UNM Philosophy/Math 415/515:
The History and Philosophy of Mathematics

Unpublished Lecture Notes

Written by Professor Paul Livingston
pmliving@unm.edu
https://paulmlivingston.academic.ws/

Philosophy/Math 415/515: Fall 2025
History and Philosophy of Mathematics

Notes: Week 1

We'll start today with something relatively simple and apparently straightforward, although like all good philosophical puzzles it leads quickly to complications. Let's pose an ordinary-seeming question: ***How many numbers are there?***

There are different ways one might think about answering this question. But it seems as if the most straightforward answer is: ***'infinitely many.'*** When we learn how to count, we learn the beginning of a series (1, 2, 3, 4 …) that we understand as having no end. There is no 'biggest' counting number, because we can always add one more. Nobody can complete the whole series by counting – our lives are not long enough! – but if we could somehow have the whole series all together, the whole set of counting numbers, the number of things in it wouldn't be ten or ten thousand or a quadrillion but rather 'infinitely many.' There would be "more" elements in this whole group, or set, than any (finite) number can count.

In this class, we'll be thinking a lot about what the word 'infinity' means and even whether we can give it a clear, mathematical meaning at all. For many centuries, as we'll see, philosophers and mathematicians thought you couldn't do much of anything with 'infinity,' that the idea was just inherently too paradoxical or confused, or that since nothing infinite 'actually' exists in space and time, there's no point to talking about it. A big part of the story of our class is how this changed, that is, how mathematicians came to have a completely well-defined set of tools for talking about, and even calculating with, infinity, and how this affects philosophical thought about the infinite and finitude. For today, we'll just do something simple with the idea of infinity that is interesting and connects to a lot of what we'll develop either.

We just asked "how many numbers are there?" and came up with the answer "infinity." However, the question as we posed it wasn't completely well-defined. For at first we considered just natural numbers, the "whole" or "counting" positive numbers 1, 2, 3, 4, 5 … But we remember from middle school math that there are lots of other types of numbers as well. There are, for instance, the *negative numbers*: -1, -2, -3… There are also the *rational numbers* – numbers that can be expressed as fractions – and the *real* numbers. The real numbers, we recall, are numbers that can be expressed as decimals, for instance 7.3, 2.92, 3.14159627… (or pi) and so forth. These include all of the rational numbers, but there are also many real numbers that are *not* rationals – i.e. that cannot be expressed as a fraction. The decimal expressions of real numbers can themselves go on forever, and in fact we define them so that they always do so. Thus 7.3 is really defined as 7.300000000…, and so forth. We also can identify the real numbers with points on a geometric number line, so that each geometric point corresponds to exactly one such point.

Now, what if we were to ask, not "how many **natural** numbers are there?", as we did before, but rather "how many **real** numbers are there?" The right answer, it seems, would once again be 'infinity'. If we

started counting real numbers, there is no way we could ever finish; again, for any group of real numbers we have so far counted, we can always find another that we haven't counted yet. Even worse, **between** any two real numbers, no matter how close, we can always find another. In fact, it seems as if there might be even "more" real numbers than natural ones. For consider all the real numbers within a certain domain, for instance all the real numbers between 1 and 2. Even within this circumscribed domain, it seems that there are *already* an infinite number of reals, just as there are an infinite number of points on the number line between 1 and 2.

We have said, then, that the number of naturals is 'infinity', and also that the number of reals is 'infinity.' But – and here we reach the really interesting question – what is the relationship between these two 'infinities?' Are they simply the same, so that we could say that there is exactly the same number of natural numbers as there is of real numbers, or are they in some sense different? *Is it possible, in fact, to say that there are "more" real numbers than natural numbers*, even though the number of both is infinite? We shall see that this is indeed possible, and even provable. This is our first introduction to the amazing world of multiple sizes, or orders, of infinity.

Before we get started, though, we need to get clearer on some of the notions that we're employing. For instance, we need to understand clearly what it means to say that there *are just as many things* of a certain type – call it type A – as there are of another type – call it type B. Suppose we're trying to determine whether or not the baseball team has the same number of players as the football team. How can we find this out? Well, in this case, we could just count, but in the case of comparing infinities, this option isn't open to us, since no one can count to infinity. Another way we could do it, though, is that we could try to find a ***one-to-one correspondence*** of baseball players with football players. That is, we could try to match up every baseball player with one and only one football player (we can put them in pairs, or draw lines between them). If we can do this, and every baseball player is matched to exactly one football player, and there's nobody left over, then we can conclude that the two teams have exactly the same number of members. If, on the other hand, we had assigned each of the baseball players to exactly one unique football player, and there were still some football players left over, we could then conclude that there were more football players than baseball players.

We can use this idea to give a general definition of equality between two groups, or sets. The idea is: two sets have the same number of members if, and only if, a one-to-one correspondence can be established between the two groups. If not – if we must leave out some members or if there is always at least one member of one of the sets left out – then the second set is bigger (has more members) than the first.

Now let's go back to numbers. We can now pose our question about the relationship between the set of naturals and the set of reals in a more precise way. We are really asking whether there is a *one-to-one correspondence* between the set of naturals and the set of reals – in which case there is the same number of members in both – or whether there can't be any such correspondence – in which case there must be more members in one of the sets than the other. Actually, we'll prove that *there is no possible*

one-to-one correspondence, and therefore that the set of reals is strictly speaking bigger than the set of naturals, even though both are "infinite."

Cantor's diagonalization argument (first form)

The proof, which was first demonstrated by Cantor during the end of the nineteenth century, is a "proof by contradiction" or *reductio ad absurdum*. This means that we'll start out by assuming a hypothesis – the opposite of what is to be proved – and show that this hypothesis leads to a contradiction, and so it cannot actually hold. In particular, let's start out with the hypothesis: there is the same number of naturals as there are of reals; that is, there is a one-to-one correspondence between naturals and reals, such that each natural is matched to just one real, and no reals are left out. To make things even easier, let's restrict the hypothesis to the reals between 0 and 1. The hypothesis – which will turn out to be false – is then that there is exactly the same number of natural numbers as there are reals between 0 and 1, or equivalently that we can establish a one-to-one correspondence between the two sets. It'll actually turn out, somewhat surprisingly, that even just between 0 and 1, there are ***already*** more numbers than in the whole domain of all the natural numbers, from 1 'up to' 'infinity.'

Proof: The hypothesis holds that there is a one-to-one correspondence between the naturals and the reals between 0 and 1. Suppose that there is such a correspondence. We can represent it by a kind of infinite table, where we correlate natural numbers, on the left side, with reals, on the right:

1 .1 2 1 2 1 2 . . .

2 .1 4 1 5 9 6 . . .

3 .4 4 4 4 4 4 . . .

4 .5 2 2 1 9 7 . ..

5 .9 3 4 3 6 8 …

.

.

.

Note that the list I've given is completely arbitrary. We're just assuming that there is **some** one-to-one correspondence; we don't care which numbers are actually matched to which.

The hypothesis just says that there is such a correspondence, but now we're going to show that this false. In particular, we're going to show that for every table we could possibly come up with, there's always going to be some real number that doesn't appear anywhere on the table. That is, there's always some real number "left over", even when we have correlated each of the natural numbers to

exactly one real. This is sufficient to show that that there are "more" real numbers – even just between 0 and 1 – than there are natural numbers at all.

How do we show this? Consider again the correspondence we suggested. (This could still be any correspondence whatsoever; we're just using it as an example. Let's mark some of the digits in the list of reals on the right, however, by putting them in bold:

1 .**1** 2 1 2 1 2 . . .

2 .1 **4** 1 5 9 6 . . .

3 .4 4 **4** 4 4 4 . . .

4 .5 2 2 **1** 9 7 . . .

5 .9 3 4 3 **6** 8 . . .

.

We've bolded the first digit of the first number, the second digit of the second, and so on.

Now we're going to create (or imagine creating) a new real number based on the list; let's call this new number **D** (for 'diagonal'). To get it, we simply take each of the bolded digits in sequence, and add one (or, if we find a 9, change it to 0) each time. Thus, in this case, we get:

D= .25527…

This is just what we got by, first adding one to the first digit of the first number, then adding one to the second digit of the second, and so forth. We could certainly imagine carrying on this procedure as long as we like, going on diagonally down the list, to generate the expansion of **D** as fully as we want.

Now – here comes the real point ***– we can show that D itself is nowhere on the list; we will never find it in the supposed one-to-one correspondence, no matter how long we look***. Why? Well, we can ask ourselves the following series of questions. First, is **D** equal to the first number on the list? No, it can't be. For it differs from this first number in at least one place, namely the *first* place. But is it equal to the second number on the list? Again, it can't be. For it differs from this second number in its *second* decimal place. And so on! We can go on and on through the list, and D will differ from each of the numbers we consider. That means that "even when the list is finished," at infinity, **D** still won't have appeared. There will be at least one real number – namely **D –** which is left over even when each of the natural numbers is correlated to exactly one real number.

That is, there are more real numbers, even just between 0 and 1, than there are natural numbers at all. By disproving the hypothesis that these two groups can be put into one-to-one correspondence, we have proven that there are "more" real numbers than there are natural numbers, in a strict and rigorous sense, even though both sets, or groups, are infinite.

This result was quite important to Cantor, but also actually revolutionary in its own right. For it showed that 'infinity' isn't just a single, undifferentiated thing, but that there are different degrees or sizes of infinity, and even that we can compare them and calculate with them. This seems to mean, as well, that we can consider different sorts of "infinite wholes" (such as the infinite wholes of all natural numbers and the real numbers between 0 and 1, which we just did consider) and consider mathematically how they are related. It turns out, as we'll see, that figuring out this calculus is deeply complicated, and leads to further paradoxes and problems in its own right. However, this first result is enough to get us started thinking about the complex and interesting issues of the infinite and sets.

Greek Mathematics and the problem of infinity

Having talked a little bit about the great discoveries of contemporary mathematics about the infinite, let's go back and trace the history of philosophical thought about the infinite briefly. We'll start with the Greeks and luckily, in talking about the history of rigorous thinking about the infinite, we can practically end there, since as we'll see there was actually very little progress made on this front between Aristotle, in the fourth century BC, and Cantor in the 19th. (For more detail on some of this history, though, please refer to Moore's book, *The Infinite*, especially chapters 1-3).

From the beginning of the historical record that reaches us, the Greeks were interested in issues of infinity and finitude. The Greek word for 'infinity' is *apeiron* – it can also be translated 'the unlimited' – and some of the earliest philosophical fragments to reach us discuss it. In particular, one of the oldest actual fragments of written philosophy, a fragment written by Anaximander sometime near the beginning of the sixth century BC, holds that:

> The principle and origin of existing things is *to apeiron*. And into that from which existing things come to be they also pass away according to necessity; for they suffer punishment and make amends to one another for their injustice, in accordance with the ordinance of time. (Quoted in Moore, p. 19)

Some of the first Greeks to think seriously about mathematics and its philosophical implications were the Pythagoreans, members of the mystical brotherhood founded by Pythagoras of Samos (born around 570 BC). The Pythagoreans combined mystical and religious beliefs about such topics as reincarnation and diet with a belief in the divinity of number and the possibility of explaining all things through numbers and ratios. Pythagoreans and Pythagoreas were associated with the famous "Pythagorean theorem" about right triangles, and the Pythagoreans were the first to discover the pitch ratios that define the eight-tone musical scale still in use today. Whereas Anaximander had thought of the apeiron, the infinite, as the source of all things, however, the Pythagoreans understood order and the limited – the peras – to define the nature of everything real and good. Based on this, the Pythagoreans appear to have thought – initially at least – that all numbers are either natural numbers (the whole numbers 1, 2, 3, etc.) or rational numbers, i.e. numbers that can be expressed as a ratio, or fraction, of two whole numbers. (These are numbers such as 1/3, 17/9, and so forth).

It was particularly disturbing for them, then, when a proof emerged that there are numbers and quantities that are not rational and cannot be expressed by any ratio or fraction of two whole numbers, but in fact can only be expressed by an infinitely long decimal (or other) expansion. The simplest of these is the square root of 2. The Pythagoreans knew that there must be such a quantity, since by the Pythagorean theorem the length of the diagonal of a 1x1 square is root 2. Nevertheless, some of the later Pythagoreans already had access to a simple and elegant proof that root 2 is not rational. Here it is:

Proof of the irrationality of √2 . This will, again, be a proof by contradiction or *reductio*. We thus start by assuming that there IS some rational expression of √2. That is, we start by assuming that there is a fraction – call it a/b – such that $a/b = \sqrt{2}$. We'll also assume that this fraction is reduced to lowest terms (this will be important later); that is, a and b have no common divisors except for 1. Squaring both sides, we have:

$a^2/b^2 = 2$ and $a^2=2\,b^2$

Now, what are a and b? We know that every whole number is either odd or even, so let's start by considering whether a and b are each odd or even. Since $a^2=2\,b^2$, a^2 itself must be even (since it is the double of another whole number). But if a^2 is even, then a itself is even, since every square of an odd is also odd and every square of an even is even (you can verify this yourself easily). So, we can conclude, $a = 2c$ for some whole number c. But now we can substitute back into the original equation above; we obtain $4c^2 = 2b^2$ or, dividing by 2 on both sides, $2c^2 = b^2$. Applying again the same principle we used before, though, we can now conclude that b^2 is even, and so that b is also. We have shown, then, *that both a and b must be even*. But this is impossible! Why? Because then the original fraction, a/b, has an even divided by an even. And a fraction that has an even over an even is *not reduced to lowest terms*. Thus our original assumptions is contradicted, and we have proven that there is no rational expression, in lowest terms, of √2.

This proof was known to the Pythagorean Hippasus, but it deeply troubled many in the brotherhood. There is a story, possibly apocryphal , that Hippasus was drowned at sea while attempting to publicize it. The proof was troubling because it showed that the rational, ordered, limited world that the Pythagoreans believed in was a myth – that infinity enters necessarily into the expression of even very simply defined and straightforward quantities.

Soon after Pythagoras, the philosopher Zeno came up with a series of paradoxes of motion and change that bear on the nature of infinity. The most famous and characteristic of these is the **paradox of the runner, or Achilles**. Suppose Achilles is supposed to run from a point A to a point B. In order to cross the distance, he must first reach the point exactly halfway between A and B; call this halfway point C. But to reach *this* point, he must first reach the point halfway between A and C; call this D. But to reach *this* point, he must first reach another halfway point … and so on and so on, to infinity. Thus, it seems that to cross even the smallest distance, Achilles must complete an infinite number of tasks. But how is

it possible to complete an infinite number of tasks in a finite amount of time? It seems that Achilles never reaches the destination, no matter how fast he runs.[1]

With considerations like those of the Pythagoreans and Zeno, the Greeks showed a deep interest in the nature of infinity and its relation to finite time and events. Plato himself was deeply interested in many of these issues, including the important issue of whether it is possible for an infinite whole or totality actually to exist. Plato seems to have thought of the ideas, or forms, as eternal existences – thus things that exist for an infinite amount of time – and in the dialogue *Timaeus* he theorized that the reality we perceive in time and space is a "moving image" of this eternity. Also, many of the classic Platonic problems about "forms" such as the idea of Being, Change, and Difference concern whether these forms are themselves limited or unlimited, a problem that Plato pursues in close connection with the problem of the One and the Many itself.

Nevertheless, despite this avid interest in the problem of infinity, there was not much progress made in thinking about the actual existence of infinity from Plato until more than two thousand years later, when Cantor first discovered the modern notion of a set. There are two reasons for this stagnation. The first is the influence of Aristotle. Aristotle considered the mathematical results known at his time, including the results of the Pythagoreans and the paradoxes of Zeno. But like the Pythagoreans themselves, he distrusted the idea of an actual infinity and thought that nothing infinite could actually exist. To support this claim and others, he drew a famous and influential distinction between *potentiality* and *actuality*. With respect to infinity, he held that infinity could only exist *potentially* and never *actually*. This means that to say that something is infinite – say the series of natural numbers or the extent of a line – just means to say that it *can* be extended indefinitely. There is no limit to either extent, which simply means that in extending it we will never run up against a limit. But this does not mean that there ever is an actually existing *completed* infinity: an "infinitely long line" or the infinite series of numbers, actually existing as a whole. This kind of reasoning led philosophers and mathematicians to think that there was no reason to examine the properties of 'actual infinities' such as the whole series of natural numbers or an infinitely long line, because (they thought, following Aristotle), there were no such things. As we've already seen, though, this line of reasoning was violently and completely rejected by Cantor in the late 19th century, who developed the theory of sets (both finite and infinite) and hence made it possible, once again, to think about infinities as actually completed wholes.

A second, related reason why philosophers did not think deeply about mathematical infinities for many centuries was the idea that there is something deeply incoherent or even unthinkable about the infinite itself. One source of this idea was theological. During the middle ages, Christian theological philosophy often identified the infinite with the Absolute, or God, by contrast with the finitude of human life and existence. Part of the suggestion was that it was just impossible for the character of the infinite to be known by a finite being such as man; so it seemed that there is, again, no point in trying to think about

[1] There is a closely related paradox concerning Achilles and the Tortoise.

actual infinities using mathematics. Another reason why mathematical infinity was thought to be incoherent was the existence of a variety of paradoxes about infinity, some of which were already known during the middle ages and Renaissance, and which seemed to show that the very idea of the infinite was incoherent. In addition to Zeno's well-known paradox, here are a few others:

Paradox of correspondence (or Galileo's paradox): Last time, we used the idea of a one-to-one correspondence to establish some interesting results about the relative size of infinities, but it also seems (or used to seem) as if the idea of one-to-one correspondence causes paradoxes for the very idea of the infinite itself. Consider, for instance, the following two sets: first, the set of all natural numbers **N**, and second, the set of all even numbers, **E**. It seems obvious that the first set is "bigger" than the second; after all, it contains everything that the second set contains, and much more besides (the odd numbers are included too). Yet it is nevertheless possible, in this case, to find a one-to-one correspondence between the two sets. All we have to do is match every number*, x,* with its double*, 2x*. Thus, 1 is paired with 2, 2 is paired with 4, 3 is paired with 6, and so forth. We therefore have a total one-to-one correspondence between **N** and **E**; hence, by last time's reasoning at least, they have the same number of elements. Yet it also seems apparent that **N** is bigger than **E**, since it includes everything that **E** includes and more. We can make the same paradox even worse (the form due to Gallileo) by running the same argument with respect to the one-to-one correspondence between N and the set of perfect squares (1, 4, 9, 16, 25..) For we can correlate each natural number with its square, and so obtain a one-one correspondence between the two sets.

Hilbert's Hotel: Suppose there exists a hotel with an infinite number of rooms. Such a hotel doesn't actually exist, but it seems conceivable that it could (at least, if the infinite itself is conceivable). It could even conceivably exist in a finite amount of space: just make sure that each floor is only half as tall as the one directly below it. Now suppose that one night, the hotel is full: that is, there is a guest in each and every room (1, 2, 3, 4, 5 …). Unfortunately, one more weary traveler shows up. Can the new guest be accommodated? In fact, they can! All we have to do is move each person in each room to the next one up; that is, we move the person in room 1 to room 2, the person in room 3 to room 4, the person in room 4 to room 5, and so on. At the end (or even the beginning) of the process, room 1 is now free and ready for occupation. Thus it seems that even though the hotel was full, it can still accept one more – and in fact, by the same reasoning, it could accept any finite number of new travelers! (For n new travelers arriving, just move the person in room 1 to room n+1, move the person in room 2 to room n+2, and so forth).

In fact, things are even worse than this. For suppose not just one or a finite number of new guests, but actually an *infinite* number of new guests shows up (perhaps they were all turned out of a similar hotel down the road). Even this infinite group can be accommodated in the already full hotel! How? We just move the person in each room, n, to room 2n. Thus the person who was in room 1 moves to room 2; the person who was in room 2 moves to room 4; the person who was in room 3 moves to room 6, and so forth. This frees up the whole infinite series of odd-numbered rooms, ready for the whole infinite group that has just arrived to move in.

Paradox of the lamp: Suppose there is a lamp that can be switched either on or off. Initially, at 2:00 it is off, but after a half hour (i.e. at 2:30), it will be switched on. Then, after another quarter of an hour (at 2:45), it will be switched off again. After an eighth of an hour, it will be switched on again; after another sixteenth of an hour, off again, and so forth, switching back and forth increasingly rapidly as the hour of 3:00 approaches. When the hour is up and 3:00 strikes, will it be on or off? Why?

These paradoxes are all predicated on the idea that there can be *actual* infinities, and for a very long time it was thought that these paradoxes, and others closely related to them, showed that this idea is just false and that the very idea of an actual infinity is incoherent. For example, Aristotle held that all infinites are only *potential* – for example, we can continue counting up higher and higher numbers as long as we like, never reaching a "highest" one, but this is the only sense in which the set of natural (counting) numbers is infinite or "unlimited".

It took hundreds of years, and many conceptual developments, to provide the basis for repudiating this assumption. But the essential breakthrough was made by Georg Cantor in the late nineteenth century, with his "discovery" or creation of the concept of a set, which he understood as a completed totality that can be either finite or infinite. This discovery or creation, in turn, played a starring role in the radical and massive transformation that philosophers and logicians soon undertook in the traditional concepts of infinity, mathematics, and even the definition of numbers themselves. We'll see more of this transformation, and begin to explore its still profound consequences, by looking more closely at logic and the foundations of set theory next week.

Philosophy 415: Spring 2023
History and Philosophy of Mathematics

Notes: Week 2

Euclid and the idea of an axiomatic system

Last week we considered some of the developments of Ancient Greek mathematics, but undoubtedly the culminating achievement of Greek mathematics is Euclid's system in his *Elements* (written about 300 BC). Euclid systematized the totality of mathematical knowledge at his time, beginning with geometry but proceeding through arithmetic and what would today be called number theory. He also put all of this in a rationally systematized form, as constructions or deductions (i.e., proofs) from a few principles stated clearly and universally at the beginning. What's important and interesting for us today about the *Elements* isn't just the math that's actually contained in it, but rather (maybe even more so) this form of a system based on a few stated premises adopted at the beginning – a (so-called) "axiomatic" system.

