semantic value of the node **loves Jill** in the tree above, we use the functional application rule. That is, if **loves Jill** $= \alpha$, then we have:

$$\llbracket \textbf{loves Jill} \rrbracket = \llbracket \textbf{loves} \rrbracket(\llbracket \textbf{Jill} \rrbracket).$$

Moreover, we know, this is going to be a function f, which takes in entities, and returns another function.

Semanticists often use a special notation when it comes denoting functions. It derives from the mathematician Alonzo Church, and is sometimes called **λ-notation**.[3]

[1] Remember: my use of **boldface** text in these notes is ambiguous for this section of the course. I use it—following Heim and Kratzer—both to distinguish between use and mention, but also when I'm (re-)introducing technical terms, as a kind of emphasis.

[2] This function sometimes goes by other names—like 'denotation function'. Apologies in advance if I accidentally switch my terminology.

[3] As a bit of history, Church is famous for, among other things, the **Church-Turing**

Consider, for example, the way Heim and Kratzer (1998) define the successor function using the following notation: $f(n) = [\lambda n : n \in \mathbb{N} \,.\, n+1]$. This reads: "$\lambda n$ is the function that maps every natural number n to its successor, $n+1$." More generally, when we see a function rendered in this so-called λ-*notation*, we read it as follows. Consider: $[\lambda\alpha : \phi \,.\, \psi]$. In words: λ is the function that maps every α such that α is in the domain specified by ϕ to its value, ψ. Thus, α is the **argument variable**, ϕ is the **domain condition**, and ψ is the **value description**.

When it's obvious, we suppress the domain condition in the λ-notion. For example, if it's clear from context that we're talking about the natural numbers, then we can define the successor function as: $[\lambda n \,.\, n+1]$. This reads: "the function that maps every natural number n to its successor, $n+1$."

With the Functional Application rule and λ-notation clearly in mind, then, here's a challenge question:

Challenge Question. Which of the following is the correct denotation of **loves**?

(1) $[\![\textbf{loves}]\!] = [\lambda x \,.\, [\lambda y \,.\, y \text{ loves } x]]$,

(2) $[\![\textbf{loves}]\!] = [\lambda x \,.\, [\lambda y \,.\, x \text{ loves } y]]$.

Having done that, can you compute the semantic value of the whole sentence **Jack loves Jill** from its most basic parts?

2. The Semantic Value of 'is'

Roughly speaking, a **copular sentence** is an **is** sentence—a sentence containing the word **is**. For our purposes, these sentences come in two varieties: (i) **identity** sentences, and (ii) **predicational** sentences.

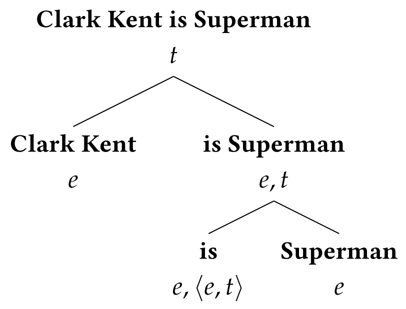
First, a copular sentence involving identity is one like 'Rhian is my sister'. Second, a copular sentence involving predication is one like 'My sister is nice'. How are we to determine whether a given copular sentence is an identity sentence or a predicational sentence? Often, you'll simply able to recognize this, *a priori*. However, here are some tests to help:

- **The Quantifier Test**. Check whether the sentence has quantifiers; if it does, then there's a good chance we're looking at a predicational sentence. (Example: 'nothing is expensive'.)

- **The Small Clauses Test**. Predicational sentences can be naturally embedded into clauses under 'consider'. (Example: 'I consider my sister nice'.)

As a rule of thumb, the identity copula functions something like the equals sign, '='. By contrast, the copula of predication is used to say that some object has a certain property.

However, why are we even talking about copular sentences and the associated tests at all? The reason is that, depending on whether 'is' is functioning as the 'is' of identity or 'is' of predication in a given sentence, it will have different semantic values. To illustrate this, consider the sentence **Clark Kent is Superman**. This is clearly the **is** of identity. Thus, the tree for the sentence, together with its semantic types, looks like this.



Plainly, since the type of **Superman** is e, the type of **is** in this case must be $e, \langle e, t\rangle$. Furthermore, notice that the semantic value of **is Superman** is:

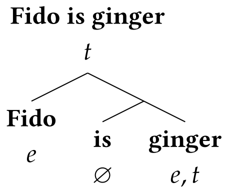
$$[\![\textbf{is Superman}]\!] = [\lambda y \,.\, y \text{ is Superman}].$$

thesis in the foundations of computer science. He also invented the (formal) notion of the computer, at the same time as Alan Turing. Later, he became Alan Turing's doctoral advisor, at Princeton.

So it makes sense to say that the semantic value of **is**, in the identity case, is:

$$[\![\textbf{is}]\!] = [\lambda x \,.\, [\lambda y \,.\, y = x]].$$

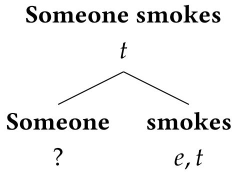
By contrast, consider **Fido is ginger**, where Fido is a dog. Here, the **is** is the predicational **is**. Rather surprisingly, semanticists generally agree that, in this case, the semantic value of **is** is null. That is, it contributes *nothing* to the meaning of the sentence **Fido is ginger**; we could just as well have written **Fido ginger**, and ended up with a sentence that is (in a formal sense) equally meaningful:



Again, maybe this seems a little surprising to you. But take another look at the Small Clauses test. Does it make more sense now? Similarly, take a look at the semantic type of **ginger**. If **is** in this case wasn't null, how would we have to change **ginger**?

5. Quantifiers

Let us now consider quantifiers. What, for example, is the semantic type of **someone**? Consider the tree on the next page.

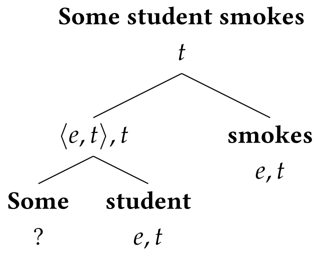


Plainly, since **smokes** is of type e, t, we want **someone** to be a word which takes in words of the same type as smokes, and outputs a truth-value. Hence, **someone** should be of type $\langle e, t\rangle, t$. Its semantic value is:

$$[\![\textbf{someone}]\!] = [\lambda f \,.\, \exists x \text{ such that } f(x) = 1].$$

More broadly, we can say that the semantic value of **Someone smokes** is 1 iff there exists x such that x smokes.

The methodology we've employed here allows us to compute the semantic type of quantifiers more generally. For instance, consider **Some student smokes**, and its associated tree:



Since we want **Some student** to be type $\langle e, t\rangle, t$, and we know that **student** is of type e, t, it follows that **some** is of type $\langle e, t\rangle, \langle\langle e, t\rangle, t\rangle$. Furthermore, the semantic value of **some** is:

$$[\![\textbf{some}]\!] = [\lambda f \,.[\lambda g \,.\, \exists x f(x) = g(x) = 1]].$$

Note that all the other quantifier phrases with which you're familiar follow the same pattern. That is, their types are $\langle e, t\rangle, \langle\langle e, t\rangle, t\rangle$. It's a good exercise to show this using other phrases, like **All students smoke**. (Problem set anyone?)

6. Predicate Modification

Now consider the sentence **Fido is a ginger dog**. We know that **ginger** is of type e, t and so is **dog**. So how do we combine **ginger** and **dog** in such a way as to yield something of type e, t, which we know we need higher up in the tree?

One answer is: we can introduce a new rule called **predicate modification**. This rule allows us to simply intersect adjectives like **black** and **dog** to give an object of semantic type e, t, thus avoiding any potential clash. The rule is formalized as follows:

> **Predicate Modification (PM)**. If α is a branching node, $\{\beta, \gamma\}$ is the set of its daughters, and moreover $\llbracket\beta\rrbracket \in D_{e,t}$ and $\llbracket\gamma\rrbracket \in D_{e,t}$, then $\llbracket\alpha\rrbracket = [\lambda x \,.\, \llbracket\beta\rrbracket = 1, \llbracket\gamma\rrbracket = 1]$.

7. Non-intersective Adjectives

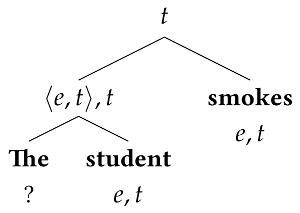
But what about the sentence **Fido is a green dog**. Or alternatively, consider **This is a fake diamond**. In these cases, predicate modification would lead to incorrect results. Why?

The answer is that predicate modification applies only to *intersective* adjectives. However, **fake** and **diamond**, for example, are *non-intersective*: the set of objects to which the word **fake** legitimately applies, and the set of objects denoted by the word **diamond**, do not overlap; they have empty intersection. Similarly the set of objects to which **dog** applies, and the set of objects to which **green** applies, have empty intersection. So predicate modification cannot be used in these cases.

8. Definite Descriptions

We have one last topic to cover before we move on to consider semantics from a slightly different angle (in Wednesday's class). This is: *definite descriptions.*

Consider the sentence **The student smokes**.



Again, **smokes** has type e, t: it's a function that takes in something of type e and returns a truth-value. Likewise, **student** is of type e, t. But what about **the**?

The tree above suggests that **the** has to be of type $\langle e,t\rangle, \langle\langle e,t\rangle t\rangle$. This is the same type as that of quantifiers, like **some**, which we considered a few sections back. Is the denotation of **the** also the following function?

$$\llbracket\textbf{the}\rrbracket = [\lambda f \,.[\lambda g \,.\, \exists x f(x) = g(x) = 1]].$$

(As a reminder, that's the same function as that denoted by **some**.) Intuitively it shouldn't be. **Some student** is true is one student smokes; but it's also true if five students smoke; or ten students smoke. But the thing picked out by **the** seems more specific. We take its semantic value to be:

$$\llbracket\textbf{the}\rrbracket = [\lambda f \,.[\lambda g \,.\, \exists! x f(x) = g(x) = 1]].$$

Here, $\exists! x$ means 'there is exactly one thing, x'.[4] So, uniqueness is an important feature of definite descriptions.

[4] If you took first-order logic with me, you'll recall that we never introduced the symbol $(\exists! x)Fx$ for 'There exists exactly one x such that Fx'. But we can think of $(\exists! x)Fx$ as a shorthand for $(\exists x)Fx \wedge (\forall y)(Fy \supset y = x$.

PHIL 210/510: Formal Methods — Yale, Spring 2025
Lecture 7: Formal Semantics III — Dr. Calum McNamara

1. Deriving Truth-conditions

Today, we're going to start by deriving **truth conditions** for various sentences. This is something we did in class last time. In particular, we derived the truth conditions for the sentence **Fred smokes**, by starting with the semantic values of the words **Fred** and **smokes**:

- $[\![\textbf{Fred}]\!] = \text{Fred}$
- $[\![\textbf{smokes}]\!] = [\lambda x.x \text{ smokes}]$.

(Remember: we read '$[\lambda x.x$ smokes$]$' as 'the function that takes in x (where x is an entity) and maps it to the value 1 iff x smokes'. Thus, the '...and maps it to 1...' is implicit. As Heim and Kratzer say (1998, p. 36), we're adopting this reading as a kind of *convention*. We could write it out more explicitly in in our λ-notation. But we won't.)

Now, to derive the truth conditions for **Fred smokes**, we use our rule of **functional application** (FA), putting Fred in as an argument to the function $[\![\textbf{smokes}]\!]$:

$$\begin{aligned}[\![\textbf{Fred smokes}]\!] &= [\![\textbf{smokes}]\!]([\![\textbf{Fred}]\!])\\ &= [\lambda x.x \text{ smokes}](\text{Fred})\\ &= 1 \text{ iff Fred smokes.}\end{aligned}$$

Let's now use the same methodology to derive the truth-conditions for more complicated sentences. (When you do this, it's a good idea to start by writing down all the semantic values for the words/phrases you're starting with, if you know them—just as we did in the case of **Fred smokes**. If you don't know them, we'll have to start by figuring them out. And to do that, we'll have to start by drawing out trees, and labelling **semantic types**.)

Challenge Question 1. Derive the truth conditions for the sentence **Jack loves Jill**.

Challenge Question 2. Derive the truth conditions for the sentence **Someone smokes**. (Note: to derive these truth conditions, you'll first have to figure out the semantic value of **Someone**—which is non-trivial. To do so, we'll start by constructing a syntactic tree.)

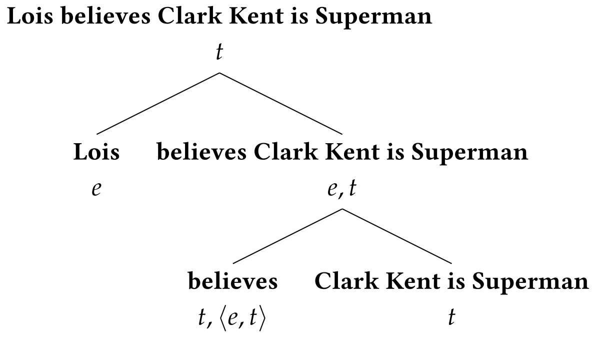
Challenge Question 3. Derive the truth conditions for **The student smokes**.

2. A Problem for Extensional Semantics

Let's try one more Challenge Problem, which will help us transition to today's main topic.

Challenge Question 4. Derive the truth conditions for **Lois believes Clark Kent is Superman**.

We'll do this last question together. First, then, let's draw out (part of) the relevant tree diagram:


Lois believes Clark Kent is Superman
t
Lois
e
believes Clark Kent is Superman
e, t
believes
t, ⟨e, t⟩
Clark Kent is Superman
t

Thus, given this tree diagram, you might think that the semantic value for **believes** should be something of type t, $\langle e, t\rangle$—that is, a function which takes in a truth value (either 0 or 1), and maps it to another function. But this immediately raises a problem. To see it, consider these two sentences:

- **Clark Kent is Superman**
- **Bruce Wayne is Batman**

Both of these sentences (let's pretend) are true true, and thus we have:

$$[\![\textbf{Clark Kent is Superman}]\!] = [\![\textbf{Bruce Wayne is Batman}]\!] = 1.$$

And so it seems like, when we're computing the truth conditions for **Lois believes Clark Kent is Superman** and **Lois believes Bruce Wayne is Batman**, our current theory will say that these truth conditions are the same. In other words, **Lois believes Clark Kent is Superman** is true iff **Lois believes Bruce Wayne is Batman** is. After all, in both cases, we're handing the function **believes** the same truth-value—namely, 1!

But this is plainly nonsense. Clearly, Lois could believe one of these sentences, without believing the other. In fact, the problem is even more rampant than you might currently appreciate. Our current semantic theory says that, *for any true sentences* φ and ψ, Lois believes one just in case she believes the other. Thus, our current semantic theory clearly fails.

3. Towards a Solution: Intensions

Our current semantic theory is an **extensional semantics**. That is, it says the semantic values of words, phrases, and sentences are just their **extensions**—that is, the things those words/phrases/sentences denote *at the actual world*. In the case of sentences, these extensions are just **truth values**. But as we just saw, this extensional view of sentences leads to problems when it comes to words like **believes**.

Thus, to rectify this, we're going to introduce the notion of an **intension**. This is a tricky notion to pin down. For example, Heim and Kratzer (1998, p. 302) give the rather unhelpful definition of an intension as "a function from indices to appropriate extensions", where the "indices" here can be people, times, and what have you.

As Heim and Kratzer also say, however, we can simplify matters by thinking of intensions just as functions from possible worlds to extensions. What exactly is meant by this will become clear as we go along.

4. Enriching our Hierarchy of Semantic Types

To get a better feel for what we mean by **intensions**, let's start as follows. First, let W be the set of *all possible worlds*.

Now, recall the set D_e which consists of all the "entities" that obtain at the actual world. (When we introduced D_e initially, we left the 'at the actual world clause' implicit.) Clearly, different entities exist at different worlds—Superman doesn't exist at our world, for example; but he exists at other possible worlds. Thus, for each possible world, w, there's a corresponding set of entities—each world has it's own D_e.

With this in mind, then, let D now be the union of all the sets D_e.[1] Thus, D is the set of all individuals, at all possible worlds. And we'll now think of e-type expressions, as those that denote entities in the set D.

Given this new notion of the set of entities, we can introduce a hierarchy of new sets:

- D is the set of all individuals (at all possible worlds),
- $D_t = \{0, 1\}$ is still the set of truth values,
- $D_{\sigma,\tau}$ is the set of all functions, which take in elements of the set D_σ and return objects in D_τ.
- $D_{s,\tau}$ is the set of all functions from W, the set of worlds, to elements of the set D_τ (whatveer that may be).

Given these sets, we can now enrich our original hierarchy of semantic types:

- e is a type,
- t is a type,
- if σ and τ are types, then so is $\langle \sigma, \tau \rangle$,
- if τ is a type, then so is $\langle s, \tau \rangle$.

[1] In formal notation, this would be written: '$D = \bigcup_e D_e$'.

Thus, for example, objects of type $\langle s, t \rangle$ are functions which take possible worlds, and maps them to truth-values. This type, as we'll see, is really going to be the key to getting the right results when it comes to sentences like **Lois believes Clark Kent is Superman**

(Also, for what it's worth, I'm not sure why the notation s is used in the case of intensions. It's traditional—but a little confusing, in my view.)

5. Characteristic Functions and Propositions

Think again about the type $\langle s, t \rangle$—the type of functions which map possible worlds to truth-values. Clearly, such a function is going to take each possible world, w, map it either to 0 or 1.

Functions like this are called **characteristic functions**. More generally: for any given set, A, its characteristic function is the function which takes each $x \in A$ and maps it to 1, and for any $x \notin A$, maps it to 0. For example, the characteristic function of the set $\mathbb{E}$ (of even natural numbers) is the function defined, for every $n \in \mathbb{N}$, by:

$$f(n) = \begin{cases} 1 & \text{if } n \text{ if even} \\ 0 & \text{otherwise.} \end{cases}$$

Every set has a corresponding characteristic function of this kind. And as it turns out, they're very, very useful.

Think, for instance, about the notion of a **proposition**. This is something we talked about early in the class. One nice thing about our present set-up is that we're now able to give a very precise characterization of this notion. Here, for instance, is how Robert Stalnaker spells this out:

> The explication of *proposition* given in formal semantics is based on a very homely intuition: when a statement is made, two things go into determining whether it is true or false. First, what did the statement say...? Second, what is the world like: does what was said correspond to it? What, we may ask, must a proposition be in order that this simple account be correct? It must be a rule, or a function, taking us from the way the world is into a truth value. But since... we may wish to consider the statement relative to hypothetical and imaginary situations, we want a function taking not just the actual state of the world, but various possible states of the world into truth values. Since there are two truth values, a proposition will be a way—any way—of dividing a set of possible states of the world into two parts: the ones that are ruled out by the truth of the proposition, and the ones that are not.

Thus, formal semantics allows us to understand the notion of a **proposition** as a characteristic function—it's the characteristic function for the set of worlds that are the way the proposition says. Alternatively, since every characteristic function corresponds to a set, we may now think of a proposition as a **set of possible worlds**.

6. Enriching our Semantic Theory

With the notion of an intension—*qua* function from worlds to extensions—now in place, we're going to enrich our semantic theory. Basically, the way in which we're going to do this is by **relativizing** the interpretation function $[\![\cdot]\!]$ to a possible world. Going forward, we make this relativization explicit by writing '$[\![\cdot]\!]^w$'. Think of this as saying 'at w... blah'.

For example, here are some new semantic entries, with the relativization to w:[2]

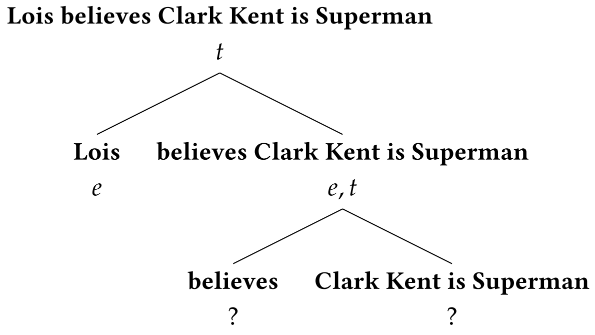
- $[\![\textbf{smokes}]\!]^w = [\lambda x.x \text{ smokes at } w]$
- $[\![\textbf{loves }]\!]^w = [\lambda x.[\lambda y.y \text{ loves } x \text{ at } w]]$
- $[\![\textbf{Fred smokes}]\!]^w = 1$ iff Fred smokes at w

But what about the word **believes** that concerned us at the outset?

[2] There's an important exception to this in the case of names. In particular, we consider names, like **Fred**, to pick out the same entities *at all possible worlds*. Thus:
- $[\![\textbf{Fred}]\!]^w = \text{Fred}$.

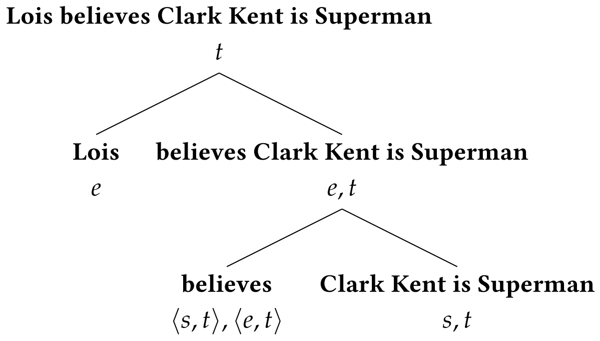
We don't need to add 'at w' to names.

7. Returning to the Example To answer this question, let's start by taking another look at the tree we looked at before:


Lois believes Clark Kent is Superman
t
Lois
e
believes Clark Kent is Superman
e, t
believes
?
Clark Kent is Superman
?

The semantic types listed all seem fine. But what about the types of **believes** and **Clark Kent is Superman**? Well, we know the former has to be a function outputting things of type e, t as its values. And we know that, *in this context*, the latter can't merely be something of type t, since, if that were the case, we'd run into the same problems we had before. So what is it?

The answer, of course, is that, in this context, **Clark Kent is Superman** must be of type s, t and thus **believes** must have type $\langle s, t\rangle, \langle e, t\rangle$. In other words, **believes** is a function which takes in, *not truth values*, but *propositions*, as its arguments. So the tree becomes the following:


Lois believes Clark Kent is Superman
t
Lois
e
believes Clark Kent is Superman
e, t
believes
$\langle s, t\rangle, \langle e, t\rangle$
Clark Kent is Superman
s, t

Given this new tree, we can now give a (rough-and-ready) semantic entry for **believes**:[3]

- $[\![\textbf{believes}]\!]^w = [\lambda p \in D_{\langle s,t\rangle}.[\lambda x \in D.$ at all the worlds w' compatible with what x believes at w, $p(w') = 1]]$.

So, the semantic value of **believes** (at a possible world w) is a function which takes in a proposition, and maps this to another function, which takes in an entity x—the *believer*—and outputs the value 1 just in case x believes that proposition.

We're now almost in a position where we can compute the truth conditions for (sentences like) **Lois believes Clark Kent is Superman**. But we need to introduce a new compositional rule, analogous to our original rule of Functional Application, to do this:

> **Intensional Functional Application (IFA)**. If α is a branching node and β, γ are its daughters, then, for any possible world w, if $[\![\beta]\!]^w$ is a function whose domain contains $\lambda w'.([\![\gamma]\!]^{w'})$, then $[\![\alpha]\!]^w = [\![\beta]\!]^w(\lambda w'.([\![\gamma]\!]^{w'}))$.

Here, then, is (part of) the derivation:

$$\begin{aligned} & [\![\textbf{believes Clark Kent is Superman}]\!]^w \\ = & [\![\textbf{believes}]\!]^w(\lambda w'.[\![\textbf{Clark Kent is Superman}]\!]^{w'}) \\ = & [\![\textbf{believes}]\!]^w(\lambda w'.\text{Clark Kent is Superman in } w') \end{aligned}$$

$$\begin{aligned} & [\![\textbf{Lois believes Clark Kent is Superman}]\!]^w \\ = & [\![\textbf{believes Clark Kent is Superman}]\!]^w)(\text{Lois}) \\ = & \ 1 \text{ iff Lois believes Clark Kent is Superman at } w \end{aligned}$$

[3] This is still a little rough. In particular, once we get to modal logic, we'll introduce the notion of an **accessibility relation** between worlds. And once we have that, we could get rid of all the English on the right-hand side of the equality. But we don't have that yet, so we'll stick with the English gloss.

Appendix. Answers to Challenge Questions

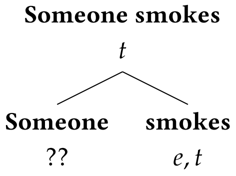
Challenge Question 1. Our basic semantic values are the following:

- $\llbracket$**Jack**$\rrbracket = \text{Jack}$
- $\llbracket$**Jill**$\rrbracket = \text{Jill}$
- $\llbracket$**loves**$\rrbracket = [\lambda x.[\lambda y.y \text{ loves } x]]$.

Now we can use these semantic values to derive the truth conditions of **Jack loves Jill**:

$$\begin{aligned}\llbracket\textbf{loves Jill}\rrbracket &= [\lambda x.[\lambda y.y \text{ loves } x](\text{Jill}) \\ &= [\lambda y.y \text{ loves Jill}] \\ \llbracket\textbf{Jack loves Jill}\rrbracket &= [\lambda y.y \text{ loves Jill}](\text{Jack}) \\ &= 1 \text{ iff Jack loves Jill}\end{aligned}$$

Challenge Question 2. Our next task is to derive the truth conditions for **Someone smokes**. To do this, however, we first need to figure out the semantic value of **Someone**. It's best, here, to start with a tree:



From the diagram, we can see that **Someone** is going to be a function that takes in things of type $\langle e,t\rangle$, and maps them to things of type t. Thus: the type of **Someone** is $\langle\langle e,t\rangle, t\rangle$ What, then, is it's semantic value? It's the following:

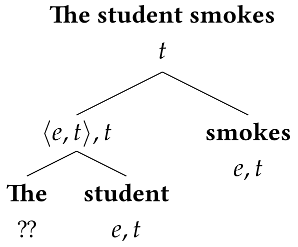
$$\llbracket\textbf{Someone}\rrbracket = [\lambda f.(\exists x)f(x) = 1].$$

Think of this as saying: The semantic value of **someone** is a function, which takes in *other* functions, f, and maps them to the value 1, just in case there exists an argument x, which we can plug into f, such that f returns the value 1.