Let's look at how Euclid's system works and how he thinks about what it is to prove a proposition that holds *universally*. If we open the book, we find a series of definitions (e.g. of point, of line, etc.) and then, famously, five "postulates" (later we will call this kind of statement an *axiom* and think about the kinds of systems that are determined by specifying axioms in advance). The idea of an axiomatic system is the idea of a system of such axioms that are in some sense "starting points" and where *all the truths* follow *logically, i.e. deductively, from the axioms.*

Euclid's five postulates are as follows:

1. To draw a straight line from any point to any point.
2. To produce a finite straight line continuously in a straight line.
3. To describe a circle with any center and distance (radius).
4. That all right angles are equal to one another.
5. That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.

We might ask some questions about the axioms. How do we know they're true? Do they HAVE to be true? Which of the axioms are "existential" (i.e. say that something exists) and which are descriptive (i.e. tell us about something which is said by the other axioms to exist)? Are they "self-evident" or obviously true, and if so, why? Are they "true by definition" or do they inform us about something (if so, what?)

You might also notice that the fifth postulate looks different than the other four. It says that if we draw two distinct lines crossing another line at less than 90 degree angles on the same side, those two lines will eventually intersect; or (equivalently) that two lines both drawn at right angles to a given line, if distinct, are parallel and will never intersect (thus, this is sometimes called Euclid's "parallel postulate".) For a long time, it seemed to mathematicians that this fifth postulate might not have the same self-evidence as the others, and many thought that it could or should be derivable from the combination of

the others. It was only in the nineteenth century that alternative systems of (so-called) non-Euclidean geometry were worked out by Lobachevsky and Riemann, by dropping the parallel postulate and allowing lines parallel at one point either to converge (elliptic geometry) or diverge (hyperbolic geometry.); and later on, by Einstein, the actual structure of spacetime was shown to be Riemannian and thus non-Euclidean. This raises questions, once more, about the Euclidean axioms, since they were shown not only to be not *obviously* or necessarily true, but in fact not true at all of actual space.

Review of Propositional Logic and Symbolism

The revolution of formal and logical thinking about mathematics in the twentieth century rests on two pillars which began to develop in the late 19th century. The first of these – which we previewed last week and will learn more about soon – is Cantor's theory of multiplicities or sets. The second is Frege's propositional (and quantificational) logic. If you've taken first-semester 'symbolic logic,' the logic that you learned there is essentially Frege's logic. Although he used a somewhat different (and rather more arcane) notation, Frege had already worked out all the essential details of the system of logic that we still use today by the time he published his groundbreaking work, *Begriffsschrift* (or: "concept-writing") in 1879. We'll talk a bit today about what is new in Frege's logic relative to all the logic that had come before (going back to Aristotle) and why his discoveries were so revolutionary. But first, we need to review some basic symbolism and also cover some variants on the symbolism that you probably learned that you might run into in various (especially older) texts and treatments. This will be a review for most people, but it's worth going through (and for those who haven't taken symbolic logic: don't worry, we won't rely on this very much at all in the rest of the course, as we can state almost all of the main results and proofs informally as well.)

Here are some symbols and common variants:

Symbol (as in *The Logic Book*)	**Variants**		**Intuitive Meaning**
P*a*			***a* (individual name) has property P**
P*x*			**x (variable) has property P**
R*xy*			**x stands in relation R to y**
&	**.**	**∧**	**And**
∨			**Or**
~	**¬**		**Not**

⊃	→		**If…Then**
≡	↔		**If and only if**
(∀x)	(*x*)		**For all x… (universal quantification)**
(∃x)			**There exists an x … such that (existential quantification)**
A⊢B			**B is derivable from A (in a system)**
A⊨B			**A entails B (i.e. if A is true, B is true)**

We also have the rules of derivation, or inference, that are familiar to you if you've taken first-semester logic. (For a summary, see the inside cover of *The Logic Book*). If you haven't taken first-semester logic, it might be worth taking a quick look over these, but don't worry about them too much, since in practice we won't generally formalize our proofs and we'll just argue in natural language.

Do note, however, the difference between **derivation** and **entailment.** Although it took philosophers a long time to disentangle them, these are very different notions. Derivation, expressed by the 'single turnstile' (⊢) is the relation that holds between two sentences if one is derivable from the other *according to the inference rules of the system that we're using*. That is, if I say that A ⊢B, or that B *can be derived from* A, this just means that B can be gotten from A by using the derivation rules of the system (for instance PL, SL, or whatever system of inference we're using). *These rules are purely mechanical and just involve the manipulation of symbols*. Derivation is a purely **'syntactic'** relationship. Entailment, expressed by the double turnstile (⊨), by contrast, is a **'semantic'** notion. For it doesn't just depend on the system of inference, and it brings in the idea of truth. If I say that A *entails* B, or A ⊨B, this means that *if A is true*, then *B must be true also*. There is *no possibility where A is true but B is false*.

It can be hard to keep track of the difference between these notions because we generally try to design our inference systems so that they match actual entailment relations. We want it to be the case that if A ⊢B then A ⊨B, and vice-versa, and it will be -- ***if*** our system is "good." In other words, we *hope* our systems will be such that, if B can be derived from A according to the rules of the system, A *really does*

entail B (soundness) and conversely, that every time A *really does* entail B we can use the system to get A ⊢B (completeness). But though we might hope this, we don't know that our systems will be sound and complete in advance! In fact, we'll spend a lot of time on this issue later.

Frege's Revolution

Let's turn to some of the interesting historical and philosophical implications of Frege's breakthrough in logic. If you've taken first-semester logic, you've probably absorbed these implications in learning the system, but you might not have noticed (or been told) how much of a revolution they represent in logical methods. To see why, we need to start with the first 'formal logic,' which was Aristotle's logic of syllogistic forms. We won't study this (now antiquated) logic in detail, but there are two things to know about it. The first is that *Aristotle understood the logical structure of sentences as a subject-predicate structure*. This means that, for him, every assertive sentence breaks up into a "subject" – what the sentence is "about" – and a "predicate" – what the sentence says about that subject. Grammatically, the subject term is a noun or pronoun ("Socrates"; "I"; "justice", etc.) and the predicate is a verb ("runs"; "walks"; "is"), sometimes together with another noun, adjective, or adverb. You may have learned in grade school to break up sentences into the "subject" part and the "predicate" part.

The second thing to know about Aristotle's logic is that he breaks down the legitimate patterns of inference between sentences into a small number of specific forms, the Aristotelian syllogisms. In fact, there are just 19 valid patterns of inference, or valid syllogisms. Here's a typical example, the syllogistic form or "mood" traditionally called "Celarent":

No A's are B's
All C's are A's
--
No C's are B's

All of Aristotle's syllogistic forms involve two premises (the major and the minor) and a conclusion, and they all relate terms according to the modifiers "no," "some," or "all." This give us some insight into the possible valid forms of reasoning, but there are actually many, many more than Aristotle's system can handle, or even adequately represent.

By the 1870s, Frege was working with symbolisms to try to find new ways to express statements and truths, particularly in mathematics. He was interested in the idea that later became the basis for the project known as *logicism*: that truths and proofs in mathematics could be reduced completely to the rules of formal logic. In this way, it would be possible to distinguish completely rigorously between good and bad proofs, and it would also be possible to show what is really at the basis of all the branches of mathematics, from arithmetic and geometry to calculus. But he soon realized that neither the traditional logic of Aristotle, nor any of the systems available at the time, was up to the task. What was needed was a totally new symbolic language – what he called the *Begriffsschrift* or "concept-writing" – which could show clearly the real logical content of every meaningful sentence, and also establish the rules for their inference and derivation from one another.

There were several respects in which the traditional logic was lacking, and where Frege sought to make improvements.

From Subject/Predicate to Concept/Object. One of the first things that Frege realized was that the subject-predicate analysis of sentences, employed since Aristotle, doesn't always do a good job of capturing their actual meaning or content. For one thing, ordinary languages (English, German, or whatever), often permit the formation of several grammatically different sentences that nevertheless say essentially the same thing. So if we use the subject/predicate analysis, or any other grammatically based analysis, we won't get at the real logical content, what the sentences are really saying. Consider, for instance, these two sentences:

The Greeks defeated the Persians.

The Persians were defeated by the Greeks.

Intuitively, these say pretty much the same thing. If there are any differences in meaning, they are less important than the content, which is the same in both. But what happens if we apply the subject/predicate analysis to both sentences? In terms of the subject/predicate analysis, the subject of the first sentence is "The Greeks;" whereas the subject of the second is "The Persians." Also, the predicate of the first sentence is "defeated the Persians," whereas the predicate of the second is "were defeated by the Greeks." On the subject/predicate analysis, then, we have two completely different subjects, and two completely different predicates; the fact that the two sentences actually say the same thing doesn't show up at all!

Look what happens, though, if we break with the subject/predicate form and treat the structure of the sentences *logically* rather than *grammatically*. Conceptually, in both sentences there is the same *relation*, the relation of "…being defeated by…", and this relates two things, the Greeks and the Persians. And they are in fact related the same way in both sentences. So with Frege's notation, we can treat "…being defeated by…" as a '*two-place*' predicate, like a function of two variables. This means that it essentially has two 'subject'-places, not just one, to fill. We can also write *both* sentences the same way: Dgp (where 'D' stands for *defeats*; 'g' stands for *the Greeks*; and 'p' stands for *the Persians*). This way, we can see clearly and almost effortlessly that the two sentences have the same actual logical content, and in fact say just the same thing.

What does it mean, though, that two sentences 'say the same thing' in this sense? Here, Frege's key idea is that logical or conceptual content does not depend on *superficial grammatical form*, but rather on *inferential role*. That is, two sentences have the same meaning (in the 'conceptual content' sense) if, and only if, *the same conclusions can be inferred from them* (together with other auxiliary premises). This is certainly the case with our two sentences about the Greeks and the Persians. Even though they look grammatically different, they have the same actual conceptual content, for we can draw exactly the same consequences in either case. This idea of content as coming from inferential role (rather than grammatical structure or even intentional 'aboutness') was revolutionary when it was proposed, and still exerts a powerful impact on much analytic philosophical thought.

Relations: Consider the straightforward (and true) mathematical sentence: "Seven is greater than five." How does this look on the subject/predicate analysis? It looks as if the subject is seven; and the predicate is 'being greater than five.' But this isn't a very clear or helpful analysis, since what the sentence asserts is clearly that a *relation* holds between *two* things, seven and five; not that a property holds of just one. Frege's analysis does much better, since (just as in the Greek/Persians case) we can write: Gsf, where 'G' stands for the *relation* 'greater than', 's' stands for seven, and 'f' stands for five.

Frege's key idea here is that sentences aren't made up of 'subjects' and 'predicates'; rather, they're composed of terms for *concepts* and terms for *objects*. Whereas objects are just things that can be named and that we say things about, concepts are essentially operators *on* objects; logically, they work like *functions* of objects to produce truth-values. What does this mean? Well, consider ordinary functions, such as you might have learned about in middle school math. We can write such functions by using variable terms; for instance $f(x) = x^2$ (the 'squared' function). This is a function of *one* variable; we put in a single value, and we get out a single value. But we can also have functions of two variables (or as many as we want). For instance f(x,y) = x+y is one such function (namely, the addition function). This function inherently has two "places", whereas the 'squared' function had only one: for addition, we need to put in *two* values in order to get one out. Frege's key idea is that concepts are just like such functions: they can have *one* place (*x* is red), *two* places (*x* is taller than *y*) *three* places (*x* is between *y* and *z*) or more. In an actual sentence, we "fill" the variable places, however many there are, with specific *values* – specific objects that go into the variable places. For instance, we might put "my apple" for x in the one-place concept (*x* is red) above. Then we obtain a normal English sentence:

My apple is red.

Or we might plug in 'me' for *x*, and 'Michael Jordan' for y in the two-place concept (*x* is taller than *y*). Then we get the sentence:

I am taller than Michael Jordan.

There is an important difference between these two sentences, however. Whereas the first one is true, the second one is false. By plugging in different values to each concept, though, we could have gotten different results. (We might, for instance, have put "my asparagus" in the first sentence, and gotten False; or put 'me' and 'Napolean Bonaparte' in the second sentence, and gotten True). Again, this is like an (ordinary) function. By plugging in different values for the variables, we get different results out Thus, concepts are again like functions, and sentences are again like values of these functions given specific inputs. The only difference is that, whereas the values in the ordinary-function case are numbers (or other mathematical objects), in the sentence case the values are either True or False. Frege called these the two "truth-values" and thought of them as something like special objects that exist in themselves. There are probably some unpalatable consequences of this view (where do these 'truth-values' actually exist? How do we mortal beings interact with them?), but it does give a clear picture of the functioning of sentences that improves greatly on the subject-predicate picture.

Multiple Quantification. Consider the following straightforward sentences, one drawn from mathematics, the other from ordinary life.

Every odd number has an even successor.

Everybody loves the president.

On the Aristotelian analysis, both of these sentences have the same form, namely: All a's are b. But actually, the form of the two is very different. The second sentence says that there is ONE individual, the president, whom everybody loves; whereas the first one says that for EACH odd number, there is some even number (not necessarily the SAME one!) that is its successor. It would be a big mistake if we understood the first sentence to mean that there is some unique even number (say 32) that is the successor of each and every odd number. But whereas Aristotle's logical syllogisms can't capture this difference at all, Frege's notation does so nicely through the device of multiple quantification. Thus we have, for the first one:

$$(\forall x)\,(\exists y)\,(Syx \,\&\, (Ox \rightarrow Ey))$$

And for the second:

$$(\exists x)\,(\forall y)\,(Lyx \,\&\, Px)$$

Note the difference in the structure of the sentences as symbolized: this shows the real logical difference between the two sentences, and makes it possible to draw very different inferences from the two.

Introduction to Set Theory

Today we'll become familiar with some of the basic ideas of Cantor's set theory, which has played a profound role in mathematical, formal, and philosophical thinking throughout the twentieth century. We'll keep it simple and intuitive at first, just trying to get a sense for the different notions and symbolism, and we'll worry about figuring out the basic *axioms* (assumptions) and rules later.

To begin with, then, what is a set? A set is any collection or grouping of objects. Intuitively, we can make a set out of any group of objects we like, and we denote the set with the curly brackets ('**{**' and '**}**') and separate elements with commas. Here are a few examples of sets:

{Barack, George, Bill**}**

{7, 18, 24, -5**}**

{15, π, the Eiffel tower**}**

The last example, in particular, shows how general the notion of a set is: intuitively, we can construct a set of *any* individuals we like, no matter how different they are from one another.

A set is completely defined by its elements. This means that if two sets have exactly the same elements, they are exactly the same set (and conversely). Thus,

{Barack, George, Bill**} = {**George, Bill, Barack**}.**

It makes no difference what order we list the elements in. Also, we can indicate a set by means of a property or description, rather than just a list. Thus we can say:

{*x* | *x* is a president of the US elected between 1992 and 2008**}.**

We read this: "The set of all x, such that x is a president of the US elected in 1992 or after." And in fact, this set is *identical* to the one we considered before:

{*x* | *x* is a president of the US elected between 1992 and 2008**} = {**Barack, George, Bill**}**

And, to take another (classic) example:

{x | x is a human being**} = {**x| x is a featherless biped**}**

This feature of sets – that they are completely defined by their elements – is called *extensionality*. It is usually thought that sets are different from concepts or ideas, since concepts and ideas don't have this feature (by contrast, they are *intensional*). Thus, the *concept* of being a featherless biped is probably different from the *concept* of being a human being – even though the extensions, and hence the sets, are the same.

In most systems of set theory, it is assumed that there exists an *empty* set, a set containing no elements. We can use extensionality, moreover, to show that the empty set is *unique* – if there is one, there is only one. The empty set can be denoted in two different ways, either just by writing the empty brackets:

{}

Or, more commonly, by the symbol:

∅

We'll use the primitive symbol '**∈**' to mean 'is an element of.' Thus:

Barack **∈ {**Barack, George, Bill**}**

And similarly,

Bill **∈ {***x* **|** *x* is a president of the US elected between 1992 and 2008**}**

There are also a couple of straightforward operations we can perform on sets to form new ones. The first of these is *union*, denoted '**∪**'. To get the union of two sets, we just put together all the elements of the first with all the elements of the second. Thus:

{Barack, George, Bill**} ∪ {**Albert, Steve**} = {**Barack, George, Bill, Steve, Albert**}**

And

{x | x is an odd natural number**} ∪ {**x|x is an even natural number**} = {**x | x is a natural number**}**

Another operation is *intersection*. This is denoted '**∩**', and the intersection of two sets is just the elements they have in common. Thus

{Barack, George, Bill} **∩** {Jimmy, Ronald, Bill} = {Bill}

And

{1, 3, 5**} ∩ {**2, 4, 6**} = ∅**

There is one more relationship to learn, and this one will turn out to be very powerful and useful. It's the relation of *subset*, which we denote with '**⊆**'. That A is a subset of B, or A **⊆** B, means that everything in A is *also* in B (though perhaps not vice-versa). Thus:

{George, Bill} **⊆** {George, Bill, Barack}

{x | x is a president elected in 2000 or after**} ⊆ {***x* **|** *x* is a president of the US elected in 1992 or after**}**

∅, the empty set, is a subset of every other set (why?) Also, every set is a subset of itself. If we want to talk about *only* the subsets of a set that are not identical to that set itself, we can use the idea of a '*proper subset'*, symbolized '⊂'.

What happens, though, if we consider *all* of the subsets of a given set? We can again group these into a new set, and we call this new set (of sets) the *power set* of the first. If we call the first set 'x', we call the power set '$\mathcal{P}$(x)'. Let's consider the power set of a few different sets. First, let's take the set {3,5}. What are its subsets? Well, we have the 'singleton' sets, {3} and {5}. We have the empty set, **∅,** which is a subset of every set. And finally, we have the initial set itself, {3,5}, since every set is a subset of itself. So we can now group these all together to form a new set; we obtain:

$\mathcal{P}$ {3, 5} = {**∅,** {3}, {5}, {3,5}}

And similarly,

$\mathcal{P}$ {Barack, George, Bill} = {**∅,** {Barack}, {George}, {Bill}, {Barack, George}, {Barack, Bill}, {Bill, George}, {Barack, George, Bill}}

Notice that in the first case, where the initial set had just two elements, the power set had four; whereas in the second case, the initial set had three elements and the power set had eight. Is there anything general we can say about the relationship between the number of elements in a set and the number of elements in its power set?

We'll work more closely with the fundamental axioms and basic rules of set theory next week. For now, though, there are just a couple of things to notice about the nature of sets. One is that, although we've dealt so far only with sets with a *finite* number of elements, there is absolutely no reason why sets must have (only) a finite number of elements. In fact, infinite sets are quite easily defined, e.g.:

{*x* | *x* is a natural number}

{*x* | *x* is red}

Both (probably) have an infinite number of elements.

Second, we've worked so far with sets that have things like people, or numbers, as elements. We can do this if we like, but we can also forms sets whose members are *simply* other sets, e.g.:

{ **∅ }**

Or { ∅, {∅}}, etc.

In mathematics, in practice, we'll usually work with the second kind of set, the kind that has only other sets as elements. This might seem abstract, but it actually makes it possible to contemplate the possibility of reducing *all* mathematical 'objects' to a *single* kind of object, namely pure sets. As we'll

see, if we add this conception to Frege's logic, we have an extremely powerful and general means for thinking about mathematics, as well as about thinking and concepts in general.

Philosophy/Math 415/515 Fall 2025
History and Philosophy of Mathematics

Notes: Week 3

Last week, we reviewed Frege's symbolic logic, talked about why it might be better – or at least more sensitive to matters of the internal structure and content of predicative sentences – than Aristotle's syllogistic logic, and began to supply the logic with "objects" in the form of sets which we can use (hopefully) to build up the whole domain of mathematics, beginning with the arithmetic of whole numbers and ratios. This week and next we'll consider Frege's best and most influential attempt actually to define number, in the *Foundations of Arithmetic* of 1884. We'll see how he plausibly defines number, starting with the numbers one and zero, in a much more clear and rigorous way than any of his philosophical predecessors, but we'll also be in a position to see why the definition that he gives ultimately fails, and what this teaches us about the foundations of mathematics more generally.

Frege's Question If arithmetic is to be a legitimate science, we had better be in a position to say what its central objects – that is, numbers – actually are. But existing philosophical definitions of numbers are by no means very clear. We can begin to see this just by starting with the question of the number one: what is this object (if such it be), or what do the signs "one" and "1" actually stand for? Clearly, these signs have some meaning that is not simply dependent on the English language (we can symbolize the same thing by using "uno" or "eins" in other languages) and whatever their meaning turns out to be, it should apparently be something objective and the same for everyone: mathematical judgments are a paradigm of objective truth. But when we try to say what object "one" stands for, it is not easy to do so. We might begin with the thought that one stands for "any object". But if we spell this out a bit, we realize that it actually just means "any one object," and we haven't provided any definition at all (but just repeated the term to be defined! Another thing we might try is to substitute for "one", in a number judgment, some (or any) *particular* object, say the moon. But this won't work either. For if we put, in place of

1 + 1 = 2

The moon + the moon = 2

This seems incorrect (and also, in a way, nonsense). For similar reasons, attempts to define "one" as "a unit" seem to fail. For if we use the indefinite article ("a" unit), we have to allow that there can be more than one of these (whatever they are). Then it won't be correct to call it "the" number one at all, and we'll also be hard pressed to say what is happening when we add one of them to another, how the first one is distinct from the second, and so forth. But if we think there is only one "unit", then adding or computing with it would seem to be like the case of "the moon" above. Clearly, we need another approach: ideally, one that respects the truth and objectivity of our real judgments about numbers, including both our judgments about the numbers of real things ("There are six birds over there") and our abstract arithmetical judgments ("3+3=6"). And since these judgments appear both universal, in that they apply with full generality to whatever we may consider, and objective (if any are), we want to show (if possible) how they are grounded in logical rules that are themselves universal and objective, the

same for everyone and at all times, and not dependent (in any sense) on subjective or psychological processes or activities.

Frege's Principles

Near the end of the "Introduction," Frege states three overarching methodological principles that he employs throughout the investigation. These principles have interesting interconnections and are philosophically both revolutionary and useful in their own right, so it is helpful to consider them in turn.

1) Always to separate sharply the psychological from the logical, the subjective from the objective;
2) Never to ask for the meaning of a word in isolation, but only in the context of a proposition;
3) Never to lose sight of the distinction between concept and object

Principle 1): Objectivity and Frege vs. Psychologism. At the time of Frege's writing – the late 19th century – the two very different sciences of logic and psychology were both, at about the same time, being founded in the form that we know them today. William James' *Principles of Psychology*, for example, was published in 1890, just six years after Frege's *Foundations*, and even before James there were many burgeoning approaches to making psychology into a rigorously scientific study of the mind. "Logic" in the nineteenth century (in philosophers such as Lotze, Wundt, Bolzano and Brentano) also took in a bewildering variety of (what we would today understand as) distinct categories, including epistemology, phenomenology, semantics, and (empirical) psychology. There was also no agreement about what the laws of psychology (if there are any) are really about: whether (for example) they govern knowledge, or processes of thinking, or of judgment, or the structure of the world, or language. Partly for these reasons, many nineteenth century logicians thought of the subject matter of logic as in some way closely related to the study of psychology as the study of "the laws of thought," understood as the laws governing how we (as human beings) actually do think. Logicians who thought in this way were the ones that Frege understood as the "psychologistic" logicians, and whose ideas about logic and mathematics he deeply opposed.

Partially following Bernard Bolzano, Frege insisted by contrast that the form of logical laws, as well as the contents of the judgments that we make, ought to be understood as rigorously objective. This extends as well to mathematics, and part of the goal of Frege's logicist project as a whole is to show how mathematics might enjoy the kind of objectivity that would be characteristic of logic itself on this kind of view. Frege polemicizes, here and elsewhere, against varieties of psychologistic and (more broadly) subjectivist accounts of mathematics. Beyond the polemics, he gives a few reasons for his rigorous separation of the subjective from the objective and his insistence that neither logic nor mathematics can be treated psychologically. One reason is that, if (for instance) numbers were a matter of our own, individual ideas (by "idea" Frege always means something individual and subjective, as opposed to "concepts," which are objective), then every person would have his or her own individual number 5, and there would be no apparent sense in which we could discuss the objective properties of the number 5 (for example the fact that it is odd, or that it is the same as 2+3), which presumably hold for everyone. For similar reasons, Frege opposes the "historicist" idea that we should seek the nature of numbers in historical facts, for instance about when and how they were discovered. If this were the case, then it

might seem possible that (for example) the Pythagorean theorem, having once been true, might become false, and then later true again. An additional target of Frege's argument is those who account for number in terms of some kind of (individual-psychological) process of counting. Thus, for instance, Kant associates the basis of numerical judgments with a subjective and temporal process of successive counting, and for this reason argues that numerical judgments require intuition and are thus (at least partly) synthetic. However, as Frege points out, we can in fact calculate rigorously with respect to even very large numbers, and we thereby derive objective results that do not simply bear on our own individual psyches or any recognizable aspect or form of our intuition.

Principle 2) – The "context principle." Frege's logic, as we have seen, is a more flexible and sophisticated tool than previous logics for dealing (exactly) with the truth and falsity of sentences, and more broadly, by putting the truth and falsity of sentences rather than the (purportedly) representational meaning of their components at the center of his logic, Frege effects a revolution in the semantics and metaphysics of meaning. The principle that we should look for the meaning of a word only in the context of a *whole* proposition – that is, something that has, at least, the structure of something that can be treated as true or false – is justly famous and is often called the "context principle." But what are the consequences of looking at things this way? One immediate consequence is that the context principle allows us to reject a kind of subjectivist account of number that had previously been popular. On the kind of account at issue, our judgments about number require that we have or entertain (or are to be explained in terms of our having or entertaining) representative *mental images* or *intuitive representations* of the numbers in question. For example, on this kind of view, our knowledge of the number five might depend upon our having (or entertaining) an illustrative mental image, for example the image of the five black spots on one side of a die. This kind of view faces obvious embarrassments when we try to generalize it: for example, our knowledge of truths involving very large numbers could hardly depend on an intuitive image of exactly these numbers, and we can also (as Frege points out) easily judge that there are 0 of some particular thing (e.g., of moons of Venus).

But if the mistake of this kind of view is thinking that each term in a number-judgment must have its own, distinct, representational meaning, the context principle shows us how to avoid this assumption. In fact, it is irrelevant to the truth of a judgment of number (either the number of a group of concrete objects or an abstract arithmetic judgment such as 78+65=143) *how* we picture or represent the individual numbers, as long as we can effective recognize the *truth* of these judgments (when they are true) and recognize the number of two groups as the same when *it (i.e. the number) is* the same. This gives, furthermore, an important clue to the positive account of number that Frege will give. In order to obtain the correct concept of number, we just need to say what it means to (and how we do) "recognize" a number "again as the same": how, that is, we "fix the sense" of numerical identities.

Principle 3) Concept and Object. As we saw last week, Frege draws a rigorous and logically based distinction between objects and concepts, insisting upon the difference in the logical structure of the terms that refer to them and also, correlatively, in their ontological structure. Whereas objects are what we can refer to with a name, or a definite description, concept-terms refer to what are, ontologically, functions, "gappy" or unsaturated things that need objects in order to determine a truth-value. This raises the problem about how we refer to concepts that we discussed last week, and to which we'll

return, but in the meantime it's worth noticing that the distinction between concept and object already clarifies something important about what judgments of number and numerical entities are about.

Do arithmetic judgments depend on intuition?

Immanuel Kant, in the *Critique of Pure Reason,* draws two distinctions which he thought helped in explaining the source and necessity of different kinds of judgments. The distinctions are: i) between judgments made (or known) *a priori* (i.e. 'before' experience or any empirical evidence) and *a posteriori* ('after' or on the basis of experience); and ii) between judgments that are *analytic* – or (for Kant – though Frege defines "analytic" somewhat differently) judgments whose truth is "contained" in the relevant constituent concepts (thus, e.g., those that are plausibly true "by definition") and those that are *synthetic* – i.e. not just contained in the concepts. Kant thought that judgments of the mathematics of his time, including (for him) geometry, algebra, and calculus was clearly *a priori*, but was impressed with their complexity and informativeness, which makes it the case that they don't seem to be deducible from "mere concepts" alone. So he classed these judgments, in terms of his distinctions, as both *synthetic* and *a priori.* One of the major questions of the *Critique of Pure Reason* is that of how such *synthetic a priori* judgments are possible at all, but in the case of geometry and arithmetic he thought that they are synthetic because they depend not only on concepts but also on *intuition*, or on the form of our *sensory* experience of the world in time and space. So, in the case of a geometric demonstration (like those we see in Euclid's "constructions"), we need to actually draw a picture, and so we see that some properties of the space that we are drawing on are relevant to the success of the proof. In the case of number, Kant thought that the relevant "intuition" is not a spatial one but rather a temporal one: for example, an intuition of repeated "moments" successive in time. This is where Kant also locates the "generation" of number and the possibility of our thinking and understanding numbers in general (although he doesn't distinguish clearly between what Frege understands as the *structure* of numbers themselves and the conditions of their origination "in us").

But Frege has several good arguments to show that arithmetic judgments can't depend on intuition. If we are able to determine that there are 0 of something, there can hardly be an intuition of 0 of them that underlies this (what would this be like?) In the case of rather large numbers – say those over 100 or so – it doesn't seem that there's anything in our sensory experience that could confirm that we're dealing with (e.g.) 1,345,210 of something rather than 1,345,209 or 1,345,211 – but something "given" in or by intuition itself would have to be able to do that if Kant is right. Moreover, we can number and count many kinds of things that aren't in space or time at all, and so don't have the necessary form that underlies all arithmetical (as well as geometric) judgments for Kant. For instance, we can enumerate principles of U.S. law or proofs of a theorem or solutions to a quadratic equation.

Since Frege, however, agrees with Kant that arithmetic judgments are *a priori* and not empirical (see below), he is led to conclude that they *a priori* and *analytic* after all. Arithmetic judgments, for him, thus owe nothing at all to intuition and we do not need anything other than pure thought to know their truth. This idea – that they are knowable by laws of pure thought alone, and true in virtue of the laws that characterize thought in a unitary way, without regard to intuitive or other objects – is a central plank of his *logicism* about arithmetic. To make it plausible, he understands "analytic" in a slightly

different way than Kant does: not as a matter of "conceptual containment," but of general principles or axioms from which can be *logically derived* the multiplicity of varied and particular properties and features of numbers. But there is still a question – which we'll come back to – about what these general laws could be, and thus *how* the variety of properties could be derived which makes mathematics so complex and fruitful, especially since (as he says in section 10), these properties are quite diverse and unique in every case, and do not all seem to follow simply from any single general statement or set of general statements.

Are judgments of (pure) arithmetic empirical judgments or generalizations of empirical judgments?

For many of the same reasons (and some other ones), that Frege thinks that arithmetic doesn't depend on intuition, he also thinks that judgments of arithmetic aren't empirical judgments. John Stuart Mill and other psychologistic logicians suppose that we derive both the concepts of individual numbers and our judgments of arithmetic truths from the observation of simple situations, e.g. when we observe piles or heaps or conglomerations of "individual" things. But this seems incorrect, for several different reasons. In general, we don't derive truths such as 1+1=2 from *observing* phenomena in the world, but rather *apply* these to phenomena in the world. If see that a room is empty, and see one student and then another student enter, and then see that there are three students in the room, I do not conclude that 1+1=3 but rather wonder where the other student came from and how they got in the room. Moreover, for larger numbers and complicated additions (such as 1,432+2,111=3,543), we know the correctness of the judgment (calculation) without ever having seen or needing to see that number of objects. And then there is also the problem of how it would ever be possible on the basis of observation to judge that there are 0 of something.

A related idea is that judgments of arithmetic are in some way *inductive* (in the scientific sense of induction) generalizations from simple individual observations of fact. We do use this inductive procedure in many kinds of empirical research. But what principles could we imagine are "generalized" in order to reach the full complexity of arithmetical judgments? And furthermore, as Frege points out, empirical generalization – where we generalize from a *number* of specific instances of a phenomenon that *all* are of that type – itself presupposes number and can't seem to be a basis for it.

Are judgments of number judgments about *objects* at all? And when I assert a truth about the number of some things, do I in fact assert that *those things* have a property?

In fact, the answer to both questions is "no", as we can see by noting some elementary facts about number judgments. Let us start with a puzzle. Consider the following two ordinary judgments:

There are four books on the table.

Four is an even number.

Obviously, the word "four" in both judgments means (in some sense) the same thing. What part of speech is it in each case? In the first case, it appears to be an adjective modifying "books"; in the second, it appears to be a noun. Yet in the sense in which the second one attributes a property to an

individual (i.e. *the* number four – note the definite article), it would also bear substitution into the first: "There are an even number of books on the table".

Let's focus in on the second kind of use, where number terms appear to be adjectives and to modify nouns. Normally, an adjective such as "green" attributes a property to a thing that is modified by it: "The leaf is green". Does the word "four" in "There are four books" attribute a property to the books? Or perhaps to the "collection" or "grouping" of the books?

Fact 1) If I judge, of the leaves on a tree, that they are green, I also judge *of each individual leaf* that it is green. But if I judge, of those leaves, that there are (say) 57 of them, I do not thereby judge that each individual leaf is 57.

Fact 2) If I point to a pack of playing cards on the table (or, in general, any object that has parts, i.e. any object) and ask, "How many?", there is no determinate answer to the question. You will (rightly) have to ask me, "are you asking how many *cards* there are (52)? Or how many *packs* there are (1)? Or how many suits (4)? (etc.)" It's only as specified in one of these ways that the question of number has any answer. But what is considered is (in some sense) nevertheless the *same* thing or group, in each case.

Fact 3) It is obviously possible to judge that the number of some type of thing is 0. For example, I can readily judge that (and it can be true that) the number of Venus's moons is 0. But there is (for obvious reasons) here no object corresponding to "Venus's moon".

These facts are sufficient to show that *no account of a judgment of number as a judgment of the properties of objects can be correct.* No matter how hard we try, we can't locate numbers in (individual) objects or their properties. Even if we think that some objects are "aggregates" or "groups," this is no help. For I still encounter the same ambiguities that we saw: If I ask the number of the group, I have to answer either *one* (since there is only one group) or I have to give some further specification of what is the relevant unit, what it is supposed to be a group "of". And then we are back in the same difficulty that produces fact 2.

Judgments of number, then, are not simply "about" objects, and it also appears to follow that no aggregative process of putting or grouping together objects or their representations "in thought" will suffice to explain them. (In any case, since judgments (e.g.) of the number of cards on the table are certainly objective, it doesn't seem to have anything to do with the *truth* of such a judgment that I do (or do not) perform such an aggregative process – and what we are interested in logically is the conditions for truth, rather than those for mental representation or conception). This shows also that it is a non-starter to try to theorize my judgment of the number of some things as a matter of *my* grouping or unifying them together, for example by "unifying them" in a "manifold" of "synthetic apperception" (Kant); or by thinking that the judgment requires that the things in the "group" *actually* (e.g. spatially) are put together in the same place, etc. So judgments of number are not judgments about (in the same sense as attributions of properties are about) external things or groups of such things.

Are judgments of number judgments of something subjective?

Given the above observation (Fact 2), we might start to think that the number of something(s) depends simply on how *we* regard it. If the same fact can sustain (as in the pack of cards example) such different number judgments, maybe number is (as Berkeley suggested – section 25) "entirely a creature of the mind." The suggestion that number is subjective in this sense sometimes also goes along with views on which the way we should clarify what numbers are is to clarify what happens in our minds or consciousness when we count, etc. But as Frege points out (section 26), no account of number that depends on a description of individual psychology could have the kind of objectivity or truth that judgments about numbers, and judgments about the world that use numbers, actually do have. We are not interested, after all, in "my" number 2 or your number 2 – we are interested in THE number 2, and its properties are as objective as those of anything. More broadly, the mere fact that something is designated by language does NOT make it or any judgments about it "subjective." If I want to know how many jellyfish there are in the North Sea, I will go and investigate the North Sea itself – not my or someone's idea of it. This is still the case, despite the fact that the boundaries of the North Sea are drawn by us, and don't necessarily correspond to anything drawn in nature. These kinds of judgments are perfectly objective, but to be objective in this sense does not necessarily mean to be about tangible or physical or concretely handleable things. For example, I can ask how many times someone has crossed the equator, and get an objective answer, without thinking that the equator is a physical or tangible thing.

Is a number a collection or agglomeration of "ones" or "units"?

Finally, a lot of philosophers (beginning perhaps with Euclid) have treated numbers in general as collections or groupings of "units" or "ones".[1] Thus we might think of the number 5 as a collection or grouping of "units". How many units? Well, five – and so we start to see how this kind of definition doesn't really help much.

But there's also the problem of the "units" themselves. Are the units that are supposed to make up five the same, or different? If they are the same, and we count "them", we can only count one. But if they are different, then we can presumably signify them with different signs – so we would write 5 (as Jevons does – section 35) as

1'+1''+1'''+1''''+1'''''

But then, as Frege points out, we might as well just have

a + b + c + d + e,

and it's not clear in what sense we're talking about "units" or "ones" at all (moreover, we have to already have the number 5 at our disposal in order to count these as five!). What we have tried to do is

[1] At the beginning of book VII of the *Elements*, Euclid gives as definitions:
1. A *unit* is that by virtue of which each of the things that exist is called one.
2. A *number* is a multitude composed of units.

to combine the idea of the sameness of the "units" with the (contradictory) idea of their difference, and this will not work. As Frege says (section 39, p. 50):

"If we try to produce the number by putting together different distinct objects, the result is an agglomeration in which the objects contained remain still in possession of precisely those properties which serve to distinguish them from one another, and that is not the number. But if we try to do it in the other way, by putting together identicals, the result runs perpetually together into one and we never reach a plurality."

The word "unit" may conceal this difficulty for a moment, but it doesn't really solve any problems.

More broadly, philosophers have for a long time talked and thought about "oneness" or "unity" in the sense of which things "are" ones or are "unified". This can mean in an ordinary sense that they are strongly integrated or hard to separate into parts, but it can't mean that they have a, or the, property of being one. First of all, if being one were a property, it would be possessed by each and every thing: "It is not easy to imagine how language could have come to invent a word for a property which could not be of the slightest use for adding to the description of any object whatsoever." (p. 40). Also, although it is a good test for any view of what numbers are to see if that view works for 1 (and 0) as well as it does for other numbers, we have here again the same problems with thinking that number judgments attribute properties. If we can say "Socrates was wise" and "Plato was wise", we can say that "Socrates and Plato were wise". But if we say "Socrates was one" and "Plato was one" we cannot say that "Socrates and Plato were one." And also, as Frege points out (section 30), if being one is treated as a property, it appears impossible to define it, at least non-circularly. Again, some philosophers have held that something's being one is an aspect our outcome of the way that *we* regard it. But this makes number subjective again, and ruins the objectivity of our judgments in arithmetic.

From all of this and the obvious fact that the number one *itself* does have objective (mathematical) properties (such as being the multiplicative identity element), Frege concludes (section 38) that the number one is a definite, particular object, and as such there are not diverse or several of them (in other words, there are no "units" or "ones"). It is a unique object, but so is the number 2 or the number 5 or any other unique number, each one with a peculiar set of properties that are uniquely its own. We still have not solved the mystery of its being, but by figuring out what it is not, we are much closer to knowing what it can – logically speaking -- be.[2]

[2] There are also deep lessons here for the critical consideration of what philosophers such as Plotinus have called – with a definite article -- "the One" and its (supposed) relationship to "the many".

Philosophy/Math 415/515 Fall 2025
History and Philosophy of Mathematics

Notes: Week 4

Last week, we considered the problem that the judgment "the books are blue" seems to attribute a property to the books, but "the number of books on the table is four" does not seem to do that (but also doesn't seem to attribute a number to the "group" or "aggregation" of books, which (if it did) should rather be one). Similarly, our number judgment with respect to the deck of cards on the table seems to depend on "how" we regard "it", and not just on "it" itself – although this doesn't stop the judgment from being as objective as one likes. Judgments of number, then, although objective, are not simply "about" objects, and it also appears to follow that no aggregative process of putting or grouping together objects or their representations "in thought" will suffice to explain them. (In any case, since judgments (e.g.) of the number of cards on the table are certainly objective, it doesn't seem to have anything to do with the *truth* of such a judgment that I do (or do not) perform such an aggregative process – and what we are interested in logically is the conditions for truth, rather than those for mental representation or conception). This shows also that it is a non-starter to try to theorize my judgment of the number of some things as a matter of *my* grouping or unifying them together, for example by "unifying them" in a "manifold" of "synthetic apperception" (Kant).

But if numbers are thus not properties of objects, we can nevertheless get a more positive clue about what they are by considering more positively the pack of cards case again. The question "how many," just directed at what is on the table, doesn't get an answer. But what *does* have a determinate and objective answer is the question along *with the specification* of a *concept*, for example "cards on the table." If I ask, not just "how many is this?" but rather, "how many cards are on the table?" I get a straightforward and objective answer. However, in Frege's logic, *card on the table* is not the term for an object, but rather for a *concept*. It doesn't, itself, stand for any object (although "the card on the table" or "that card on the table" would), but it *does* make sense to ask, of any individual object, whether it falls under that concept (i.e., whether it is a card on the table). When I ask the question that gets the (correct) answer "52", though, I am not asking about the cards themselves, either individually or as a (i.e. one) group. Rather, as is shown by the need to specify, I am asking a question about the *concept* "card on the table:" asking (as it were) how many "times" that concept is itself instantiated. Something similar is the case when I judge that there are (e.g.) 0 moons of Venus: even though there is no object, I am judging (as it were) about the concept "moon of Venus" that it is *not* instantiated, that nothing falls under it.

This leads Frege to his first positive determination about what numbers are. If the content of a statement of number is (as this suggests) an assertion about a concept, then we might say that judgments of number are (in some sense) judgments about concepts rather than judgments about objects. However, this can't be exactly right, for several related reasons. First, if we were to just say that judgments of number are judgments *about* (or of the properties of) concepts, we'd obviously be violating Frege's own insistence on keeping objects and concepts grammatically and logically distinct. Given this distinction, as we saw last week, we can't put concept terms in the object "place" in a

sentence, so we can't say that a concept has a particular property. Even if, then, we want to say something like "52 is the number of cards on the table," thereby treating 52 as a property of the concept "cards on the table," we can't do that. If we can get some important insight out of the thought that number-assertions are "about" concepts, then, we'll have to do it somewhat more indirectly than just by saying that they assert properties of them. In fact, there is a crucially important issue here, and on any account we'll have to find a way to talk "about" conceptual groupings instead of just the objects grouped, in order to get the correct results about number judgments, but (as we saw above) somehow we have to do this without just talking about these groups themselves as objects.

Another reason why the idea that numbers are properties of concepts won't work is that numbers are, fairly evidently, not properties at all. Rather, grammatically they are themselves specific, individual (Frege says, "self-sufficient") objects: for example, we use the definite article ("the" number 1, or "the" number 52) to refer to them, and they themselves have many properties (e.g. being odd or even, being prime, etc.) that we can specify and describe.

Another way to put the problems which we have discovered in the first half of the book with historical attempts to "define" number is that – whether we define "a number" as an aggregate or a grouping or an act or a collection of "units", we are still looking for a definition of the form:

A number is ……

(Thus, e.g., "A number is an aggregate"; "A number is a collection of units"; etc.)

That is, these historical philosophers are looking for a *general definition* of what *any* number, or what all numbers are. We might be able to get something like this in a very general "type" sense, but if we want to be able to define *the* numbers (i.e. the number one, the number two, the number three, etc.) in such a way that provides a basis for deducing (proving) their unique properties, it's probably that no such "general" definition can work. (As we saw with the exercise of proving that each natural number is either even or odd, it's unlikely that even very simple number-theoretic properties can be proven in this general, once-and-for all, kind of way). What we rather will need to do is define *each* number in terms of its position in a general series, for whose definition we have also provided.