Now we can derive the truth-conditions for **Someone smokes**:

$$\begin{aligned}\llbracket\textbf{Someone smokes}\rrbracket &= [\lambda f.(\exists x)f(x) = 1](\llbracket\textbf{smokes}\rrbracket) \\ &= 1 \text{ iff there exists an entity that smokes}\end{aligned}$$

Challenge Question 3. Now let's try **The student smokes**. We'll use the same methodology we used in the last question.



From the diagram, it's clear that **Some** is going to have type $\langle e,t\rangle, \langle\langle e,t\rangle, t\rangle$. What about it's semantic value? I'll list all the basic semantic values below:

- $\llbracket$**smokes**$\rrbracket = [\lambda x.x \text{ smokes}]$
- $\llbracket$**student**$\rrbracket = [\lambda x.x \text{ is a student}]$
- $\llbracket$**some**$\rrbracket = [\lambda f.[\lambda g.(\exists x)f(x) = g(x) = 1]]$

(Why is this the right semantic value for **some**? Compare it that of **someone**!) Now we can compute the other semantic values:

$$\begin{aligned}\llbracket\textbf{some student}\rrbracket &= [\lambda f.[\lambda g.(\exists x)f(x) = g(x) = 1]](\llbracket\textbf{student}\rrbracket) \\ &= [\lambda g.(\exists x)x \text{ is a student and } g(x) = 1] \\ \llbracket\textbf{Some student smokes}\rrbracket &= [\lambda g.(\exists x)f(x) = g(x) = 1](\llbracket\textbf{smokes}\rrbracket) \\ &= 1 \text{ iff there exists a student who smokes}\end{aligned}$$

PHIL 210/510: Formal Methods — Yale, Spring 2025
Lecture 8: The Semantics of Conditionals — Dr. Calum McNamara

1. Intensional Semantics Completed

Last time, we met a problem for our Fregean, **extensional** semantic theory—namely, it seemed to give bogus results in cases involving words like **believes**. To solve this problem, we began introducing the notion of an **intension**. For the purposes of this course, you can think of an intension as a **function from possible worlds to truth values**. And these, we also said, can be thought of as **propositions**.

Introducing intensions meant we had to enrich our type hierarchy, however:

- e is a type,
- t is a type,
- if σ and τ are types, then so is $\langle \sigma, \tau \rangle$,
- if τ is a type, then so is $\langle s, \tau \rangle$.

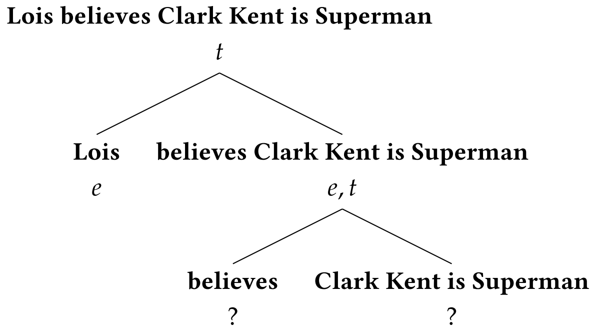
(Again, when it comes to the last clause, we'll focus exclusively on the type $\langle s, t \rangle$.)

Now, with the notion of an intension—*qua* function from worlds to extensions—now in place, we need next to enrich our semantic theory. Basically, the way in which we're going to do this is by **relativizing** the interpretation function $\llbracket \cdot \rrbracket$ to a possible world. Going forward, we make this relativization explicit by writing '$\llbracket \cdot \rrbracket^w$'. Think of this as saying 'at w... blah'.

For example, here are some new semantic entries, with the relativization to w:[1]

- $\llbracket$**smokes**$\rrbracket^w = [\lambda x. x$ smokes at $w]$
- $\llbracket$**loves** $\rrbracket^w = [\lambda x.[\lambda y. y$ loves x at $w]]$
- $\llbracket$**Fred smokes**$\rrbracket^w = 1$ iff Fred smokes at w

But what about the word **believes** that concerned us at the outset? To answer this question, let's start by taking another look at the tree we looked at before:


Lois believes Clark Kent is Superman
t
Lois
e
believes Clark Kent is Superman
e, t
believes
?
Clark Kent is Superman
?

The semantic types listed all seem fine. But what about the types of **believes** and **Clark Kent is Superman**? Well, we know the former has to be a function outputting things of type e, t as its values. And we know that, *in this context*, the latter can't merely be something of type t, since, if that were the case, we'd run into the same problems we had before. So what is it?

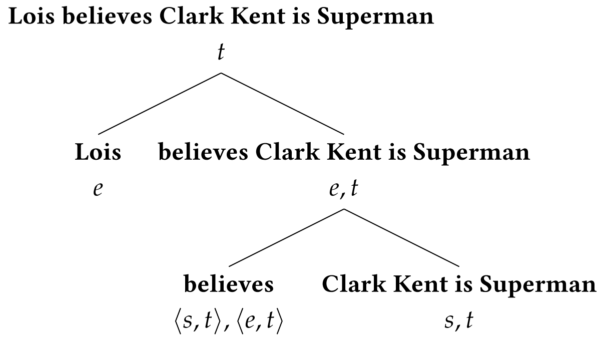
The answer, of course, is that, in this context, **Clark Kent is Superman** must be of type s, t and thus **believes** must have type $\langle s, t \rangle, \langle e, t \rangle$. In other words, **believes** is a function which takes in, *not truth values*, but *propositions*, as its arguments. So the tree becomes the following:

Given this new tree, we can now give a (rough-and-ready) semantic entry for **believes**:[2]

[1] There's an important exception to this in the case of names. In particular, we consider names, like **Fred**, to pick out the same entities *at all possible worlds*. Thus:
- $\llbracket$**Fred**$\rrbracket^w =$ Fred.

We don't need to add 'at w' to names.

[2] This is still a little rough. In particular, once we get to modal logic, we'll introduce the notion of an **accessibility relation** between worlds. And once we have that, we could get rid of all the English on the right-hand side of the equality. But we don't have that yet, so we'll stick with the English gloss.


Lois believes Clark Kent is Superman
t
Lois
e
believes Clark Kent is Superman
e, t
believes
⟨s, t⟩, ⟨e, t⟩
Clark Kent is Superman
s, t

- $[\![\textbf{believes}]\!]^w = [\lambda p \in D_{\langle s,t\rangle}.[\lambda x \in D.$ at all the worlds w' compatible with what x believes at w, $p(w') = 1]]$.

So, the semantic value of **believes** (at a possible world w) is a function which takes in a proposition, and maps this to another function, which takes in an entity x—the *believer*—and outputs the value 1 just in case x believes that proposition.

We're now almost in a position where we can compute the truth conditions for (sentences like) **Lois believes Clark Kent is Superman**. But we need to introduce a new compositional rule, analogous to our original rule of Functional Application, to do this:

> **Intensional Functional Application (IFA)**. If α is a branching node and β, γ are its daughters, then, for any possible world w, if $[\![\beta]\!]^w$ is a function whose domain contains $\lambda w'.([\![\gamma]\!]^{w'})$, then $[\![\alpha]\!]^w = [\![\beta]\!]^w(\lambda w'.([\![\gamma]\!]^{w'}))$.

Here, then, is (part of) the derivation:

$$\begin{aligned} & [\![\textbf{believes Clark Kent is Superman}]\!]^w \\ = \; & [\![\textbf{believes}]\!]^w(\lambda w'.[\![\textbf{Clark Kent is Superman}]\!]^{w'}) \\ = \; & [\![\textbf{believes}]\!]^w(\lambda w'.\text{Clark Kent is Superman in } w') \end{aligned}$$

$$\begin{aligned} & [\![\textbf{Lois believes Clark Kent is Superman}]\!]^w \\ = \; & [\![\textbf{believes Clark Kent is Superman}]\!]^w)(\text{Lois}) \\ = \; & 1 \text{ iff Lois believes Clark Kent is Superman at } w \end{aligned}$$

2. Conditionals: Why Not a Truth-functional Analysis?

Let's now turn to an entirely different issue. In this lecture, we're going to be interested in the truth-conditions of **conditionals**—statements involving 'if'.

The meaning of 'if'—and conditional statements more generally—is one of the longest-standing problems in philosophy (and semantics). Indeed, the ancient Greek poet Calimachus famously said about this problem "Even the crows on the rooftops are cawing about the meaning of conditionals".

But wait! You may wonder: 'Hasn't this problem been solved? We know the meaning of statements like 'If A, then B' from ordinary propositional logic!' Recall that in propositional logic, we define 'If A, then B'—written $A \supset B$—using the following truth table:

A	B	$A \supset B$
1	1	1
1	0	0
0	1	1
0	0	1

Thus, according to the truth table, 'If A, then B' is false only when the antecedent is true, or the consequent is false. But now here's a question: Does this truth table do a good job of capturing the way we use 'if' in ordinary

English?

Among semanticists and philosophers alike, the answer to this question is a resounding 'No'. While the first two lines of the truth table seem okay, the idea that a conditional is automatically true whenever its antecedent is false seems to licsense bizarre reasoning.

For example, consider that the material conditional $A \supset B$ is true whenever the antecedent, A, is false. In general, however, it doesn't seem right to say that we can know that a conditional is true, just because we know that its antecedent is false. For example, it doesn't seem right that just because I know 'It's raining' is false (at the present moment), I thereby know the truth of 'If it's raining, then the Mets will win the world series'.

Additionally, if 'If A then B' means $A \supset B$, then a negated conditional is equivalent to the following conjunction: $A \wedge \neg B$. But with that in mind, consider the sentence 'If Patch is a rabbit, then she's a rodent.' That sounds true. But the negation of that sentence doesn't sound at all equivalent to the following 'Patch is a rabbit and she's not a rodent.'

More generally, it doesn't seem like *any* possible truth-functional analysis of the conditional works. (To see why, play around the changing the last two lines of the truth table. If you change the last line from 1 to 0, for example, then $A \supset B$ is equivalent to $A \wedge B$. But that clearly isn't right!)

3. Another Problem: Indicatives vs. Subjunctives

There's another problem with the "material conditional" analysis of conditionals, however. To see it, consider the following sentences:

(1) If Shakespeare didn't write Hamlet, then someone else did.

(2) If Shakespeare hadn't written Hamlet, then someone else would have.

Here, the first sentence seems to be true, but the second seems to be false. It's not clear, however, how we can distinguish this, if 'If A, then B' just means $A \supset B$'. (In particular, notice that the A and B in this case would have to be the same 'hadn't written Shakespeare' doesn't make sense on its own—the past-tense morphology seems only to make sense when embedded within the conditional context.)

Statements like (1) are often referred to as **indicative conditionals**, while statements like (2) are referred to as **subjunctive conditionals** or **counterfactuals**.[3] It's widely agreed that natural language conditionals come in these two kinds—but there's only one "kind" of conditional in propositional logic, namely, the material conditional.

4. A Better Theory—the Variably Strict Theory

To get around the problems just mentioned, a number of alternative accounts of 'if' statements have been proposed. For example, an early attempt was to say that 'If A, then B' in English really means something more like '*Necessarily*, if A, then B'. In other words: 'There's no possible world in which A is true, and B is false'. This theory is sometimes called the **strict conditional** theory of the conditional. I don't like it much.

Another theory was put forward by Angelika Kratzer, based on some remarks from David Lewis. Kratzer's idea was that 'if' acts as a kind of quantifier: to say 'If A' is to quantify over a domain of worlds—namely, worlds at which A is true. It follows (on Kratzer's view) that 'If A, then B' is true, just in case every world in this restricted domain is a B-world. This is called the **restrictor** theory. I like it more than the strict conditional theory—but I still have problems with it.

For the purposes of this course, then, we're going to focus on a theory due to Robert Stalnaker (1968) and David Lewis (1979). It's usually known as the **variably strict theory**.

To motivate it, let's start with some famous remarks from the philosopher Frank Ramsey (1929):

> "[i]f two people are arguing 'If A will B?' and are both in doubt as to A, they are adding A hypothetically to their stock of knowledge and are arguing on that basis about B.

[3] We'll use both terms in this course.

This is sometimes known as the *Ramsey test* for conditionals—they idea is, when we consider 'If A, then B', we first *suppose* the antecedent, A, and then ask whether B.

The problem with this, however, is that Ramsey's remarks are about *belief*, but what we want is a set of *truth conditions* for conditionals. So can we translate Ramsey's idea from one involving "belief conditions" to truth conditions? Robert Stalnaker (1968) provides a very compelling answer:

> How do we make the transition from belief conditions to truth conditions[?]... The concept of a *possible world* is just what we need to make this transition, since a possible world is the ontological analogue of a stock of hypothetical beliefs. The following set of truth conditions, using this notion, is a first approximation of the account that I shall propose: Consider a possible world in which A is true *and which otherwise differs minimally from the actual world. 'If A, then B' is then true (false) just in case B is true (false) at that possible world.*

More generally, then, Stalnaker's idea is that 'If A, then B' is true at a possible world w just in case B is true at the A-world that's "minimally different" to w. In the jargon, we say that this A-world is the **closest** or **most similar** A-world to w.

Can we make this more precise, however? We can—and we'll do so a little bit today, but also a little bit in our section on logic, later in the course. For today, all we'll do is start by introducing a function f, which Stalnaker calls a **selection function**.

Formally, a selection function $f : Pow(W) \times W \to W$ is a function which takes a proposition A, and a world w, and maps these to an A-world—namely, the closest A-world to w. (Recall that we're thinking of propositions as sets of possible worlds—or, equivalently, as functions from worlds to truth-values. Thus, if W is the set of all possible worlds, then the power set of W, $Pow(W)$, is the set of all propositions.)

5. What makes a world 'closest'?

David Lewis (1979) gave a very similar semantics for conditionals to Stalnaker. (One major difference is that Lewis thinks there can be multiple "equally close" A-worlds to w, while Stalnaker denies this. Later on in the course, we'll speak more about this idea.) Unlike Stalnaker, however, Lewis gives a pretty elaborate account of what we mean by 'closest A-world'.

As he says, 'closest' here can't mean anything like our pre-theoretic notion of 'closest'. After all, consider the following issue, raised by Kit Fine:

> The counterfactual 'If Nixon had pressed the button there would have been a nuclear holocaust' is true or can be imagined to be so. Now suppose that there never will be a nuclear holocaust. Then that counterfactual is, on Lewis's analysis, very likely false. For given any world in which antecedent and consequent are both true it will be easy to imagine a closer world in which the antecedent is true and the consequent false. For we need only imagine a change that prevents the holocaust but that does not require such a great divergence from reality.

Fine's point here is that a world in which there's all out nuclear world is very *dis*similar (viz., not very close), in an intuitive sense, to the actual world. But of course, the counterfactual 'If Nixon had pressed the button there would have been a nuclear holocaust' seems intuitively true.

Faced with this sort of issue, Lewis proposes the following account of what should make a world count as 'closest':

(i) It should match the actual world up until a time shortly before A,

(ii) It should match the actual world's laws of nature.

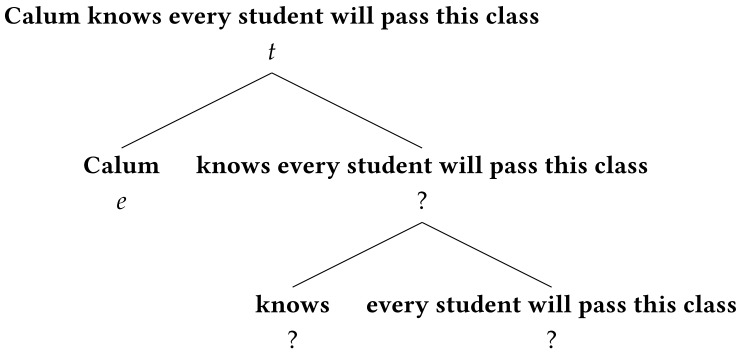
If the laws of nature are deterministic, however, then no A-world can satisfy perfectly (assuming the actual world is not an A-world).

Lewis thus says that in this case, we should allow that the closest A-world(s) is (are) ones in which a *miracle* occurs: a small violation of the actual world's laws, sufficient to bring A about.

PHIL 210/510: Formal Methods — Yale, Spring 2025
Lecture 8: Modal Logic I — Dr. Calum McNamara

1. Intensional Semantics—Another Computation

To segue into our new topic, let's start with another exercise in intensional semantics. Consider:


Calum knows every student will pass this class
t
Calum
e
knows every student will pass this class
?
knows
?
every student will pass this class
?

We're going to derive the truth conditions for this sentence. But to do so, let's start by replacing the question marks with the relevant types.

Challenge Question. Replace the question marks with the correct semantic types.

Next, we need to figure out the semantic value for **knows**.

- $[\![\textbf{knows}]\!]^w = ?$

Challenge Question. Give an entry for the semantic value of **knows**.

Finally:

Challenge Question. Use the semantic value for **knows** to compute the truth conditions for **Calum knows every student will pass this class**.

2. Modal Logic—What?

If you succeeded in giving a semantic value for **knows**, you'll have seen that, in order to specify this semantic value, you had to talk about relations that hold *between possible worlds*. To see what I mean more clearly, consider the following sentence: 'It's possible that Jones smokes'. Clearly, this sentence could be true, even if Jones doesn't in fact smoke. After all, it doesn't seem like it would violate any natural or metaphysical laws if Jones was a smoker—which is why we say that it's *possible* that she could smoke, even if she doesn't in fact smoke. Saying this, however, requires us to say something about things are like *in another possible world*, from the perspective of the actual world. For it to be *possible* for Jones to smoke at the actual world, it must be that she *does* in fact smoke in another possible world.

As a final example, consider again the Stalnaker-Lewis semantics for conditionals:

Challenge Question. Give the Stalnaker-Lewis "viarably strict" semantics for sentences like 'If A, then B'.

In a sense, **modal logic** is the study of relations between possible worlds, which we need to make sense of these ideas. That's going to be our topic for the next couple of weeks.

Broadly speaking, modal logic—at least of the kind we'll be studying here—is an extension of standard propositional logic, one which allows us to reason about statements and arguments involving words like 'necessary' and 'possible', but also 'knows', 'ought', and so on. It's a powerful tool for philosophers to have in their toolkit. For example, it's used all over the place in metaphysics, epistemology—even in ethics. Modal logic, in its modern form, was initiated by Saul Kripke in two mid-twentieth century papers (Kripke, 1959, 1963).[1]

3. Flavors of Modality

Very roughly, a **modal** is a word that qualifies the truth of some statement. For example, consider again the statement 'It's possible that Jones smokes'. Notice again that this statement could be true, even if Jones doesn't *in fact* smoke.

[1] Amazingly, the first of these papers was published while Kripke was still a teenager; and the second was published only a year after he completed his undergraduate degree at Harvard. For a nice overview of Kripke's ideas, see Ahmed 2007.

Likewise, imagine I say 'It's necessary that Jones smokes'. If this sentence were true, it would say something stronger than merely 'Jones smokes'—it would say, in addition, something like: she couldn't have failed to smoke.

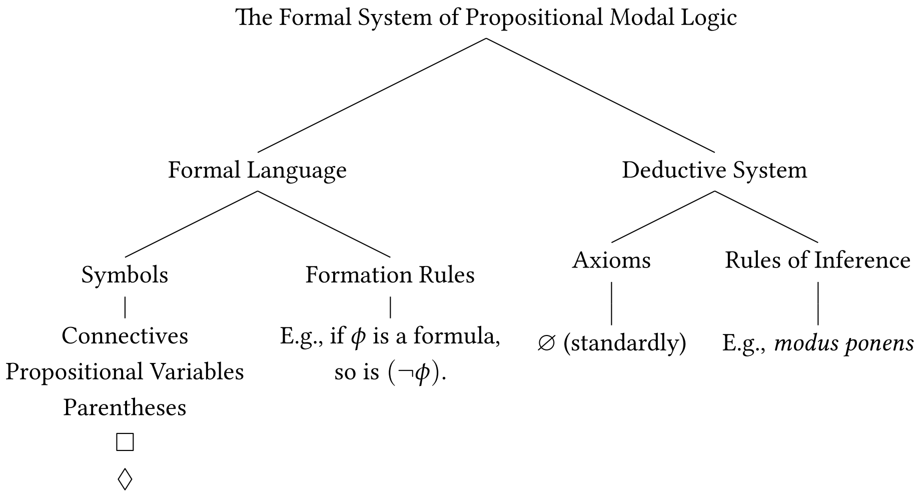
Modals like these come in many different "flavors". For instance, there is the **metaphysical** flavor of modality, which has to do (roughly) with possibility and necessity in the broadest sense. There is also the **deontic** modality, which has to do with the notions of obligation and permission. (Examples: 'You *may* have a slice of cake'; 'You *ought* to keep your promises'. Incidentally, this is the kind of modality Guillermo was talking about in his presentation.) And then there is the **epistemic** modality, which has to do with knowledge and consistency with one's evidence. ('Jones *might* be in her office'.) There are many other flavors of modality besides these (e.g., there is also the **nomic** flavor, which has to do with what's possible or necessary *according to the laws of nature*). But the flavors just mentioned have been important in philosophy, historically speaking, and make for a good selection for us to start with. So they'll be the ones on which we'll focus, for the most part, in this course.

4. The Language of Propositional Modal Logic

Modal logic studies the behavior of modals (like those we encountered in the previous section) in logical contexts. To do this, we need to start by developing a *formal language.* Recall the formal language of propositional logic, that we encountered earlier in the course. In the present case, we enrich this language with two new symbols, $\Box$ and $\Diamond$ (pronounced "box" and "diamond", respectively), which stand in place for various modals. Harking back to the diagram we looked at earlier in the course, this enrichment gives us something like the formal system on the next page.

More precisely, in propositional modal logic, we take the **formal language** to consist of the following symbols:

- An infinite stock of **atomic sentences**, p, q, r, etc. (with and without numerical subscripts);
- The **truth function connectives**, $\neg, \wedge, \vee, \supset$;[2]



- The symbols $\Box$ and $\Diamond$;
- Parentheses: $($, $)$.

We also introduce the following **formation rules**. These rules tell us which are the "gramatically correct" strings of the foregoing symbols. We call the gramatically correct strings **well-formed formulas** (or just **wff**s):

(i) Every atomic sentence, p, q, r is a wff;

(ii) If ϕ is a wff, then so is $\neg\phi$;

(iii) If ϕ and ψ are wffs, then so are $(\phi \wedge \psi)$, $(\phi \vee \psi)$, and $(\phi \supset \psi)$;

(iv) If ϕ is a wff, then so are $\Box\phi$ and $\Diamond\phi$;

(v) Nothing else is a wff.

Challenge Question. Why do we need clause (v)?

Just as we could have defined certain of the propositional connectives in terms

[2] I refuse to include the symbol for 'if and only if'. If you ever need to use it, you can introduce it as a shorthand for $(\phi \supset \psi) \wedge (\psi \supset \phi)$.

of other connectives (e.g., we can define '$\supset$' in terms of '$\neg$' and '$\vee$'),[3] we could also have defined $\Diamond$ in terms of $\Box$, or vice versa.

Challenge Question. Show how we can define $\Box$ using $\Diamond$ and the other truth-functional connectives. (Hint: think of $\Box$ as saying 'Necessarily' and $\Diamond$ as saying 'possibly'. Another hint: think about how we can define $\forall$ using $\exists$ and the truth-functional connectives.)

If you managed to crack the previous question, then you'll have seen that $\Box$ and $\diamond$ act like quantifiers. In particular, they act like special kinds of quantifiers with a very specific domain: they quantity, not over things we denote with names, but over possible worlds.[4] After all, if we think of $\Box$ as pertaining to metaphysical necessity, then $\Box\phi$ says (very roughly!—see below) that ϕ is true in all possible worlds. Similarly, if we interpret $\Diamond$ as pertaining to metaphysical possibility, then $\Diamond\phi$ says (roughly) that ϕ is true in at least one possible world. This isn't quite right, as we'll say in a moment. But it's close enough to get us started.

Note, however, that the particular interpretation we give to $\Box$ and $\Diamond$ depends upon the flavor of modality we're studying. For example, if it's epistemic modality rather than metaphysical modality that we're interested in, then we can plausibly think of $\Box\phi$ as saying 'ϕ is known', and $\Diamond\phi$ as 'ϕ is consistent with one's evidence.' Thus, the modal operators, $\Box$ and $\Diamond$ are *context-sensitive*. As we'll see later, this fact plays an important part in the theory of modal frames and modal models—to which we now turn.

5. Modal Frames and Modal Models

Recall the truth-tables from propositional logic. These tell us how to assign truth-values to well-formed formulas (wffs) based on the truth-values given to the atomic formulas, like p, q, etc. For example, in the last class, we saw the truth table for the material conditional:

A	B	$A \supset B$
1	1	1
1	0	0
0	1	1
0	0	1

In this section, we're going to introduce an analogous notion for modal logic. This is called a **modal model**. (Good news for those of you took Phil 115 with me: modal models are easier to get your head around than the models we study in first-order predicate logic!)

Formally, a modal model is a triple $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, \mathcal{I} \rangle$, where $\mathcal{W}$ is a set of possible worlds, $\mathcal{R}$ is a relation between the worlds in $\mathcal{W}$, and $\mathcal{I}$ is a function mapping atomic sentences (p, q, r, etc.) to truth-values—namely, their particular truth-values at the possible worlds in $\mathcal{W}$. Some of these ideas might take some getting used to. Let's start off by unpacking the various definitions a bit.

The first element in the set, $\mathcal{W}$, is something you're already familiar with from the previous chapter. Sometimes this set is called the **universe** of the model. I'll occasionally use that terminology in the sequel.

The second element, the relation $\mathcal{R}$, by contrast, is known as the **accessibility relation**. It is a *binary* relation between pairs of worlds (so, it's a subset of the Cartesian product of $\mathcal{W} \times \mathcal{W}$), and helps to characterize the particular "flavor" of modality that we're working with. When some world w_1 is $\mathcal{R}$-related to another world w_2, I will sometimes speak metaphorically, and say that w_1 "sees" w_2. Watch out for that lingo—when you see it, you should read it non-metaphorically as "w_1 is $\mathcal{R}$-related to w_2". Usually, we'll write '$w_1 \mathcal{R} w_2$' when w_1 sees w_2 (and similarly for other possible worlds).