Hume's Principle and Statements of Identity (of Number) In order to solve all of these problems, Frege proposes that we need to find a way of talking about numbers as self-sufficient entities, but as in some way bearing on or determining concepts (such as the concept "cards on the table"), and as also being the entities they are only within, and relative to, the whole *series* which in fact defines them. In fact, as we've already seen suggested above, given the context principle it will be sufficient to determine the *sense* of statements of number – and thus, to determine *what* numbers are – if we can determine precisely what is meant by saying that two numbers (e.g., the numbers of two concepts) *are the same*. If, in other words, we can make sense of the propositions which express our recognition of a number as the same (again) in diverse contexts, we will have in a sense "defined" what a number is, in general. And to do this, it is sufficient to define in general the sense of the proposition "The number belonging to concept F is the same number as that belonging to concept G." In doing this, we will be giving, as Frege

says, a "general criterion" for the identity of numbers, and so will be able to settle on this basis all relevant questions about what they are.

Accordingly, Frege proceeds to give what is called a "contextual definition" of number. Contextual definitions differ from ordinary ones in that, rather than defining the concept by breaking it down into more basic or primary components, we simply specify a criterion for the **identity of the concept's objects and leave it at that**. If we can say when two F's are the same, and when they are different, we have importantly done at least much of the work of defining "what it is" to be an F. So we will try to do this with numbers, starting with reflecting on the phrase "the number of F's is…".

So going back to the concept of number: if we can settle all questions about the judgments of identity or difference that we make with respect to the concept, we have settled in a rigorous way what these judgments are *about,* and thus defined the concept. But is there a criterion for the identity of the number of two collections or groups? In fact there is, and it was originally stated by Hume:

Hume's Principle: The number of F's is the same as the number of G's if, and only if, the F's can be put into one-to-one correlation with the G's.

In fact we have already met the criterion of one-to-one correlation: it is a powerful tool for defining number *in general*, up to and including the infinite. Here, we are going to use it make sense of our judgments about number in general, and once we have done that we can proceed to define more specifically 0, and then 1, etc. by just pointing to a representative group or concept for the 0 (Frege suggest using the concept "not identical with itself", but he also says that any concept that has an empty extension will do) and proceeding from there.

There is a partial analogy here, as Frege points out, to a procedure we might well use in defining "directions", given only that we have access (perhaps intuitive access) to lines on the plane. We here fix the criterion of sameness of direction by stipulating that two lines will have the same direction if and only if they are parallel to one another. The "direction" is then just what parallel lines share with one another; or we can just identify the concept of a specific direction with the concept of being parallel with some specific line, or, indeed, with the "equivalence class" of *all* lines that are parallel with one another. In a similar fashion, we are here choosing a representative group, and then saying that whatever can be put into one-to-one correspondence with this group will be something with the same number. In fact, if we want a determinate and specific sense for the phrase "the Number x," where G is a concept having the number x, we might as well, Frege suggests, just identify "the Number x" with the *extension* of the concept "equinumerous with G" – that is (in set-theoretical language) the set containing *all* sets whose members can be put into one-to-one correspondence with the members of G.[1]

[1] Of course, when x is finite, this set – the extension – will have many more than x members (typically, it will have infinitely many members). But since we are not saying that the extension *has* the number but rather that the extension *is* the number, this is immaterial. All that matters is that we supply an object that plausibly underlies the sense of our judgments of numerical identity and allows for the self-identical sense of "the number G" wherever it is given, and this definition does that.

This leads Frege to the explicit contextual definition of number that he gives in section 68:

The Number which belongs to the concept F is the extension of the concept "equinumerous[2] to the concept F"

In this way, Frege has succeeded in providing a definition of number that vindicates all judgments of the identity of any particular number across the various conceptual contexts in which it is given, as well as treats the number itself as a particular and self-sufficient object with determinate mathematical properties. He has also cleared up the legacy of millennia of confusion about the proper logical definition of number, and has done so with an absolute minimum of non-logical or extra-logical apparatus. All, in fact, we need beyond basic symbolic logic is the minimal assumption that concepts have, in general, extensions: that is (in set-theoretical terms) that for any well-defined concept, there is a determinate range of individuals falling under it, which we can treat collectively and consider as a group. As we'll see next week, however, it is just this assumption that will prove fatal, not only to Frege's attempt to define number in this way, but to the whole project of logicism; and the resulting failure will lead us directly to the deepest problems of predication, meaning, and completeness, and back to the ancient question of the one and the many, especially as it (inevitably) intersects with the problem of the sense or meaning of concepts, and of their determination of a (potential or actual) infinity of their instances.

It might be thought to be a strange feature of Frege's definition that the number (say) 3 turns out to be – in set theoretical terms – the extension of the concept "equinumerous with anything having three members", and so turns out to be – again, in set-theoretical terms, taking extensions to be sets – something like a set of sets, each of which contains three members, but of which there may be vastly more (or even infinitely many). Thus, thought of as a set, 3 will have many more than 3 (maybe even infinitely many) members. We'll see how to avoid this, if it seems inelegant, a bit later, by taking "pure sets" as proxies for these extensions, so that we only need one proxy for each number, and each proxy will have exactly as many elements as the number itself.

[2] In the sense of being able to be correlated with it by Hume's Principle

Building the Series

So far, then, we've got a working definition which fixes the sense of the *identity* of numbers in the sense that it shows what we mean – in ordinary language and logically – when we say that *the number of F's is the same as the number of G's.* In this sense, we already have the logical framework to show what we are doing when we "recognize the same number again" in whatever group or empirical instance we find it; and this gives us, in an important sense, much of what we need to understand what a number "is". This is good, but it's obviously not enough, since we still don't have what we need to say, for example, "the number of books is five", or to say "5=3+2". However, since we have the idea of taking extensions as the *objects* that numbers "are", we can do what we need to do, and even "define" the numbers individually, if we can say *which* extensions they are on the basis of logical definitions. And in order to do this, we need to build up the actual natural numbers (0, 1, 2, etc.) one-by-one, in a regular way.

Right after giving the contextual definition of "number belonging to the concept F" in section 73, Frege accordingly turns to the definition of the first actual number that he defines: namely, 0. We need a way of defining 0 as the number of some concept, and obviously it should be a concept with an *empty* extension – that is, that has *nothing* in its extension. Frege wants to do this in a logical way and so accordingly – although he says we could also do this with *any* concept that has an empty extension (such as "round square" or "wooden iron"), we can do it in a more logically pure way by choosing "not identical with itself". Since *everything* (Frege is assuming) is identical with itself, we can be sure at the outset that this will be a concept with empty extension, and accordingly that its number will be 0. So now we have defined, not only the sense of "the number of F's is…." in general, but also the sense of "The number of F's is 0". And we can also begin to talk of the properties of 0 itself, etc.

Of course, we still need more. So next we'll define the sense of the expression (remember we are still, in accordance with the context principle, working with whole sentences rather than with individual phrases or terms such as "successor of") "*n* follows in the series of natural numbers directly after *m*". Effectively, we'll show how to do this in order to get from 0 to 1, and then we'll show that we have the method in place to go from *any arbitrary m* to its successor in the series, i.e. *m*+1. How, though? Since we already know that we "have" a unique 0 – it's the extension that we defined in the last paragraph – we can now take 1 to be the number of the concept "identical with 0", i.e. the number of '0's, i.e. one! This means – remember – that the number one is the extension of the concept "equinumerous to the concept of being a 0" and since *that* concept is instantiated only once – there is only one thing in its extension (since 0 is unique) – we will have what we need. And we have – as Frege points out – a *purely logical* definition of both zero and one, neither of which (p. 90) requires anything like an observation of a fact or a provision of something intuitive.

So, given the definitions of 0 and 1, all we need now is a definition of the succession of numbers, so that we can go from an arbitrary n to n+1. First, in section 79-80, Frege defines membership in a series ending with *n* in general, so that we will have a concept, i.e. "membership in a series of natural numbers ending with *n"*, whose number is n+1 (since in the series 0, 1, 2, 3…n, there are n+1 members). (in general, as he says in section 55, a number n+1 belongs to a concept F, if there is an object, *a*, falling under F and such that the number *n* belongs to the concept "falling under F, but not *a*". Here, the

relevant *a* is just n). Then he sketches the outline of a proof that the series of natural numbers, thus defined, has no last member, since the idea is that we can always just take, for each n, the concept of being a member of the series ending in n. Later on we'll restate this with the general axioms of Peano arithmetic, and also give it a set-theoretical underpinning that's convenient. But note that we already have the basic idea we need for *all* of the natural numbers and have thus given them a completely logically rigorous definition that is also completely workable for mathematical purposes – thus avoiding any kind of empirical or intuitive or operationalist underpinning.

First glimpse of the mathematical infinite, rigorously defined

Thus we have *all* of the finite natural numbers -- but wait – there's even more that we can get from Frege's methods for defining "the number of" in general and the natural numbers in particular! In section 84, Frege (following Cantor's set theoretical discoveries from a few years earlier) provides the first logically rigorous definition of the mathematical infinite or of (what we'll come to deal with later) as the "first" infinity, the number called (by Cantor) omega (ω). In fact, this is easily done, given Frege's definitions. Given that we have a concept of a finite number, ω is just the number of that concept! That is, ω (or the "first" infinite number) is just the number of the series of *all* the finite numbers, taken together. In a certain way, given Frege's definitions, it is also the *first* number that "follows after" *all* the finite numbers, in the ordered series (later on we'll call these the series of ordinals). Given our definitions and analysis, the infinite becomes perfectly mathematically tractable, and is just as mathematically and logically well-defined as *any* of the familiar finite numbers, and we can (given the definitions of arithmetic operations that we'll get in a minute) even work with it mathematically perfectly as well as with finite numbers. As Frege points out, this is so *even if* we think that we have no intuition corresponding to the infinite, or cannot possibly imagine it, or think it "goes beyond" somehow our powers of understanding or comprehension.

Firming up the structure definitions – the Peano Axioms

Given what we've done so far, we're now in a position to define the whole series of natural numbers with five general axioms, and we'll also be able to define in these terms the operations of addition (+) and multiplication. The Peano axioms were first given by Giuseppe Peano in 1889, building on work by Dedekind and Pearce, five years after the *Foundations* was published by Frege in 1884. But they plausibly capture exactly the same structure that Frege is talking about when he talks about the "series of natural numbers" in the *Foundations.*

For Peano arithmetic (**PA**), we'll add to basic logic (with equality) two things: a single special element, called "**0**", and an operation, **S**() (intuitively, the "successor of" operation). Then we can define **PA** with the following rules:

1. **0** is a natural number
2. For every natural number n, **S**(n) is a natural number
3. There is no n for which **S**(n)=0

4. For all natural numbers m and n, if **S**(m)=**S**(n) then m=n

5. For any one-place predicate ϕ(x), if

    a. ϕ(**0**) and

    b. if ϕ(x) then ϕ(**S**(x))

    then ϕ(x) holds for all natural numbers

Given these rules, we can easily define addition and multiplication. In doing so, we use something called a *recursive definition*; we define how it works for the "base case" of 0 and then give rules that allow us to define the general case based on that.

Addition:

For any natural numbers m, n,

1. m+**0** = m

2. m + **S**(n) = **S** (m+n)

And **multiplication:**

For any natural numbers m, n,

1. m X **0** = **0**

2. m X **S**(n) = m + (m X n)

The recursive definitions might seem "circular" in a way – after all, the second rule in each case uses the very notion we're trying to define! However, given that the definition will always "bottom out" in a basic case (the case of 0), this is ok. It doesn't introduce any actual circularity.

The system of Peano arithmetic lets us "define" the natural numbers, in a way, and introduce the notions of addition and multiplication. This is a good start, but we'll want to do a lot more. We're going to want to talk about all the operations on numbers, all kinds of numbers (not just naturals), and everything else that we refer to in mathematics. Also, the fifth Peano axiom is a bit problematic, since it uses the notion of an *arbitrary* predicate or property (which we need to do proofs by "mathematical induction", such as the proof that all numbers are either even or odd). If we want to be as rigorous as possible about this, and accomplish a *general* foundation not only for arithmetic but for mathematics as a whole, we'll need to switch to the much more powerful language of sets.

Philosophy/Math 415-515
History and Philosophy of Mathematics

Notes: Week 5

Building the Numbers using sets

Last week, we followed in detail both Frege's contextual definition of the phrase "the number of…" and his specific definitions, in terms of extensions, of the individual numbers starting with 0, and going all the way up to the "first" infinite number, ω, and we got a look at Peano's axioms for arithmetic. This is all helpful and will give us a lot, but we'll also want to know how to define other kinds of numbers besides the natural numbers, and also to use the more powerful language of set theory, which can be used to define mathematical "entities" generally. To begin with, let's define the natural numbers again, this time in terms of sets. There are actually several ways we could do this, but the most standard way (following von Neumann) is to identify 0 with the empty set, and then identify each succeeding number with the set containing *all* of its predecessors. Thus:

$0 = \emptyset$

$1 = \{0\} = \{\emptyset\}$

$2 = \{0,1\} = \{\emptyset,\{\emptyset\}\}$

$3 = \{0, 1, 2\} = \{\emptyset, \{\emptyset\}, \{\emptyset,\{\emptyset\}\}\}$

Etc.

Given this, we can also just define the *successor* of X, using our set-theoretical notion of union, as $X \cup \{X\}$, and we get the first four Peano axioms "for free". If there is an infinite set (as there is, according to the axiom of infinity, below), we also get a set-theoretical translation of the fifth Peano axiom.

This way of doing the reduction has several other nice features. It gives us a way to 'build' the natural numbers hierarchically, starting with (almost literally) nothing, the empty set. Because of this rigorous rule, we can show that sets 'built' this way are *well-ordered*: that is, every number has exactly one successor, and there are no two different numbers that have the same successor. The sets that we're using for numbers are all ***transitive*** sets (a set X is *transitive* if and only if: whenever $A \in B$ and $B \in X$, $A \in X$). At each stage, moreover, the number *n*, thought of as a set, has exactly *n* members – which is exactly what we want to be the case. In a certain sense, each number just *is* the set of all the numbers that precede it. And since we can always, at each stage, build a new stage, this set of predecessors will always exist, at least for any finite n.

What about when we get to infinity, though? We've talked a lot already about **N**, the set of all finite natural numbers. Well, identifying numbers with sets this way, what happens when we continue the process of "building" numbers up to here, and identify this set itself with a number? We get a new

number, which is often called **ω**. Like any of the finite natural numbers, **ω** just consists in the set of all of its predecessors -- in this case, the set of all finite natural numbers. In fact, **ω** – the first 'infinite' number – is just our old friend **N** in a new guise, this time considered as a number somewhere in the ordered series of numbers, rather than *just* as a set.

Now we have the basic materials we need to build up *all* types of numbers out of sets. It's easy to define the set of rationals, **Q**, and the set of integers, **I** (negative as well as positive whole numbers) using *ordered pairs* of natural numbers.[1] For instance, we build the fractional number a/b with the ordered pair <a,b>. We can build the integer -3 using the ordered pair (interpreted as a difference) <2, 5>.[2]

Russell's Paradox

Over the past weeks, we learned some of the basic ideas underlying Cantor's theory of sets (or, as it was sometimes called at that time, theory of classes). We also saw how Frege's attempt to define number – in terms of which numerical judgments are grounded in judgments of the *equinumerousity* (possible 1-1 correlation of the elements) of the extensions of concepts – depends on the notion of sets as specifiable by means of their defining properties. In other words, for Frege's attempt to work, it has to be the case that, in general, we can go from a concept to its extension: that is, from a concept to the set (or class) of individuals that are characterized as falling under that concept. And these extensions have to be, in general, real and determinate in order for the theory to work. We'll talk today about how this theory led to the discovery of a profound and interesting paradox that, as we'll see, has deep implications for how we should think not only about mathematics and logic, but also about language, truth, and the universe as such.

We've seen that the basic idea of a set is the idea of a multiplicity that can be grouped as a unity: a many that can be grouped as a one. This is a very general idea, and Cantor himself was very enthusiastic about it. He even thought that we might understand his notion of a set as a new and rigorous way of thinking about what Plato called "ideas" and what we sometimes today call 'concepts.' Intuitively, a concept is a grouping of *many* as *one*: if I have learned the concept "red," for instance, I can use that one concept to refer to any one of a huge (probably infinite) collection of individuals. In this respect, we might think of the relationship of set membership (symbolized by the '∈') as capturing the underlying structure of predication itself (so that given a concept, such as "being green," the things determined by that concept, or of which the concept can *truly* be predicated – the *extension* of that concept – will just be the elements of the set).

---

[1] The *ordered pair* <m,n> can be identified as the *set* { {m}, {m,n}}.

[2] There is a little wrinkle for both of these definitions, since ordered pairs don't necessarily determine *unique* rational or integral numbers (since, for instance, 3/2=6/4 or 2-5 = 1-4). To deal with this, we actually identify the rationals and the integers with *equivalence classes* of ordered pairs of naturals. (For details, see Maddy or Enderton).

Cantor therefore thought he had formulated the relationship between concepts and what they are predicated of; and as we'll see, there are many senses in which this is actually true. However, the formalization comes with a price, or at any rate it seems to show that there are fundamental limitations to the very nature of concepts themselves.

To see why, let's ask ourselves: what sets are really possible? Is it possible to make a "one" out of any "many," or are there some totalities that "just can't" be formed? And why? At first, it seems like we should be able to make a *single* set out of *any number of individuals* whatsoever. As we saw last time, formally there's no problem with making a set out of individuals that are very different from one another. And we saw that there's no problem with infinite sets either; in practice, very many of the sets we'll be interested in while doing mathematics are going to be infinite, in fact. We also saw that we can specify a set either by giving the whole list or, where this is impractical, by giving a concept or definition that identifies just those things that are in the set.

So it seems at first that we should have a set corresponding to *any* coherent definition that we can give in natural language. This was probably the intuition that Cantor had, and it was certainly Frege's own idea. In the book *Basic Laws of Arithmetic* (*Grundgesetze der Arithmetik*), Frege formulated a basic axiom that involves this principle. It was the fifth one in his system, so he called it "Basic Law V." The underlying principle is also sometimes called (for reasons that will become clear) the "universal comprehension principle"

***Universal Comprehension Principle**:[3] For any property definable in language, there is a set consisting of all and only elements that have that property.

Let's think a little bit about the sets that this principle allows us to form. To begin with, we can use it to form sets like the ones we've discussed:

{x|x is red} or {x|x is a president elected after 1992}

Both of these are perfectly well-defined properties, so there should be no problem with forming the sets, and indeed there is not.

We can also consider sets that are members of themselves. For instance, consider the set M:

M= {x|x is a set with more than three elements}

As we've seen, sets can be elements of other sets, and M is just an example of this: it groups together all the sets with more than three elements. So {0,1} is not an element of M, but {1, 2, 9, 347} is an element of M. What about M itself? Well, does M have more than three elements? Yes, assuredly it does; there

[3] (More technically: Basic Law V is: any two predicates, F and G, have the same extension iff every F is a G, i.e. iff Fx ≡ Gx for all x.) We'll put an asterisk before the name of this principle because it turns out not to be true, at least according to now-standard formulations of set theory. Also, when Frege originally wrote he spoke in terms of "value ranges" or extensions rather than sets or classes; but they are equivalent.

are many more than three sets that have more than three elements each. (In fact there are an infinite number). So, we can conclude, M ∈ M. We can call such sets – ones that are elements of themselves – *reflexive* sets.

This might seem counterintuitive at first, but there are lots of other examples of plausibly reflexive sets we could give. Consider, e.g.,

P = {x|x is a set discussed on this page}

Is P discussed on this page? Of course it is – we're doing it right now! So, along with a few other sets (M and a couple of others), P itself is an element of P.

Or consider: E={x|x is a set definable in English}.

Is E definable in English? Of course it is – we just did define it! So, E itself is an element of E.

So far, none of this is any big problem. We have preserved Frege's basic law V and allowed sets to be defined quite generally, including self-membered sets. But now enter Russell. Since the last years of the 19th century, Russell had been thinking about paradoxes and the foundations of mathematics, and sometime in 1901 he discovered the paradox that still bears his name. We just saw that some sets, intuitively, have the property of being elements of themselves. Now, consider the set of all sets that do *not* have this feature; i.e. the set of all sets that are *not* elements of themselves (call it the Russell set, R)

R = {x| x is *not* an element of x}

Or another way to put it:

R={x|x is *not* reflexive}

There are lots of perfectly normal sets that are elements of R. In fact, *most* sets are, since most sets aren't reflexive. (For instance, the set of all presidents elected since 1992 isn't reflexive, since it itself is (clearly) not a president elected since 1992). But what about R itself? Either R is an element of itself or it is not. Well, suppose R *is* an element of itself (i.e. R∈ R). Then it *is* reflexive, so by the definition of R, it isn't an element of R. But then we have R ∉ R, which contradicts the assumption! So let's try the other possibility, R ∉ R. Then it is *not* reflexive, so by the definition of R, it *is* an element of R. But then we have R∈ R: again, a contradiction! So either way, whether R is or is not assumed to be an element of itself, we have a contradiction. It seems to follow that there can't be a set R. Even though the 'definition' seems perfectly good, there is no set – there can't be any set – that corresponds to it, as Frege's UCP requires.

Russell communicated the paradox to Frege in a letter in 1902, and Frege's response shows how seriously he took the issue here, even though it pretty much ruined his own system by ruining his Basic Law V. It's important to note, as well, that this isn't just a 'minor' or 'side' issue, since it has a deep and central impact on what we should think about the relationship of language and concepts to the world

itself. Frege's Universal Comprehension Principle expresses the intuition that there is a natural link between language and what exists, in the sense that whatever can be talked about – whatever property we can coherently define – can exist as an actual object of thought. If this *isn't* true, as Russell's paradox seems to show, then we'll want to know why. Is it a problem with language, or with the world itself? If the Russell set indeed "doesn't exist," the reason doesn't seem to be the same as the reason why, for instance, unicorns or Santa Claus doesn't exist. These things don't exist but they *could have* existed; in fact, we can indeed think about them and say perfectly well-defined things about them (such as "a unicorn has just one horn,") so in a certain sense they do exist, at least as objects of thought. It seems, by contrast, that the Russell set *necessarily* doesn't exist – there is no way we can even so much as "really" think about it or make it an object of thought. On some views, there are other "things" that don't exist in this sense as well, things like a "round square" or a "wooden iron" whose very *concept* seems to be contradictory. But this is again different from the Russell set itself; for whereas these "things" don't exist because their *concept* is contradictory, the concept defining the Russell set (that of not being reflexive) is not (obviously, at least) contradictory in itself. Moreover, if it *is* contradictory, and hence must be ruled out from the outset, then we will have to rule out lots of other seemingly legitimate concepts as well. For instance, if the concept of *not being reflexive* is contradictory, then the concept of *being reflexive* is contradictory as well. Additionally, we'll have to reconsider the very idea that language can make reference to itself, or to the totality of the universe in which it itself exists. But these seem to be very general and central capacities of language, so it's difficult to see what could justify this kind of restriction.

More generally, Russell's paradox is interesting because it bears on two general and interrelated issues, the issues of *reflexivity* and *totality*. Something like Russell's paradox seems to arise whenever we consider the possibility of a bit of language – say a sentence or a name – making reference to the *whole* of a totality in which that bit of language itself is an element or part. Some such totalities are: the set of all propositions, the totality of language, the set of all propositions uttered by me, the totality of whatever is true, the totality of speech acts, everything that is sayable, everything that is thinkable, or the universe itself. In fact, with all of these cases, the issues of totality and reflexivity are in fact intimately connected to one another. For as soon as we refer to, think of, or acknowledge any of a number of different kinds of totalities, we seemingly have to (implicitly at least), make reference to *the very act* of referring, thinking, or acknowledging as involved in that very totality.

As Russell points out, as well, the same general issues underlie a variety of other closely related paradoxes. Some of these had long been recognized, while some were just coming to light around the time Russell wrote. But before Russell's discovery and elucidation of this common set of issues, none of them were thought to be particularly significant or deep; the ones that were known already were, for the most part, treated as minor puzzles or amusements. It took Russell's work and its recognition by the leading theorists of set theory to see that they have deep and important implications for metaphysics, ontology, and semantics as well.

1) The Liar Paradox. This paradox, in many ways the simplest of the group, was first suggested by the (minor) ancient Greek philosopher Epimenides (and there's a version of it in the New Testament as well). Consider the Cretan who says:

"All Cretans lie constantly and never tell the truth."

Is what he said true or false? If it is true, then as a Cretan, he must be lying. Therefore it must be false. Conversely, if it is false, then Cretans do not lie constantly, and it is at least possible that it is true. This is not yet an actual paradox, since on the second assumption (that what the Cretan says is false), we might consistently hold that Cretans sometimes tell the truth and sometimes lie, and that this one is indeed lying now. We can rule this out and create an actual paradox, however, by considering:

"This sentence is false" or

"I am now lying."

2) Grelling's Paradox. Call an adjective "autological" if and only if it describes itself. For example, "short" is autological since it is a short word, but "long" is not. If an adjective is not autological, call it "heterological". Is "heterological" heterological?

3) Berry's Paradox. Consider the number of syllables it takes to name finite numbers in English. Thus, 5 has a one-syllable name ("five") and 17 has a three-syllable name: ("seventeen"). There's some finite number of distinct syllables in the English language (probably a few hundred); therefore as we count up, we'll exhaust all the numbers nameable by one syllable, and then the numbers nameable in two syllables or less, then the numbers nameable in three syllables or less, etc. Therefore, there must be a number, *n,* which is "the least integer not nameable in fewer than nineteen syllables." (In fact, this is 111,777: 'one hundred and eleven thousand, seven hundred and seventy-seven'). But consider the description in the double quotes itself: it has just 18 syllables! But it is itself a name for *n*; thus the least integer not nameable in fewer than 19 syllables can be named in 18 syllables . This is a contradiction.

4) Burali-Forti's Paradox (intuitive version). This paradox was discovered in 1897, even before Russell's paradox, by a student of Cantor's. There is actually some evidence that Cantor himself discovered it even earlier, but didn't write or speak about it.

As we saw last week, we can define the series of "ordinal" numbers as a series in which each member, *n*, has exactly *n* predecessors in the series. The series of natural numbers starting with 0 has this property; thus 1 has exactly one predecessor (namely, 0) and 3 has three predecessors (0, 1, 2), etc. Each of these series of predecessors itself has an (ordinal) number, a unique number of elements in the series. (Again, we'll have to show this more rigorously later, but maybe we can take it on faith now). The series of ordinals *up to and including n*, then, always exceeds *n* by one: thus the series of elements *up to and including* 3 has 4 elements, etc. Now consider the series of *all* ordinals (**IF this series exists**); we can call the number of elements in this series **Ω**. Since **Ω** is an ordinal, the series of all ordinals also includes **Ω**. But then, this is the series of all ordinals *up to and including* **Ω**; so by the previous reasoning, its number

of elements is **Ω** +1. This is a contradiction: the number of the series of *all* ordinals is both **Ω** and **Ω** +1. (Question: does the parenthetical remark in bold above point to at least one way of avoiding the contradiction?)

5) Cantor's Paradox. This one was also discovered by Cantor, though again he never wrote about it publicly.

We can define a 'pure' set as one that has only other sets as elements (thus no elements such as persons or physical objects), including possibly the empty set. Now consider the *totality of all such sets*; call it **V.** Now, we know intuitively (and we'll soon prove formally) that for any set A, the power set of A (also called $\mathcal{P}$(A)), has more elements than does A itself. (In fact, for finite sets at any rate, we can easily show that if A has *n* elements, then $\mathcal{P}$(A) has 2^n elements.) Now, if **V** is a set, then it has a power set, $\mathcal{P}$(**V**), and $\mathcal{P}$(**V**) has more elements than **V** itself does; that is, there is some y $\in \mathcal{P}$(**V**) that is not an element of **V**. But **V** was supposed to include *all* the sets. This is a contradiction.

Confronting the Paradoxes

Russell's paradox thus seems to formulate some very general issues and problems. Not only will we, obviously, have to clear up these problems if we want to have a well-formulated and consistent set theory, but since they seem to bear on linguistic reference itself in the ways we've discussed, it seems we must clear them up if we are going to have a good picture of how language (or thinking) relates to the world at all. So, what should we do? Historically, Russell's paradox was deeply influential on the first set of problems, for it was in the attempt to resolve it that the axioms of set theory were largely formed and agreed upon. More recently, philosophers have begun to consider more directly its bearing on the second set of problems, on the relation of language to the world and the possibility of considering (thinking or speaking of) different sorts of totalities, and on the nature of reflexivity or self-reference itself.

As Russell notices clearly in the 1908 paper where he developed the consequences of his paradox in detail, all of the paradoxes we discussed above have a common character. In particular:

> In all the above contradictions (which are merely selections from an indefinite number) there is a common characteristic, which we may describe as self-reference or reflexiveness. The remark of Epimenides must include itself in its own scope … In each contradiction something is said about *all* cases of some kind, and from what is said a new case seems to be generated, which both is and is not of the same kind as the cases of which all were concerned in what was said.[4]

[4] Russell (1908), p. 154.

Another way to put this, as we saw above, is to say that each of the paradoxes involves the interrelated issues of *totality* and *reflexivity*. In each case, we make reference to a totality; but in so doing, we also make reference (implicitly at least) to the act of making reference itself. This reference turns out in each case to be problematic, in that it immediately generates a new case (e.g. the Russell set, the definition of "111,777", **Ω,** or $\mathcal{P}$(**V**)) which "both is and is not" a member of the totality.

How, then, should we deal with the problem in its general form? Since the problem obviously seems to arise from the combination of the issues of *reflexive reference* and *reference to totalities*, we might simply try to restrict or 'ban' either kind of reference, or both. This is going to have counter-intuitive consequences, of course (such as that we can't refer to "this very sentence" or "all propositions"), but perhaps this is the price we have to pay for a solution to the paradox. We might start by just banning reflexivity. For instance, we could just *stipulate* that there is no such thing as a reflexive set; that is, there is no set A such that A ∈ A. This is already quite counter-intuitive, since it requires that lots of sets that seem intuitively to exist (e.g. the set of sets with more than 5 elements; the set of sets discussed in these notes; all that is sayable) don't actually exist. But in any case, this restriction alone won't give us a solution to the paradox. For even if we make the stipulation, we still must consider whether there is a *universal* set (call it U) or a set of all sets, and if so, whether this set is an element of itself. Supposing that there is such a set U, it is an element of the set of all sets (since it is itself a set). But then U is an element of itself: U ∈ U. But just this is what we had attempted to prohibit.

Again, we have a contradiction, so it looks like just banning reflexivity alone won't suffice. Accordingly, we might next try to ban reference to *totalities*, or at any rate those totalities that lead to Russell-style reflexive problems. This was Russell's own first idea, and he formulates it in a principle that has come to be called the *vicious-circle* principle (VCP):

> This leads us to the rule: 'Whatever involves *all* of a collection must not be one of the collection'; or, conversely: 'If, provided a certain collection had a total, it would have members only definable in terms of that total, then the said collection has no total.'[5]

We can put the principle even more simply:

> VCP: No totality can contain members definable only in terms of itself.

The idea is that any such totality, which contains members defined in terms of itself, creates a kind of vicious circle. For if, say, the set T contains an element, t, which is in turn definable only in terms of T, then T is (extensionally) defined by t, but t is defined by T, so we might indeed object to both as circular. The correlate of this for propositions and concepts is:

> VCP: No concept can refer to a totality of objects definable only in terms of itself.

[5] Russell (1908), p. 155.

As Russell actually notes, this principle poses problems for a number of the totalities we refer to in ordinary language and discourse. For instance, consider *language*, defined as the totality of propositions. Since "propositions" must itself be defined in terms of language, if the VCP holds, it is apparently impossible to refer to this totality, or to language as a whole. Similarly, various statements about limits are going to turn out to be impossible. Consider a proposition defining the limits of the thinkable, for instance "Everything thinkable (and nothing unthinkable) can be written down." Since this limit to the totality of the thinkable is apparently itself thinkable, it violates the VCP; therefore, if the VCP is correct, it is impossible to refer to this limit or say anything about it. But it seems we just did so! If it is indeed possible to think about and refer to such totalities, then, it seems that the VCP is incorrect, at least if applied to ordinary language and reference. [6]

There is, moreover, another problem, which, as Russell also notes, we run into as soon as we try to use the VCP actually to decide which sets actually *do*, and which *don't*, exist. The problem is that in order to "ban" any particular set from existing, we must first apparently refer to it! But we can't refer to a set that doesn't exist. Thus:

> The above principle [i.e. the VCP] is, however, purely negative in its scope. It suffices to show that many theories are wrong, but it does not show how the errors are to be rectified. We can not say: 'When I speak of all propositions, I mean all except those in which 'all propositions' are mentioned'; for in this explanation we have mentioned the propositions in which all propositions are mentioned, which we cannot do significantly. It is impossible to avoid mentioning a thing by mentioning that we won't mention it. One might as well, in talking to a man with a long nose, say: 'When I speak of noses, I except such as are inordinately long', which would not be a very successful effort to avoid a painful topic. Thus it is necessary, if we are not to sin against the above negative principle, to construct our logic without mentioning such things as 'all propositions' or 'all properties', and without even having to say that we are excluding such things. The exclusion must result naturally and inevitably from our positive

[6] Russell argues (in sections II and III) of Russell (1908) that we can solve this problem, at least to a certain extent, by drawing a clear distinction between speaking of "any" and speaking of "all." For instance, when I say that "All men are mortal," I am not really asserting anything of the totality "all men," (i.e., I am *not* saying that humanity itself will one day die out); rather, I'm saying that any arbitrary man – whichever one you like – is mortal. Maintaining this distinction, we might agree with Russell (p. 162) that, for instance, when we assert the law of the excluded middle: "All propositions are either true or false," we're not really saying anything about the totality of *all* propositions, but just saying that any individual proposition is either true or false. Russell then denies that it's even possible at all to speak of the totality of *all* propositions. However, this only solves the problem to a certain extent. For even if, in these cases, we are not really talking about totalities, there certainly are cases where we do at least seem to talk about these totalities, and we really are talking about *all* (or the whole) of a certain type: consider, e.g. "Mankind is a noble species" or "Language (the systematic totality of propositions) is a structure of regular signs."

> doctrines, which must make it plain that 'all propositions' and 'all properties' are meaningless phrases.[7]

Suppose, for instance, we want to prohibit the existence of the totality of all propositions. Then we must say: "The totality of all propositions doesn't exist. There is no such totality that we can consider or make reference to." But just in saying this, we have already considered and made reference to this very totality! It seems that there is no straightforward way to positively formulate the restriction we want, without violating that very restriction, since we can't prohibit reference to the requisite totalities without making reference to them ourselves in the very act of prohibition.

Another possibility, the one that Russell next pursues, is to define the principles and axioms of set theory in such a way that the problematic sets are effectively ruled out, even if none of these axioms actually *mention* the sets that we are thereby ruling out. If we can do this, we can overcome the last problem, and still avoid the problematic sets that lead to paradox and contradiction. This is in fact the strategy that was pursued by most of the founders of set theory, and it played a key role in ideas about what the underlying axioms and assumptions of set theory should be. Russell's own solution (for a time, at least) was the notorious *theory of types*. According to the theory of types, the whole universe of sets is inherently stratified into a hierarchical series of levels or "types." There is, moreover, a single formal rule that specifies their possible relationships: a set of type-level *n* can only include elements of levels less than *n*. Thus, a set of level 4 (for instance) can only include sets that are at "earlier" levels (1, 2, or 3); it can never include another set of level 4 or higher. This effectively prohibits self-membership, since in order to include itself a set would have to include another (itself) at the same type-level as itself, which is prohibited. It also formulates a (somewhat) natural intuition about how sets might actually be formed or produced. The idea is that at any time, we only have "available" sets of a certain degree of complexity; we can form a new set, which will itself have a higher degree of complexity, but we never have a set that is *more* (or just as) complex antecedently available to play a role in forming a set that is *less* complex. We might start, for instance, with the empty set at level one; then the set containing the empty set ($\{\emptyset\}$) will be at level two; the set containing this set will be at level three, etc. Then we'll have a strictly ordered hierarchy, and we'll never have any set containing itself. We'll also never 'form' the "set of all sets," since at each level we can always just create one more.

Another, closely related possibility is the one actually pursued by Zermelo and Fraenkel, the philosophers who laid down the foundations of set theory as it's most commonly theorized and discussed today. (In fact, standard set theory is sometimes called 'ZF' set theory, for Zermelo and Fraenkel.) This is the solution of laying down *axioms* – absolute first principles – that effectively prohibit the existence of reflexive sets, or of the set of all sets. The completed ZF set theory does this by means of two distinct axioms (there are several other axioms, as well, which we'll learn next week). The first of these is the Axiom of Foundation (sometimes called the Axiom of Regularity).

[7] Russell (1908), p. 155.

Axiom of Foundation: Any non-empty set A contains at least one element, B, such that the *intersection* of B and A is empty.

This effectively prohibits any self-membered set from existing, but we need to think a little about why. To see this, suppose that there is some self-membered set, S. Then S ∈ S. Now consider the set T = {S}. (This set exists, by another axiom, the axiom of *pairing*, which we'll learn about later). Now T contains only one element, namely S. But the intersection of T with S is just S – therefore it's not empty, and the Axiom is violated. Accordingly, the Axiom of Foundation prohibits that there will ever be any self-membered sets. It does so, intuitively, by establishing that for *any* set, we'll be able to "decompose" it into elements, and then decompose its elements into elements, successively, and that at some point we'll reach something that's not further decomposable. In other words, we'll never run into the kind of vicious circle that Russell was worried about.

This by itself, however, doesn't resolve the status of the Russell set, or the Universal set of all sets. For we might, as far as the axiom of foundation goes, still have a set of all sets – it would just have to be decomposable ultimately into simple elements. And we'd still face the tricky question of whether this set, being a set, is an element of itself (or if it isn't, why it isn't.) Also, the Axiom of Foundation doesn't tell us positively which sets do exist – it just implies about certain sets that they *don't* exist. To solve this problem, Zermelo and Fraenkel introduced another axiom, the Axiom of Separation.[8] This axiom is also sometimes called the Axiom of Limited Comprehension, and it amounts to a kind of limitation of Frege's original Universal Comprehension Principle, or Basic Law V (which has, of course, been rejected owing to Russell's result).

Axiom (schema) of Separation*: Given any set B* and any well-defined property **ϕ**, the set A= {x ∈ B| x is **ϕ**} exists.

In other words, the Axiom of Separation says that given any 'already existing' set, B, we can *separate out* just those elements of B that have a certain definable property, **ϕ**, thus creating the set A as a subset. This is a restricted form of Frege's original comprehension principle. For it doesn't say that we can just create a set corresponding to any property (as the original principle did): just that if we already "have" a set to draw on, we can *separate out* the elements from that set that have the property. The important consequence is that this appears to establish that we can't 'create' a set of all sets. For in order to do so, according to the axiom, we'd have to already have a set to separate this out of; but the only set that we could use would apparently be the set of all sets itself. So we can't 'build' this set unless we already have it, which seems to give good reason to think that we can in fact never 'build' or form it in any way.

These two axioms are, as noted, enshrined in the most standard "system" of set theory, the ZF system (though there are indeed other systems of set theory that suspend one or both of them) and effectively

[8] This is actually not a single axiom, properly speaking, but an *axiom schema*. That's because it's defined not just for one property **ϕ**, but for any such property; there are thus an infinite number of "instances" corresponding to the infinite number of properties we can put in for **ϕ**. We'll worry more about this wrinkle later.

formulate a kind of intuition or "picture" of what sets exist as a whole. A more general version of what is common to this "picture" and also to Russell's theory of types is sometimes called the "iterative conception" of sets. This is as close as anything comes to the "standard" picture of the set-theoretical "universe" and it takes the form of a big open "V". (See the picture from Moore's *The Infinite*, p. 157). The idea is that at the very bottom are the simplest sets, the 'singleton' sets each containing just one element that is not itself a set (if we allow sets to have non-set members) or perhaps, even more simply, the single, unique empty set. Then a bit higher up (or at the next 'level' in Russell's hierarchy), we get sets containing these base-level sets. There are more of these than at the first level; and then at the next level, we get sets containing those at the second level (of which there are still more), and so on. We know that we can form each higher level by taking the power set of the sets at the lower level, so it seems that there is no limit to how far we can go. However, because of the consequences of Russell's paradox, there's no stage at which we finish; the top of the V is open-ended, and has no limit. Also, we never, at any point, find a self-membered set; for such a set, as we've seen, can never be "formed" out of simpler ones. It's important to notice, though, that the fact that each set has a determinate level or "size" doesn't mean that some (indeed many!) of the sets aren't infinite. In fact, with the help of an additional axiom (the axiom of infinity, which we'll meet later), we can insure that not only is there an infinite set, but there are infinitely many infinite sets, of increasing levels of "size" or complexity.

The iterative conception solves the problem presented by Russell's paradox pretty nicely, *if* we accept the intuitions underlying it and formulated in the standard axioms. If we do accept these and work within them, there will be no self-membered sets and there will also be no universal sets (such as the set of all sets), so there's no way that the whole set of problems that Russell raised can arise. But this is true only if we accept these intuitions. As I've tried to mark with the use of scare quotes in the discussion above, the intuitions underlying the iterative conception are, in general, highly "constructivist." That is, they more or less all depend on the idea that sets are "made" or "constructed" out of other sets, so that we can only "form" a set at a certain level if there are already "available" other sets "out of which" we can form it. This is what seemingly prevents us from (ever) making a set of all sets: for we'll never have the "raw materials" available out of which to form it if we cannot presuppose it already.

These intuitions are somewhat plausible with respect to "pure sets" themselves, at least if we think of such sets as things that inherently can *only* be formed by means of grouping together or separating out from other such things. But how plausible are they when applied to the more general issues of the nature of concepts and linguistic reference? The answer is: *not very plausible*. Consider all of the following normal-sounding statements:

Everything that is thinkable can be expressed in language.

Language (as a whole) is a system of signs.

The universe contains many wonderful things.

There is more in heaven and earth than is dreamed of in your philosophy.

These claims might themselves turn out to be true or false, but it does seem that they are meaningful. If the constructivist intuitions enshrined in standard set theory and in the iterative conception of the set are correct, though, we can't even (so much as *express*) them meaningfully; for they all refer to totalities which, according to those intuitions, can't exist even as objects of thought. This is because they all refer to infinite totalities (the thinkable, language, the universe, all meaningful propositions) that indeed probably can't be *constructed* by grouping together separate elements. But does this mean that the totalities indeed don't exist? It doesn't seem so. It is, at any rate, far from evident that we must have the ability to "put together" these (seeming) sets from separate elements before we can have any concrete idea of them, or any ability to refer to them meaningfully.

Thus, the implications of the standard conception of sets for ordinary language and reference are indeed counterintuitive and surprising. But the standard conception was itself, as we've seen, largely formulated in response to Russell's paradox, and was meant to be a satisfying way of dealing with it, or at least preventing it from ever coming up in the formal theory itself. If it is indeed unsatisfying in dealing with ordinary cases of concepts and language, however, then this suggests that we might have to revisit Russell's paradox, and think about other ways of responding to it or accommodating it without imposing the constructivist intuitions of the standard, iterative conception of the set. We might, for instance, just have to bite the bullet and live with paradox and contradiction in ordinary life (if not in mathematics), allowing that paradoxes like Russell's paradox (and the Liar) will always arise due to problems of ordinary language, and there is little we can do to formalize them away or rule them out. Or another possibility – one that is quite interesting and has only recently begun to be explored – is that we might try to *formalize the paradox itself*, and see what consequences (including, perhaps, contradictory ones!) might flow from it. That is, we might try out the hypothesis that *in mathematics as well as in ordinary language*, there are certain specific cases where this kind of paradox will arise – cases, for instance, in which there genuinely *are* self-membered sets, or self-referring concepts, or cases in which we genuinely (and meaningfully) *do* talk about totalities in which the very act of speaking is included. We might indeed have to acknowledge certain necessary contradictions here, but this doesn't mean that there will be contradictions everywhere, or that the acknowledgment will ruin any possibility of formalizing thought and language at all. In fact, we might use just this acknowledgment as a way of studying something that is in fact very important to philosophy historically, and continues to define many of its problems today: the problem of thinking about the real existence of contradictions, and especially those contradictions that arise at the limits of thought and language, where we (as philosophers) try to grasp and express what delimits the totality of all that can be thinkable or sayable, and thus seem consigned to take a position "beyond" these very limits themselves.