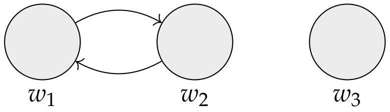
Finally, the **interpretation function**, $\mathcal{I}$ tells us, as I've already said, which atomic sentences are true at which worlds, and which aren't, respectively. For example, it will tell us whether a given atomic formula p is true at w_1; whether it's true at w_2; and so on. (It won't, however, tell us whether, e.g., $(p \wedge q)$ is true at w_1—more on that in a moment.) Thus, truth, in modal logic, is a **world-bound** notion. Formulas are true (or false) *at possible worlds*.[5]

[3] See note 3 in chapter 1.

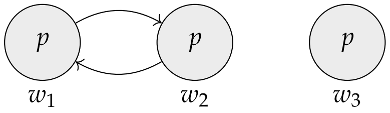
[4] So far as I am ware, this interpretation of $\Box$ and $\diamond$ is due to David Lewis.

[5] If you think about this, especially with the notion of propositions as set of possible worlds in your mind, then this makes sense. Recall from the previous chapter that we

The set of worlds $\mathcal{W}$ and the accessibility relation $\mathcal{R}$ together form what we call the **frame** of the model. For example, suppose $\mathcal{W} = \{w_1, w_2, w_3\}$. And suppose that $\mathcal{R} = \{w_1 \mathcal{R} w_2, w_2 \mathcal{R} w_3\}$. (That is, the accessibility relation relates w_1 to w_2 and w_2 to w_3. I'll continue to use this notation in what follows.) Then, diagramtically, the frame $\langle \mathcal{W}, \mathcal{R} \rangle$, looks something like this:



Now suppose that we let the interpretation function maps the atomic sentence p to 1 (i.e., true) at each world, w_1, w_2, and w_3. Then we'd get the beginnings of a modal model:



In order to specify a modal model completely, however, we need to specify the truth-value of *all* propositions at the worlds in the frame $\mathcal{M}$. This said, in practice we'll usually only specify the truth-values of a few propositions, so you need not worry about this being an arduous task.

Now, in addition to the interpretation function $\mathcal{I}$, we also have something called a *valuation function* $\mathcal{V}$.[6] This function takes the model, and then maps the "complex formulas" ($(p \wedge q)$, etc.), including modal formulas, to their truth values at worlds, based on the the set of worlds, accessibility relation, and truth value of each atomic sentence at each world. (Again, let me stress the point: in modal logic, formulas are true *at possible worlds.*) For example, consider the fact that world w_1 in the above Figure "sees" only one world, w_2. And at that world, the atomic proposition p is true. It follows from this that, because p is true at every world that w_1 sees, the valuation function $\mathcal{V}$ maps $\Box p$ to 1 at the world w_1. Thus, paraphrasing $\Box$ as 'necessarily', we get that p is necessarily true *at world* w_1 since every world accessible from w_1 is p-world.

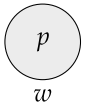
A bit more formally, here is how our valuation function assigns truth values to sentences of the our language:

- If p is an atomic sentence, then $\mathcal{V}(p, w_i) = \mathcal{I}(p, w_i)$.
- $\mathcal{V}(\neg\phi, w_i) = 1$ iff $\mathcal{V}(\phi, w_i) = 0$.
- $\mathcal{V}((\phi \wedge \psi), w_i) = 1$ iff $V(\phi, w_i) = 1$ and $V(\psi, w_i) = 1$.
- $\mathcal{V}((\phi \vee \psi), w_i) = 0$ iff $V(\phi, w_i) = 0$ and $V(\psi, w_i) = 0$.
- $\mathcal{V}((\phi \supset \psi), w_i) = 0$ iff $V(\phi, w_i) = 1$ and $V(\psi, w_i) = 0$.
- $\mathcal{V}(\Box\phi, w_i) = 1$ iff, for all w_j such that $w_i \mathcal{R} w_j$, $\mathcal{V}_{\mathcal{M}}(\phi, w_j) = 1$.
- $\mathcal{V}(\Diamond\phi, w_i) = 1$ iff, for some w_j such that $w_i \mathcal{R} w_j$, $\mathcal{V}_{\mathcal{M}}(\phi, w_j) = 1$.

That is: $\Box\phi$ is true at world w_i (in model $\mathcal{M}$ according to the valuation function $\mathcal{V}_{\mathcal{M}}$) just in case ϕ is true at all the worlds w_j such that w_i sees w_j. The same thing goes for $\Diamond\phi$, except we replace 'all worlds' with 'some world'.

We will say more about the valuation function $\mathcal{V}$ in due course. For now, however, let's do a little practice.

Challenge Question. Is the formula $\Box p$ true at w_2 in the above figure? Is $\Box p$ true at the world w below? How about $\Diamond p$?



said (on one view) propositions are sets of possible worlds. If a particular formula p is true at w_1, then, what that means is that w_1 is a member of the set p.

[6] Another way to write this: $[\![\cdot]\!]$. Thus, our valuation function is none other than the denotation function from our study of intensional semantics!

PHIL 210/510: Formal Methods — Yale, Spring 2025
Lecture 8: Modal Logic I — Dr. Calum McNamara

1. Modal Frames and Modal Models

Last time, we were introduced to **modal frames** and **modal models**. Formally, a frame is a pair, $\langle \mathcal{W}, \mathcal{R} \rangle$, where $\mathcal{W}$ is a set of possible worlds, and $\mathcal{R}$ is an **accessibility relation**—a binary relation defined on the set $\mathcal{W}$. A model adds to this an **interpretation function**, $\mathcal{I}$, which specifies the truth-value of each atomic sentence, p, q, etc., at each world in the model.

Once we have a modal model in place, we can specify the truth-values for various sentences of our formal language—including modal sentences. These are given by the **valuation function**, $\mathcal{V}$, which extends the interpretation function $\mathcal{I}$:

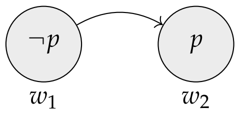
- If p is an atomic sentence, then $\mathcal{V}(p, w_i) = \mathcal{I}(p, w_i)$.
- $\mathcal{V}(\neg\phi, w_i) = 1$ iff $\mathcal{V}(\phi, w_i) = 0$.
- $\mathcal{V}((\phi \wedge \psi), w_i) = 1$ iff $V(\phi, w_i) = 1$ and $V(\psi, w_i) = 1$.
- $\mathcal{V}((\phi \vee \psi), w_i) = 0$ iff $V(\phi, w_i) = 0$ and $V(\psi, w_i) = 0$.
- $\mathcal{V}((\phi \supset \psi), w_i) = 0$ iff $V(\phi, w_i) = 1$ and $V(\psi, w_i) = 0$.
- $\mathcal{V}(\Box\phi, w_i) = 1$ iff, for all w_j such that $w_i \mathcal{R} w_j$, $\mathcal{V}_{\mathcal{M}}(\phi, w_j) = 1$.
- $\mathcal{V}(\Diamond\phi, w_i) = 1$ iff, for some w_j such that $w_i \mathcal{R} w_j$, $\mathcal{V}_{\mathcal{M}}(\phi, w_j) = 1$.

You'll notice here that the truth-values of modal sentences, like $\Box\phi$ and $\Diamond\phi$ depend, not just on what's true at the given world w, but also on what's true at worlds *related* to w. This is the key role played by the accessibility relation $\mathcal{R}$.

But you might wonder: why do we need the accessibility relation? Instead of having, e.g., $\Box\varphi$ be true at w just in case all the worlds that are $\mathcal{R}$-related to w are ϕ-worlds, why can't we just say the following: $\mathcal{V}(\Box\phi, w) = 1$ iff, for all worlds $w' \in \mathcal{W}$, $\mathcal{V}(\phi, w') = 1$.

The answer is that we want the apparatus of modal logic to allow us to study *all* kinds of modals, not just specific ones. And the accessibility relation $\mathcal{R}$ allows us to characterize the flavor of modality we're interested in.

To see what I mean, consider an example. Suppose we interpret $\Box$ as saying 'It is known that'. Now consider the following modal model:



Challenge Question. Is the formula $\Box p$ true at w_1? If it is, why is that weird?

If you managed to answer the previous question, then you might agree that, when it comes to the interpretation of $\Box$ as 'It is known that', we shouldn't allow that a sentence like $\Box\phi$ can be true at a world, without ϕ itself being true at that world. After all, in the jargon, knowledge is a **factive attitude**.

Contrast that, however, with the interpretation of $\Box$ as 'It is morally obligatory that'. Now above model doesn't look so weird—it's often the case that something isn't morally obligatory, without us actually doing it. So, it might be that $\Box\phi$ is true, even if ϕ isn't.

As we're going to see this session—and the next one, and the one after—we use different accessibility relations when we're interpreting $\Box$ (and $\Diamond$) in different ways. This is really the heart of modal logic.

2. Validity in a Model

First, however, let's take a bit of a detour.

Recall that in introductory propositional logic, you learned about a special class of wffs called **tautologies**. In more advanced logic books, these are sometimes called **valid** formulas. A valid formula, remember, is one which is true no matter what truth-values its atomic formulas are assigned. (That's putting it roughly, of course; but the rough parse should do to jog your memory.) For example, $(p \vee \neg p)$ is a valid formula. This can be seen by looking at its truth-table:

p	$(p \vee \neg p)$
1	1
0	1

As you can see, if p is assigned the truth-value 1, then $(p \vee \neg p)$ gets assigned truth-value 1. Similarly, if p gets assigned the truth-value 0, then $(p \vee \neg p)$ still gets assigned truth-value 1. No matter what, then, $(p \vee \neg p)$ gets assigned truth-value 1, which means that it is a tautology (valid formula).

Now, in propositional modal logic, we have a corresponding notion of a valid formula. Here is the (informal) definition:

> **Validity in a Modal Model**. If ϕ is formula of modal propositional logic, then ϕ is **valid in modal model** $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, \mathcal{I} \rangle$ iff, for every $w \in \mathcal{W}$, ϕ is true at w.

So a formula is valid in a modal model iff it's true at *every world in the model.*

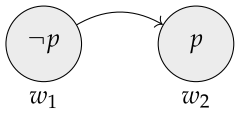
Notice how this is an even more demanding notion than a proposition's being necessary. In modal propositional logic, recall, a proposition is necessary at a world w iff it's true at all the worlds that w *sees.* This, plainly, is a weaker than validity.

Now, it's an interesting fact that, in modal logic, we can sometimes tell whether a formula is valid in lots of different models just by figuring out whether it's valid in one model in particular. The reason for this is that, sometimes, models share certain properties in virtue of having an accessibility relation, $\mathcal{R}$, that itself satisfies certain constraints. For example, the following formula is valid in all models for which the accessibility relation is transitive: $\Box p \supset \Box\Box p$. It will thus be useful for us to classify models based on their accessibility relations. Let's start doing this now.

3. Classes of Models: K, T, S4, S5

To get the ball rolling, let's consider a slight variant of our earlier example. Suppose we're interested in the wff $\Box p \supset p$. If you read this as "If It is known that p, then p", then it seems strange to say that a modal logic could deny this proposition's truth—as we heard before, knowledge is factive. So how could a proposition be known, without being true?

As we saw, however, the system of modal logic we've set up so far allows that this proposition could be false. Once more, here's a model which makes that so:

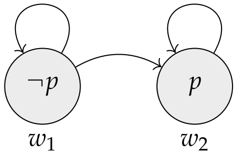


As you'll see, $\Box p$ is true at the world w_1, since every world that w_1 sees (namely w_2) is a world at which p is true. However, the formula p is false at w_1, since $\neg p$ is true there. Thus, as you'll remember from propositional logic, a material conditional $\phi \supset \psi$ is false only in the case in which its antecedent is true and its consequent is false. So in this case, it follows that $\Box p \supset p$ is false at w_1.

A question naturally arises: What conditions do our models have to satisfy in order for $\Box p \supset p$ to turn out valid (i.e., true at every world in the model)? This is the kind of question we'll be interested in in this section.

A quick note before we get started, however. We're going to assume that *all* our models are so-called *K-models.* (Here, 'K' is for 'Kripke'—who, as I noted in a footnote on the last handout, gave the semantics for modal logic that we're currently working with.) At the moment, you can think of the K-models as those for which the accessibility relations can be anything whatsoever. This isn't exactly accurate; as we'll see when we come to the study of characteristic axioms, assuming our models are K-models is a non-trivial assumption, since the K-models are characterized by certain substantive axioms. But in practice, most of the logics modal logicians are interested in are K-models anyway, so we can put further discussion of what this means to the side for the moment.

Let's now think again about the wff we considered above: $\Box p \supset p$. By inspecting the figure above, it should be reasonably clear that the formula would be valid if we altered the model as follows:



Now $\Box p \supset p$ is *true* at all the worlds in the model. Take world w_1 first. Since w_1 now sees itself, as well as w_2, it sees a world at which p is false (namely w_1). So the antecedent of $\Box p \supset p$ isn't true, which (by standard propositional logic) means the whole conditional *is* true. Similarly, the only world w_2 see is itself—a world at which p is true. Thus it follows that $\Box p$ is true at w_2. And since p is true at w_2, it then follows that the conditional $\Box p \supset p$ is true at w_2. In short, then: $\Box p \supset p$ is valid in the model above.

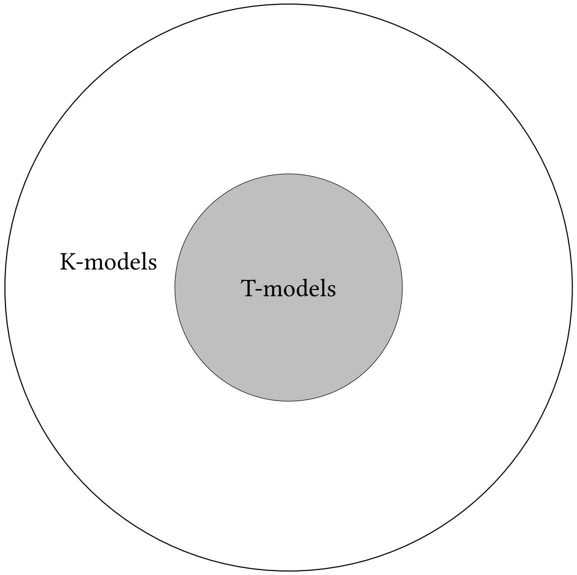
What we did to change the model to make the accessibility relation $\mathcal{R}$ **reflexive**. Recall that a reflexive relation is a relation such that, for every x in the domain of the relation, we have $x\mathcal{R}x$. More carefully, in the special case of accessibility relations:

> **Reflexive Accessibility Relation**. An accessibility relation $\mathcal{R}$ on a set of worlds $\mathcal{W}$ is *reflexive* iff, for all $w \in \mathcal{W}$, $w\mathcal{R}w$.

The relation in the present case is thus reflexive since, for every world w in the model above, w sees itself—i.e., wRw holds.

We call the class of models with reflexive frames the **T-models**. ('T' is for 'truth'. The reason for this name will become clear later, though we should also note that some of the names for classes of models we'll be interested in here are adopted for historical reasons, and so don't always have nice heuristic names like this one. Unfortunately, you'll just have to memorize them.) As it turns out, every model with a reflexive frame—i.e., every T-model—is one in which the wff $\Box p \supset p$ is valid. More generally, every T-model is one in which instances of the following schema are valid: $\Box \phi \supset \phi$.

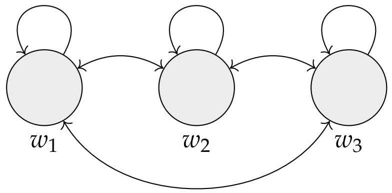
An important thing to note about the relationship between K-models and T-models is that the T-models are a *subset* of the K-models. In other words, every T-model is a K-model (since *all* the models we're interested in here are K-models), but the reverse isn't true: some K-models aren't T-models, because the T-models all have reflexive frames and not all the K-models do. The relationship can be represented pictorially as follows:



As we'll see, the relationships between different classes of models often follow this pattern. For example, some models have reflexive *and* **transitive** frames, which means they're a subset of the T-models. Recall that a transitive (accessibility) relation is defined like this:

> **Transitive Accessibility Relation**. An accessibility relation $\mathcal{R}$ on a set of worlds $\mathcal{W}$ is *transitive* iff, for all $w_i, w_j, w_k \in \mathcal{W}$, if $w_i\mathcal{R}w_j$ and $w_j\mathcal{R}w_k$, then $w_i\mathcal{R}w_k$.

In fact, let's now consider the class of such models. Here is a pictorial representation of the frame of one such model in particular:



As you can see, this frame is both reflexive and transitive. Every world w_i, for $i = 1, 2, 3$, is such that $w_i \mathcal{R} w_i$ holds. Moreover, the frame is transitive since, for every i, j, k, if we have $w_i \mathcal{R} w_j$ and $w_j \mathcal{R} w_k$, then we also have $w_i \mathcal{R} w_k$.

The class of models for which the frames are both reflexive and transitive are called the **S4-models**. An important schema whose instances hold in this class of models is the following: $\Box\phi \supset \Box\Box\phi$. Instances of this schema do not always hold in T-models, however.

As a final example of an important class of models, consider the class whose frames are reflexive, transitive, *and* symmetric. A **symmetric** accessibility relation is defined like this:

> **Symmetric Accessibility Relation**. An accessibility relation $\mathcal{R}$ on a set of worlds $\mathcal{W}$ is *symmetric* iff, for all $w_i, w_j \in \mathcal{W}$, if $w_i \mathcal{R} w_j$, then $w_j \mathcal{R} w_i$.

An example of the frame of such a model is, believe it or not, given in the figure immediately above. I sneakily made the accessibility relation in this frame symmetric, as well as reflexive and transitive, by making all the arrows between worlds double-headed. Models whose frames are reflexive, symmetric, and transitive are called **S5-models**.

S5-models are important since they are usually considered to be the class of models corresponding to *metaphysical modality*, one of the "flavors" of modality that we considered in the last class. In fact, the different flavors of modality can all be claimed to have their own associated class of models, characterized by a particular accessibility relation.

Challenge Question. We already decided that, when $\Box$ is interpreted as 'It is known that', the corresponding accessibility relation should be reflexive. What about symmetry and transitivity? Do you think either of these should hold when it comes to this interpretation of $\Box$?

Challenge Question. Same question, but now interpret $\Box$ as 'It is morally obligatory that'.

4. Bonus! Partitions[1]

A relation on a set that's reflexive, symmetric, and transitive is called **equivalence relation**. Momentarily, let's denote such a relation by $\sim$. So, if S is a set, $\sim$ is a relation $\sim \subseteq S \times S$ that's reflexive, symmetric, and transitive.

Now, choose some arbitrary element $x \in S$. The set of all elements y in S that $x \sim y$ is called the **equivalence class** of x. More carefully, the equivalence class of x—usually written '$[x]$'—is the set:

$$[x] = \{y \in S : x \sim y\}.$$

The set of all equivalence classes of S is called the **quotient set** of S.

You'll notice that every element $x \in S$ is a member of the equivalence class $[x]$. And two equivalence classes $[x]$ and $[y]$ are either equal or disjoint. (Why?) Therefore, the set of all equivalence classes of S forms a **partition** of S—where a partition, you'll recall, is a set of subsets of S that are mutually exclusive and jointly exhaustive.

This is cool. It says that, really, equivalences relations and partitions are more-or-less the same kind of thing. Moreover, given that metaphysical modality is usually thought of as requiring an accessibility relation that's reflexive, symmetric, and transitive, you can think of metaphysical modality as imposing a *partition* on the set of all possible worlds.

[1] This section is just for the mathematically inclined.

PHIL 210/510: Formal Methods — Yale, Spring 2025
Lecture 10: Modal Logic I — Dr. Calum McNamara

1. Practice with Validity Proofs

Last time, we were introduced to the notion of a *valid formula.* Recall that a formula ϕ is **valid in a model** $\mathcal{M}$ iff ϕ is true at every world in $\mathcal{M}$.

Likewise, we heard that a formula ϕ is **valid in a class of models** iff ϕ is valid in every model in that class.

We also looked at ways in which we can constrain our models—namely, by implementing constraints on the accessibility relation, $\mathcal{R}$. For example, demanding that the accessibility relation be **reflexive** gives us a class of models called **T-models**. Likewise, if we demand that the accessibility relation be both reflexive and **symmetric**, then we get a class of models called **S4-models**. And so on. As we heard, moreover, it's these different constraints on accessibility relations that help us characterize different "flavors" of modality. In the case in which we're interpreting $\Box$ as 'It is known that', for instance, it seems like we want the class of models to be (at least) **T-models**.

Finally, I claimed that certain formulas of modal logic come out as valid in certain classes of models, but not in others. For example, the following formula is (I claim) valid in all **T-models**—in any model with a reflexive accessibility relation:

$$\Box\varphi \supset \varphi \tag{T}$$

Formulas like this are important, because—as we'll see in a moment—they help to characterize different **modal logics**. What we're going to do for the first part of today, then, is *prove* claims like the one above. Let's start with this claim.

Claim 1. *Let $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, \mathcal{I} \rangle$ be a modal model. Then, if the accessibility relation $\mathcal{R}$ is reflexive, it follows that all instances of the schema $\Box\phi \supset \phi$ are valid.*

Let's now prove this claim together.

Proof. To make this proof crystal clear, I'll work through the steps in more detail than is strictly necessary. To start off, notice that we are trying to prove a conditional claim—a claim of the form "If such-and-such, then so-and-so". As we heard in Lecture 3, when one is trying to prove a claim like this, it's usually a good idea to begin by *assuming* the antecedent of the conditional one is trying to prove.[1] So let's do that: suppose that the relation $\mathcal{R}$ in the modal model $\mathcal{M}$ is reflexive. Then it follows, by the definition of a reflexive relation, that each world $w \in \mathcal{W}$ "sees" itself: $w\mathcal{R}w$.

Now consider the schema $\Box\phi \supset \phi$. Our goal is to show that, given the hypothesis that $\mathcal{R}$ is reflexive, it follows that $\Box\phi \supset \phi$ is valid. Now, in order for an instance of this schema to be *false* at arbitrary world w in $\mathcal{W}$, we know that its antecedent needs to be true and its consequent needs to be false at that world. (That follows from the truth-table for $\supset$.) If that's *not* the case, then the instance of $\Box\phi \supset \phi$ is true at w. We're going to show that such an instance *can't*, in fact, be false at w if the accessibility relation $\mathcal{R}$ is reflexive.

We'll do this by using another technique which we discussed in Lecture 3—namely, proof by contradiction. That is, we'll assume that, even though $\mathcal{R}$ is reflexive, an instance of the schema $\Box\phi \supset \phi$ *is* false at w. We'll then show that this assumptions leads to a contradiction, which lets us reject the assumption and thus conclude that any instance of $\Box\phi \supset \phi$ must be true at w. Furthermore, since our choice of the world w was arbitrary, we'll be able to conclude that any instance of $\Box\phi \supset \phi$ is true at *any* world in $\mathcal{W}$—i.e., that such an instance is valid in the modal model $\mathcal{M}$. Finally, since our choice of the model $\mathcal{M}$ itself was arbitrary, except for our assumption that its accessibility relation is reflexive, we'll be able to conclude that any such modal model will make all instances of the schema $\Box\phi \supset \phi$ valid.

So that's the plan. Let us now execute it. To start, let w be an arbitrary world

[1] As an analogy, think of $\supset$-Introduction rule that one often uses in natural deduction-style proofs in propositional logic. The rule works as follows. Suppose you're trying to prove the claim $p \supset q$. Then, one starts off, in a sub-proof, by *assuming* that the antecedent, p, is true. One then reasons one's way to the consequent, q. And then, having done that, one can "discharge" the assumption, closing the sub-proof and concluding simpliciter that $p \supset q$ is true.

in $\mathcal{W}$, and suppose that (some instance of) $\Box\phi \supset \phi$ is false at w.[2] By the truth-table for '$\supset$', that means that $\Box\phi$ is true at w, but ϕ is false at w. By the definition of '$\Box$', we know that $\Box\phi$ is true at w iff all the worlds that w sees are worlds at which ϕ is true. And since $\mathcal{R}$ is reflexive (by assumption), we know that w sees itself. But from this it follows that ϕ is true at w. And now that contradicts our assumption that $\Box\phi \supset \phi$ is false at w, since (as we said a moment ago) this implies that ϕ is false at w.

As we hoped, then, we have arrived at a contradiction. So $\Box\phi \supset \phi$ must, after all, be true at w. And since our choice of w was arbitrary, it follows that $\Box\phi \supset \phi$ is true at each world w in $\mathcal{W}$. Moreover, since all we assumed about the model $\mathcal{M}$ was that its accessibility relation was reflexive, it follows that $\Box\phi \supset \phi$ is valid in all models with such relations. In other words, *if* a modal model has a reflexive accessibility relation, then it follows that all instances of $\Box\phi \supset \phi$ are valid, which is what we were trying to show. □

As I said, the above proof includes much more detail than is necessary. But if you're seeing proofs like this for the very first time, it can help to spell out your reasoning in great detail, like I did above, to make each step clear.

Now, let's look at another important formula of modal logic:

$$\Box\phi \supset \Box\Box\phi \tag{4}$$

Last time, I said that this formula is true in any model with a transitive accessibility relation. We're now going to prove that, too. Here we go:

Claim 2. *Let $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, \mathcal{I} \rangle$ be a modal model. Then, if the accessibility relation $\mathcal{R}$ is transitive, it follows that all instances of the schema $\Box\phi \supset \Box\Box\phi$ are valid.*

This time, rather than giving you all the gory details, I'll prove you with a kind of template, so you can fill in the details for yourself.

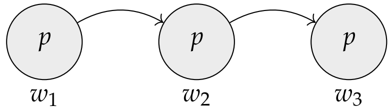
Proof. To start off, you should assume (as we did in the previous proof) that $\mathcal{R}$ is a transitive relation, since we're trying to prove a conditional claim. Next, it's worth recalling what it means for a relation $\mathcal{R}$ to be transitive. Once you've done that, proceed as we did before. Choose an arbitrary world w in $\mathcal{W}$, and assume that some instance of $\Box\phi \supset \Box\Box\phi$ is false at w. Then, use the truth table for $\supset$ to spell out what it takes for such a conditional to be false at w. Show that this assumption leads to a contradiction. This will allow you to reject the assumption that the instance of $\Box\phi \supset \Box\Box\phi$ is false at w after all. You can then proceed via the same steps we did in the previous proof: since w was arbitrary, the instance of $\Box\phi \supset \Box\Box\phi$ is true at all worlds in $\mathcal{W}$, i.e., such instances are valid; and since the choice of model itself was arbitrary, apart from the assumption the model had a transitive frame, you can conclude that all instances of $\Box\phi \supset \Box\Box\phi$ are valid in models with transitive frames. From this you can conclude the conditional you were trying to prove is true. □

The proofs above made use two techniques, which it's useful to have in your toolkit, namely: (i) assuming the antecedent of a conditional, if a conditional is what you're trying to prove; and (ii) proof by contradiction. You don't *have* to use these two techniques in your proofs, of course—many other techniques are available. But in the case of proof by contradiction in particular: if anything is going to work to establish the claim you're trying to prove, then *this* is!

One more thing before we move on. Consider Claim 1 again. That claim, as we already noted, is a conditional: it says that *if* a model has a reflexive accessibility relation, *then* all instances of $\Box\phi \supset \phi$ are valid in that model. Now, you might be tempted to conclude that the reverse of this is true as well, i.e., that, if $\Box\phi \supset \phi$ is valid in a model, then its accessibility relation is reflexive. But unfortunately, it isn't. To see this, consider a diagram we've already encountered:

In this model, $\Box p \supset p$ is valid. (Can you say why?) But it isn't the case that the frame is reflexive. Thus, we have a counterexample to the conditional claim if

[2] I'll sometimes speak of the schema $\Box\phi \supset \phi$ as being true at a world, or valid in a model, etc. But you should bear in mind that this is strictly incorrect. $\Box\phi \supset \phi$ can't be true, or valid, because it isn't even a formula. Rather, its *instances* are formulas. But again, I'll continue to speak informally of the schema as being true, valid, etc., in what follows.



$\Box\phi \supset \phi$ is valid in a model, then its accessibility relation is reflexive. This fact, and others like it, are bearing in mind throughout our study of modal logic.

2. The Logics K, T, S4, and S5

Above, I referred to classes of models with names like 'T-models', 'S4 models', and so on, where these names were given by features of the associated accessibility relation. Each of these classes of models has an associated logic. What do I mean by this?

Recall the way we originally set up the formal system of propositional logic, way back in the second lecture, and also the system of modal propositional logic at the start of this chapter. In the case of (standard) propositional logic in particular, I was careful to say that one way we could set up the system was to have no axioms but lots of inference rules; but another way was to have a small set of axioms, together with only one inference rule (*modus ponens*). If we opted for the latter, then one set of axioms we could use is the following:

- $(\phi \supset (\psi \supset \phi))$,
- $(\phi \supset (\psi \supset \chi)) \supset ((\phi \supset \psi) \supset (\phi \supset \chi))$,
- $((\neg\phi \supset \neg\psi) \supset (\psi \supset \phi))$.

For the moment, let us take the latter view: let us suppose, that is, that we've set up the deductive system of propositional logic using the above axioms, rather than an extensive list of rules of inference and an empty set of axioms. As we saw, in the case of modal logic, we extend the language of propositional logic with two new symbols, $\Box$ and $\Diamond$, to get the language of modal propositional logic. But how we should we extend the deductive system to account for the new formulas we can form using these two new symbols?

Well, to start with, we can extend the list of axioms a bit. The most conservative extension of the above three axioms for modal propositional logic is the following:

- $(\phi \supset (\psi \supset \phi))$,
- $(\phi \supset (\psi \supset \chi)) \supset ((\phi \supset \psi) \supset (\phi \supset \chi))$,
- $((\neg\phi \supset \neg\psi) \supset (\psi \supset \phi))$.
- $\Box(\phi \supset \psi) \supset (\Box\phi \supset \Box\psi)$

The axiom we've added to the list here is called the **K-axiom** (again, after 'Kripke'; as you can probably tell, Kripke made quite the impact on modal logic). In order to do anything useful with this axiom, however, we need an associated rule of inference to go with it. Alongside the rule *modus ponens*, then, let us add the following rule:

Necessitation. From ϕ, infer $\Box\phi$.

At a first pass, this rule looks wrong. After all, couldn't (an instance of) ϕ be true (at a world) without being *necessarily* true? That seems plausible. But consider how Sider, 2010 justifies necessitation's legitimacy:

> [S]o long as we're careful how we use our axiomatic system, [necessitation] won't get us into trouble.... In a proof [in modal propositional logic], each line must be either i) an axiom or ii) a wff that follows from earlier lines in the proof by a rule; in a proof from [a set of formulas] a line may also be iii) a member of [that set of formulas] (i.e., a "premise"). A theorem is defined as the last line of any proof. So every line of every proof is a theorem. So whenever one uses necessitation in a proof—a proof simpliciter, that is—one is applying it to a theorem. And necessitation does seem appropriate when applied to theorems: if ϕ is a theorem, then $\Box\phi$ ought also to be a theorem. Think of it another way. The worry about necessitation is that it doesn't preserve truth: its premise can be true when its conclusion is false. But necessitation does preserve *logical truth*. So if we're thinking of our axiomatic definition of theoremhood as being a (proof-theoretic) way to represent logical truth, there seems to be no trouble with its use of necessitation. (205)

Hopefully, that clears up any doubts you might have about necessitation.

Once we've added the K-axiom and the rule of necessitation to our formal system, we get a very weak modal logic which is (unsurprisingly, at this point) called **K**. (Can you guess why the logic is called this? Also, as an aside, note that this is why our assumption that our models were K-models was non-trivial—because there is a close relationship between the class of models, K, and the logic K. And K includes a new axiom and rule of inference.) The K-axiom is the so-called **characteristic axiom** of the modal logic K.

Claim 3. Show that the K-axiom is valid in the class of all K-models.

Adding further axioms to our list gets us stronger modal logics. For example, if we add $\Box\phi \supset \phi$ to our list of axioms, we get a modal logic called **T**. This schema $\Box\phi \supset \phi$ is called the **T-axiom**, and is the characteristic axiom of the logic T.

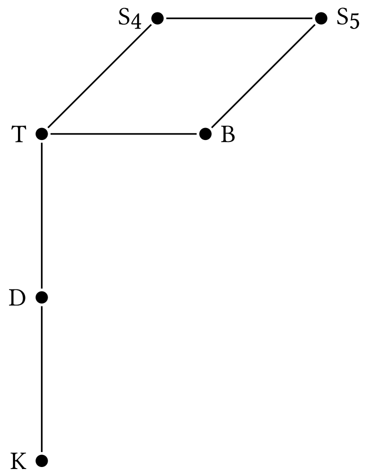
Similarly, a logic which includes the T-axiom and the following characteristic axiom, $\Box\phi \supset \Box\Box\phi$, gets us the logic known as **S4**. (That axiom is called the **4-axiom**, by the way.)

Now the pattern starts to look relatively clear. Earlier, we considered an additional class of models—the S5 models. This was the class of models for which the accessibility relation was an equivalence relation, i.e., reflexive, symmetric, and transitive. Does S5 then have its own characteristic axiom? The answer is 'Yes'. It's the following, which we call the **5-axiom**: $\Diamond\phi \supset \Box\Diamond\phi$. In words, this might be read: "If it's possible that ϕ, then it's necessarily possible that ϕ". A logic which includes the K-axioms, the T-axiom, the 4-axiom, and the 5-axiom, is called the modal logic **S5**. Arguably, it's the most important of the modal logics that we'll be studying.

As you can see, then, the study of modal logic doesn't consist just in studying a single logical system—like in the case of propositional logic or (first-order) predicate logic. Rather, it consists in the study of a rich *family* of logics, each characterized by their own axioms and corresponding classes of models.

Moreover, the family we've been introduced to so far—which includes the logics K, T, S4, and S5—isn't the end of it. Next class, we'll see that there are additional logics—with names like **B** and **D**—that have their own characteristic axioms and classes of models. Some of these classes of models, moreover, require very interesting constraints on the accessibility relation—constraints with names like **Euclideanness** and **Seriality**. We'll see that next class.

But in the meantime, since this handout is shorter than usual, here's a pretty picture, which should help you to visualize the relationships between different kinds of models.

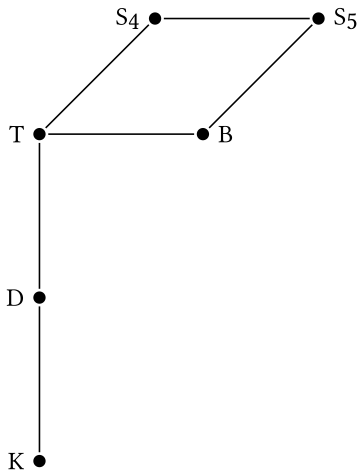


(Please appreciate how long this picture took to code.)

PHIL 210/510: Formal Methods | Yale, Spring 2025
Lecture 10: Modal Logic I | Dr. Calum McNamara

1. Further Systems of Modal Logic: B and D

Last time, I included the following pretty picture on the hand-out, showing the connections between some of the systems of modal logic we've looked at so far in this class.



The systems we've studied so far have been: **K**, **T**, **S4**, and **S5**. You'll recall that K is the system of modal logic we get by adding the following axiom to our list of axioms of propositional logic, $\Box(\phi \supset \psi) \supset (\Box\phi \supset \Box\psi)$, alongside the rule of inference **Necessitation**: If ϕ, infer $\Box\phi$. Similarly, T is the system we get when we add the T-axiom to our system: $\Box\phi \supset \phi$. S4 is what we get when add the 4 axiom to *that*: $\Box\phi \supset \Box\Box\phi$. And finally, S5 is what we get when we add the 5 axiom to T: $\Diamond\phi \supset \Box\Diamond\phi$.

Today, we're going to think about two more systems of modal logic, which have been historically important. The first is a system called **B** and the second is a system called **D**.

Starting with the former: D is the system that results when we add the following characteristic axiom—the **D axiom**—to the axioms of the logic K:

$$\Box\phi \supset \Diamond\phi.$$

Secondly, the modal logic we get by adding the following axiom—the **B-axiom**—to the axioms of the logic T gets us a modal logic called B:

$$\phi \supset \Box\Diamond\phi.$$

It's worth pausing to think about what these axioms *mean*. Start with the D-axiom: reading $\Box$ as 'necessarily' and $\Diamond$ as 'possibly', as we've been doing, $\Box\phi \supset \Diamond\phi$ seems to say: "If ϕ is necessary, then it's possible". That sounds plausible. But recall that, at the very start of this chapter, we said that there could be alternative readings of $\Box$ and $\Diamond$. Here's an alternative reading: let $\Box$ denote (moral) obligation and $\Diamond$ (moral) permission. Then $\Box\phi \supset \Diamond\phi$ seems to say "If ϕ is obligatory, then ϕ is permitted".[1] That also seems extremely plausible.

In fact, the modal logic D—which, as we said, is characterized by the D-axiom—is often thought to be the quintessential *deontic logic* (whence it gets its name). Given the reading of the D-axiom in terms of obligation and permission, that seems to make sense.

What about the logic B? Does it characterize a particular flavor of modality? Not so obviously. But here's one attempt. Historically speaking, the modal logic B is named after the great Dutch mathematician Brouwer, who founded what's called **intuitionistic logic.** Brouwer was among those mathematicians who thought mathematics should be *constructive*—that is (roughly), that mathematical *truth* is inseparable from mathematical *proof.* To put this plainly: take a statement like the continuum hypothesis, which we discussed earlier in the course. As we heard, this hypothesis *cannot* be proved from the standard ax-

[1] Note that this is *not* the same thing as Kant's famous dictum "ought implies can". That confusion is often made with regards to the D-axiom. But in fact, ought implies can mixes two different flavors of modality: something like the deontic flavor, on the one hand, and the metaphysical or nomological flavor, on the other.

ioms of set theory. Still, you might think—indeed, I *do* think—it nevertheless has a truth-value. Just because we can't *prove* it, doesn't mean it isn't true (or false, for that matter). Brouwer, however, would deny this. On his view—and the view of the intuitionists more generally), statements which haven't yet been proved are strictly speaking neither true nor false. They are, in other words, indeterminate. (Notice that this view involves denying the famous *law of excluded middle*: for any ϕ, $\phi \vee \neg\phi$).

Thus, one way to read the B-axiom—influenced by Brouwer—is as pertaining to the notion provability. As an extremely rough parse, you might be able to read the B-axiom as: "If ϕ is true, then ϕ necessarily has a proof". That's worth dwelling on a bit.

In any case, one interesting thing about the B axiom is that, if we add *it* to the logic S4, then we get back the modal logic S5. So, another way we can axiomatize S5 is by adding by 4 and B to the logic T, rather than the 5 axiom. Indeed, it's a good exercise to prove this. (Problem set?)

Challenge Question. Show that any model which validates K, T, 4 and B also validates 5.

2. Frame Conditions

The foregoing might make you wonder. We know, at this point, that the T axiom is valid in the class of all models whose frames are reflexive. Similarly, we know that the 4 axiom is valid in the class of all models whose frames are transitive. In a sense, then, the T-axiom *characterizes* the models with reflexive frames, just as the 4 axiom characterizes models with transitive frames. So: do the D axiom and B axiom likewise clarify classes of models? That is, are the D-axioms and B-axioms valid in all models whose frames have certain properties?

Challenge Question. Which class of models does the B axiom characterize? (Hint: think about the logic S5.) If you get it, *show* that B is valid in the class of models with the requisite kind of frame.

What about the D-axiom? What kind of frame does it characterize? As it turns out, D is valid in all models whose frames are **serial**, i.e., have a serial accessibility relation. A serial accessibility relation is the following:

> **Serial Accessibility Relation**. For all worlds $\mathcal{W}$, there exists a world w' (possibly equal to w) such that $w\mathcal{R}w'$ holds.

In a slogan, then, we can think of seriality like this as: for each world w, there is some world that w *sees*. Alternatively: there are no "blind" worlds.

This condition also helps us to see why the logic D sits above K, but below T, in the diagram from the first page: Why's that? Because the T axiom also requires that there be no blind worlds. But in a sense, it requires a special case of this. In particular, it requires that every world see *itself*.

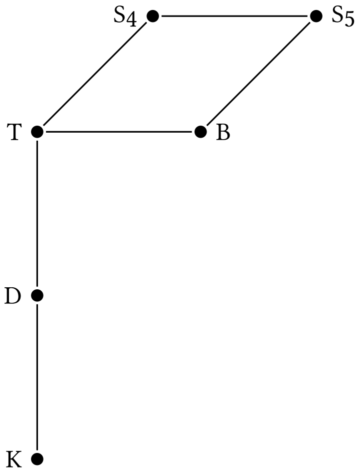


Figure1: Relationships between different modal logics

On the next problem set, you'll be introduced to yet further logics, and you'll be asked to show that their characteristic axioms are valid in certain classes of models. It'll be fun! Good luck!

3. Soundness and Completeness

We are now going to very briefly talk about the **soundness** and **complete-**

ness of the systems of modal logic that we've been considering. First, however, what does it *mean* for a logic to be sound? Complete?

Here are very rough characterizations of the foregoing notions (more formal characterizations follow shortly). Recall the valuation function, $\mathcal{V}$, which we discussed extensively before. As we said there, $\mathcal{V}$ assigns a truth-value to every formula in a model, based on things like the accessibility relation in the model, and the truth-values that the interpretation function, $\mathcal{I}$, assigns to the atomic formulas. Now consider some formula ϕ. Above, we said that such a formula is valid in a model just in case $\mathcal{V}(\phi) = 1$ at all worlds in the model.

There is a more general notion of validity in modal logic even than this (indeed, we touched on this previously). In particular, we say that a formula ϕ is *valid* (*simpliciter*)[2] if it is valid in *every* modal model. (We can also relativize this notion to the various sytems of modal propositional logic that we've studied so far. For example, a formula is K-valid if it is valid in every K-model.) In other words, if, no matter what accessibility relation or interpretation function, etc., we choose, $\mathcal{V}(\phi) = 1$, then ϕ is a valid formula (in modal propositional logic). We write this $\vDash \phi$.[3]

A related notion is that of a formula ϕ being a *theorem* of the system of logic we're interested in. Recall the various axioms and deductive rules that we've considered. (The latter are *modus ponens* and necessitation.) A *proof* of ϕ (in a system of modal propositional logic) is a finite list of formulas, each line of which either (i) is an axiom or (ii) follows from earlier wffs in the proof by one of our rules of inference; and (iii) the last line of the proof is ϕ. If ϕ can be derived in this way, we say that it is a *theorem* of the system of logic in question. (In fact, in a proof, every line can be considered a theorem.) When this is the case, we write $\vdash \phi$.

Soundness and completeness state relationships between validity and theoremhood. In particular, soundness says the following. Let S be a system of modal propositional logic (e.g., K, T, S4, etc.). Then we have:

> **S-Soundness**. If $\vdash_S \phi$, then $\vDash_S \phi$.

A very rough, heuristic way to think about this is the following: If there is a *proof* of ϕ in the system S, then ϕ is also a valid formula. In other words, our axioms and rules of inference don't allow us to prove invalid formulas. In a slogan: no false positives.

Completeness for the system S is the converse:

> **S-Completeness**. If $\vDash_S \phi$, then $\vdash_S \phi$.

Again, very roughly, if ϕ is a valid formula (in system S), then we can prove that this is so, given our axioms and rules of inference. Our axioms and rules of inference, in other words, don't allow any valid formulas to "slip through the net".

It is good news that *all* of the systems of modal logic that we've looked at—K, T, S4, S5, D, and B—are both sound and complete. Believe it or not, however, not every formal system has these two properties. (Higher-order logic, as I've mentioned—in which one is allowed to quantify, not only over objects, but also over properties, operators, etc.—is not complete; there are valid formulas for which no proof of theoremhood exists.) This is a pretty striking claim. If you find it interesting, you might want to do a bit of reading about Kurt Gödel's well-known incompleteness theorems. Note that, importantly—indeed, this was one of the main upshots of Gödel's results—is that the formal system we use in the case of ordinary arithmetic is incomplete.

Here, however, we won't consider those. Instead, we'll briefly sketch the proof of soundness for the system K. Proving soundness is more straightforward than you might imagine. Once we've proved soundness for K, it's a fairly straightforward exercise to prove soundness for the other systems of modal propositional logic that we've considered. Proving completeness, however, is far more demanding, and so we do not attempt to do that here. If you're interested in the completeness proofs for any of the systems of modal propositional logic that we've considered, then you might want to consult Sider 2010, or better yet, Cresswell and Hughes 1996.

[2] Really, we should say something like 'MPL-valid' where MPL stands for 'modal propositional logic'. But I'll ignore that bit of verbiage here.

[3] Strictly, we should subscript the symbol '$\vDash$' with the system of modal logic we're concerned with, e.g., '$\vDash_K$'. Similarly for the symbol '$\vdash$' considered below.

Here, then, is what we're going to show:

Theorem 1 (Soundness of K). *If* $\vdash_K \phi$*, then* $\vDash_K \phi$.

Proof. We proceed by induction.[4] That is, we first show that the axioms of K are valid (the base case), and then we show that our rules of inference, *modus ponens* and necessitation, preserve theoremhood (the inductive step).

Base Case. First, then, recall that the axioms of K are just the axioms of standard propositional logic, plus the K-axiom:

- $\Box(\phi \supset \psi) \supset (\Box\phi \supset \Box\psi)$.

It is a fairly straightforward exercise to establish that the axioms of standard propositional logic are valid. So here we consider only the first of these axioms, and leave the others for the reader to prove as an exercise. Thus, consider the axiom schema:

- $\phi \supset (\psi \supset \phi)$.

We need to show that all its instances are valid. So, suppose (for reductio) that some instance of the axiom is not valid in an arbitrary model $\mathcal{M}$. Then, by the truth-tables, we know that ϕ is valid in $\mathcal{M}$, but $\psi \supset \phi$ is invalid in $\mathcal{M}$. Again, by the truth-tables, $\psi \supset \phi$ is invalid only if ψ is valid and ϕ is invalid in $\mathcal{M}$. But a moment ago, we said that ϕ had to be valid in $\mathcal{M}$. Contradiction! Since our choice of model was arbitrary, we thus conclude that there can be no such invalid instance of $\phi \supset (\psi \supset \phi)$.

The proof that all instances of the K-axiom are valid involves similar reasoning. Assume (again, for reductio) that some instance of $\Box(\phi \supset \psi) \supset (\Box\phi \supset \Box\psi)$ is invalid in an arbitrary model $\mathcal{M}$. By the truth-tables, this implies that $\Box(\phi \supset \psi)$ is valid and $\Box\phi \supset \Box\phi$ is invalid in $\mathcal{M}$. Since $\Box\phi \supset \Box\phi$ is invalid in $\mathcal{M}$, it must be that $\Box\phi$ is valid in $\mathcal{M}$ but $\Box\psi$ is invalid in $\mathcal{M}$. Since $\Box\psi$ is invalid in $\mathcal{M}$, there must be a world w_1 such that w_1 "sees" a world w_2 at which ψ is not true. But since w_1 sees w_2, ϕ must be true at w_2; otherwise, $\Box\phi$ would be false at w_1, contrary to what we showed a moment ago, i.e., that $\Box\phi$ is valid in $\mathcal{M}$. Now notice: w_2 is a world at which ϕ is true and ψ is false. From this it follows that $\phi \supset \psi$ is false at w_2. And since w_1 sees w_2, it follows that $\Box(\phi \supset \psi)$ is false at w_1. But this contradicts our earlier claim, that $\Box(\phi \supset \psi)$ is valid in $\mathcal{M}$. We thus conclude that the K-axiom must be valid in $\mathcal{M}$ after all. And since our choice of $\mathcal{M}$ was arbitrary, it follows that the K-axiom is valid in all models. This concludes our proof of the base case.

Inductive step. We now need to show that our two rules of inference, *modus ponens* and necessitation, preserve validity. Thus, let us assume that we have a proof whose first n lines (for $n \geq 0$) are valid wffs. We need to show that $n+1$st line is valid. There are three cases to consider: (i) the $n+1$st line is an axiom; (ii) the $n+1$st line follows from earlier lines via *modus ponens*; or (iii) the $n+1$st line follows from earlier lines via the rule of necessitation. Consider case (i) to start. If the $n+1$st line of the proof is an axiom, then it follows immediately from our proof of the base case that it is a valid wff. So turn now to the second case. If the $n+1$st line is a formula ψ derived from earlier lines by *modus ponens*, then it follows that we must have wffs ϕ and ψ on (separate) earlier lines in the proof. By assumption, ϕ and $\phi \supset \psi$ are both valid. Hence ψ is valid, since if it were not, then $\phi \supset \psi$ would not be valid, contradicting our assumption. Lastly, then, consider case (iii). The $n+1$st line of the proof is a wff of the form $\Box\phi$. Since this line follows from an earlier line by the rule of necessitation, we must have ϕ on a line earlier in the proof. Again, ϕ is assumed to be valid. By our definition of a valid formula, ϕ is true at all worlds (in every model $\mathcal{M}$). So it follows immediately that $\Box\phi$ is also valid: for any world w in the model, either (a) w sees no worlds, in which case $\Box\phi$ is vacuously true; or (b) w at least one world. And by our assumption that ϕ is valid, it follows that every world that w sees is a world at which ϕ is true. We conclude, then, that the rule of necessitation preserves validity. Moreover, we have shown more generally that K is sound, as desired. □

As I said, it's a relatively easy exercise to generalize this proof to the other systems of modal propositional logic that we've considered. And doing so is also a lot of fun. Good luck!

[4] Remember that we briefly talked about proofs by induction in Class No. 3! If you've forgotten how that works, it's worth taking another look at the hand-out from that class.

PHIL 210/510: Formal Methods | Yale, Spring 2025
Lecture 13: Conditional Logic | Dr. Calum McNamara

1. Lewis-Stalnaker Semantics—Again!

A few sesions back, we talked about the **semantics of conditionals** (in natural language). In particular, we focused on the Lewis-Stalnaker **variably strict** theory of natural language conditionals, according to which (very roughly) a conditional sentence 'If ϕ, then ψ' is true at a possible world, w, just in case ψ is true at the "closest" ϕ-world(s) to w.

This gloss of conditionals makes clear that—just like 'Necessarily ϕ', etc.—their truth-conditions depend, not just on how things are at a world, but on how things are *at other possible worlds*. Thus, conditionals in English are widely thought to be **modal sentences**.

For this reason, you might suspect that we can give a formal account of conditionals—more formal than the account we previously gave—using something like our modal models. And indeed, it turns out we can.

2. Stalnaker Models

Recall our account of a modal model, $\mathcal{M}$. This is a triple, consisting of (i) a set $\mathcal{W}$ of possible worlds; (ii) an accessibility relation $\mathcal{R}$ between the worlds in $\mathcal{W}$; and (iii) an interpretation function $\mathcal{I}$. So: $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, \mathcal{I} \rangle$.

Now, on some accounts of conditionals, models like the ones we already have are sufficient to give a semantics for English language conditionals. For example, **strict conditional theorists** say that sentences of the form 'If ϕ, then ψ' in English correspond to *necessitated* material conditionals: $\Box(\phi \supset \psi)$.[1]

However, in this course, we're focusing on the Lewis-Stalnaker semantics, which depends on relations of closeness. And our current definition of the accessibility relation $\mathcal{R}$ doesn't seem sufficient to capture this notion. After all, intuitively, we want the notion of closeness to be something like a **ternary relation**: world w is closer to w' than w''. But the accessibility relation $\mathcal{R}$ is merely a **binary relation**. So again, it doesn't look like we have the resources to capture closeness, in the relevant sense, in our present models.

One option, at this point, would be for us to simply introduce a new, ternary relation into our modal models, of the kind just described. (Indeed, this is how David Lewis (1973) proceeds.) However, we're going to come at things in a slightly different way. As we did previously, we're going to follow Stalnaker in introducing a special kind of *function* into our models. Stalnaker calls this is a **selection function**.

To spell this out, let $\mathcal{W}$ (again) be our set of possible worlds. Then, let $\mathcal{P}(\mathcal{W})$ be the set of all subsets of $\mathcal{W}$ (the power set of $\mathcal{W}$). As we did in the semantics portion of the course, we think of subsets of $\mathcal{W}$ as corresponding to **propositions**. Thus, the set $\mathcal{P}(\mathcal{W})$ is the set of all propositions.

Now, the set $\mathcal{P}(\mathcal{W}) \times \mathcal{W}$ is the **Cartesian product** of $\mathcal{P}(\mathcal{W})$ and $\mathcal{W}$. Intuitively, it's the set of all *pairs*, where the first element is a proposition, and the second element is a world. We then define a selection function as follows (below, I use ϕ, ψ, ambiguously, both for propositions, and for the sentences that express them):

$$f(\phi, w) : \mathcal{P}(\mathcal{W}) \times \mathcal{W} \to \mathcal{P}(\mathcal{W}).$$

So, a selection function is a function that maps a pair consisting of a proposition and a world, to a set of possible worlds. Intuitively, this set of possible worlds is the set of *closest* ϕ-worlds to w. (We'll precisify this idea in a moment.)