Cantor's Theorem

As we've already seen, for finite sets at any rate, the power set of a given set is always larger – contains more elements – than the set itself. (In fact, we showed in class that if a finite set is of size n, then its power set is of size 2^n). But what about infinite sets? As we saw already on the first day of class, one of Cantor's most profound results was to show that there are *many* infinite sets with different sizes. We proved a particular case of this with the "diagonalization" argument that shows that there are more real

numbers than natural numbers. But this is just a specific form of a more general result that Cantor had reached by 1891. The result is that for *any* set x (finite or infinite), the power set, $\mathcal{P}(x)$, has strictly more members than x itself. This means that for any set, finite or infinite, we can generate an infinite series of larger ones; and there is no evident stopping point.

Cantor's Theorem: For *any* set x, the power set, $\mathcal{P}(x)$, has strictly more members than x itself.

Proof: First, we can easily establish that $\mathcal{P}(x)$ has at least as many elements as x itself. To see this, note that we can establish a one-to-one correlation between all the elements of x and some (not necessarily all!) elements of $\mathcal{P}(x)$. Just correlate each $y \in x$ to the singleton set {y}. Since $\mathcal{P}(x)$ is the power set of x, it contains all the subsets of x, and thus it contains each of these singleton sets.

Now we need to establish that there is no one-to-one correlation between all the elements of x and all the elements of $\mathcal{P}(x)$. Once again, we'll do a proof by *reductio*; that is, we'll assume (contrary to fact) that there is some such correlation, and then we'll derive a contradiction. So suppose there is some function, f, that is a one-to-one correlation from elements of x to elements of $\mathcal{P}(x)$. Now, on this assumption, f is going to map each element, y, in x to some set, f(y), in $\mathcal{P}(x)$. Sometimes, y will map to a set in which y itself is an element; but this may also not be the case. In particular, let's now consider z, which is the set of all the y's that are mapped to sets that they are *not* elements of:

$Z = \{ y \in x \mid y \notin f(y)\}$.

Now, Z itself is a *subset* of x, so it's also an *element* of $\mathcal{P}(x)$. Therefore (since we're using f to map onto each element of $\mathcal{P}(x)$), there is some element of x, call it w, such that Z=f(w). Now we can ask: is w itself an element of Z? Well, if it is an element of Z, then, because of the way Z is defined, it's *not* an element of f(w). (After all, Z is defined as the set of all the elements that *aren't* in the set that f correlates them to). But Z=f(w)! So if it is an element of Z, it is not (CONTRADICTION). What, then, if w is *not* an element of Z? Well, then $w \notin f(w)$, so by the definition of Z above, it is in Z. Again, we have a CONTRADICTION. Therefore, if w is an element of Z, it is not, and if it is not an element of Z, it is. Therefore there is no one-to-one correlation f.

Cantor's Theorem is interesting because its proof closely resembles two things we've seen before. First, the way the proof itself works closely resembles Russell's paradox (note that, just as in Russell's paradox, we need to consider a set (here, the set Z) of all elements that *don't* have a crucial property, and then we derive the contradiction that if something has the property it doesn't, and vice versa). In fact, Russell later said that it was in thinking about Cantor's theorem that he first came up with the paradox itself. But there are also differences: Cantor's theorem doesn't deal with the totality of *all* sets, but only with the relations between particular sets and their power sets; and because of this difference, Cantor's theorem isn't a paradox, but rather just a result.

Second, though, the proof of Cantor's theorem is just a generalization of the "diagonal argument" about the naturals and the reals that we discussed the first day of class. To see the connection, notice that we can represent each real number between 0 and 1 (let's again stick with these for ease) as an (infinite) *set*

of natural numbers. How? Well, first, let's re-write them in binary notation. Thus, we'll have, e.g., *a* = .1 0 1 0 0 … (this corresponds roughly to .625 in digital, but since we don't have the whole expansion, we don't know exactly what it will figure out to ultimately). Now all we have to do is number the places of this expansion with natural numbers: 1 for the first place, 2 for the second, etc., and identify the real number with the set of places for which it has a '1'. So, for instance, our *a* will be identified with the set containing 1, 3, and whatever other numbers correspond to the places in a where there is a 1 rather than a 0. So we can think of each real number as just a set of naturals; and then the set of *all* reals (between 0 and 1) will just correspond to the set of subsets of the set **N** of all naturals, or in other words the *power set* of **N**. If we want to show that this power set is bigger than the original set, we just diagonalize on the list, as we did on the first day; if we're representing the reals in binary notation, as we did above, we just alter the diagonalization rule so that we always change a '1' to a '0' and vice versa (rather than changing n to n+1, or 9 to 0, as we did the first day). This will give us exactly the same result, and it's also exactly the same as what we did in the general proof of Cantor's theorem. There, as in the numerical diagonalization proof, we assumed that there is a one-to-one correlation in order to produce a contradiction; and then, through our diagonal set (the set Z) we produced just such a contradiction.

Actually, these two similarities provide clues to a deeper homology, as we'll see next week. In fact, just as it is the general operation of diagonalization that underlies Cantor's theorem, and hence allows us to form ever-larger sets, including infinite ones, from sets already presented, diagonalization also structurally underlies Russell's paradox itself. It is thus highly interesting, and suggestive, that Russell's paradox leads directly to contradiction, whereas the (only slightly different) application of diagonalization in Cantor's theorem, or in relation to the relative sizes of sets of numbers, does *not* produce contradiction, but rather the vast (infinite) hierarchy of infinite (or transfinite) sets and numbers.

Philosophy 415: Fall 2025
History and Philosophy of Mathematics

Notes: Week 6

Last week, we considered the foundations of set theory and some of the paradoxes that set theoreticians quickly encountered concerning totality and reflexivity. As we saw, this demonstrated that it's impossible to hold something like a universal comprehension principle – Frege's basic law V – without contradiction, and so that if we are going to put set theory on a "solid" foundation, we'll have to come up with alternative ways to axiomatize it and think about its structure. We also saw that the paradoxes of totality and reflexivity are closely connected with issues about the infinite, and in particular about which kinds of infinite sets can, and which can't, exist without contradiction. But does this mean that the infinite is, after all, simply paradoxical? Luckily for set theory as well as for the foundations of mathematics, it does not.

As we saw last week in thinking about Russell's own suggestion that the paradoxes result from a particular kind of combination between *totality* and *reflexivity*, we can't really just prevent the paradoxes from arising by saying or stipulating that there can't be any reflexivity in our language, or that there can't be any totalities. If we were to ban all totalities, we would ban just the kind of generality we need for general mathematical results; and we would also wind up depending on the totalities that we were banning, just by stating the ban. Another possibility, the one that Russell next pursues, is to define the principles and axioms of set theory in such a way that the problematic contradictory sets are effectively ruled out, even if none of these axioms actually *mention* the sets that we are thereby ruling out. If we can do this, we can overcome the last problem, and still avoid the problematic sets that lead to paradox and contradiction. This is in fact the strategy that was pursued by most of the founders of set theory, and it played a key role in ideas about what the underlying axioms and assumptions of set theory should be. Russell's own solution (for a time, at least) was the notorious *theory of types*. According to the theory of types, the whole universe of sets is inherently stratified into a hierarchical series of levels or "types." There is, moreover, a single formal rule that specifies their possible relationships: a set of type-level *n* can only include elements of levels less than *n*. Thus, a set of level 4 (for instance) can only include sets that are at "earlier" levels (1, 2, or 3); it can never include another set of level 4 or higher. This effectively prohibits self-membership, since in order to include itself a set would have to include another (itself) at the same type-level as itself, which is prohibited. It also formulates a (somewhat) natural intuition about how sets might actually be formed or produced. The idea is that at any time, we only have "available" sets of a certain degree of complexity; we can form a new set, which will itself have a higher degree of complexity, but we never have a set that is *more* (or just as) complex antecedently available to play a role in forming a set that is *less* complex. We might start, for instance, with the empty set at level one; then the set containing the empty set ($\{\emptyset\}$) will be at level two; the set containing this set will be at level three, etc. Then we'll have a strictly ordered hierarchy, and we'll never have any set containing itself. We'll also never 'form' the "set of all sets," since at each level we can always just create one more.

Another, closely related possibility is the one actually pursued by Zermelo and Fraenkel, the philosophers who laid down the foundations of set theory as it's most commonly theorized and discussed today. (In fact, standard set theory is sometimes called 'ZF' set theory, for Zermelo and Fraenkel.) This is the solution of laying down *axioms* – absolute first principles – that effectively prohibit the existence of reflexive sets, or of the set of all sets. The completed ZF set theory does this by means of two distinct axioms (there are several other axioms, as well, which we'll learn next week). The first of these is the Axiom of Foundation (sometimes called the Axiom of Regularity).

Axiom of Foundation: Any non-empty set A contains at least one element, B, such that the *intersection* of B and A is empty.

This effectively prohibits any self-membered set from existing, but we need to think a little about why. To see this, suppose that there is some self-membered set, S. Then S $\in$ S. Now consider the set T = {S}. (This set exists, by another axiom, the axiom of *pairing*, which we'll learn about later). Now T contains only one element, namely S. But the intersection of T with S is just S – therefore it's not empty, and the Axiom is violated. Accordingly, the Axiom of Foundation prohibits that there will ever be any self-membered sets. It does so, intuitively, by establishing that for *any* set, we'll be able to "decompose" it into elements, and then decompose its elements into elements, successively, and that at some point we'll reach something that's not further decomposable. In other words, we'll never run into the kind of vicious circle that Russell was worried about.

This by itself, however, doesn't resolve the status of the Russell set, or the Universal set of all sets. For we might, as far as the axiom of foundation goes, still have a set of all sets – it would just have to be decomposable ultimately into simple elements. And we'd still face the tricky question of whether this set, being a set, is an element of itself (or if it isn't, why it isn't.) Also, the Axiom of Foundation doesn't tell us positively which sets do exist – it just implies about certain sets that they *don't* exist. To solve this problem, Zermelo and Fraenkel introduced another axiom, the Axiom of Separation.[1] This axiom is also sometimes called the Axiom of Limited Comprehension, and it amounts to a kind of limitation of Frege's original Universal Comprehension Principle, or Basic Law V (which has, of course, been rejected owing to Russell's result).

Axiom (schema) of Separation*: Given any set B* and any well-defined property ϕ, the set A= {x $\in$ B| x is ϕ} exists.

In other words, the Axiom of Separation says that given any 'already existing' set, B, we can *separate out* just those elements of B that have a certain definable property, ϕ, thus creating the set A as a subset. This is a restricted form of Frege's original comprehension principle. For it doesn't say that we can just create a set corresponding to any property (as the original principle did): just that if we already "have" a

[1] This is actually not a single axiom, properly speaking, but an *axiom schema*. That's because it's defined not just for one property ϕ, but for any such property; there are thus an infinite number of "instances" corresponding to the infinite number of properties we can put in for ϕ. We'll worry more about this wrinkle later.

set to draw on, we can *separate out* the elements from that set that have the property. The important consequence is that this appears to establish that we can't 'create' a set of all sets. For in order to do so, according to the axiom, we'd have to already have a set to separate this out of; but the only set that we could use would apparently be the set of all sets itself. So we can't 'build' this set unless we already have it, which seems to give good reason to think that we can in fact never 'build' or form it in any way.

These two axioms are, as noted, enshrined in the most standard "system" of set theory, the ZF system (though there are indeed other systems of set theory that suspend one or both of them) and effectively formulate a kind of intuition or "picture" of what sets exist as a whole. A more general version of what is common to this "picture" and also to Russell's theory of types is sometimes called the "iterative conception" of sets. This is as close as anything comes to the "standard" picture of the set-theoretical "universe" and it takes the form of a big open "V". (See the picture from Moore's *The Infinite*, p. 157). The idea is that at the very bottom are the simplest sets, the 'singleton' sets each containing just one element that is not itself a set (if we allow sets to have non-set members) or perhaps, even more simply, the single, unique empty set. Then a bit higher up (or at the next 'level' in Russell's hierarchy), we get sets containing these base-level sets. There are more of these than at the first level; and then at the next level, we get sets containing those at the second level (of which there are still more), and so on. We know that we can form each higher level by taking the power set of the sets at the lower level, so it seems that there is no limit to how far we can go. However, because of the consequences of Russell's paradox, there's no stage at which we finish; the top of the V is open-ended, and has no limit. Also, we never, at any point, find a self-membered set; for such a set, as we've seen, can never be "formed" out of simpler ones. It's important to notice, though, that the fact that each set has a determinate level or "size" doesn't mean that some (indeed many!) of the sets aren't infinite. In fact, with the help of an additional axiom (the axiom of infinity, which we'll meet later), we can insure that not only is there an infinite set, but there are infinitely many infinite sets, of increasing levels of "size" or complexity.

The iterative conception solves the problem presented by Russell's paradox pretty nicely, *if* we accept the intuitions underlying it and formulated in the standard axioms. If we do accept these and work within them, there will be no self-membered sets and there will also be no universal sets (such as the set of all sets), so there's no way that the whole set of problems that Russell raised can arise. But this is true only if we accept these intuitions. As I've tried to mark with the use of scare quotes in the discussion above, the intuitions underlying the iterative conception are, in general, highly "constructivist." That is, they more or less all depend on the idea that sets are "made" or "constructed" out of other sets, so that we can only "form" a set at a certain level if there are already "available" other sets "out of which" we can form it. This is what seemingly prevents us from (ever) making a set of all sets: for we'll never have the "raw materials" available out of which to form it if we cannot presuppose it already.

These intuitions are somewhat plausible with respect to "pure sets" themselves, at least if we think of such sets as things that inherently can *only* be formed by means of grouping together or separating out from other such things. But how plausible are they when applied to the more general issues of the

nature of concepts and linguistic reference? The answer is: *not very plausible*. Consider all of the following normal-sounding statements:

Everything that is thinkable can be expressed in language.

Language (as a whole) is a system of signs.

The universe contains many wonderful things.

There is more in heaven and earth than is dreamed of in your philosophy.

These claims might themselves turn out to be true or false, but it does seem that they are meaningful. If the constructivist intuitions enshrined in standard set theory and in the iterative conception of the set are correct, though, we can't even (so much as *express*) them meaningfully; for they all refer to totalities which, according to those intuitions, can't exist even as objects of thought. This is because they all refer to infinite totalities (the thinkable, language, the universe, all meaningful propositions) that indeed probably can't be *constructed* by grouping together separate elements. But does this mean that the totalities indeed don't exist? It doesn't seem so. It is, at any rate, far from evident that we must have the ability to "put together" these (seeming) sets from separate elements before we can have any concrete idea of them, or any ability to refer to them meaningfully.

Thus, the implications of the standard conception of sets for ordinary language and reference are indeed counterintuitive and surprising. But the standard conception was itself, as we've seen, largely formulated in response to Russell's paradox, and was meant to be a satisfying way of dealing with it, or at least preventing it from ever coming up in the formal theory itself. If it is indeed unsatisfying in dealing with ordinary cases of concepts and language, however, then this suggests that we might have to revisit Russell's paradox, and think about other ways of responding to it or accommodating it without imposing the constructivist intuitions of the standard, iterative conception of the set. We might, for instance, just have to bite the bullet and live with paradox and contradiction in ordinary life (if not in mathematics), allowing that paradoxes like Russell's paradox (and the Liar) will always arise due to problems of ordinary language, and there is little we can do to formalize them away or rule them out. Or another possibility – one that is quite interesting and has only recently begun to be explored – is that we might try to *formalize the paradox itself*, and see what consequences (including, perhaps, contradictory ones!) might flow from it. That is, we might try out the hypothesis that *in mathematics as well as in ordinary language*, there are certain specific cases where this kind of paradox will arise – cases, for instance, in which there genuinely *are* self-membered sets, or self-referring concepts, or cases in which we genuinely (and meaningfully) *do* talk about totalities in which the very act of speaking is included. We might indeed have to acknowledge certain necessary contradictions here, but this doesn't mean that there will be contradictions everywhere, or that the acknowledgment will ruin any possibility of formalizing thought and language at all. In fact, we might use just this acknowledgment as a way of studying something that is in fact very important to philosophy historically, and continues to define many of its problems today: the problem of thinking about the real existence of contradictions, and especially those contradictions that arise at the limits of thought and language, where we (as

philosophers) try to grasp and express what delimits the totality of all that can be thinkable or sayable, and thus seem consigned to take a position "beyond" these very limits themselves.

Cantor's Theorem

As we've already seen, for finite sets at any rate, the power set of a given set is always larger – contains more elements – than the set itself. (In fact, we showed in class that if a finite set is of size n, then its power set is of size 2^n). But what about infinite sets? As we saw already on the first day of class, one of Cantor's most profound results was to show that there are *many* infinite sets with different sizes. We proved a particular case of this with the "diagonalization" argument that shows that there are more real numbers than natural numbers. But this is just a specific form of a more general result that Cantor had reached by 1891. The result is that for *any* set x (finite or infinite), the power set, $\mathcal{P}(x)$, has strictly more members than x itself. This means that for any set, finite or infinite, we can generate an infinite series of larger ones; and there is no evident stopping point.

Cantor's Theorem: For *any* set x, the power set, $\mathcal{P}(x)$, has strictly more members than x itself.

Proof: First, we can easily establish that $\mathcal{P}(x)$ has at least as many elements as x itself. To see this, note that we can establish a one-to-one correlation between all the elements of x and some (not necessarily all!) elements of $\mathcal{P}(x)$. Just correlate each $y \in x$ to the singleton set {y}. Since $\mathcal{P}(x)$ is the power set of x, it contains all the subsets of x, and thus it contains each of these singleton sets.

Now we need to establish that there is no one-to-one correlation between all the elements of x and all the elements of $\mathcal{P}(x)$. Once again, we'll do a proof by *reductio*; that is, we'll assume (contrary to fact) that there is some such correlation, and then we'll derive a contradiction. So suppose there is some function, f, that is a one-to-one correlation from elements of x to elements of $\mathcal{P}(x)$. Now, on this assumption, f is going to map each element, y, in x to some set, f(y), in $\mathcal{P}(x)$. Sometimes, y will map to a set in which y itself is an element; but this may also not be the case. In particular, let's now consider z, which is the set of all the y's that are mapped to sets that they are *not* elements of:

$Z = \{ y \in x \mid y \notin f(y)\}$.

Now, Z itself is a *subset* of x, so it's also an *element* of $\mathcal{P}(x)$. Therefore (since we're using f to map onto each element of $\mathcal{P}(x)$), there is some element of x, call it w, such that Z=f(w). Now we can ask: is w itself an element of Z? Well, if it is an element of Z, then, because of the way Z is defined, it's *not* an element of f(w). (After all, Z is defined as the set of all the elements that *aren't* in the set that f correlates them to). But Z=f(w)! So if it is an element of Z, it is not (CONTRADICTION). What, then, if w is *not* an element of Z? Well, then $w \notin f(w)$, so by the definition of Z above, it is in Z. Again, we have a CONTRADICTION. Therefore, if w is an element of Z, it is not, and if it is not an element of Z, it is. Therefore there is no one-to-one correlation f.

Cantor's Theorem is interesting because its proof closely resembles two things we've seen before. First, the way the proof itself works closely resembles Russell's paradox (note that, just as in Russell's

paradox, we need to consider a set (here, the set Z) of all elements that *don't* have a crucial property, and then we derive the contradiction that if something has the property it doesn't, and vice versa). In fact, Russell later said that it was in thinking about Cantor's theorem that he first came up with the paradox itself. But there are also differences: Cantor's theorem doesn't deal with the totality of *all* sets, but only with the relations between particular sets and their power sets; and because of this difference, Cantor's theorem isn't a paradox, but rather just a result.

Second, though, the proof of Cantor's theorem is just a generalization of the "diagonal argument" about the naturals and the reals that we discussed the first day of class. To see the connection, notice that we can represent each real number between 0 and 1 (let's again stick with these for ease) as an (infinite) *set* of natural numbers. How? Well, first, let's re-write them in binary notation. Thus, we'll have, e.g., *a* = .1 0 1 0 0 … (this corresponds roughly to .625 in digital, but since we don't have the whole expansion, we don't know exactly what it will figure out to ultimately). Now all we have to do is number the places of this expansion with natural numbers: 1 for the first place, 2 for the second, etc., and identify the real number with the set of places for which it has a '1'. So, for instance, our *a* will be identified with the set containing 1, 3, and whatever other numbers correspond to the places in a where there is a 1 rather than a 0. So we can think of each real number as just a set of naturals; and then the set of *all* reals (between 0 and 1) will just correspond to the set of subsets of the set **N** of all naturals, or in other words the *power set* of **N**. If we want to show that this power set is bigger than the original set, we just diagonalize on the list, as we did on the first day; if we're representing the reals in binary notation, as we did above, we just alter the diagonalization rule so that we always change a '1' to a '0' and vice versa (rather than changing n to n+1, or 9 to 0, as we did the first day). This will give us exactly the same result, and it's also exactly the same as what we did in the general proof of Cantor's theorem. There, as in the numerical diagonalization proof, we assumed that there is a one-to-one correlation in order to produce a contradiction; and then, through our diagonal set (the set Z) we produced just such a contradiction.

Ordinality and Cardinality: First look at the transfinite hierarchy

As we've already seen, for finite sets at any rate, the power set of a given set is always larger – contains more elements – than the set itself. (In fact, we showed in class that if a finite set is of size n, then its power set is of size 2^n). But what about infinite sets? As we saw already on the first day of class, one of Cantor's most profound results was to show that there are *many* infinite sets with different sizes. We proved a particular case of this with the "diagonalization" argument that shows that there are more real numbers than natural numbers. But this is just a specific form of a more general result that Cantor had reached by 1891. The result is that for *any* set x (finite or infinite), the power set, $\mathcal{P}(x)$, has strictly more members than x itself. This means that for any set, finite or infinite, we can generate an infinite series of larger ones; and there is no evident stopping point.