Now that we have this definition in hand, we can extend our previous models. The result we get is a kind of model known as a **Stalnaker model**: $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, f, \mathcal{I} \rangle$. This kind of model simply adds a selection function to our previous kind of model.

We can then extend our definition of the valuation function, from modal logic:

- If ϕ is an atomic formula, then $\mathcal{V}(\phi, w) = \mathcal{I}(\phi, w)$,
- $\mathcal{V}(\neg\phi, w) = 1$ iff $\mathcal{V}(\phi) = 0$,

[1] If you're taking the Keith DeRose/Tim Williamson epistemology seminar this semester: Williamson thinks this about counterfactuals.

- $\mathcal{V}(\phi \wedge \psi, w) = 1$ iff $\mathcal{V}(\phi, w) = 1$ and $\mathcal{V}(\psi, w) = 1$ (and analogously for the other truth functional connectives, $\vee$, $\supset$, etc.),
- $\mathcal{V}(\Box\phi, w) = 1$ iff $\forall w'$ such that $w\mathcal{R}w'$, $\mathcal{V}(\phi, w') = 1$ (and analogously for $\Diamond\phi$),
- $\mathcal{V}(\phi > \psi, w) = 1$ iff $f(\phi, w) \subseteq \psi$.

Note that, according to Stalnaker, the foregoing applies equally to both indicatives and subjunctives. That's why I've used the $>$ symbol in the definition of conditionals. It's a place-holder, which can be filled in by either $\rightarrow$ (the indicative conditional) or $\Box\!\!\rightarrow$ (the counterfactual conditional). Lewis thinks the foregoing semantics only works for counterfactual conditionals. In fact, he's one of the very few that think indicative conditionals are just ordinary material conditionals. (You might wonder: How, then, does Stalnaker distinguish between indicatives and subjunctives/counterfactuals? That question, it turns out, is related to the next point).

Note also that we assume all the worlds in $f(\phi, w)$ are accessible from w. In other words, all the worlds "selected" by f, when given ϕ and w as arguments, are worlds, w', such that $w\mathcal{R}w'$.

3. Constraining the Selection Function

The foregoing definition gives us a formal—if, admittedly, abstract—semantics for conditionals. However, there's an obvious issue with it. How do we know the selection function f is really going to select anything like the *closest* ϕ-worlds to w? At the moment, nothing about the definition of f guarantees that this function will respect anything like an intuitive *closeness relation*: the formal definition merely says this functions maps propositions and worlds to sets of worlds.

We heard last time about Lewis's account of closeness in terms of histories, laws of nature, and miracles. Stalnaker, however, comes at this issue from a slightly different angle. (Actually, Lewis basically agrees with him about all this—it's in his 1973 book, if you're interested.) In particular, rather than giving a "metaphysical" account of closeness between worlds, Stalnaker imposes abstract constraints on the selection function. These constraints don't allow us to pin down this relation precisely. But they do guarantee that anything that could possible function as a relation of closeness between worlds will satisfy certain properties.

Let's start with two obvious constraints:

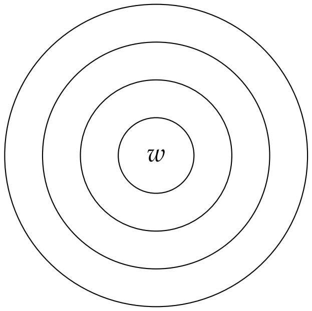
- **Success**. $f(\phi, w) \subseteq \phi$ (the closest ϕ-worlds to w should *be* ϕ-worlds).
- **Centering**. If $w \in \phi$, then $w \in f(\phi, w)$ (if w itself is a ϕ-world, then w should be among the closest ϕ-worlds to itself).

Those two constraints seem completely obvious. Indeed, it's hard to know what 'closest ϕ-world' would even mean, if f didn't satisfy them. (That said, the Centering constraint turns out to be slightly controversial. We may talk about that when we get to the section of the course on probability.)

The next constraint is less intuitive. But we need it, if we want the notion of closeness to form an *ordering*.

- **Reciprocity**.[2] If $f(\phi, w) \subseteq \psi$ and $f(\psi, w) \subseteq \phi$, then $f(\phi, w) = f(\psi, w)$.

These constraints are sufficient to ensure that relations of closeness form what Lewis calls a **system of spheres**:



Our next constraint is very important, but highly controversial. It's really the chief point of disagreement between David Lewis and Stalnaker:

[2] Reciprocity was originally referred to using the acronym 'CSO'. But no one seems to know what that actually stands for.

- **Uniqueness**. $|f(\phi, w)| \leq 1$.

Very roughly, the Uniqueness constraint says that, for any ϕ and world w, there's a *unique* closest ϕ-world to w. (If ϕ is logically inconsistent, then there are no such worlds at all. But let's set that case aside.)

Stalnaker is committed to the Uniqueness constraint. But Lewis rejects it. He says that it's metaphysically implausible to say that, for any ϕ and w, there's a *unique* closest ϕ-world to w. After all, consider this sentence:

(1) If I had flipped the coin yesterday at noon, it would have landed heads.

If Stalnaker is right, then there's a *unique* closest flip-world to actuality. And it's a heads-world or a tails-world. But then again—and this, again, is Lewis's complaint—this arguably seems strange. After all, we're assuming that the coin is fair. So what could make, say, a heads-world closer to actuality than a tails-world?

Faced with this sort of issue, Lewis rejects Stalnaker's Uniqueness constraint. However, Stalnaker himself says that we should think about things differently. In English, the notion of closeness in play when we utter conditional sentences is a *context-sensitive* notion. And usually, context doesn't pin down this notion very precisely. Rather, it makes different precisifications of 'closest ϕ-world' *admissible*.

Challenge Question. If this is right, what will Stalnaker say about the truth-value of a sentence like (1) above? Does this remind you in any way of Al Hájek's view?

One last thing. I said above that Stalnaker thinks his semantics applies to *both* indicative conditionals and counterfactuals. But you might wonder: how, then, are we supposed to distinguish between these conditionals, using this semantics? The answer is that Stalnaker (1975) says that, in the case of indicative conditionals, we require the selection function to satisfy an additional constraint:

- **Indicative Constraint**. $f(\phi, w)$ must be an epistemically possible world. Formally, if $B(w)$ is the set of all worlds you *b*elieve could be actual at w, then we require that $f(\phi, w) \subseteq B(w)$.

That makes sense if you think about it. After all, constrast the following:

(2) a. # It's raining, and if it's not raining, the streets aren't wet.
    b. It's raining, and if it hadn't been raining, the streets wouldn't be wet.

4. Stalnaker's Logic

At this point, you might be wondering: Why did we impose *those* constraints on selection functions in particular? The answer is that they give rise to a beautiful *logic* for conditionals. In other words, remember how, in some sense, the idea that reflexive accessibility relations "correspond" to the axiom $\Box\phi \supset \Box\Box\phi$? Well, it turns out that our constraints on selection functions similarly correspond to certain axioms.

In particular, Stalnaker's logic for conditionals—which, historically, is called **C2**—has the following axioms:

- **PC**. All axioms of propositional logic
- **Identity**. $\phi > \phi$
- **MP**. $(\phi > \psi) \supset (\phi \supset \psi)$
- **Reciprocity**. $(((\phi > \psi) \wedge (\psi > \phi)) \wedge (\phi > \chi)) \supset (\psi > \chi)$
- **Conditional Excluded Middle (CEM)**. $(\phi > \psi) \vee (\phi > \neg\psi)$

Here, the Identity axiom corresponds to the Success constraint on selection functions; the MP axiom corresponds to Centering; Reciprocity corresponds to Reciprocity (obviously); and CEM corresponds to Uniqueness.

Lewis's logic for conditionals is basically identical to Stalnaker's, except that he rejects CEM. This is really the key difference between Stalnaker and Lewis, at the level of logic.

In fact, Lewis rejects CEM for reasons other than that it corresponds to the Uniqueness constraint on selection functions, and this (he thinks) gives rise to implausible metaphysical implications. Another reason is that Lewis wanted to introduce a second conditional operator, $\Diamond\!\!\rightarrow$, alongside $\Box\!\!\rightarrow$, in the same way we introduce both $\Box$ and $\diamond$. The former connective is supposed to correspond to the English language 'might'-counterfactual:

(3) If I had flipped the coin yesterday, it might have landed heads.

And just as we have $\Box\phi \dashv\vdash \neg\Diamond\neg\phi$, Lewis wanted the following principle to hold for 'might'-counterfactuals:

- **Duality**. $\phi \mathrel{\Box\!\!\rightarrow} \psi \dashv\vdash \neg(\phi \mathrel{\Diamond\!\!\rightarrow} \neg\psi)$.

To illustrate, the Duality says the following are equivalent:

(4) a. If I had flipped the coin yesterday at noon, it would have landed heads.
b. It's not the case that, if I had flipped the coin yesterday at noon, it might not have landed heads.

Intuitively, Duality is plausible. But it turns out that any logic that validates it, together with CEM, entails that 'might'-counterfactuals *entail* corresponding 'would'-counterfactuals. That is, the first sentence below, entails the second:

(5) a. If I had flipped the coin yesterday at noon, it might have landed heads.
b. If I had flipped the coin yesterday at noon, it would have landed heads.

That, surely, isn't right.

Challenge Question. Show that Duality, together with CEM, entails the collapse of 'might'-counterfactuals into 'would'-counterfactuals. (You may assume classical logic.)

5. Why the logic C2?

Once you get into conditional logic, you realize that C2 is a very *beautiful* logic for conditionals. Why's that? Well, partly because it's just about the strongest logic for conditionals we can come up with, which doesn't entail that English language conditionals—either indicatives or subjunctives—just *are* material conditionals, $\phi \supset \psi$.

To see what I mean, here's an example. Consider the following, additional axiom we might impose on conditionals (which a lot of people like for indicative conditionals):

- **Import-Export (IE)**. $\phi \rightarrow (\psi \rightarrow \chi) \dashv\vdash (\phi \wedge \psi) \rightarrow \chi$.

Example: The following sentences are equivalent (according to IE):

(6) a. If I flip the coin, then if it lands heads, I'll win the bet.
b. If I flip the coin and it lands heads, then I'll win the bet.

This principle seems *extremely* plausible, on a first pass. But it turns out, if you add it to C2, then the logic says indicative conditionals just *are* material conditionals. Observe:

(i) Suppose $(\phi \supset \psi) \rightarrow (\phi \rightarrow \psi)$. (This is an instance of the first part of IE above.)

(ii) Then, by IE, (i) is equivalent to the following $((\phi \supset \psi) \wedge \phi) \rightarrow \psi$.

(iii) But by classical logic, the antecedent $((\phi \supset \psi) \wedge \phi)$ is equivalent to $(\phi \wedge \psi)$ (you can make a truth-table to show this). So we get that $(\phi \wedge \psi) \rightarrow \psi$.

(iv) But now it follows from Identity and classical logic that $(\phi \wedge \psi) \rightarrow \psi$ is a logical truth. So, this implies that our original formula $(\phi \supset \psi) \rightarrow (\phi \rightarrow \psi)$ is a logical truth.

(v) But the axiom MP tells us that $(\phi \rightarrow \psi) \supset (\phi \supset \psi)$ is a logical truth as well. So $\phi \rightarrow \psi$ and $\phi \supset \psi$ are equivalent.

PHIL 210/510: Formal Methods — Yale, Spring 2025
Lecture 14: The Logic of Conditionals Continued — Dr. Calum McNamara

1. Indicatives vs. Subjunctives

Last time, we heard about **Stalnaker models**, which give the truth-conditions for English language conditionals—both **indicatives** and **subjunctives**. Recall the distinction between these kinds of conditionals.

(1) If Oswald didn't shoot Kennedy, someone else did. (Indicative)

(2) If Oswald hadn't shot Kennedy, someone else would have. (Subjunctive)

On Stalnaker's theory, indicatives and subjunctives have a common semantics: 'If ϕ, then ψ' is true at a world w, just in case $f(\phi, w) \subseteq \psi$. And recall that $f(\phi, w)$ is the *closest* ϕ-world to w. Here, f is a **selection function**.

In order to ensure that f captures something like a *closeness* relation between worlds, we impose constraints on the selection function. The constraints we looked at last time are:

- **Success**. $f(\phi, w) \subseteq \phi$ (the closest ϕ-worlds to w should *be* ϕ-worlds).
- **Centering**. If $w \in \phi$, then $w \in f(\phi, w)$ (if w itself is a ϕ-world, then w should be among the closest ϕ-worlds to itself).
- **Reciprocity**.[1] If $f(\phi, w) \subseteq \psi$ and $f(\psi, w) \subseteq \phi$, then $f(\phi, w) = f(\psi, w)$.
- **Uniqueness**. $|f(\phi, w)| \leq 1$. (There's a *unique* closest ϕ-world for each w).

These constraints together ensure that closeness is a **total order** on the set of worlds. That is, for each world w, there's a unique **sequence** of possible worlds, $\langle w, w_1, w_2, ...\rangle$, which tells us that w is the closest world to itself, w_1 is the next closest, and so on.

These constraints then give rise to analogous constraints on the **logic** of conditionals. Stalnaker's own logic is called C2. It comprises the following axioms.

- **PC**. All axioms of propositional logic
- **Identity**. $\phi > \phi$
- **MP**. $(\phi > \psi) \supset (\phi \supset \psi)$
- **Reciprocity**. $(((\phi > \psi) \wedge (\psi > \phi)) \wedge (\phi > \chi)) \supset (\psi > \chi)$
- **Conditional Excluded Middle (CEM)**. $(\phi > \psi) \vee (\phi > \neg\psi)$

We then take the closure of these axioms under our old rule of modus ponens, as well as a new rule—which is basically the version of modus ponens applied to English language conditionals:

- **Detachment**. If ϕ and $\phi > \psi$, infer ψ.

Challenge Question. Last time, we defined the language with $\Box$ as a primitive. It turns out you can define $\Box$ (and therefore $\Diamond$) using just the conditional symbol $>$. Can you say how?

One lingering issue at the moment is that it's not clear how we can distinguish between indicatives and subjunctives in the present set-up. But intuitively, there's a strong distinction between those kinds of conditionals.

To rectify this, Stalnaker (1975) says that, on top of the constraints on the selection function given above, this function must obey an additional constraint in the case of indicative conditionals. This is: it must select only **epistemically possible** worlds—worlds consistent with your knowledge or evidence. More formally:

- **Indicative Constraint**. Let B_w be the set of all my "belief worlds" at w. Then, $f(\phi w) \subseteq B_w$.

This makes sense, if you think about it. After all, contrast the following:

(3) ?? The coin landed heads and if it landed tails, I won the bet.

[1] Reciprocity was originally referred to using the acronym 'CSO'. But no one seems to know what that actually stands for.

(4) The coin landed heads, and if it had landed tails, I would have won the bet.

Here, by asserting 'The coin landed heads', the speaker (implicitly) tells us they *believe* it landed heads. The clash we get in (3) then comes about because the antecedent of that conditional—an indicative—is inconsistent with the speaker's beliefs, violating the indicative constraint. In contrast, such a clash does not occur in the case of (4). And that's because—unlike indicatives—the consequent of a counterfactual needn't be consistent with your beliefs (as the name implies).

In a slogan, then: indicative conditionals speak about **epistemic** possibilities. In contrast, counterfactuals speak about **metaphysical** (or perhaps **causal**) possibilities.

2. The Revenge of the Material Conditional[2]

Once you get into conditional logic, you realize that C2 is a very *beautiful* logic for conditionals. Why's that? Well, partly because it's just about the strongest logic for conditionals we can come up with, which doesn't entail that English language conditionals—either indicatives or subjunctives—just *are* material conditionals, $\phi \supset \psi$. (Remember: when we covered conditionals a couple of weeks ago, one of our key data was that English language conditionals *aren't* material conditionals.)

More carefully, what I mean is that there's almost nothing we can *add* to Stalnaker's logic, which doesn't result in $>$ collapsing back to $\supset$. Your problem set this week has a question of this nature. It asks you to show that if we add the following principle to the logic of conditionals, then English language conditionals are equivalent to material conditionals:

- **Or-to-if**. $\phi \vee \psi \vdash \neg\phi > \psi$.

Here's another example (you can use the proof below as a template for the Or-to-if question). Consider the following, additional axiom we might impose on conditionals (which a lot of people like for indicative conditionals—that's why I've used the $\rightarrow$ symbol):

- **Import-Export (IE)**. $\phi \rightarrow (\psi \rightarrow \chi) \dashv\vdash (\phi \wedge \psi) \rightarrow \chi$.

Example: The following sentences are equivalent (according to IE):

(5) a. If I flip the coin, then if it lands heads, I'll win the bet.
b. If I flip the coin and it lands heads, then I'll win the bet.

This principle seems *extremely* plausible, on a first pass. But it turns out, if you add it to C2, then the logic says indicative conditionals just *are* material conditionals. This fact was first proved by Allan Gibbard, in the 1980s. Observe:

Proof. Consider the following conditional: $(\phi \supset \psi) \rightarrow (\phi \rightarrow \psi)$. By Import-Export, this is equivalent to $((\phi \supset \psi) \wedge \phi) \rightarrow \psi$. However, by classical logic, the antecedent, $(\phi \supset \psi) \wedge \phi$, is equivalent to $\phi \wedge \psi$. After all, consider the following truth-table:

ϕ	ψ	$\phi \wedge \psi$	$(\phi \supset \psi) \wedge \phi$
1	1	1	1
1	0	0	0
0	1	0	0
0	0	0	0

So $((\phi \supset \psi) \wedge \phi) \rightarrow \psi$ collapses to $(\phi \wedge \psi) \rightarrow \psi$.

However, now notice that $(\phi \wedge \psi) \rightarrow \psi$ is a logical truth. After all, think about our Stalnaker semantics. It says that $(\phi \wedge \psi) \rightarrow \psi$ is true at a world w just in case the closest $\phi \wedge \psi$-world is a ψ-world. But by classical logic, $\phi \wedge \psi$ is true at a world only if ψ is true at that world. So, $(\phi \wedge \psi) \rightarrow \psi$ is true at every world.

But if $(\phi \wedge \psi) \rightarrow \psi$ is a logical truth, then so must be $((\phi \supset \psi) \wedge \phi) \rightarrow \psi$. After all, a moment ago, we showed that these two formulas are equivalent.

[2] Mostly copy-pasted from last time.

Similarly, if $((\phi \supset \psi) \wedge \phi) \rightarrow \psi$ is a logical truth, then so must be $(\phi \supset \psi) \rightarrow (\phi \rightarrow \psi)$, by Import-Export.

But now, consider $(\phi \supset \psi) \rightarrow (\phi \rightarrow \psi)$. Since it's a logical truth (as we've just shown), and our MP axiom tells us that indicative conditionals entail material conditionals, it follows that $(\phi \supset \psi) \supset (\phi \rightarrow \psi)$ is a logical truth as well. In other words, the material conditional $\phi \supset \psi$ entails the indicative conditional $\phi \rightarrow \psi$. But then, since the MP axiom tells us things go in the opposite direction, too, it follows that indicatives and material conditionals are equivalent. Q.E.D. □

3. The Story so Far

So much for the logic of conditionals—you'll have a chance to play with proofs like the one above on Problem Set 3, due after the break.

After the break, we'll be moving on to two of my favorite topics: probability and decision theory—both areas in which I've done my own research.

For now, however, let's try to get a grip on the story so far. We began this course by thinking about set theory, as well as related concepts like functions, and relations. We were also interested in the abstract idea of a formal system—of which propositional logic can be viewed as a species. We heard about some ideas having to do with the mathematics of infinity. And we then got introduced to a few techniques for *proving* things—as well as to a few important notions, like use and mention; types and tokens; and—most importantly—*possible worlds*.

Even at this early stage in the course, the ideas we developed raise really interesting philosophical issues. We talked about some of those already—issues about *truth* and *falsity* or mathematical claims, for example. But believe it or not, there are more "applied" issues, to which these topics give rise, too. Here are some examples.

Let's start with the use/mention distinction. This distinction turns out to be really important in the study of *slurs* in natural language. In particular, some linguists—like Luvell Anderson and Ernie Lepore—think that one of the defining features of slurs is that's inappropriate even to *mention* them. Everyone agrees, after all, that it's offensive to *use* slurs in conversation. But Anderson and Lepore argue that what makes slurs "special", in natural language, is that even *mentioning* them is ill-advised. (Whether or not you agree with that is something that we can discuss; my point is just to show that a seemingly abstract, formal distinction turns out to have important uses in linguistics—and maybe even in ethics.)

Similarly, consider our conversation about *possible worlds*. Should we admit possible worlds into our ontology? David Lewis gave a striking argument that we should. Why? Because—he says—they turn out to be so incredibly *useful* to our theorizing. This is a bit like how we admit the existence of numbers (or other mathematical objects). Science, you might think, couldn't be done without admitting the existence of numbers. And Lewis thinks much the same thing goes for, e.g., theorizing in linguistics, philosophy—even economics. The things we take to exist are precisely the things required by our best theories.

That's a strong argument, in my view. But maybe it rules out the existence of things you might want to admit. Religious people might be inclined to think, for example, that Lewis's methodological approach to what exists there rules out the existence of God. What do you think?

After this section of the course, we moved on to formal semantics. Formal semantics aims to tell us how the *meanings* of whole sentences can be computed from the meanings of the parts. The study of formal semantics begins with Frege's conjecture:

Frege's Conjecture. Semantic composition is functional application.

We were then able to use this conjecture to show how, starting with a stock of "primitive" semantic types and values, we could work our way up to more sophisticated types and values. For example, it turned out that on this way of thinking about things, the semantic value of a word like 'smokes' is a certain *function*.

One interesting thing about this approach to semantics is that it's pretty radically *externalist*. An externalist theory of meaning is one on which the meanings of terms in natural language are largely independent of what we *think* their meanings are. As Putnam famously put it: “Meaning just ain’t in the head”.

To illustrate, let’s consider a famous example. In the late 18th century, it was discovered that water is H_2O. The ancient Greeks certainly didn’t know that. Nor did Isaac Newton. So, when they used the term ‘water’ (or the Greek translation thereof), did the *mean* the same thing as us? The internalist says ‘No’; the externalist says ‘Yes’. What do you think?

Here’s another example. Consider a world—twin earth—where everything is the same as it is on (actual) earth, except, instead of being made of H_2O, water is made of XYZ. Aside from that, it’s exactly the same. It looks the same; tastes the same; people use it to wash dishes and clothes; they bathe in it; and they keep themselves hydrated with it. When twin earthers say ‘water’ do they mean the same thing as us?

After this part of the course, we moved on to modal logic. Modal logic, you might think, is the logic of *modal* expressions—things like ‘might’, ‘possibly’, ‘necessarily’, ‘believes’, ‘ought’, and so on.

We saw that we could study *all* the logical uses of these modals by building a semantics based on *possible worlds* and relations between them. In particular, the *accessibility relations* we added to our modal models allowed us to characterize different “flavors” of modality. We then saw that different constraints we can place on accessibility relations correspond to various modal *axioms*. And in turn, these give rise to various modal logics.

The ability to pin down various concepts that are of interest in philosophy by thinking of the *logic* of those concepts is a really useful skill. For example, take a concept that’s a bit of a philosophical mystery: causation. One thing you might want to do, for example, is try to characterize the relation of causation by thinking about the logic of expressions like ‘A is a cause of B’. Is that relation *transitive*? *Symmetric*? *Reflexive*? None of the answers here are entirely obvious.

4. Where We’re Going

After the break, we’ll start thinking about *probability*, and then *decision theory*. To start with, we’re going to think about questions like ‘What *is* probability?’ If you think about—and despite how ubiquitous probability is in our day-to-day lives—it’s not at all obvious what kind of thing probability *is*. Indeed, all the main theories of probability are subject to serious problems.

Nevertheless, we can develop a rich, mathematical theory of probability—a theory which tells us that, no matter what kind of thing probability is, it has to *behave* in a certain way. We’ll then think about various philosophical issues arising from the mathematical theory of probability. For example, we’ll ask questions like: ‘Why is it *irrational* to be, say, 70% confident that it will rain tomorrow, and also 70% confident that it won’t rain?’ Why, in other words, should our *degrees of belief* behave like probabilities.

Similarly, we’ll ask questions like: ‘If there are *objective* probabilities—what philosophers call *chances*—why should be degrees of belief match them?’ Why, in other words, would it be irrational to believe that a certain coin had a 50% chance of landing heads on the next toss, and yet, at the same time, be 70% confident that it’ll land heads on the next toss.

Finally, we’ll think about the probabilities of *conditionals*. This, it turns out, is one of the thorniest topics in philosophy—full stop!

After that, we’ll think about decision-making. We’ll see that theory of probability in part gives rise to a plausible, mathematical theory of decision-making. And we’ll see—among other things—that the semantic theory of conditionals we developed in the first part of the course turns out to be very important to that theory of decision-making.

PHIL 210/510: Formal Methods — Yale, Spring 2025
Lecture 15: Probability I — Dr. Calum McNamara

Today, we're going to make a start on one of my favorite topics: probability theory. Specifically, we'll look at the mathematical foundations of probability, and we'll derive some useful results. Before we do that, however, we're going to start with a philosophical question—namely, what kind of thing *is* probability?

1. Objective vs. Subjective Probability

Our use of the word 'probability' is ambiguous. On the one hand, it sometimes refers to an *objective* quantity. When scientists say, for example, that quantum mechanics is a probabilistic theory, they (usually) mean that *nature itself* is probabilistic. It's an objective fact that particles have probabilities of following certain trajectories, etc.[1]

On the other hand, 'probability' can sometimes be used to refer to something like *subjective uncertainty*. When you ask me, for instance, whether the Mets will win the World Series this year, and I respond "Probably not", it's plausible that I'm expressing something like my *low degree of confidence* that the Mets will win the World Series.

Philosophers usually reserve the word 'chance' for the objective kind of probability, and 'credence' for the subjective kind. There are, of course, connections between these two kinds of probability (for instance, if you know the *chance* that a certain coin will land heads, if tossed, is 50%, then it seems like your *credence* that it'll land heads, if tossed, should also be 50%). For the most part, however, we're going to be focused on the subjective kind of probability in this lecture. We'll say more about the objective kind of probability—chance—in subsequent lectures. (For what it's worth, I regard the metaphysical question of what kind of thing chance *is* as one of the hardest questions in metaphysics. Almost every prima facie plausible view is open to serious objections.)

2. Traditional vs. Bayesian Epistemology

The study of knowledge, as well as related concepts like belief, justification, etc., forms a subfield of philosophy known as *epistemology*. Traditionally, epistemologists asked questions like 'What does it take to believe/know a proposition?' For them, belief was an on/off matter—you either believed something or you didn't. But plausibly, this view is too coarse-grained.

Beginning in the early twentieth century, with the work of Frank Ramsey (1903–1930), epistemologists started to work with more fine-grained notions of belief—in particular, credence. For them, the salient questions weren't about what it takes to believe or disbelieve a proposition, but rather what *degree of confidence* one should have in a proposition, and how these degrees of confidence should hang together. Notably, for example, the laws of probability came to be viewed as *constraints* on rational credences. In other words, if you were (ideally) rational, then your subjective degrees of belief (credences) would behave like probabilities. This approach to epistemology is known as *formal* or *Bayesian* epistemology.[2]

We'll see examples of what I mean here as we go forward. But first, let's make a few background assumptions crystal clear. From now on, we'll assume that credence (like "full belief") is an attitude one takes towards a *proposition*. In particular, we'll assume that credence is the *degree of confidence* one has in a proposition—i.e., the degree of confidence one has that the actual world is a member of that proposition. (Recall: we're thinking of propositions as *sets of possible worlds*.) We'll also assume that these degrees of confidence are real numbers in the interval $[0, 1]$. And finally, we'll assume that credence 1 in a proposition A denotes full confidence in A; credence 0 in A denotes minimal confidence in A (or full confidence in $\neg A$); and credence 0.5 in A means that one is just as confident in A as one is in its negation. These assumptions seem obvious. But they're actually non-trivial. Note also that there's another, closely

[1] Of course, as a matter of interpretation, quantum mechanics (QM) is highly controversial, and there are deterministic versions of this theory. I think it's fair to say, however, that the textbook formulation of QM is probabilistic.

[2] 'Bayesian' after the Rev. Thomas Bayes (c. 1701-1761). Why, then, is this approach to epistemology named after Bayes? It's tempting to think that it has something to do with *Bayes's theorem*, which we'll encounter later today. But in fact, it's because Bayes was one of the first to discuss probability in terms of degrees of belief, in his letters. The first real proponent of this view, however, was Pierre-Simon Laplace.

related interpretation of 'credence' that we'll look at in a moment—namely the interpretation of credence as *expectation of truth-value.*

3. A Return to Set Theory

A moment ago I said that we'll be assuming propositions are sets of possible worlds. This is an idea we've discussed at several points in the course already. If you've ever taken a statistics class, you might have heard the word 'events' used in place of 'propositions'. For example, you may have heard things like "The probability of the event that the dice lands 6 is $1/6$". Rest assured: when statisticians say 'event', they're referring to exactly the same thing that we are when we say 'proposition' (interpreted as a set of possible worlds). It's not that we're departing in any way from the kind of probability theory done in other university departments.

So, with that remark having been made, let's assume (purely for simplicity) that there are only finitely-many possible worlds.[3] We can collect these worlds into a set, $\mathcal{W} = \{w_1, ..., w_n\}$. And since propositions are just sets of possible worlds (on our interpretation), any subset $A \subseteq \mathcal{W}$ is itself a proposition.

If we collect all the subsets of $\mathcal{W}$ into another set, we get the *power set* of $\mathcal{W}$, denoted $\mathcal{P}(\mathcal{W})$. This is the set of all propositions. Note that, since $\mathcal{W}$ is a subset of itself, it's a proposition.[4] In particular, it's the necessarily true proposition. Conversely, the empty set, $\varnothing$, is the necessarily false proposition, that is true at *no* world.

Now consider two arbitrary propositions, A and B. If the intersection of A and B is the empty set, i.e., $A \cap B = \varnothing$, then we say that A and B are *mutually exclusive.* We can interpret this as: there is no world w in $\mathcal{W}$ at which both A and B are true (or alternatively: there is no world w which is a member of both A and B).

We need the notion of mutually exclusive sets in order to define another important notion, which will come up frequently: that of a *partition.* A partition, $X_1, ..., X_n$, is a set of subsets of $\mathcal{W}$ that satisfies two important properties. First, each proposition X_i in the partition is mutually exclusive; i.e., there is no world w in $\mathcal{W}$ that's an element of both X_i and X_j (for distinct i and j). Second, the X_i are jointly exhaustive: every world w in $\mathcal{W}$ is an element of some X_i. Or, in other words, the union of all the X_is, $X_1 \cup ... \cup X_n$, just is the set of all possible worlds, $\mathcal{W}$.

As an example, the set $\{X, \neg X\}$ is an easy example of a partition: every world w is an element of one of X or $\neg X$, and no world w is an element of both.

Another important notion we'll need is the notion of an *algebra* of propositions. An algebra of propositions is a set of subsets of $\mathcal{W}$, $\mathcal{F}$, that satisfies three properties:

(i) $\mathcal{F}$ contains $\mathcal{W}$,

(ii) $\mathcal{F}$ is closed under complements. In other words, if $\mathcal{F}$ contains a proposition X, then it also contains $\neg X$,

(iii) $\mathcal{F}$ is closed under unions. So, if contains two propositions X and Y, then it also contains their union, $X \cup Y$.

Here are some examples of algebras. First, let $\mathcal{F} = \{\varnothing, \mathcal{W}\}$. This algebra, which we call the *trivial algebra*, satisfies the properties (i)–(iii). After all, it satisfies (i) because it contains the trivial proposition, $\mathcal{W}$. And it satisfies property (ii), because it contains $\varnothing$, which is the complement of . Finally, $\mathcal{F}$ satisfies property (iii) because $\mathcal{W}$ and $\varnothing$ are mutually exclusive propositions, and $\mathcal{F}$ contains their union: $\mathcal{W} \cup \varnothing = \mathcal{W}$.

A second example: the power set of $\mathcal{W}$, $\mathcal{P}(\mathcal{W})$, is an algebra. It's a good exercise to verify that this is so.

As a final point before we move on, I should mention a closely related notion: that of a *σ-algebra.* A σ-algebra is just like an algebra except that we extend

[3] If we assume that $\mathcal{W}$ is instead *countably* infinite, then nothing much changes. But if we assume that $\mathcal{W}$ is *uncountably* infinite, then we need to replace summations with integrals in much of what follows. For the most part, we can ignore this complication here. As I said, however, the only reason we assume that there are only finitely-many possible worlds is to keep things simple, particularly the mathematics. The mathematical theory we develop below easily extends to infinite cases.

[4] Why is the claim that $\mathcal{W}$ is a subset of itself legitimate? It's a good exercise to figure out why this is, using the definition of 'subset'.

the definition of property (iii), as follows. First, suppose momentarily that $\mathcal{W}$ is a countably infinite set, rather than a finite one. Then:

(iii′) $\mathcal{F}$ is closed under countable unions. That is, if $X_1, X_2, \ldots$ is a countably infinite family of propositions in $\mathcal{F}$, then $X_1 \cup X_2 \cup \ldots$ is also in $\mathcal{F}$.

Since we can safely assume, for the most part, that $\mathcal{W}$ is finite in what follows, we can stick with simple algebras rather than σ-algebras.

4. Probability Axioms

Suppose that we have some particular algebra of propositions, $\mathcal{F}$. A *credence function*, c is a function that takes each proposition A in the algebra $\mathcal{F}$, and maps it to a real number. Intuitively, if c is *your* credence function, then these real numbers represent your degree of confidence in each proposition in $\mathcal{F}$.

Formally, *any* credence function can represent a system of degrees of belief. But intuitively, some credence functions are *better* at representing the world than others. For example, imagine your credence function assigns the real number .7 to the proposition that it's raining, but *also* assigns .7 to the proposition that it's not raining. Intuitively, tere's something wrong with that—but what?

Well, to start with, those credences do not satisfy the *axioms of probability*. These are:

Probability Axioms.

(1) **Normality**. $c(\mathcal{W}) = 1$.

(2) **Non-negativity**. For all propositions X, $c(X) \geq 0$.

(3) **Additivity**. If X and Y are two disjoint propositions, i.e., $X \,\&\, Y = \emptyset$, then $c(X \vee Y) = c(X) + c(Y)$.

If a credence function satisfies these axioms, then we call it, simply, a *probability function.*

A few further points about the axioms. First, you might notice that, according to the way I defined credence functions initially, the first two axioms are trivially satisfied. After all, I said that credences are just real numbers in the interval $[0, 1]$. That's fine; but it does show that the normative content of the axioms is really contained all in third axiom, additivity.

Secondly, we haven't really said anything to justify *why* it seems credences should satisfy the probability axioms. In other words: What's wrong with having .7 credence that it will rain, but also .7 credence that it won't? In the next lecture, we'll see that there are very good arguments to the effect that you should satisfy these axioms. But for now we'll simply take this for granted.

5. Practice with the Probability Axioms

Believe it or not, even though we've only got three probability axioms, almost *every* truth about probability can be derived from these alone. Let's look at some examples of how interesting facts can be derived from these axioms. This should give you some practice writing mathematical proofs using the probability axioms, too.

Challenge Question. Show that: $c(A) + c(\neg A) = 1$.

Intuitively, this seems true. After all, if my credence in *It's raining* is 0.7, then it looks like I ought to have credence 0.3 in *It's not raining*. And these credences sum to 1. But how are we to *prove* the general statement, using only our axioms? Here's how:

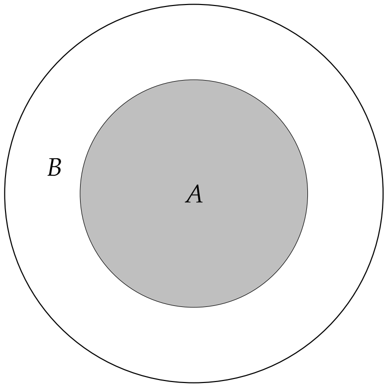
Proof. First, since A and $\neg A$ are disjoint propositions, we know, by Axiom 3 above, that $c(A) + c(\neg A) = c(A \vee \neg A)$. But $A \vee \neg A$ just is the set of all possible worlds, $\mathcal{W}$. Thus, by Axiom 1, we know that $c(\mathcal{W}) = 1$. So it follows that $c(A \vee \neg A) = 1$. But just a second ago, we showed that $c(A) + c(\neg A) = c(A \vee \neg A)$. So it follows from this that $c(A) + c(\neg A) = 1$. □

Challenge Question For any proposition A, $c(A) \leq 1$.

How do we prove this one? Consider the following:

Proof. Here, we can actually just use the fact we proved a moment ago to help us. First, choose an arbitrary proposition A. Then, consider that, by the thing we proved in the first Challenge Question, $c(A) + c(\neg A) = 1$. Now, using a bit of algebra, it follows that $c(A) = 1 - c(\neg A)$. Consequently, $c(A) \leq 1$. In particular, $c(A) = 1$ only if $c(\neg A) = 0$. □

Finally, consider the situation in which one proposition, A, *entails* another proposition, B. In terms of possible worlds, we can think of this situation as follows: if A entails B, then every world at which A is true is a world at which B is true. For example, the proposition *It's raining and the coin landed heads* entails the proposition *It's raining*, since every world at which the first is true is one at which the second is true. Visually, we can think of this situation as follows:



What we want to prove now is the following fact.

Challenge Question. Show that, if A entails B, then $c(A) \leq c(B)$.

I'll leave you to prove this one on your own. (Hint: you might want to first derive a lemma, namely that, for any A and B, $c(A \vee B) = c(A) + c(B) - c(A \wedge B)$.)

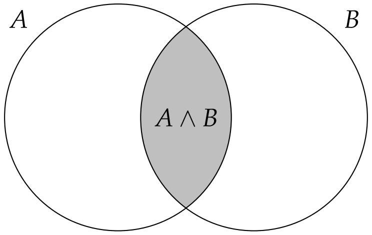
6. Conditional Probability

We're now going to introduce the notion of *conditional probability*: roughly, the probability that some proposition is true, given that some other proposition is true. More precisely, we define the conditional probability of B given that A, written '$c(B \mid A)$', as follows:

$$c(B \mid A) = \frac{c(A \wedge B)}{c(A)},$$

so long as $c(A) > 0$. Thus, conditional probability is defined as a ratio of unconditional probabilities: in this case, the probability that A and B are both true, divided by the probability that A is true.

There's a nice, visual way to think about conditional probability. To see it, consider this Venn diagram:



Thus, we can think of the conditional probability $c(B \mid A)$ as follows: $c(B \mid A)$ is the probability that one is in the B-region, given that one is in the A-region. It's the *proportion* of the B-region that lies within the A-region.

A brief digression: some philosophers—Ramsey is one if them—think that we should take conditional probability as primitive, and define unconditional probability in terms of conditional probability.[5] There are attractive reasons to take this approach; for example, it avoids some awkwardness concerning probability-zero events. But for now, let's set that issue aside. We can stick with the so-called "ratio definition" of conditional probability going forward.

[5] Specifically, we can define the unconditional probability of a proposition A as the conditional probability $c(A \mid \mathcal{W})$—the probability that A is true, given that the trivial proposition is true.

PHIL 210/510: Formal Methods — Yale, Spring 2025
Lecture 15: Dutch Books, Accuracy — Dr. Calum McNamara

1. Expectation

Before we get started today, I want to introduce you to one last important notion about probability. This is the notion of **expectation**.

Let's start with an example—this example is actually the example that gave rise to probability theory in the first place.

The Problem of the Points. Imagine we're playing a game in which a fair die is tossed. You get a point whenever the die lands on 1, 2, or 3, and I get a point whenever it lands on 4, 5, or 6. At the moment, you have 1 point and I have 2. The first to get three points wins a prize of $12. Just as we're about to roll the die for a fourth time, the game is interrupted—the police burst into the room, and we have to scurry away. (Unlicensed gambling is illegal!) Later on, we meet up, and the game's arbitrator decides we'll split the pot of money, rather than risk being caught again. The question now is: How should we split the pot, to make things fair?

Should I take all the money? That doesn't seem fair—even though I was ahead, you still had a *chance* of winning.

Should we split the money evenly? That doesn't seem fair either—after all, I had more points than you!

Challenge Question. How should we split the pot of money, so as to make the distribution fair?

In a famous correspondence, two mathematicians—Pascal and Fermat—solved this problem, and in turn, effectively founded probability theory. Here is how they did so. There are three possible ways this game could end, given the point we've arrived at. These are:

- The die could land on 4, 5, or 6 on the next toss, in which case I win.
- The die could land on 1, 2, or 3 on the next toss. We'd then have to roll again, since we'd both have two points at that point in the game. It could then land on 4, 5, or 6 on the toss after that, in which case I win again.
- The die could land on 1, 2, or 3 on the next toss, and then 1, 2, or 3 on toss after that.

In two out of three of these scenarios, I win the pot of money, while in one of the scenarios, you win. Notice, however, that—since the die is fair—the probability of getting a scenario of the first kind is $1/2$, while the probability of getting a scenario of the second two kinds is $1/4$. Thus, Pascal and Fermat surmised that the fair distribution here is to split the pot three-quarters/one-quarter: I get $9, you get $3. That does, I think, seem fair.

In coming up with this solution, Pascal and Fermat effectively invented the idea of a *random variable*. Formally, a **random variable** $X : \mathcal{W} \to \mathbb{R}$ is a function, which maps worlds to real numbers—*any* function that maps worlds to real numbers. Informally, however, it's sometimes worth thinking of a random variable as a definite description. In the case above, for example, we can think of the relevant definite description as 'My earnings at the end of the game'. In that case, X maps every world of the first kind to the value 12; it maps every world of the second kind to the value 12; and it maps every world of the third kind to the value 0. Thus, the *expected value* of this random variable is:

$$\mathbb{E}_p[X] = 1/2 \cdot 12 + 1/4 \cdot 12 + 1/4 \cdot 0 = 9.$$

(Here, '$\mathbb{E}_p$' is the notation we use for the *expectation* of a random variable X, according to a probability function p.)

More formally, the **expected value** (or **expectation**) of a random variable X is a probability weighted average:

$$\mathbb{E}_p[X] := \sum_w p(w) \cdot X(w).$$

In this case, $p(w)$ is the probability that world w is actual—according to some fixed probability function—and $X(w)$ is the value that the random variable

X takes at w. As we'll see in the next two parts of the course, the notion of expectation is *extremely* important.

2. Dutch Books

For example, one way in which it's important is when it comes to *justifying* the interpretation of probability as a set of constraints on rational credences. In the last class, I mentioned that this is how **Bayesians** think of probability: according to them, probabilities correspond to ideally rational degrees of belief.

That *seems* quite natural—it would be weird, after all, to say that the probability it will rain today is .7, but the probability that it won't rain is also .7. Something's not right about that. But can we make this idea precise? In particular, can we say that having credences like those just mentioned is somehow *normatively* defective.

Arguably we can. In the 1920's, Frank Ramsey proposed an argument which came to be known as the *Dutch book argument*—an argument which *justifies* the idea that probabilities correspond to rational degrees of belief. In rough terms, Ramsey's thought was that, if you have credences that *don't* satisfy the probability axioms, then we can always cook up a set of bets, each of which you'll be inclined to think is acceptable, but which jointly *guarantee* you a sure loss of money.

To see how this kind of argument works, let's start by introducing some definitions, and by making an assumption. First, a **unit bet** on a proposition A is a bet that pays \$1 if A is true, and \$0 if A is false. Formally, we can think of a unit bet as a random vaariable: it maps a world w to 1 if the A is true at w, and it maps a world w to 0 if A is false at w.

Now, your **fair price** for a unit bet on A is, we'll assume, just your credence in A. For example, if your credence in A is .7, then your fair price for the bet is \$0.70. Roughly speaking, the idea is that your fair price is the price at which you'd be completely indifferent between having the bet, and having the amount of money corresponding to your fair price: If I offered you a choice between \$0.70 and a unit bet on A, then you'll be indifferent between the two options iff 0.70 is your fair price—viz., iff .7 is your credence in A. At any price lower than \$0.70, you'll think having the bet is better; and at any price higher than that, you'll think having the money is better. (Of course, all this involves significant idealization. Maybe you're risk-averse, etc.—in real life, people often are. There are ways to handle that kind of thing; but we're simplifying things dramatically, and assuming risk-aversion, etc., away.)

Now, consider again the case in which someone has .7 credence in a proposition A, and .7 credence in a proposition $\neg A$. Then, here is how we can construct a Dutch book argument against them.

- We offer them a bet on A, for a price of \$0.70. Since their credence in A is .7, the price for the bet is also their fair price, and they should find the bet acceptable—at the very least, they don't find it *unacceptable*.
- Likewise, we offer them a bet on $\neg A$ for a price of \$0.70. Since their credence in $\neg A$ is .7, the price for the bet is their fair price, and they should find *this* bet acceptable, too.
- By taking both bets, however, they'll have paid \$1.40 in total.
- But the most they stand to win from these bets is \$1, since if A is true, $\neg A$ is false, and vice versa.
- The upshot is that, by taking these bets, they're guaranteed to lose \$.40, *no matter which of A or* $\neg A$ *is true*. In other words, they're *guaranteed* a sure loss of money.

More generally, Ramsey was able to prove a theorem which says that, if you violate *any* of the probability axioms, then we can *always* propose you a set of bets like the one above. (Sometimes, we'll offer to *sell* you the bets; sometimes, we'll offer to *buy* them from you. But the point remains: if you violate probability axioms, you're *Dutch-bookable*.) Thus, since (we assume) having money is something you value, the argument is supposed to show that probabilistic incoherence—violating the probability axioms—results in "giving away" something you value, with no possible compensatory gain.

Let me pre-empt and objection here. You might think the Dutch book argument doesn't show very much. After all, in some sense, it depends on the existence of a *bookie*—someone who can buy or sell the relevant bets—and who knows your fair prices. But maybe the existence of such a bookie is an unrealistic assumption.

That, however, is the wrong way to think about the argument. Here is how (our hero!) David Lewis (1999) describes things:

> [T]he point [of a Dutch book argument is to show] that if you are vulnerable to a Dutch book... that means that *you have two contradictory opinions about the expected value of the very same transaction*. To hold contradictory opinions may or may not be risky, but it is in any case irrational. (p. 405, emphasis added)

Note: look out for questions about Dutch books on the upcoming problem set!

3. Accuracy

Thus, the Dutch book argument provides a compelling justification for the Bayesian claim that the probability axioms are *normative*, in the sense that, if you're *rational*, then your degrees of belief will satisfy them.

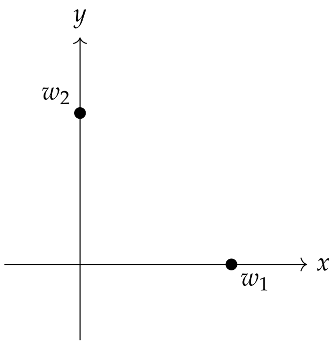
Nevertheless, there's another kind of complaint we can make against this argument—a different complaint to the one just sketched. Here is how Jim Joyce (1999)—my advisor!—describes things:

> [W]hen called upon to defend the claim that rational degrees of belief must obey the laws of probability [Bayesians] generally present some version of the Dutch Book Argument... which establishes conformity to the laws of probability as a norm of prudential rationality by showing that expected utility maximizers whose partial beliefs violate these laws can be induced to behave in ways that are sure to leave them less well off than they could otherwise be. *This overemphasis on the pragmatic dimension of partial beliefs tends to obscure the fact that they have properties that can be understood independently of their role in the production of action.* Indeed, [Bayesians] have tended to pay little heed to the one aspect of partial beliefs that would be of most interest to epistemologists: namely, their role in representing the world's state. My strong hunch is that this neglect is a large part of what has led so many epistemologists to relegate partial beliefs to a second-class status.

(When Joyce was writing this, in 1998, Bayesian epistemology hadn't yet really taken off—hence, the last sentence in the above quotation. Now, however—and thanks almost entirely to Joyce's paper—that's no longer true. Bayesian epistemology is *at least* as prominent as "traditional" epistemology.) In short, then, Joyce's complaint is that the Dutch book argument shows that violating the probability axioms is *practically* defective. But, as epistemologists, it would be nice if we could show that it was somehow also *epistemologically* defective, too.

Joyce himself gives an argument to this effect, inspired by earlier work of the statistician De Finetti. One of the beautiful things about Joyce's argument is that it lends itself to a lovely *geometric* interpretation: we can draw pictures to show why it has to be true.

To do so, however, we have to start by showing how we can represent possible worlds geometrically. Thus, start with the simplest case, in which we're interested in a proposition A and its negation, $\neg A$. We can represent these worlds as points in the plane: let w_1—the world at which A is true—be the point $(1, 0)$, and let w_2, the world at which $\neg A$ is true be the point $(0, 1)$:

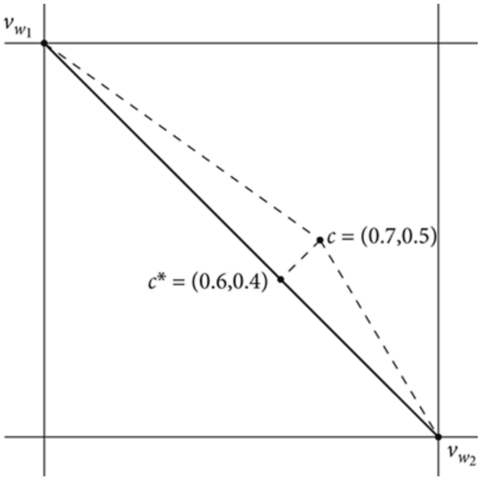


(For the mathematically inclined, we can think of worlds as *vectors*, with each component of the vector being the truth value assigned to a given proposition.)

Now, consider the set of all probability functions defined over the set $\{A, \neg A\}$. (This isn't an algebra—but we'll ignore the other elements needed to form an algebra, $\mathcal{W}$, $\varnothing$, for simplicity.) Any probability function defined on $\{A, \neg A\}$ can be expressed as a weighted average of the truth-values of A and $\neg A$. If $v_{w_i}(A)$ is the truth value of A at w_i, then every probability function can be expressed in the form:

$$p(A) := \lambda \cdot v_{w_1}(A) + (1 - \lambda) \cdot v_{w_2}(A).$$

Geometrically, what this means is that the set of all probability functions on $\{A, \neg A\}$ is really just the straight line that runs between $(1, 0)$ and $(0, 1)$.



For example, in the figure below, the point labeled 'c^*' on the line corresponds to a probability function that assigns .6 to A and .4 to $\neg A$.

But now consider the point *not* on the line, labelled 'c'. This corresponds to a credence function that violates the probability axioms—it assigns credence .7 to A, but credence .5 to $\neg A$ (and this doesn't sum to 1). But notice: by Pythagoras's theorem, the point labelled 'c' is further from w_1 than c^* is; and it's further from w_2 than c^* is, too. Thus, no matter which of w_1 or w_2 is actual—viz., no matter which of A or $\neg A$ is true—c is "further from the truth" than the function c^*.

Of course, this geometric result is only metaphorical. It helps us to get at the core of Joyce's argument. But in order to prove the argument formally, Joyce introduces a class of functions—so-called *inaccuracy measures*—which measure a credence function's (in)accuracy at a possible world. One famous example of such a function is called the *Brier score*. It says that the inaccuracy of a credence function c at a world w is defined like this:

$$\mathfrak{I}(c, w) := \sum_i (c(A_i) - v_w(A_i))^2.$$

(Note that we can think of $\mathfrak{I}(c, w)$ as a random variable!) He's then able to prove the following, beautiful result, which generalizes the geometric argument we just saw.

First, some terminology:

Weak Accuracy-dominance A credence function c is said to *weakly accuracy-dominate* another credence function, c', relative to the inaccuracy measure $\mathfrak{I}$, just in case, $\mathfrak{I}(c, w) \leq \mathfrak{I}(c', w)$ for all $w \in \mathcal{W}$, and for some $w \in \mathcal{W}$, $\mathfrak{I}(c, w) < \mathfrak{I}(c', w)$.