Since, as we now know, the sizes of infinite sets are different, we need some symbolism for talking about the sizes of sets (intuitively, the 'number' of elements they contain). First, though, we need a *conceptual* distinction between two "concepts" or "aspects" of numbers, ordinality and cardinality. It's

sometimes difficult at first to keep these two notions apart, since they always go together in the case of finite numbers (however, as we'll see, they can and often do come apart for infinite numbers!) "Ordinality," roughly, means "position in an order." "Cardinality," roughly, means "quantity" in the sense of "how many?". To see the difference, consider the following. If I give you a well-ordered sequence, for instance the series of letters in the alphabet A,B,C…, it's not natural or intuitive to use them to answer a "how many?" question. For instance, if I knew that all the rooms along a hallway were "lettered" in order, A, B, C, etc., it wouldn't be immediately obvious (without doing some thinking) how far down I have to go in order to get to room Q. To figure this out, I'd have to figure out the *cardinality* of this ordinal sequence, i.e. how many letters there are between A and Q, inclusive (in fact there are 17). Cardinality, is in fact, a much more general notion. For all kinds of sets have cardinality, even if they don't have any natural ordering at all. Thus, e.g., the set {George, 9, the Eiffel tower} has cardinality 3. In general, two sets have the same cardinality if and only if there is a one-to-one correspondence between them.

Let's use vertical strokes to denote, in general, the size of sets in the sense of 'cardinality'. Thus: |{0, 14, 23, 38}| = 4, for instance. Based on what we've established so far, we can establish some preliminary results about the size of sets. First, obviously, we know that the size or cardinality of any infinite set, for instance **N** (the set of all naturals) is greater than any finite number. That is, for any finite n, |**N**| > n. What's more surprising, though, is what we already established on the first day, that *the set of real numbers has greater cardinality, or size, than the set of natural numbers*. Thus |**R**| > |**N**|. And now we know, from Cantor's theorem, that this is just a specification of the more general result: for any set X, $|\mathcal{P}(X)| > |X|$.

There are also, though, some rather surprising results about *equality* of size as well. To begin with, we've proven various results about $\mathbf{R}_{(0,1)}$, the set of reals between 0 and 1. But we can establish a one-to-one correlation between this set and the set of *all* reals, **R.** How? Just consider a line segment from 0 to 1 that's bent into a semicircular arc (see diagram in Moore, p. 118). Then we can correlate *each* point on this bent line segment with a point on the line containing all positive and negative reals. Thus $|\mathbf{R}_{(0,1)}| = |\mathbf{R}|$ (and so all of our diagonalization results about $\mathbf{R}_{(0,1)}$hold for **R** as well). Also somewhat surprisingly, the cardinality of the set of all rationals, **Q**, is the same as the cardinality of the set of (whole) natural numbers, **N**. How can we show this? We can create a one-to-one correlation as follows. Arrange the rationals in a table, with numerators along the horizontal axis and denominators along the vertical axis:

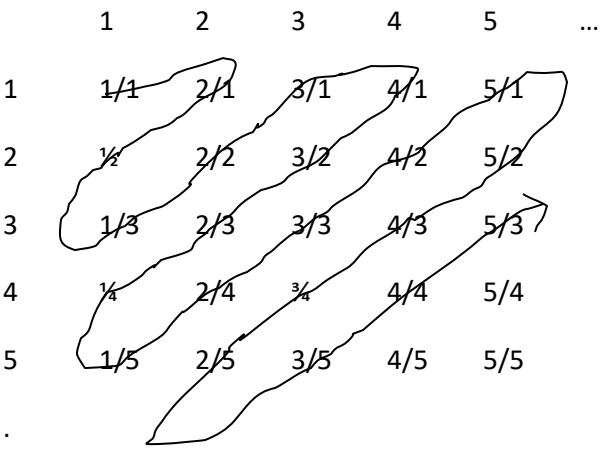
	1	2	3	4	5	…
1	1/1	2/1	3/1	4/1	5/1	
2	½	2/2	3/2	4/2	5/2	
3	1/3	2/3	3/3	4/3	5/3	
4	¼	2/4	¾	4/4	5/4	
5	1/5	2/5	3/5	4/5	5/5	

.

.

Now, we just pass through the table using a “criss-cross” method (see diagram on the next page). Assign 1 to the top left entry. Then go one over – to 2/1 (assign this to 2). Then go to the first entry in the second row, ½ (and assign this to 3). Now go to the first entry in the third row, assigning this to 4, etc. In this way, we can pass through the whole table one by one, assigning each and every rational number to a unique natural number.

We can use a similar "criss-cross" method to show that the cardinality of the set of *ordered pairs* of natural numbers – **N x N** – is also the same as that of **N** itself.



We can also show that $\mathbf{R}_{(0,1)}$ x $\mathbf{R}_{(0,1)}$ – or the set of points in a 1X1 *square* – has the same cardinality as just $\mathbf{R}_{(0,1)}$ itself (and hence, by our earlier result, as **R** itself). Each of the points in the square can be identified by its x and y coordinates, each of which will be a real number. Suppose a certain point <x,y> has the coordinates x= $0.a_1a_2a_3a_4a_5\ldots$ and y= $0.b_1b_2b_3b_4b_5\ldots$ Then we just correlate this ordered pair with a new number, c, by interweaving the decimal places, using the places of x as the odd-numbered and the places of y as the even-numbered places in the new number. Thus, c=0. $a_1\ b_1a_2\ b_2a_3\ b_3a_4\ b_4a_5\ b_5\ldots$ This gives the needed one-to-one correspondence.

So, to summarize, we have the following results about relative sizes of sets:

$|\mathbf{N}| > n$ where n is any finite number.

$|\mathbf{R}| > |\mathbf{N}|$

$|\mathbf{Q}| = |\mathbf{N}|$

$|\mathbf{N} \times \mathbf{N}| = |\mathbf{N}|$

$|\mathbf{R}_{(0,1)} \times \mathbf{R}_{(0,1)}| = |\mathbf{R}_{(0,1)}| = |\mathbf{R}|$

Transfinite cardinality holds many surprises!

Building up the numbers

Last week, we saw how we can "build" up the finite numbers from pure sets, using the standard construction from von Neumann, whereby we take the empty set to be 0, and go from there by identifying each succeeding number with the set containing *all* of its predecessors. Thus:

0 = ∅

1 = {0} = {∅}

2 = {0,1} = {∅,{∅}}

3= {0, 1, 2} = {∅, {∅}, {∅,{∅}}}

Etc.

What about when we get to infinity, though? We've talked a lot already about **N**, the set of all finite natural numbers. Well, identifying numbers with sets this way, what happens when we continue the process of "building" numbers up to here, and identify this set itself with a number? We get a new number, which is often called **ω**. Like any of the finite natural numbers, **ω** just consists in the set of all of its predecessors -- in this case, the set of all finite natural numbers. In fact, **ω** – the first 'infinite' number – is just our old friend **N** in a new guise, this time considered as a number somewhere in the ordered series of numbers, rather than *just* as a set.

Now we have the basic materials we need to build up *all* types of numbers out of sets. As we saw last week, it's easy to define the set of rationals, **Q**, and the set of integers, **I** (negative as well as positive whole numbers) using *ordered pairs* of natural numbers.[2] For instance, we build the fractional number a/b with the ordered pair <a,b>. We can build the integer -3 using the ordered pair (interpreted as a difference) <2, 5>.[3] We can then show, as well, that both of these sets are well-ordered.

Finally, what about the real numbers, **R**? This is an interesting issue, since as we've seen the fact that there are many more (infinitely more) real numbers than rational ones was a huge shock to the ancient Greeks, and also gives us, as we've seen, the first infinite set that has a bigger cardinality than that of **N**. There are a few different ways to define the real numbers. For instance, we could do it in terms of converging series of sums of rationals (that's essentially what we presupposed in using the "binary" representation above). But by far the simplest and (in many ways) most elegant way was given by Richard Dedekind and is called the **Dedekind cut**. Intuitively, the idea is that we can consider each real number to be a "cut" of the whole set of rational numbers, so that *every* rational number is either below or above the "cut". In fact, we can just identify the cut – i.e. the real number – with the set of all

[2] The *ordered pair* <m,n> can be identified as the *set* { {m}, {m,n}}.

[3] There is a little wrinkle for both of these definitions, since ordered pairs don't necessarily determine *unique* rational or integral numbers (since, for instance, 3/2=6/4 or 2-5 = 1-4). To deal with this, we actually identify the rationals and the integers with *equivalence classes* of ordered pairs of naturals. (For details, see Maddy or Enderton).

rational numbers less than it. This sounds circular, but it needn't be, since we can just define the Dedekind cuts as sets of rationals:

A Dedekind cut is a subset x of **Q** (the set of rationals) such that:

a) x is not ∅ or **Q** itself

b) x is "closed downward." That is,

for all rationals q, r, if q ∈ x and r < q, then r ∈ x

c) x has no largest member

Then we can just identify each real number with a Dedekind cut, and the set of all real numbers, **R**, with the set of all Dedekind cuts. We also have a nice ordering relation: where a and b are reals, a < b iff a is a proper subset of b.

Axioms of Set Theory

So far, we've motivated and learned set theory just intuitively, talking about which kinds of sets can and can't exist. But the essence of set theory, like any other mathematical theory, is in its precise definitions and rules. These are captured in a system of axioms which are the "starting points" for all formal set theory (very much as Euclid's axioms are the "starting points" for geometry). Similarly, also, to geometry, there are several different systems of axioms which yield different results but are nevertheless consistent. Today, however, we'll learn the most standard axiomatization, the so-called ZFC axiomatization based on the work of Zermelo and Fraenkel, plus the Axiom of Choice (the 'C' of ZFC). This axiom set is more or less canonical, but this does not mean that there are not deep and profound philosophical issues lurking beneath virtually each of them. Also, the choice of which axioms to adopt and whether the ZFC system, as opposed to any of the other available systems, is the "right" one, is itself a philosophically significant question; we'll talk later about the interesting and important philosophical issues that are sensitive to this choice.

To begin with, there are several axioms that we've already – implicitly at least – 'met' in our informal explorations of set theory so far:

1. Axiom of Extensionality: If two sets have exactly the same elements, they are exactly the same set.

This is the axiom that lets us say that a set is defined by its elements.

2. Axiom of Pairing: If A and B exist, then {A,B} exists.

This is a general (and quite simple) principle of set formation. We've appealed to it, implicitly at least, in our discussions ever since we started talking about what sets are.

3. Axiom of Union: Given a set containing various sets, the set containing all and only the elements of these contained sets exists.

In other words, supposing we have a set such as {{a,b}, {c,d}, {e}}, the Axiom of Union lets us put what's inside all of these sets together to construct the union set: {a,b,c,d,e}. This is the axiom underlying our informal use of the operation of union since we started talking.

4. Axiom of Power Set: Given a set A, the set $\mathcal{P}$(A), defined as the set containing all and only subsets of A, exists.

This is the axiom underlying our discussion of power sets.

Now we have a couple of axioms that, while perhaps more philosophically controversial in certain ways, have also been presupposed in our discussions so far.

5. Axiom of Empty Set: There is an empty set, i.e. a set that has no elements.[4]

6. Axiom of Infinity: There is an infinite set. More technically, if we let S(X) stand for X ∪ {X}, then there is a set that has ∅ as an element, and that is such that for each x in it, S(x) is also in it.

In terms of the way we defined natural numbers, the axiom of infinity thus says that there's a set that contains 0 and all of its successors. This guarantees the existence, at least, of **N** (or **ω**).

Now we have the two axioms that we discussed last week; both of these were introduced to help ensure that Russell's paradox, and related paradoxes, can't arise.

7. Axiom of Foundation (or regularity): Every non-empty set, x, contains an element, y, such that x and y have no common elements.

Recall that (and how) the Axiom of Foundation functions to prohibit self-membered sets. It also guarantees that each set can be "decomposed" into its subsets in such a way that we'll never get caught in a circle; thus it imposes a kind of hierarchy on the universe of sets itself.[5]

---

[4] Some lists of the axioms omit the Axiom of Empty Set, since (on some views of logic and identity at least) it can be 'derived' from the Axiom of Separation, for instance by taking from any existing set the subset of elements that have both property P and ~P.

[5] Why does this prohibit *circles* of foundation (such as the situation where A ∈ B, B ∈ C, and C ∈ A)? In such a situation (by the axiom of replacement), it is possible to group all the members of the circle into a set, say S = {A, B, C}. But then S is not disjoint from A, since A contains C; it is not disjoint from B, since B contains A; and it is not disjoint from C, for C contains B. Therefore it has no element with which it is disjoint, and the condition expressed by the axiom fails. By a slightly more complicated argument, the Axiom of Foundation also effectively prohibits infinite descending chains of foundation.

8. Axiom (schema) of Separation: Given any set A and any well-defined property ϕ, the set B of all elements of A having the property ϕ exists.

Recall that this is the replacement for Frege's original law of universal comprehension. It doesn't say, as Frege's law did, that we can create a set corresponding to just *any* property. Rather, it says that given any *already existing* set, we can 'separate out' a set containing just those elements that have the property. Also remember that this isn't a single axiom, properly speaking. Since there are an infinite number of properties ϕ, this is just a general schema that refers to an infinite number of individual axioms (one for each property).

There's another useful general axiom schema which is closely related to the Axiom (schema) of Specification, the Axiom (schema) of Replacement. In fact, since the Axiom (schema) of Specification can be derived from the Axiom (schema) of Replacement together with the Axiom of Empty Set, the former is sometimes omitted from presentations of the axioms that include the Axiom (schema) of Replacement.

9. Axiom (schema) of Replacement. If F is a function defined on the set A, its image under the function, F(A), also exists.

The Axiom (schema) of Replacement just says that if we have a set and a well-defined function (to give a full specification of the Axiom schema, we'd need to spell out in more detail what being well-defined amounts to), then the result of evaluating the function for each member of the set and putting all the results together is also a set. More intuitively, if we have a set consisting of a number of elements and a well-defined function from those elements to others, it allows us to 'replace' (hence the name) each element of the first set with a unique element of the other and thereby create a new set. For instance, if we have the set {1, 2, 3} and a function that takes 1 to a, 2 to b, and 3 to c, it allows us to say that the set {a, b, c} exists.

Like the Axiom Schema of Separation, this is also a schema, rather than a single axiom, since there are an infinite number of functions, and strictly speaking we need a different axiom for each one.

Finally, we have (on some axiomatizations, at least) the famously controversial Axiom of Choice:

10. Axiom of Choice. Suppose X is a set, all of whose element are non-empty. Then there is a "choice function," f, that selects one element from each of the elements of X, and a 'choice set' that groups these all together.

The idea of the Axiom of Choice is straightforward: suppose I have a big bucket containing smaller buckets, and in each of the smaller buckets is a number of items. The choice function just takes one item from each of the smaller buckets, and groups all the selected items together. For finite sets, this is no problem, and in fact we can derive the axiom of choice for finite sets from the other axioms. However, where it's thought to be more problematic is where we have an infinite set of sets, and we again need to choose one from each of an infinite number of buckets (which *may* themselves contain

infinitely many items). The reason this is thought to be problematic is that the Axiom itself gives us no "selection criterion" – it doesn't tell us *how* to choose an item out of each one, just that we *can* do so. Bertrand Russell drew a famous analogy for why this is problematic: If we have an infinite number of pairs of shoes, and have to select one shoe from each pair, we can easily come up with a rule for doing so: just pick the left shoe in each case. But if we have to perform the same task with an infinite number of pairs of socks, we are at a loss to specify any such rule, since socks don't come sorted into left and right. The Axiom of Choice, if it is true, essentially puts us in the latter situation, the situation of selecting from socks: we have to know that we can make an infinite number of selections, without having any rule for doing so. Thus, the Axiom has been called "non-constructive", in that it (plausibly at least) doesn't actually tell us *what is* the set whose existence it asserts.

The Continuum Hypothesis

Let's summarize some of our cardinality results so far:

$|\mathbf{N}| > n$ where n is any finite number.

$|\mathbf{R}| > |\mathbf{N}|$

$|\mathbf{Q}| = |\mathbf{N}|$

$|\mathbf{N} \times \mathbf{N}| = |\mathbf{N}|$

$|\mathbf{R}_{(0,1)} \times \mathbf{R}_{(0,1)}| = |\mathbf{R}_{(0,1)}| = |\mathbf{R}|$

We know from Cantor's theorem that there are various sizes of infinity, and we've already seen that there are in fact at least two such sizes (since we know by diagonalization that $|\mathbf{R}| > |\mathbf{N}|$).

We also distinguished between "ordinality" (which one? in an order) and "cardinality" (how many?) For the finite natural numbers, ordinality and cardinality coincide; this is what lets us, after all, use the normal ordinal counting sequence ("1", "2", "3", "4", etc.) to answer questions of cardinality too. However, when we get to infinity, things (as usual!) get more tricky. Using the definition of numbers in terms of pure sets that we adopted earlier, where

$0 = \emptyset$

$1 = \{0\} = \{\emptyset\}$

$2 = \{0,1\} = \{\emptyset,\{\emptyset\}\}$

$3 = \{0, 1, 2\} = \{\emptyset, \{\emptyset\}, \{\emptyset,\{\emptyset\}\}\}$

Etc.,

we identify each number with the whole sequence of its predecessors; thus for instance, as we saw, 3 = {0,1,2}; 4={0,1,2,3}, etc. These sets all have the nice property of *transitivity*: that is, for any one of them,

(say X), whenever a is an element of b, and b is an element of X, a is also an element of X. (Thus, for instance, 2 is an element of 3; and 3 is also an element of 4; so 2 is an element of 4, as it should be). In fact, we can use this property to define *ordinal numbers*: an *ordinal number* is just a transitive set consisting only of transitive sets. This is in fact just a fancy way to say that each ordinal number is the set grouping together all of its predecessors.

We also know from the Axiom of Infinity that there is an (at least one) *infinite* ordinal. This is the set consisting of all the finite ordinals, and we have already come up with a symbol for this ordinal: **ω**. We can also keep using our rule for the generation of ordinals here (we just include the ordinal we have along with all of its predecessors) to generate the very next ordinal, the successor of **ω,** or **ω+1.** As we've seen, for all the finite numbers at least, we can just identify ordinals with cardinals (the question of which one? in the series can be just identified with the question of how many? up to that one). But here things start to get tricky, and we have to begin to distinguish ordinals from cardinals. The trouble is that the cardinality of **ω+1** is in fact equal to that of **ω** – not "one greater", as we might like. Why? Well, recall the issues about infinity that we discussed a few weeks ago; and in particular, remember Hilbert's hotel, which (although full) could still accommodate one more weary traveler. The point of that example was effectively that we can put the set of **ω+1** elements in one-to-one correspondence with the set of only **ω** elements. (More formally, just map each finite *n* in **ω+1** to *n*+1 in **ω**, and then we can just map the "extra" element of **ω+1** (this element is actually **ω**) to 0). Using similar considerations, in fact, we've established that:

| **ω+n**| = | **ω**| for any finite **n**

and also even that

| **ω + ω| = |2ω**| = | **ω**|

In fact, using similar considerations (criss-cross through an *n*-dimensional array), we can even show that for any finite **n**

|n x ω | = | ω|

And that for any finite **n**:

| ω^n **| = | ω|**[6]

In fact, we can do even more while remaining within the domain of sets that have the same cardinality as **ω**. For example, | ω^{ω} **| = | ω| !** This may seem strange, but the real reason is that is that ω^{ω} is just what we get by multiplying **ω** by itself **ω** times. Since there are **ω** iterations, this result can be put into

[6] **IMPORTANT NOTE:** Here we are using so-called "ordinal exponentiation". Ordinal exponentiation is defined in terms of multiplication of ordinals, which is well-defined even for infinite ordinals such as **ω.** BUT it is a very different notion from "cardinal exponentiation," which is defined in terms of power sets. We will switch to "cardinal exponentiation" below when we discuss the cardinality of $\mathcal{P}$ (ω).

one-to-one correspondence with **ω** itself – thus it has the same cardinality. In general, in fact, anything that we can get by doing some arithmetical operation **ω** times can be put into one-to-one correspondence with **ω**. In fact, this way we can even get the fantastically large

$$\varepsilon_0 = \omega^{\omega^{\omega^{\cdot^{\cdot^{\cdot}}}}} = \sup\{\omega, \omega^\omega, \omega^{\omega^\omega}, \omega^{\omega^{\omega^\omega}}, \ldots\}$$

This number also has the "funny" feature that $\varepsilon_0 = \boldsymbol{\omega}^{\varepsilon 0}$. Because of this, it is also the "largest" number that has a straightforward expression in terms of purely arithmetical operations. But it still has the same cardinality as does **ω** itself.

Given all of this, we might suspect that there is only one infinite cardinality, the cardinality of **ω**. This would be true, if all infinite sets could be put in one-one correspondence with each other, and they all could be put in one-one correspondence with **ω.** But in fact, we already know that this *isn't* true. For recall that (by Cantor's theorem, or diagonalization),

$|\mathcal{P}(x)| > |x|$ for any x

And in particular,

$|\mathbf{R}| = |\mathcal{P}(\boldsymbol{\omega})| > |\boldsymbol{\omega}|$

We also saw that for finite sets at least, $|\mathcal{P}(\mathbf{x})| = 2^{|x|}$; and since $|x| = x$ for finite *ordinals*, that for these ordinals at least, $|\mathcal{P}(\mathbf{x})| = 2^x$. If we like, we can extend this idea to power sets of infinite sets, and then we have, by definition:

$|\mathcal{P}(\boldsymbol{\omega})| = 2^{\boldsymbol{\omega}} = |\mathbf{R}|$.[7]

We now know, at any rate, that there are *at least two* infinite cardinalities, and that one is larger than the other. Let's fix some terminology. To talk about cardinals rather than ordinals (since many ordinals, as we've seen, have the same cardinality), we'll use (following Cantor) the series of *alephs*, $\aleph_0, \aleph_1, \aleph_2$, etc.