Strict Accuracy-dominance A credence function c is said to *strictly accuracy-dominate* another credence function, c', relative to the inaccuracy measure $\mathfrak{I}$, just in case, $\mathfrak{I}(c, w) < \mathfrak{I}(c', w)$ for all $w \in \mathcal{W}$.

Theorem 1. Let $\mathcal{F}$ be an algebra of propositions, $\mathfrak{I}$ an inaccuracy measure, and c a credence function defined on $\mathcal{F}$. Then the following are equivalent:

(i) c is not weakly accuracy-dominated.

(ii) c is not strictly accuracy-dominated.

(iii) c satisfies the probability axioms.

PHIL 210/510: Formal Methods — Yale, Spring 2025
Lecture 15: Conditionalization, Chance — Dr. Calum McNamara

1. Conditionalization and Bayes' Rule

Two sessions ago, I briefly mentioned that Bayesianism can be characterized by *three* core commitments. The first is descriptive. And the second two are *normative*. They are:

(1) **Gradationality**. Our Beliefs come in *degrees*; they're not just on/off. (That's the descriptive claim.)

(2) **Probabilism**. Rational degrees of belief, at a given time, obey the probability axioms.

(3) **Conditionalization**. When you learn the truth of a proposition, A, you should change your credences by the rule known as conditionalization.

Conditionalization is a **diachronic** norm on rational credences. That is, it describes a relationship between your credences *at different times*. (In contrast, probabilism is a **synchronic** norm on rational credences. It describes how your credences should hang together *at a particular time*.)

Here is the rule of conditionalization, which you've already seen:

> **Conditionalization**. Let c be your credence function before a learning event (your **prior** credence function); and let c_A be your credence function after learning A, and nothing stronger (your **posterior** credence function). Then, your new credence in any proposition B should be:
>
> $$c_A(B) = c(B \mid A) : \frac{c(A \wedge B)}{c(A)},$$
>
> provided $c(A) > 0$.

Notice a few things about conditionalization. First, focus on the ratio definition of conditional probability:

$$c(B \mid A) = \frac{c(A \wedge B)}{c(A)}.$$

Now, if we multiply both sides by $c(A)$, then we get:

$$c(B \mid A) \cdot c(A) = \frac{c(A \wedge B)}{c(A)} \cdot c(A) = c(A \wedge B).$$

Thus, what we've just derived is that $c(A \wedge B) = c(B \mid A) \cdot c(A)$. Of course, using an exactly parallel argument, we can derive that $c(A \wedge B) = c(A \mid B) \cdot c(B)$. This allows us to write conditionalization in a slightly different way:

> **Bayes' Rule**. Let c be your credence function before a learning event (your **prior** credence function); and let c_A be your credence function after learning A, and nothing stronger (your **posterior** credence function). Then, your new credence in any proposition B should be:
>
> $$c_A(B) = c(B \mid A) = \frac{c(A \mid B) \cdot c(B)}{c(A)},$$
>
> provided $c(A) > 0$.

The Bayes' rule version of conditionalization is really useful, in practice. Why? Well, imagine you're trying to compute the probability that some hypothesis is true, given that you've observed some evidence, for or against it. Then, Bayes' rule tells you how probable the hypothesis is, given the evidence, as a function of your prior credences about the probability of the evidence, given the hypothesis.

2. Total Probability/Jeffrey Conditionalization

To see more precisely how this works, let's introduce an important law of probability—the so-called **law of total probability**. There are two different ways we can write this.

First, imagine that you have a *partition* of propositions. (Remember: this is a set of propositions that's mutually exclusive and jointly exhaustive.) Let this partition be $A_1, ..., A_n$. Then the probability of any proposition B is:

$$c(B) = \sum_i c(A_i \cap B).$$

Then, given what we showed above, about the probability of a conjunction, this can be written alternatively as:

$$c(B) = \sum_i c(B \mid A_i) \cdot c(A_i).$$

So: the probability of B (for any B), is the sum over the probabilities that B is true, *given* that each A_i is true, then weighted by your credence that A_i itself is true.

Why is this useful with respect to Bayes' rule/conditionalization? Well, because it allows us to write Bayes' rule yet another way:

$$c(B \mid A) = \frac{c(A \mid B) \cdot c(B)}{c(A \mid B) \cdot c(B) + c(A \mid \neg B) \cdot c(\neg B)}.$$

And this is probably the most useful version of the rule when it comes to *applying* Bayes' rule.

Challenge Question. Imagine you're entertaining two hypotheses about a coin—namely, that it's fair (H) and that it's biased 75% towards heads ($\neg H$). You've flipped the coin once, and observed it land on heads. What's the probability that the coin is fair, given that you've made this observation. Imagine you start fifty/fifty between the two hypotheses.

Incidentally, the law of total probability allows us to state a *generalization* of conditionalization, too. To see what it is, first notice something important about conditionalization—it tells you how to change your credences when you learn a proposition *with certainty*. Arguably, however, there are very few learning experiences where we become completely certain of the truth of some proposition. For example, suppose I tell you that a certain die landed on an even number. Do you thereby become *certain* that it landed on 2, 4, or 6? In some cases, the answer is 'Plausibly not'. After all, maybe you think I have bad eyesight, or can't remember what an even number is, etc. In such cases, conditionalization seems to fall silent. And yet—given that you think I'm not *completely* unreliable in my report—it still seems like my telling you the die landed even should at least cause you to *raise* your credence that it did so.

The great Bayesian philosopher Richard Jeffrey noticed this, and proposed the following generalization of conditionalization, to handle these kinds of cases. Imagine $A_1, ..., A_n$ is again a partition of propositions, and you have a learning experience which causes your credences in the A_i to shift from $c(A_i)$ to $c^*(A_i)$. Then:

> **Jeffrey Conditionalization.**[1] After a learning experience of the kind described, your new credence in any proposition B should be:
>
> $$c^*(B) = \sum_i c(B \mid A_i) \cdot c^*(A_i).$$

So, your new credence in any propositition B is equal to the sum of your old conditional credence in B given A_i, weighted by your new credence in A_i. In effect, this just *is* the law of total probability. Notice that conditionalization is just the special case of Jeffrey conditionalization in which some partition element A_i gets the new probability $c^*(A_i) = 1$.

Notice also two other things about the update rules we've looked at. First, in order for the rules to work, it needs to be that learning experiences don't affect your *conditional* credences. (This is called the *rigidity* property.) Is this plausible?

Additionally, notice that, in the case of both conditionalization and Jeffrey conditionalization, the evidence you might learn forms a *partition.* Is *this* plausible in all cases?

If you answered 'Yes' to the latter, then momentarily think back to the third problem set, where I asked you if the following three conditions on evidence are plausible.

- **Factivity**. If ϕ is part of your evidence, then ϕ is true.

[1] Jeffrey himself called this rule *probability kinematics.*

- **Positive Introspection**. If ϕ is part of your evidence, then it's part of your evidence that ϕ is part of your evidence.
- **Negative Introspection**. If ϕ is *not* part of your evidence, then it's part of your evidence that ϕ is not part of your evidence.

As it happens, these three conditions are what we need to guarantee that evidence forms a partition. But maybe you didn't think they were all plausible...

3. Chance

Let's now change gears. So far in this unit, we've been focused on the **subjective** interpretation of probability—viz., as **rational credence**. We also said, however, that there's an **objective** interpretation of probability, which philosophers refer to as **chance**. What, then, is chance? And how does it relate to rational credence?

Clearly, there has to be *some* relation between chance and rational credence. For example, imagine you knew that the chance a certain could would land heads, if tossed, was .5. But suppose also that your rational credence that it'd land heads, if tossed, was only .25. There's something very weird about this—prima facie, it seems like your credences should *match* the chances.

Indeed, David Lewis (again!) thought that—whatever kind of thing chance is—it has to be something that *guides* credence. As he himself put it: "Don't call any alleged feature of reality 'chance' until you've shown that you have something, knowledge of which could constrain rational credence". Thus, let's now try to spell out this idea more precisely. To do so, we'll need to introduce a bit notation.

Let $\mathcal{W} = \{w_1, ..., w_n\}$ be the set of all possible worlds (To keep the math simple, we're pretending there are only finitely many worlds.) Now, for each possible world, w, let $ch_{w,t}$ be the *chance function* at time t. Intuitively, this function is a probability function which gives the chance of any proposition A, at w, and at time t.

Now, let Ch be a definite description for 'the objective chances at time t'. Formally, Ch is a function that takes any possible world, w, and maps it the function $ch_{w,t}$. Given this function, we can form a *proposition*—namely, the proposition that the chances at t are given by the function $ch_{w,t}$. To do this, we collect together all the worlds w, at which the t-chances are given by that function:

$$\langle Ch = ch_t \rangle = \{w \in \mathcal{W} : Ch = ch_{w,t}\}.$$

Now we can state a first-pass-principle, saying how chance should contain credence.

Miller's Principle. For any proposition A:

$$c(A \mid \langle Ch = ch_t \rangle) = ch_{w,t}(A).$$

Intuitively, here's what this principle says: conditional on the proposition that the t-chances are given by the function $ch_{w,t}$, your credence in any proposition A should match the chance of A, according to $ch_{w,t}$. Or, to frame things in terms of conditionalization: if you were to *learn* the proposition $\langle Ch = ch_t \rangle$, then your credences should match the chances, according to $ch_{w,t}$.

This principle is a good start; but it can't possibly be right. Why? Well, imagine that c is your *current* credence function; imagine that $Ch = ch_t$ is the proposition that the chances *yesterday at noon* were given by the function $ch_{w,t}$; and imagine that at midnight you saw a certain coin land heads—let A be the proposition that it did so. Then, chances yesterday at noon might say that A has chance .5; but since you *saw* the coin land heads, your credence that it did so should be 1!

Faced with cases like this one, David Lewis famously argued for a version of the principle above, that doesn't commit us to weird upshots like this one. To state it, let's first introduce a definition: let c_0 be your *ur-prior credence function*—i.e., the credence function you have before you've received *any* evidence at all (in life!). Let E be some evidence. And let's assume that you update your credences by conditionalization. Then, if you were to learn E at the beginning of your epistemic life, your new credence in any A is: $c_E(A) = c_0(A \mid E)$. Now we can state Lewis's principle, about chance and rational credence:

> **Principal Principle**. Let c_0 be an ur-prior credence function, let A be any proposition, let E be an evidence proposition, and let $\langle Ch = ch_t \rangle$ be the proposition that the t-chances are given by the function $ch_{w,t}$. Then, if E is *admissible* with respect to $\langle Ch = ch_t \rangle$, we have: $c_0(A \mid E \wedge \langle Ch = ch_t \rangle) = ch_{w,t}(A)$.

The "admissibility" clause here is crucial. What is an *admissible* piece of evidence? As Lewis thinks of it, it's a piece of evidence that doesn't itself contain any information about the present chances. For example, imagine that you look into a crystal ball, and see the coin land heads tomorrow. Then, this piece of information is *inadmssible* with respect to the chance of the coin landing heads tomorrow. In contrast, if E is the proposition that you had corn flakes for breakfast yesterday, then this is perfectly admissible.

4. The Big Bad Bug

Lewis called the principle above the *Principal Principle* because, as he put it, it tells us *everything we know* about the concept of objective chance. What kind of thing could chance *be*, in other words, unless it wassomething such that, if you knew it, it would guide your (rational) credences? That seems right, if you think about it.

This is an important part of Lewis's philosophical methodology: if you want to know *what* something is, then it's a good idea to start by asking 'Well, what kind of *role* does that thing play in our lives?' That said, Lewis also had a grand, metaphysical vision about how the world works known as *Humean Supervenience.* It'd take a whole class to explain *exactly* what this vision is. But here is how Lewis describes it, in a beautiful (and famous!) passage:

> Humean supervenience is named in honor of the great denier of necessary connections. It is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another… We have geometry: a system of external relations of spatio-temporal distances between points. Maybe points of spacetime itself; maybe point-sized bits of matter or aether of fields; maybe both. And at those points we have local qualities: perfectly natural intrinsic properties which need nothing bigger than a point at which to be instantiated. For short, we have an arrangement of qualities. And that is all. There is no difference without difference in the arrangement of qualities. All else supervenes on that.

The idea, then, is that "nomic concepts"—like chance, laws of nature, etc.—all "supervene" on the arrangement of matter, past, present, and future. For example, in the case of chance, Lewis's view is (very roughly!) a frequentist one: we say that a certain coin has a 50% chance of landing heads because, of all the times the coin is ever flipped, 50% of those flips come up heads—the chances supervene on the "categorical" facts about how the coin lands.

Sadly, this view about chance turns out to be *inconsistent* with Lewis's own Principal Principle—a fact Lewis himself was well aware of. Lewis spent more than a decade trying to come up with a solution to this problem. But it's not at all clear that he succeeded. Indeed, the search continues for many people—they want to find a principle *like* the Principal Principle that isn't inconsistent with Humean Supervenience. Maybe you can think of one…

5. Generalizing the Principal Principle

One last thing. You'll notice that, although when we stated the Principal Principle (or Miller's Principle), the proposition $\langle Ch = ch_t \rangle$ referred to a chance function, nothing really *commits* us to this interpretation. Abstractly, a principle like the Principal Principle just says you should *defer* to a certain probability function. And we can interpret this function in different ways.

For example, here's another interpretation. The philosopher of science Bas van Fraassen gave the following principle:

> **Reflection**. Let p be your prior credence function, and let $\langle P = p^* \rangle$ be the proposition that your future credences are given by the function p. Then, for any A:
>
> $$p(A \mid \langle P = p^* \rangle) = p^*(A).$$
>
> If you know your future credence in A is going to be $p^*(A)$, then you should have that credence right now.

PHIL 210/510: Formal Methods Yale, Spring 2025
Lecture 18: Probabilities of Conditionals Dr. Calum McNamara

1. Left-overs: Generalizing the Principal Principle

Last time, we discussed Lewis's *Principal Principle*. You'll notice that, although when we stated the Principal Principle (or Miller's Principle), the proposition $\langle Ch = ch_t \rangle$ referred to a *chance* function, nothing really commits us to this interpretation. Abstractly, a principle like the Principal Principle just says you should *defer* to a certain probability function. And we can interpret this function in different ways.

For example, here's another interpretation. The philosopher of science Bas van Fraassen gave the following principle:

> **Reflection**. Let p be your prior credence function, and let $\langle P = p^* \rangle$ be the proposition that your future credences are given by the function p. Then, for any A:
>
> $$p(A \mid \langle P = p^* \rangle) = p^*(A).$$
>
> If you know your future credence in A is going to be $p^*(A)$, then you should have that credence right now.

Notice that, if for each value p^* might take, we have $p(A \mid P = p_i^*)$, for $i = 1, ..., n$, then, using the law of total probability, we can re-state Reflection equivalently:

$$p(A) = \sum_i p(P = p_i^*) \cdot p_i^*(A).$$

Thus, in effect, Reflection says that your current credences should be your *expectation* of your future credences. That seems completely right. (For those of you with some familiarity with probability theory, effectively, Reflection tells you that updating should form a martingale process.)

Challenge Question. If we assume your evidence forms a partition, then we can prove that the "expectational formulation" of Reflection is equivalent to the claim that you should udpate by Conditionalization. Can you prove this?

The general point we want to take away from this is that, really, the Principal Principle and Reflection are part of a very large class of principles, called *expert deference principles*.

2. Probabilistic Independence

Anyway, let's now turn to today's topic. In order to get to today's key results, I need to introduce you to one more concept from probability theory—the concept of *independence*.

Intuitively, two propositions are probabilistically independent just in case learning that the one is true gives you no evidence about the probability of the other. For example, intuitively, learning that a die was rolled and landed on an even number is probabilistically independent of whether the Mets will win the World Series this year.

Formally, we can define probabilistic independence in terms of conditional probabilities. Two propositions A and B are independent just in case:

$$p(B \mid A) = p(B).$$

So, A and B are independent just in case the probability of B, conditional on A is the same as the probability of B alone.

You may have also seen the following definition of independence:

$$p(A \wedge B) = p(A) \cdot p(B).$$

(In words: the probability of A and B is the *product* of the probability of A and the probability of B.) On the latest problem set, you're asked to *derive* this definition from the previous one. Good luck!

3. Probabilities of Conditionals

We can now finally turn to today's main topic—probabilities of *conditionals*.

In particular, what we're going to think about are the probabilities of indicative and subjunctive conditionals. Since—at least on the semantic views we've explored—these are propositions, they should have well-defined probabilities. And indeed, people have very strong intuitive ideas about what these probabilities should be.

Let's start with an example. Suppose I'm about to roll a fair, six-sided die, and I say:

(1) If the die lands on an even number, then it will land on 2.

How confident should you be of this sentence? Here, almost everyone says your credence should be $1/3$. Why's that? Well, plausibly because $1/3$ is also the *conditional probability* of getting a 2, *given* that I roll an even number.

Intuitions like this are surprisingly robust. (In fact, there have been empirical studies, which show that intuitions like the one just mentioned are *very* widespread indeed.) For example, consider yet another sentence:

(2) If I flip this fair coin, then it'll land on heads.

How confident are you of *that* sentence? Intuitively, the answer is $1/2$. And once again, that's just the conditional probability of getting a heads, given that the coin is flipped.

Intuitions like these motivate a general principle, about the probabilities of conditionals. It was first introduced by Stalnaker (1970), from whom we get the name:[1]

> **Stalnaker's Thesis**. Let $A \rightarrow B$ be an indicative conditional, and let p be a rational credence function. Then:
>
> $$p(A \rightarrow B) = p(B \mid A),$$
>
> provided this is well-defined.

So, Stalnaker's Thesis says: *in general* your credence in an indicative conditional $A \rightarrow B$ should match your conditional credence in its consequent, given its antecedent.

There's a corresponding principle for subjunctive credences. To get a feel for it, suppose I say the following: 'Yesterday at noon, I *didn't* flip this fair coin. But:

(3) If I had flipped the coin, it would've landed heads.'

This sounds quite a lot like Stalnaker's Thesis. But we clearly can't have that your credence in the subjunctive matches your conditional credence in the antecedent, given the consequent. That won't work because you give the antecedent credence 0. Thus, the corresponding principle is this (named after Skyrms (1980)):

> **Skyrms's Thesis**. Let $A \mathrel{\Box\!\!\rightarrow} B$ be a subjunctive conditional, and let p be a rational credence function. Then:
>
> $$p(A \rightarrow B) = ch_t(B \mid A),$$
>
> where ch_t gives the chances just before the antecedent takes place (and provided this is well-defined).

So, Skyrms's Thesis says: your credence in a subjunctive conditional should match the conditional *chance*, at a relevant time, of the consequent given the antecedent. Intuitively, that also seems right.

4. Lewisian Triviality

Both Stalnaker's Thesis and Skyrms's Thesis seem intuitively compelling. Indeed, some philosophers have regarded them as *obvious*—barely even in need of justification.

In his 1976, however, David Lewis showed that Stalnaker's Thesis in particular

[1] Stalnaker's thesis sometimes goes by other names in the literature: 'the thesis', 'the equation', and sometimes 'Adams thesis'. (For what it's worth, I think the latter is a misnomer—but I won't get into the reasons for that.)

faces very serious challenges. (It's easy to extend Lewis's result to Skyrms's Thesis, and we'll do so in a moment.) Indeed, the philosopher Branden Fitelson once told me that, historically speaking, Lewis's results were *so* damning, that they effectively brought an end to the study of the probabilities of conditionals for almost 40 years.

To see how Lewis's results work, we only need three assumptions. The first is that you satisfy the probability axioms; the second is that you update your credences by conditionalization. (Those are the standard Bayesian assumptions, so they're hardly worth questioning. We'll assume them henceforth.) The third assumption is that Stalnaker's Thesis should hold for *all* rational credence functions—since the thesis is supposed to be a normative thesis, that also seems right.

Now consider the following. Choose some particular $A \to B$. Then:[2]

$$\begin{aligned}
p(B \mid A) &= p(A \to B) && \text{(ST)}\\
&= p(A \to B \mid B) \cdot p(B) + p(A \to B \mid \neg B) \cdot p(\neg B) && \text{(Total)}\\
&= p_B(A \to B) \cdot p(B) + p_{\neg B}(A \to B) \cdot p(\neg B) && \text{(Cond)}\\
&= p_B(B \mid A) \cdot p(B) + p_{\neg B}(B \mid A) \cdot p(\neg B) && \text{(ST)}\\
&= p(B \mid A \wedge B) \cdot p(B) + p(B \mid A \wedge \neg B) \cdot p(\neg B) && \text{(Cond)}\\
&= 1 \cdot p(B) + 0 \cdot p(\neg B) && \text{(Ratio)}\\
&= p(B)
\end{aligned}$$

So, what we've established, then, is that, if Stalnaker's Thesis holds for all rational credence functions, then A and B must be probabilistically independent: $p(B \mid A) = p(B)$. And that's the case for *any* A and B. But of course, that's absurd! In fact, Lewis himself shows that this can hold only if every rational credence function gives positive probability to only two possible worlds!

Another way to see the absurdity is the following. Consider: 'If the die lands on an even number, then it'll land on 2' and 'The die lands on 2'. Intuitively, your credence in the former should be $1/3$ (as we said above), and your credence in the latter should be $1/6$. But Lewis's results imply that you must assign these sentences *exactly the same credence*. And that goes for any indicative conditional you can think of.

We can extend arguments like Lewis's to Skyrms's Thesis. To do so, we only need one extra assumption—namely, that you should satisfy the *Principal Principle*. Then we have:

$$\begin{aligned}
ch(B \mid A) &= p(A \to B) && \text{(Skyrms)}\\
&= ch(A \to B) && \text{(Principal Principle)}\\
&= ch(A \to B \mid B) \cdot ch(B) + ch(A \to B \mid \neg B) \cdot ch(\neg B) && \text{(Total)}\\
&= ch_B(A \to B) \cdot p(B) + ch_{\neg B}(A \to B) \cdot ch(\neg B) && \text{(Cond)}\\
&= ch_B(B \mid A) \cdot p(B) + ch_{\neg B}(B \mid A) \cdot ch(\neg B) && \text{(Skyrms)}\\
&= ch(B \mid A \wedge B) \cdot ch(B) + ch(B \mid A \wedge \neg B) \cdot ch(\neg B) && \text{(Cond)}\\
&= 1 \cdot ch(B) + 0 \cdot ch(\neg B) && \text{(Ratio)}\\
&= ch(B)
\end{aligned}$$

This result was first proved by Robbie Williams (2012). (Although "proved" is a bit of a stretch here—the result is really just a straightforward corollary of Lewis's results.) It shows that, if Skyrms's Thesis holds, then for any A and B, the *chances* of A and B must be probabilistically independent. But again, that's absurd.

Results like these are often known as *triviality* results, because they show, not that something like Stalnaker's Thesis is inconsistent—in the sense of being logically contradictory—but because they show that it can hold only in "trivial" cases. Again: the triviality here results because, if Stalnaker's Thesis is completely right—then there are only two epistemically possible worlds. And that's obviously false.

5. More on Triviality

One of the amazing things about Lewis's triviality results is that they're actually very simple. They make use only of tools from Bayesian epistemology and probability theory that you've learned in the last two weeks. Nevertheless,

[2] In this derivation, I use the following shorthands: 'ST' stands for 'Stalnaker's Thesis', 'Total' stands for 'Law of Total Probability', 'Cond' stands for 'Conditionalization', 'Ratio' stands for 'Ratio formula for conditional probability'.

they show something *extremely* surprising. (By the way, for anyone who's worried about using formal tools in their own work, Lewis's results should quell your fears: you don't need *that much* mathematical background to prove things that are extremely deep, surprising, and beautiful.)

Nevertheless—like all mathematical results—Lewis's triviality results rely on some assumptions. One of them, for example, is that Stalnaker's Thesis should hold for *all* credence functions in a class of credence functions "closed" under conditionalization. But maybe that's not right—maybe we should think that conditionalization isn't *always* the right way to revise your credences. (Incidentally, in 1986, Lewis proved that the result still goes through if you assume the class of rational credence functions is closed under Jeffrey conditionalization.)

If we reject this assumption, then, maybe Lewis's triviality results are not so worrying. Not so fast! Since Lewis's original paper, philosophers (and mathematicians) have showed that you can prove results like Lewis's in *lots* of different ways, using *lots* of different assumptions. The triviality results, it seems, are *very* difficult to escape.

For example, consider a result proved by our friend Alan Hájek (1989) (in his PhD dissertation). In Lewis's result, the quantification is over probability functions: suppose *all rational credence functions* satisfy Stalnaker's Thesis. In Hájek's case, in contrast, the quantification goes over indicative conditionals themselves: $A \rightarrow B$. What he shows is that, in any model with countably many worlds, we cannot generally find a probability function p such that, for all propositions $A \rightarrow B$, $p(A \rightarrow B) = p(B \mid A)$.

Hájek's result is quite mathematically sophisticated. (Hájek himself had training as a statistician, before he turned to philosophy.) But we can illustrate his result with a very simple example. Thus, let $\mathcal{W} = \{w_1, ..., w_6\}$ be a set of six worlds, where each world w_i corresponds to a world where a certain fair die lands on i, after 1 toss. So: w_1 is the world where the die lands on 1; w_2 is the world where the die lands on 2; etc. Now, since the die is fair, suppose you give equal credence to each of the worlds $w_1, ..., w_6$, so $p(w_i) = 1/6$, for $i = 1, ...6$. Next, consider the following:

(4) If the die doesn't land on 1, it will land on 2.

What's your credence in this sentence? Stalnaker's Thesis tells us it should be $1/5$, since that's the probability of the consequent, given the antecedent. But notice that there can be no *proposition*—i.e., subset of the worlds $w_1, ..., w_6$—whose credence in $1/5$ in this case. After all, since you give probability $1/6$ to each world w_i, any set of these worlds must credence equal to some multiple of $1/6$. And no multiple of $1/6$ is equal to $1/5$! So, with respect to the sentence just above, Stalnaker's Thesis must be wrong.

Effectively, what Hájek's result shows is that, for *any* model with (only) countably many worlds, we can always construct examples like this one, given the choice of a probability function. We thus have another reason to doubt Stalnaker's Thesis.

6. Reactions

Lewis's triviality results show, very clearly, that some of our cherished assumptions about indicative conditionals have to go. Maybe, for example, we have to jettison our intuitions about the probabilities of conditionals; or maybe we have to go for something even more radical.

One extreme reaction, for instance, is noted by philosophers like Dorothy Edgington (1995). Edgington—and these days, many others—think that the triviality results should lead us to question whether conditionals are really *propositions* at all. Edgington thinks instead that conditionals are merely "expressive" of our conditional credences—they don't in other words, have truth conditions. That's a hard view to maintain for a number of reasons—but not an uncommon one. (For example, if Edgington is right, then how can it make sense to say things like 'The die landed even, and if the coin was flipped it landed heads'. According to Edgington, that sentence is a conjunction of a proposition and a non-proposition. What?!)

More recently, there's been a movement which says that we can get around the triviality results by embracing a *contextualist* view about conditionals. (Indeed, I myself have contributed to this literature.) Using this view to combat the triviality results isn't as straightforward as it might seem, however…

PHIL 210/510: Formal Methods — Yale, Spring 2025
Lecture 19: Decision Theory I — Dr. Calum McNamara

By this point in the course, you're well acquainted with the **Bayesian** view in epistemology. As we heard, Bayesians think that beliefs are *degreed* attitudes, called **credences**; and if you're rational, these credences obey the probability axioms.

This is a standalone view. But historically, it grew out of attempts to describe betting behavior. Indeed, we saw an example of this in our first session on probability—namely, the *problem of the points*. And we also saw that one prominent attempt to justify the Bayesian view appeals to betting behavior too (viz., the Dutch book argument).

So, historically, there's been a close connection between the Bayesian view of rational belief, and *betting behavior*. More broadly, there's been a close connection, historically speaking, between the Bayesian view, and theories of *rational action*.

Today, we're going to start thinking about rational decision-making. And we'll see that the Bayesian view plays an important role in this discussion.

1. Expected Utility Theory

Recall our notion of an **expectation**. The expectation of a random variable X is a probability-weighted average of X's possible values:

$$\mathbb{E}_p[X] = \sum_w p(w) \cdot X(w).$$

As an example, let's calculate an expectation:

Challenge Question. Let X be a random variable corresponding to the definite description, 'The outcome of a fair die roll'. X's possible values are $1, \ldots, 6$, and the probability of each outcome is $1/6$. What is the *expected value* of X?

We can use this idea of the expectation of a random variable to say how you should *choose* between different options.

To see how, imagine you have a (finite) collection of **options** $O_1, \ldots, O_n$. Intuitively, you can think of the O_i as standing for different actions that you can perform, or different strategies that you might take. Collectively, the collection of the O_i, written '$\mathcal{O}$', is called your **decision problem**. In the context of EU theory, we can think of options as *functions* (random variables) from worlds to real numbers. The real number outputted by a random variable O, when given w as an argument, $O(w)$, is the amount **utility** you'd receive by choosing O at w.

Now, the orthodox theory of rational decision-making is **expected utility theory** (or 'EU theory'). It says that when you're choosing from among a collection of options, you should choose the option O that maximizes expected utility, defined as follows:

$$\text{EU}(O) = \sum_w p(w) \cdot O(w).$$

Thus, you should choose the option that you *expect* to have the best results. (Notice that the right-hand side is just your expectation of a random variable. So EU theory says that you should choose the random variable that does best, in expectation.)

Let's look at an example. Suppose you can choose weighted to take a certain bet, O_1, or decline it, O_2. The bet pays \$1 if a fair coin lands heads, but loses \$2 if the coin lands tails. Assuming you value dollars linearly, with the obvious choice of units—a point we'll return to below—what option should you choose here.

Well, we have:

$$\begin{aligned}\mathrm{EU}(O_1) &= \sum_{w \in H} p(w) \cdot O_1(w) + \sum_{w \in \neg H} p(w) \cdot O_1(w) \\ &= .5 \cdot 1 + .5 \cdot -2 \\ &= -.5\end{aligned}$$

And:

$$\begin{aligned}\mathrm{EU}(O_2) &= \sum_{w \in H} p(w) \cdot O_2(w) + \sum_{w \in \neg H} p(w) \cdot O_2(w) \\ &= .5 \cdot 0 + .5 \cdot 0 \\ &= 0\end{aligned}$$

So you should choose the second option, to *decline*.

3. Justifying EU Theory: Long-run Arguments

EU maximization *sounds* sensible. But what we'd really like is some more formal way to *justify* it. How, then, can we do so?

One reason for maximizing EU is that it makes for good policy in the **long run**. To see this, we need to sketch two mathematical facts about probabilities: the strong law of large numbers, and the weak laws of large numbers. Both these facts concern sequences of independent, identically distributed trials—the sort of setup that results from repeatedly betting the same way on a sequence of coin tosses, for example. Both the weak and strong laws of large numbers say, roughly, that over the long run, the average amount of utility gained per trial is overwhelmingly likely to be "close" to the expected value of an individual trial.

More precisely: the **weak law of large numbers** states that, where each trial has an expected value of x, for any arbitrarily small real numbers $\epsilon > 0$ and $\delta > 0$, there is some finite number of trials n, such that for all m greater than or equal to n, with probability at least 1δ, your average payoff for the first m trials will fall within ϵ of x. In other words, if you were to repeat some decision problem over and over, then the average gain per trial is highly likely to become arbitrarily close to your expected value within a finite amount of time. So in the finite long run, the average value associated with a gamble is overwhelmingly likely to be close to its expected value.

The **strong law of large numbers** states, on the other hand, that, where each trial has an expected value of x, with probability 1, for any arbitrarily small real number $\epsilon > 0$, as the number of trials increases, your average payoff per trial will fall within ϵ of x. In other words, then, as the number of repetitions of a decision situation approaches infinity, the average gain per trial will become arbitrarily close to your expected value with probability 1. So in the long run, the average value associated with a gamble is *virtually certain* to equal its expected value.

Thus, long run arguments justify expected utility theory—it's said—because, in the long run your *actual* payoffs would, with probability 1, be proportional to your *expected* payoffs on any given trial. It makes sense, then, to choose an option in a decision problem that *maximizes* EU.

4. Justifying EU Theory: Representation Theorems

But then again, so what? In most cases, the decisions we face are one-off events. So it's cold comfort to say that, if you *were* to face the decision problem infinitely many times, you'd do best by following EU theory. That, you might think, is a pretty slight justification.

So let's look at another way EU theory can be justified. Unlike the long-run-style arguments, these new arguments are much more recent. The first argument of this form was given by Ramsey (1926). And similar arguments were later given by von Neumann and Morgenstern (1947) and Savage (1967).

The kind of argument I have in mind is known as an **argument from representation theorems**. The rough idea is that, rather than thinking of EU theory as something which *itself* has to be justified, we instead show that it follows from more basic premises. And then we give arguments for those premises instead. In particular, the goal of this kind of argument is to show that, if you have **preferences** between outcomes (worlds) that satisfy certain conditions, then you can always be represented *as if* you were choosing by maximizing EU. In this sense, then, the normative content of EU theory is *not* located in the claim that you should choose by maximizing EU. Rather, it's found in your more basic preferences.

Proving theorems of this kind is quite difficult. But to get a feel for how they work, let me introduce you to the **axioms of preference**, given by von Neumann and Morgenstern (1947). First, then, let $\succeq$ represent 'weak preference'. That is, $wsucceqv$ means you'd weakly prefer if w were the actual world to v being the actual world. Strict preference is then defined as $w \succeq v$ but not $v \succeq w$. And indifference is defined as $w \succeq v$ and $v \succeq w$. (In what follows, I'll write indifference as '$\sim$'.) Now, here are the axioms:

- **Completeness**. For all w, v, $w \succeq v$ or $v \succeq w$ (or both).
- **Transitivity**. For all w, v, x, if $w \succeq v$ and $v \succeq x$, then $w \succeq x$.
- **Continuity**. For all w, v, x, if $w \succeq v$ and $v \succeq x$, then there exists a probability $p \in [0, 1]$ such that $v \sim (p \cdot w) + (1 - p) \cdot x$
- **Independence**. For all w, v, x, if $w \sim v$, then for every probability $p \in [0, 1]$, $(p \cdot w) + (1 - p) \cdot x \sim (p \cdot v) + (1 - p) \cdot x$.

Completeness just says that every world features in your preference-ranking. Transitivity says that your preferences don't form "cycles". Independence says that, if you're indifferent between w and v, then you're indifferent between a gamble that gives you w with probability p, and x with probability $(1 - p)$; and a gamble that gives v with probability p, and x with probability $(1 - p)$. Continuity is really the only difficult axiom here. Here's roughly what it says: if you prefer w to v and v to x, then there exists a probability p such that you'll be indifferent between v and a gamble that combines w and x with probabilities p and $1 - p$, respectively. The rough idea is that, no matter how much you prefer w to v and v to x, we can always find a probability such that weighting p by this probability, and weighting x by $(1 - p)$ makes you indifferent between that gamble, and having v for sure.

I'll say more about Continuity in a moment. The important thing for us to note now is that, if you satisfy the axioms above—and your choices reflect your preferences—then you *will* always choose as if you were maximizing EU. Once again, then, representation theorem arguments seek to say that the normative content of EU theory is found in the above axioms. It's your *preferences* that can be rational or irrational; and EU theory is merely a consistency condition on your preferences.

5. Digression: Pascal's Wager

One important thing to note about the Continuity axiom is that it implies utility is *bounded* (above and below). In other words, there is some maximum, and minimum, value for the utility of a world. Now, with that in mind, let's consider a famous decision problem. Here's how it goes:

> God is, or He is not. But to which side shall we incline? Reason can decide nothing here. There is an infinite chaos which separated us. A game is being played at the extremity of this infinite distance where heads or tails will turn up... Which will you choose then? Let us see. Since you must choose, let us see which interests you least. You have two things to lose, the true and the good; and two things to stake, your reason and your will, your knowledge and your happiness; and your nature has two things to shun, error and misery. Your reason is no more shocked in choosing one rather than the other, since you must of necessity choose... But your happiness? Let us weigh the gain and the loss in wagering that God is... If you gain, you gain all; if you lose, you lose nothing. Wager, then, without hesitation that He is.

Here, Pascal is essentially proposing you the following gamble:

	God exists	God doesn't exist
Believe in God	Infinite happiness	Status quo
Don't believe in God	Misery	Status quo

What, then, does EU theory say about this gamble?

Many writers—a shocking number, in fact—have argued that EU theory says you should choose to believe in God in this case. (Indeed, that's effectively Pascal's argument.) After all, if we take the utility of infinite bliss to be ∞, we take the utility of misery to be some finite (or infinite) negative number (say, -100), and we take the utility of the status quo to be 0, then:

$$\begin{aligned} \text{EU(Believe in God)} &= p(\text{God exists}) \cdot \infty + p(\text{God doesn't exist}) \cdot 0 \\ &= \infty \\ \text{EU(Don't believe in God)} &= p(\text{God exists}) \cdot -100 + p(\text{God doesn't exist}) \cdot 0 \\ &= p(\text{God exists}) \cdot -100. \end{aligned}$$

Thus, the expected payoff of believing in God is infinite! But according to the representation theorem argument for EU theory, this argument *makes no sense*. That's because the Continuity axiom implies that utility is bounded. So there's no way the utility of God's existence can be infinite. Thus: Pascal's wager involves a violation of the Continuity axiom!

6. The Value of Information

Let me now introduce you to a very cool consequence of EU theory—the so-called *value of information* theorem. Roughly, the Value of Information Principle says that, in expectation, it's always at least as good to learn free information, before making a decision, as it is to make that decision straight away. Formally, if $E_1, ..., E_n$ form a partition, then:

$$\sum_i p(E_i) \cdot \max_j \sum_s p(s \mid E_i) \cdot O_j(s) \geq \max_j \sum_s p(s) \cdot O_j(s).$$

To see this, just notice that it follows, by the law of total probability, that the right-hand side of the inequality above can be re-written as follows:

$$\begin{aligned} \max_j \sum_s p(s) \cdot O_j(s) &= \max_j \sum_s p(s \mid E_i) \cdot p(E_i) \cdot O_j(s) \\ &= \max_j \sum_i p(E_i) \cdot \sum_s p(s \mid E_i) \cdot O_j(s). \end{aligned}$$

Now compare the last line here to the left-hand side to the foregoing inequality. The first is a maximum of averages, while the second is an average of maxima. The former can never exceed the latter, on general mathematical grounds. So EU theory says that learning new information *always* leaves you better off, in expectation.

7. Risk-aversion

Historically speaking, one of the chief criticisms of EU thoery is that it doesn't pay sufficient attention to an agent's attitude to *risk*. To illustrate what I mean by this, consider the following case—the now famous *Allais paradox*. Imagine you have a choice between the following gambles:

(1) \$5, 000, 000 with probability 0.1, or \$0 with probability 0.9,

(2) \$1, 000, 000 with probability 0.11, or \$0 with probability 0.89.

Presented with these options, most people report that they'd rather take the first. But now consider the following two gambles instead:

(1*) \$1, 000, 000 with probability 0.89, \$5, 000, 000 with probability 0.1, and \$0 with probability 0.01

(2*) \$1, 000, 000 with probability 1.

Unlike in the first case, most people say that they'd rather take the second gamble here, namely (2*). However, these preferences are inconsistent with the claim that you should maximize EU. Specifically, there is *no* assignment of utility values to dollars such that (1) has a higher expected utility than (2), but (2*) has a higher expected utility than (1*). But since these choices both seem intuitively rational, this looks like a problem for EU theory. What do you think?

PHIL 210/510: Formal Methods Yale, Spring 2025
Lecture 20: Causal vs. Evidential Decision Theory Dr. Calum McNamara

Today we're going to be looking at *generalizations* of expected utility theory. But first, we need to talk about two things left over from last time.

1. The Value of Information

EU theory has a very cool consequence—the so-called *value of information* theorem. Roughly, the Value of Information Principle says that, in expectation, it's always at least as good to learn free information, before making a decision, as it is to make that decision straight away. Formally, if $E_1, ..., E_n$ form a partition, then:

$$\sum_i p(E_i) \cdot \max_j \sum_s p(s \mid E_i) \cdot O_j(s) \geq \max_j \sum_s p(s) \cdot O_j(s).$$

To see this, just notice that it follows, by the law of total probability, that the right-hand side of the inequality above can be re-written as follows:

$$\begin{aligned} \max_j \sum_s p(s) \cdot O_j(s) &= \max_j \sum_s p(s \mid E_i) \cdot p(E_i) \cdot O_j(s) \\ &= \max_j \sum_i p(E_i) \cdot \sum_s p(s \mid E_i) \cdot O_j(s). \end{aligned}$$

Now compare the last line here to the left-hand side to the foregoing inequality. The first is a maximum of averages, while the second is an average of maxima. The former can never exceed the latter, on general mathematical grounds. So EU theory says that learning new information *always* leaves you better off, in expectation.

7. Risk-aversion

Historically speaking, one of the chief criticisms of EU thoery is that it doesn't pay sufficient attention to an agent's attitude to *risk*. To illustrate what I mean by this, consider the following case—the now famous *Allais paradox*. Imagine you have a choice between the following gambles:

(1) $\$5,000,000$ with probability 0.1, or $\$0$ with probability 0.9,

(2) $\$1,000,000$ with probability 0.11, or $\$0$ with probability 0.89.

Presented with these options, most people report that they'd rather take the first. But now consider the following two gambles instead:

(1*) $\$1,000,000$ with probability 0.89, $\$5,000,000$ with probability 0.1, and $\$0$ with probability 0.01

(2*) $\$1,000,000$ with probability 1.

Unlike in the first case, most people say that they'd rather take the second gamble here, namely (2*). However, these preferences are inconsistent with the claim that you should maximize EU. Specifically, there is *no* assignment of utility values to dollars such that (1) has a higher expected utility than (2), but (2*) has a higher expected utility than (1*). But since these choices both seem intuitively rational, this looks like a problem for EU theory. What do you think?

3. Dominance and Act-State Dependence

Like I said, risk-aversion has historically played an important role in criticisms of EU theory. However, there's been another kind of criticism that's been at least as important. And it stems from a principle called **dominance**.

Dominance is best illustrated with an example. Thus, suppose I offer you a bet that pays $10 if a coin lands heads, or $5 if it lands tails. You can either accept this bet or decline it. What should you do?

Obviously, you should accept the bet. After all, no matter how things turn out—no matter whether the coin lands heads or tails—you gain money by accepting the bet, whereas you gain nothing by declining. In short, then, we say that accepting the bet *dominates* declining.

More generally, one option O_1 **weakly dominates** another, O_2, iff $O_1(w) \geq O_2(w)$ for all possible worlds w, and for some some possible world w,

$O_1(w) > O_w(2w)$. Furthermore, O_1 **strictly dominates** O_2 iff $O_1(w) > O_2(w)$ for all w.

As it turns out, dominance is a *consequence* of EU maximization. That is, if O_1 dominates O_2, then O_1 must have greater expected utility than O_2 as well. You might even think that dominance is the *degenerate case* of EU-maximization.

But in fact, the situation surrounding dominance turns out to be more complicated than you might realize. To see why, let's look at an example.

> *Drunk Driving.* You've been drinking all day, and now it's time to return home to sleep it off. You can either drive your own car home, or take a taxi. You may or may not get in a car accident on the way home. But if you drive your own car home and get in an accident, then at least you'd have saved on the taxi fare. What, then, should you do: take a taxi? Or drive yourself?

To make things concrete, let's say that getting in an accident is worth $-\$1,000$ to you (maybe, for example, that's how much you'd expect to pay on medical bills), while the taxi fare you expect to pay is a flat \$10. Here, then, is a decision matrix, which summarizes your situation:

	$\neg C$	C
D	\$0	$-\$1,000$
T	$-\$10$	$-\$1,010$

In this matrix, D denotes the option that you *d*rive your own car home, and T is the proposition that you instead take a *t*axi. Meanwhile, C and $\neg C$ are the propositions that you get in a *c*rash, and don't get in a crash, respectively. So, in this case, D and T are your options, while C and $\neg C$ are the relevant states of the world.

Now, a straight forward application of EU theory seems to deliver the verdict that you should drive your own car home. After all:

$$\begin{aligned} \mathrm{EU}(D) &= p(C) \cdot u(\$0) + p(\neg C) \cdot u(-\$1,000) \\ &= p(C) \cdot 0 + p(\neg C) \cdot -1,000 \\ &= p(\neg C) - 1,000. \end{aligned}$$

Furthermore:

$$\begin{aligned} \mathrm{EU}(T) &= p(C) \cdot u(-\$10) + p(\neg C) \cdot u(-\$1,010) \\ &= p(C) \cdot -10 + p(\neg C) \cdot -1,010 \\ &= -10 + p(\neg C) - 1,000. \end{aligned}$$

Thus, it seems like driving your own car home leaves you better off, no matter whether you crash or not.

But of course, that's absurd. The issue here is that driving yourself home *makes* it more likely that you'll get in a crash, while taking a taxi makes this likely. More generally, EU theory doesn't take account of how the state of the world can *depend* on your choice of an option.

How, then, should we fix this?

4. Evidential Decision Theory

In the 1960s, this was an open question. But the philosopher Richard Jeffrey proposed a brilliant answer to it. Rather than using *unconditional* probabilities, he said, we should use *conditional* probabilities, when calculating EU. In particular, we should use the probability of various outcomes, *conditional* on your choice of an option.

To see what this means, we need to slightly change our initial formalism. Earlier (recall), we thought options as random variables: functions from worlds to real numbers. Now we're going to think of them as *propositions*—namely, the finest-grained propositions you believe you can *make* true by deciding. Thus, to *make* a proposition true is to perform the action which that proposition expresses. For example, if I want to make it true that I take a taxi, rather than drive myself home, I simply perform the action of taking the taxi. (We need to think of options in this way if they're going to be used when calculating EU—after all, the arguments to probability functions are propositions.)

We also need to introduce a function, u—your *utility function*—which maps worlds to real numbers. Intuitively, this function takes in a possible world, and maps it to a real number, $u(w)$, that represents how good things would be, for you, if that world were actual. For present purposes, the exact numbers we use are unimportant. So long as your preferences between worlds satisfy the rationality axioms we talked about last time, we can simply choose a world w to act as the zero-point (i.e., $u(w) = 0$), and we can choose another world v to act as the unit (i.e., $u(v) = 1$), and then all our numbers simply fall out of that. As we say, the utility function is unique up to *positive affine transformation.*

Now, here's what Jeffrey's theory says. Rather than choosing options by calculating EU, you instead choose options that maximize the following quantity—*evidential expected utility*:

$$\text{EEU}(O) = \sum_w p(w \mid O) \cdot u(w).$$

So, the idea here is: you should choose on option that maximizes EU, *conditional* on that option being chosen.

It's fairly easy to see that Jeffrey's theory delivers the right answer in the *Drunk Driving* problem. To see this, let's imagine that $p(\text{Crash} \mid \text{Drive}) = .8$, and $p(\text{No Crash} \mid \text{Taxi}) = .8$. Then:

$$\begin{aligned}\text{EEU(Drive)} &= p(\text{Crash} \mid \text{Drive}) \cdot -1,000 + p(\text{No Crash} \mid \text{Drive}) \cdot 0\\ &= .8 \cdot -1,000 + .2 \cdot 0\\ &= -800\end{aligned}$$

And the EEU of the option B—taking *b*oth boxes—is:

$$\begin{aligned}\text{EEU(Taxi)} &= p(\text{Crash} \mid \text{Taxi}) \cdot -1,000 + p(\text{No Crash} \mid \text{Taxi}) \cdot 0\\ &= .2 \cdot -1,000 + .8 \cdot -10\\ &= -208\end{aligned}$$

Since -208 is greate than -800, EDT says you choose to take the taxi—the right answer.

More generally, by calculating expectations using probabilities $p(w \mid O)$, EDT takes account of *correlations* between your choice of an option and the state of the world. Moreover, it proves a plausible restriction of the dominance principle. When is dominance reasoning valid? *Only* when your choice of an option is *probabilistically independent*—viz., uncorrelated—with what state of the world obtains. Call this the *evidential dominance principle.*

5. Newcomb

Jeffrey's EDT provides a neat solution to the problem of dominance, which plagues standard EU theory.

> *Newcomb.* In front of you are two boxes, labelled 'A' and 'B', respectively. You can take either just box A, or both boxes. Box B is transparent and contains what you can see to be a $\$1,000$ bill. Box A, in contrast, is opaque, and you don't know what's inside it. The contents of box A were determined yesterday, on the basis of a prediction I made about your behavior. If I predicted that you'd take only box A, then I put $\$1,000,000$ inside that box. But if I predicted you'd take both boxes, then I left the opaque box empty. Note that I'm a highly reliable predictor of your behavior. So, with that in mind: what is your choice?

In this case, EDT says that you should take just the opaque box. The reason is that the contents of that box *depend*, evidentially, on your choice of a particular option. More precisely, there's a strong *correlation* between choosing one box and getting the million dollars; and there's a strong correlation between choosing both boxes and getting only a thousand dollars. It follows that the

EEU of the option O—taking only *o*ne box—is:

$$\begin{aligned}\text{EEU}(O) &= p(\$1{,}000{,}000 \mid O) \cdot 1{,}000{,}000 + p(\$1{,}001{,}000 \mid O) \cdot 1{,}001{,}000 \\ &\approx 1 \cdot 1{,}000{,}000 + 0 \cdot 1{,}001{,}000 \\ &= 1{,}000{,}000.\end{aligned}$$

And the EEU of the option B—taking *b*oth boxes—is:

$$\begin{aligned}\text{EEU}(B) &= p(\$0 \mid B) \cdot 0 + p(\$1{,}000 \mid O) \cdot 1{,}000 \\ &\approx 0 \cdot 0 + 1 \cdot 1{,}000 \\ &= 1{,}000.\end{aligned}$$

Then, since EEU(O) is strictly greater than EEU(B), EDT says that you should choose the former.

But that *also* seems absurd—at least to me, and to many other philosophers. (With that said, however, the Newcomb problem remains controversial.) After all, by assumption, your choice of an assumption cannot *affect* what's in the opaque box. I made the prediction yesterday, and either put the money in the opaque box then, or didn't. So, even though your choice is *correlated* with the contents of the opaque box, you cannot *cause* the contents to be different than what they are: true, taking both boxes gives you good *evidence* that the opaque box is empty. But since the contents of that box are fixed, taking both gives you $\$1{,}000$ more than taking one box would *no matter what's inside the opaque box*.

6. Causal Decision Theory

What the Newcomb problem illustrates, I think, is that, sometimes, the correlations EDT pays attention to are mere *spurious* correlations—correlations that don't give you any evidence about what your choice of an option can *cause*. What we need, then, arguably is a *causal* decision theory—one that restricts the dominance principle, not to cases in which your choices are *probabilistically* independent of the state of the world, but where they're *causally* independent.

The first theory of this kind was given by Stalnaker (in a letter, to David Lewis). And it was developed in detail by Allan Gibbard and Bill Harper, Lewis, Brian Skyrms, and my advisor, Jim Joyce. Their idea—cashed out in different ways—was to use probabilities of *counterfactuals*, rather than conditional probabilities, to calculate expected utilities. That is, we have the following:

$$\text{CEU}(O) = \sum_{w} p(O \mathrel{\Box\!\!\rightarrow} w) \cdot u(w).$$

Roughly, this equation says that you should choose an option, O, whose outcome you expect to be best *if you were to choose* O. Then, since there's a tight connection between counterfactuals and causation—for example, when you want to know whether A caused B, you'll often ask yourself questions like 'Suppose A hadn't happened. Then, *would* B have happened?'—we can think of this theory as saying that you should choose an option that you expect to *cause* the best results.[1]

Causal decision theory (CDT) also gets the right results in the Newcomb problem—i.e., it tells you to take both boxes. To see why, think about Lewis's miracles account of similarity. That account says you should hold the past fixed, up until a moment shortly before the counterfactual's antecedent. So: since the contents of the opaque box are fixed by the time you take the opaque box, your credence in $O \mathrel{\Box\!\!\rightarrow} \$1{,}000{,}000$ is just your credence that I predicted you to take only the opaque box. Similarly for $B \mathrel{\Box\!\!\rightarrow} \$1{,}001{,}000$. And mutatis mutandis for the other counterfactuals. Thus, in this case, CDT says *dominance* reasoning applies—taking both boxes leaves you better off, no matter what.

[1] For more on the relation between subjunctive supposition and causation, and how this relates to CDT, see Lewis (1981), Joyce (1999), Hitchcock (2013), Kment (2023), McNamara (2023), and especially Gallow (2024). Note that not every philosopher agrees that Stalnakerian CDT is best interpreted as a *causal* theory. See, e.g., Dorr (2016) and Hedden (2023) for more on this.