By definition, let's let $\aleph_0$ (pronounced "aleph-null") be the cardinality of the first infinite set, **ω.** Thus,

$|\boldsymbol{\omega}| = \aleph_0$

All the sets that can be put in one-to-one correspondence with **ω** (for instance: **ω** +1, 2 **x ω**, etc.) will also have cardinality $\aleph_0$. Because these sets can be put in one-to-one correspondence with the set of all natural (counting) numbers, these sets are sometimes called *countable* sets (or "countably infinite" sets). Then $\aleph_1$, the *very next* cardinality, will be the cardinality of the *first* set that is "bigger" than all of these in the sense that it *cannot* be put in one-to-one correspondence with **ω** or any of these. And $\aleph_2$

[7] **IMPORTANT NOTE:** Now we are using exponentiation in the sense of "cardinal exponentiation" – NOT "ordinal exponentiation." Cardinal exponentiation is *defined* in terms of the power set, rather than in terms of multiplication (see note 1 above). From now on, exponentiation involving infinite sets will always be "cardinal exponentiation" unless otherwise noted.

will be the (cardinality of the) first infinite set that can't be put in one-to-one correspondence with any set having cardinality $\aleph_1$ or $\aleph_o$, and so forth.[8]

We are now in a position to pose one of the most famous unsolved (and, as we now know, unsolvable) problems in all of mathematics. We know that $|\mathbf{R}|$ -- the cardinality of the set of real numbers (or, as it this set is often called because of its conceptual connection with the set of points on a continuous line, the *continuum*) is *greater than* the cardinality of the set of naturals, which equals (by definition) $\aleph_0$. But what is $|\mathbf{R}|$, actually? In particular, is it true $|\mathbf{R}|$ is equal to $\aleph_1$, *the very next* transfinite cardinal after $\aleph_0$? The hypothesis that this *is* true – that $|\mathbf{R}|=\aleph_1$, is known as Cantor's Continuum Hypothesis. Cantor thought it was true, and himself labored in vain throughout the last years of his life to prove it from the ZFC axioms. But he didn't succeed, and no one else did either.

Since we know that $|\mathbf{R}|$, whatever it is, is equal to $|\mathcal{P}(\boldsymbol{\omega})|$, and that (by definition), this = $2^{|\omega|}$ or $2^{\aleph 0}$, we can put the hypothesis very simply:

Cantor's Continuum Hypothesis (CH)

$2^{\aleph 0} = \aleph_1$

Clearly, the truth or falsity of the Continuum Hypothesis bears fundamentally on the question of what sets there "actually" are, or at any rate which sets there are according to the ZFC axioms. For instance, if the CH is true, then there is no set of real numbers that is larger than (of greater cardinality than) the set of naturals but smaller than set of *all* real numbers; if it is false, then there is such a set. In fact, we can generalize the CH in such a way that it bears on the shape of the *whole* hierarchy of finite and transfinite sets, the whole "set-theoretical universe" V:

Cantor's Continuum Hypothesis – Generalized Form (GCH)

For every ordinal *a*,

$2^{\aleph a} = \aleph_{a+1}$

Seen another way, the issue of the truth or falsity of the GCH is really just the question of the "power" of the "power set" operation. By just how much does the power set of a given set exceed that set itself? A set of size $\aleph_1$ has, remember, the very next size that it is possible for any set to be, beyond $\aleph_0$. The question is whether this size is just as big as the power set itself; whether, in other words, the hierarchy

[8] More rigorously, just as we defined the ordinals in terms of pure sets, we'll want a definition of the transfinite cardinals in terms of pure sets as well. What, then, actually are the cardinals? Well, since we know that, for each cardinality, there is *at least one* ordinal that has that cardinality (although of course there are usually *many* ordinals with the same cardinality), and also that the for any set of ordinals, the set has a least member, we can just identify each cardinality with the least ordinal that has that cardinality. Thus, the cardinals really are (some) ordinals (for instance, $\aleph_o$ "actually" equals not only $|\boldsymbol{\omega}|$ but also $\boldsymbol{\omega}$ itself, since $\boldsymbol{\omega}$ is the smallest ordinal that has that cardinality) and we just use the two notations to talk about the same things under different aspects or guises.

of alephs (or possible sizes of sets) "measures" the excess of the power sets themselves. If so – that is, if the GCH holds – then the total universe of sets is relatively ordered and well-behaved. We can (as it were) use what we "know" from each level of set creation to "predict" what will happen when we form the power set. By contrast, if this is not so and the GCH is false, then the power set operation, we might say, introduces a kind of "immeasurable" excess in the hierarchy or series of sets; there's literally no telling how much "bigger" (in terms of the series of alephs) a power set will be than its original set.

What, then, is the status of the CH? We now know (perhaps sadly or perhaps suggestively) that it can't be either *proven* or *disproven* from the ZFC axioms. In other words, it is literally *independent* of the axioms: the axioms, and the system they define, just don't say whether it's true or false. This result is the combination of two very important results in set theory (and its meta-theory). First, in 1938, Kurt Gödel showed that the GCH can't be *disproven* in ZFC (this was Gödel's second most famous result, after the incompleteness theorem, which we'll study later). That is, it's impossible to prove the *negation* of the GCH in ZFC; we might put this as saying that it's *possible* (according to ZFC) that the GCH is true. However, in 1963, P. J. Cohen, by means of a profound technique called "forcing," showed that the GCH can't be *proven* in ZFC either. Thus (as we might say), it's equally possible (according to ZFC) that the GCH is false. This has sometimes been taken to suggest that the axioms actually under-describe the real universe of sets, that they don't really settle the question of what sets actually exist. We could take this to mean that the set theoretical universe is just inherently vague and ill-defined. Or we might think (as Gödel himself did) that the GCH has a determinate truth value (Gödel thought it was "actually" false) but we just haven't come up with the full (or correct) set of axioms to settle it.

Philosophy 415: Fall 2025
History and Philosophy of Mathematics

Notes: Week 7

Building up the Reals

We've already seen how to "build" – in terms of sets – all of the following: the naturals, the (positive and negative) integers, and the rational numbers. Finally, what about the real numbers, **R**? This is an interesting issue, since as we've seen the fact that there are many more (infinitely more) real numbers than rational ones was a huge shock to the ancient Greeks, and also gives us, as we've seen, the first infinite set that has a bigger cardinality than that of **N**. There are a few different ways to define the real numbers. For instance, we could do it in terms of converging series of sums of rationals (that's essentially what we presupposed in using the "binary" representation above). But by far the simplest and (in many ways) most elegant way was given by Richard Dedekind and is called the **Dedekind cut**. Intuitively, the idea is that we can consider each real number to be a "cut" of the whole set of rational numbers, so that *every* rational number is either below or above the "cut". In fact, we can just identify the cut – i.e. the real number – with the set of all rational numbers less than it. This sounds circular, but it needn't be, since we can just define the Dedekind cuts as sets of rationals:

A Dedekind cut is a subset x of **Q** (the set of rationals) such that:

a) x is not ∅ or **Q** itself

b) x is "closed downward." That is,

for all rationals q, r, if q ∈ x and r < q, then r ∈ x

c) x has no largest member

Then we can just identify each real number with a Dedekind cut, and the set of all real numbers, **R**, with the set of all Dedekind cuts. We also have a nice ordering relation: where a and b are reals, a < b iff a is a proper subset of b.

Axioms of Set Theory

So far, we've motivated and learned set theory just intuitively, talking about which kinds of sets can and can't exist. But the essence of set theory, like any other mathematical theory, is in its precise definitions and rules. These are captured in a system of axioms which are the "starting points" for all formal set theory (very much as Euclid's axioms are the "starting points" for geometry). Similarly, also, to geometry, there are several different systems of axioms which yield different results but are nevertheless consistent. Today, however, we'll learn the most standard axiomatization, the so-called ZFC axiomatization based on the work of Zermelo and Fraenkel, plus the Axiom of Choice (the 'C' of ZFC). This axiom set is more or less canonical, but this does not mean that there are not deep and profound philosophical issues lurking beneath virtually each of them. Also, the choice of which axioms

to adopt and whether the ZFC system, as opposed to any of the other available systems, is the "right" one, is itself a philosophically significant question; we'll talk later about the interesting and important philosophical issues that are sensitive to this choice.

To begin with, there are several axioms that we've already – implicitly at least – 'met' in our informal explorations of set theory so far:

1. Axiom of Extensionality: If two sets have exactly the same elements, they are exactly the same set.

This is the axiom that lets us say that a set is defined by its elements.

2. Axiom of Pairing: If A and B exist, then {A,B} exists.

This is a general (and quite simple) principle of set formation. We've appealed to it, implicitly at least, in our discussions ever since we started talking about what sets are.

3. Axiom of Union: Given a set containing various sets, the set containing all and only the elements of these contained sets exists.

In other words, supposing we have a set such as {{a,b}, {c,d}, {e}}, the Axiom of Union lets us put what's inside all of these sets together to construct the union set: {a,b,c,d,e}. This is the axiom underlying our informal use of the operation of union since we started talking.

4. Axiom of Power Set: Given a set A, the set $\mathcal{P}$(A), defined as the set containing all and only subsets of A, exists.

This is the axiom underlying our discussion of power sets.

Now we have a couple of axioms that, while perhaps more philosophically controversial in certain ways, have also been presupposed in our discussions so far.

5. Axiom of Empty Set: There is an empty set, i.e. a set that has no elements.[1]

6. Axiom of Infinity: There is an infinite set. More technically, if we let S(X) stand for X ∪ {X}, then there is a set that has ∅ as an element, and that is such that for each x in it, S(x) is also in it.

In terms of the way we defined natural numbers, the axiom of infinity thus says that there's a set that contains 0 and all of its successors. This guarantees the existence, at least, of **N** (or **ω**).

Now we have the two axioms that we discussed last week; both of these were introduced to help ensure that Russell's paradox, and related paradoxes, can't arise.

[1] Some lists of the axioms omit the Axiom of Empty Set, since (on some views of logic and identity at least) it can be 'derived' from the Axiom of Separation, for instance by taking from any existing set the subset of elements that have both property P and ~P.

7. Axiom of Foundation (or regularity): Every non-empty set, x, contains an element, y, such that x and y have no common elements.

Recall that (and how) the Axiom of Foundation functions to prohibit self-membered sets. It also guarantees that each set can be "decomposed" into its subsets in such a way that we'll never get caught in a circle; thus it imposes a kind of hierarchy on the universe of sets itself.[2]

8. Axiom (schema) of Separation: Given any set A and any well-defined property ϕ, the set B of all elements of A having the property ϕ exists.

Recall that this is the replacement for Frege's original law of universal comprehension. It doesn't say, as Frege's law did, that we can create a set corresponding to just *any* property. Rather, it says that given any *already existing* set, we can 'separate out' a set containing just those elements that have the property. Also remember that this isn't a single axiom, properly speaking. Since there are an infinite number of properties ϕ, this is just a general schema that refers to an infinite number of individual axioms (one for each property).

There's another useful general axiom schema which is closely related to the Axiom (schema) of Specification, the Axiom (schema) of Replacement. In fact, since the Axiom (schema) of Specification can be derived from the Axiom (schema) of Replacement together with the Axiom of Empty Set, the former is sometimes omitted from presentations of the axioms that include the Axiom (schema) of Replacement.

9. Axiom (schema) of Replacement. If F is a function defined on the set A, its image under the function, F(A), also exists.

The Axiom (schema) of Replacement just says that if we have a set and a well-defined function (to give a full specification of the Axiom schema, we'd need to spell out in more detail what being well-defined amounts to), then the result of evaluating the function for each member of the set and putting all the results together is also a set. More intuitively, if we have a set consisting of a number of elements and a well-defined function from those elements to others, it allows us to 'replace' (hence the name) each element of the first set with a unique element of the other and thereby create a new set. For instance, if we have the set {1, 2, 3} and a function that takes 1 to a, 2 to b, and 3 to c, it allows us to say that the set {a, b, c} exists.

---

[2] Why does this prohibit *circles* of foundation (such as the situation where A ∈ B, B ∈ C, and C ∈ A)? In such a situation (by the axiom of replacement), it is possible to group all the members of the circle into a set, say S = {A, B, C}. But then S is not disjoint from A, since A contains C; it is not disjoint from B, since B contains A; and it is not disjoint from C, for C contains B. Therefore it has no element with which it is disjoint, and the condition expressed by the axiom fails. By a slightly more complicated argument, the Axiom of Foundation also effectively prohibits infinite descending chains of foundation.

Like the Axiom Schema of Separation, this is also a schema, rather than a single axiom, since there are an infinite number of functions, and strictly speaking we need a different axiom for each one.

Finally, we have (on some axiomatizations, at least) the famously controversial Axiom of Choice:

10. Axiom of Choice. Suppose X is a set, all of whose element are non-empty. Then there is a "choice function," f, that selects one element from each of the elements of X, and a 'choice set' that groups these all together.

The idea of the Axiom of Choice is straightforward: suppose I have a big bucket containing smaller buckets, and in each of the smaller buckets is a number of items. The choice function just takes one item from each of the smaller buckets, and groups all the selected items together. For finite sets, this is no problem, and in fact we can derive the axiom of choice for finite sets from the other axioms. However, where it's thought to be more problematic is where we have an infinite set of sets, and we again need to choose one from each of an infinite number of buckets (which *may* themselves contain infinitely many items). The reason this is thought to be problematic is that the Axiom itself gives us no "selection criterion" – it doesn't tell us *how* to choose an item out of each one, just that we *can* do so. Bertrand Russell drew a famous analogy for why this is problematic: If we have an infinite number of pairs of shoes, and have to select one shoe from each pair, we can easily come up with a rule for doing so: just pick the left shoe in each case. But if we have to perform the same task with an infinite number of pairs of socks, we are at a loss to specify any such rule, since socks don't come sorted into left and right. The Axiom of Choice, if it is true, essentially puts us in the latter situation, the situation of selecting from socks: we have to know that we can make an infinite number of selections, without having any rule for doing so. Thus, the Axiom has been called "non-constructive", in that it (plausibly at least) doesn't actually tell us *what is* the set whose existence it asserts.

The Continuum Hypothesis

Let's summarize some of our cardinality results so far:

$|\mathbf{N}| > n$ where n is any finite number.

$|\mathbf{R}| > |\mathbf{N}|$

$|\mathbf{Q}| = |\mathbf{N}|$

$|\mathbf{N} \times \mathbf{N}| = |\mathbf{N}|$

$|\mathbf{R}_{(0,1)} \times \mathbf{R}_{(0,1)}| = |\mathbf{R}_{(0,1)}| = |\mathbf{R}|$

We know from Cantor's theorem that there are various sizes of infinity, and we've already seen that there are in fact at least two such sizes (since we know by diagonalization that $|\mathbf{R}| > |\mathbf{N}|$).

We also distinguished between "ordinality" (which one? in an order) and "cardinality" (how many?) For the finite natural numbers, ordinality and cardinality coincide; this is what lets us, after all, use the

normal ordinal counting sequence ("1", "2", "3", "4", etc.) to answer questions of cardinality too. However, when we get to infinity, things (as usual!) get more tricky. Using the definition of numbers in terms of pure sets that we adopted earlier, where

0 = ∅

1 = {0} = {∅}

2 = {0,1} = {∅,{∅}}

3= {0, 1, 2} = {∅, {∅}, {∅,{∅}}}

Etc.,

we identify each number with the whole sequence of its predecessors; thus for instance, as we saw, 3 = {0,1,2}; 4={0,1,2,3}, etc. These sets all have the nice property of *transitivity*: that is, for any one of them, (say X), whenever a is an element of b, and b is an element of X, a is also an element of X. (Thus, for instance, 2 is an element of 3; and 3 is also an element of 4; so 2 is an element of 4, as it should be). In fact, we can use this property to define *ordinal numbers*: an *ordinal number* is just a transitive set consisting only of transitive sets. This is in fact just a fancy way to say that each ordinal number is the set grouping together all of its predecessors.

We also know from the Axiom of Infinity that there is an (at least one) *infinite* ordinal. This is the set consisting of all the finite ordinals, and we have already come up with a symbol for this ordinal: **ω**. We can also keep using our rule for the generation of ordinals here (we just include the ordinal we have along with all of its predecessors) to generate the very next ordinal, the successor of **ω,** or **ω+1.** As we've seen, for all the finite numbers at least, we can just identify ordinals with cardinals (the question of which one? in the series can be just identified with the question of how many? up to that one). But here things start to get tricky, and we have to begin to distinguish ordinals from cardinals. The trouble is that the cardinality of **ω+1** is in fact equal to that of **ω** – not "one greater", as we might like. Why? Well, recall the issues about infinity that we discussed a few weeks ago; and in particular, remember Hilbert's hotel, which (although full) could still accommodate one more weary traveler. The point of that example was effectively that we can put the set of **ω+1** elements in one-to-one correspondence with the set of only **ω** elements. (More formally, just map each finite *n* in **ω+1** to *n*+1 in **ω**, and then we can just map the "extra" element of **ω+1** (this element is actually **ω**) to 0). Using similar considerations, in fact, we've established that:

| **ω+n**| = | **ω**| for any finite **n**

and also even that

| **ω + ω| = |2ω**| = | **ω**|

In fact, using similar considerations (criss-cross through an *n*-dimensional array), we can even show that for any finite **n**

|n x ω | = | ω|

And that for any finite **n**:

| ω^n | = | ω|[3]

In fact, we can do even more while remaining within the domain of sets that have the same cardinality as **ω**. For example, **| ω^{ω} | = | ω| !** This may seem strange, but the real reason is that is that **ω^{ω}** is just what we get by multiplying **ω** by itself **ω** times. Since there are **ω** iterations, this result can be put into one-to-one correspondence with **ω** itself – thus it has the same cardinality. In general, in fact, anything that we can get by doing some arithmetical operation **ω** times can be put into one-to-one correspondence with **ω**. In fact, this way we can even get the fantastically large

$$\varepsilon_0 = \omega^{\omega^{\omega^{\cdot^{\cdot^{\cdot}}}}} = \sup\{\omega, \omega^{\omega}, \omega^{\omega^{\omega}}, \omega^{\omega^{\omega^{\omega}}}, \ldots\}$$

This number also has the "funny" feature that $\varepsilon_0 = \omega^{\varepsilon_0}$. Because of this, it is also the "largest" number that has a straightforward expression in terms of purely arithmetical operations. But it still has the same cardinality as does **ω** itself.

Given all of this, we might suspect that there is only one infinite cardinality, the cardinality of **ω**. This would be true, if all infinite sets could be put in one-one correspondence with each other, and they all could be put in one-one correspondence with **ω.** But in fact, we already know that this *isn't* true. For recall that (by Cantor's theorem, or diagonalization),

$|\mathcal{P}(x)| > |x|$ for any x

And in particular,

$|\mathbf{R}| = |\mathcal{P}(\omega)| > |\omega|$

We also saw that for finite sets at least, $|\mathcal{P}(\mathbf{x})| = 2^{|x|}$; and since $|x| = x$ for finite *ordinals*, that for these ordinals at least, $|\mathcal{P}(\mathbf{x})| = 2^{x}$. If we like, we can extend this idea to power sets of infinite sets, and then we have, by definition:

$|\mathcal{P}(\omega)| = 2^{\omega} = |\mathbf{R}|$.[4]

---

[3] **IMPORTANT NOTE:** Here we are using so-called "ordinal exponentiation". Ordinal exponentiation is defined in terms of multiplication of ordinals, which is well-defined even for infinite ordinals such as **ω.** BUT it is a very different notion from "cardinal exponentiation," which is defined in terms of power sets. We will switch to "cardinal exponentiation" below when we discuss the cardinality of $\mathcal{P}(\omega)$.

[4] **IMPORTANT NOTE:** Now we are using exponentiation in the sense of "cardinal exponentiation" – NOT "ordinal exponentiation." Cardinal exponentiation is *defined* in terms of the power set, rather than in terms of multiplication (see note 1 above). From now on, exponentiation involving infinite sets will always be "cardinal exponentiation" unless otherwise noted.

We now know, at any rate, that there are *at least two* infinite cardinalities, and that one is larger than the other. Let's fix some terminology. To talk about cardinals rather than ordinals (since many ordinals, as we've seen, have the same cardinality), we'll use (following Cantor) the series of *alephs*, $\aleph_0, \aleph_1, \aleph_2$, etc.

By definition, let's let $\aleph_0$ (pronounced "aleph-null") be the cardinality of the first infinite set, **ω.** Thus,

$|\boldsymbol{\omega}| = \aleph_0$

All the sets that can be put in one-to-one correspondence with **ω** (for instance: **ω** +1, 2 **x ω**, etc.) will also have cardinality $\aleph_0$. Because these sets can be put in one-to-one correspondence with the set of all natural (counting) numbers, these sets are sometimes called *countable* sets (or "countably infinite" sets). Then $\aleph_1$, the *very next* cardinality, will be the cardinality of the *first* set that is "bigger" than all of these in the sense that it *cannot* be put in one-to-one correspondence with **ω** or any of these. And $\aleph_2$ will be the (cardinality of the) first infinite set that can't be put in one-to-one correspondence with any set having cardinality $\aleph_1$ or $\aleph_o$, and so forth.[5]

We are now in a position to pose one of the most famous unsolved (and, as we now know, unsolvable) problems in all of mathematics. We know that $|\mathbf{R}|$ -- the cardinality of the set of real numbers (or, as it this set is often called because of its conceptual connection with the set of points on a continuous line, the *continuum*) is *greater than* the cardinality of the set of naturals, which equals (by definition) $\aleph_0$. But what is $|\mathbf{R}|$, actually? In particular, is it true $|\mathbf{R}|$ is equal to $\aleph_1$, *the very next* transfinite cardinal after $\aleph_0$? The hypothesis that this *is* true – that $|\mathbf{R}|=\aleph_1$, is known as Cantor's Continuum Hypothesis. Cantor thought it was true, and himself labored in vain throughout the last years of his life to prove it from the ZFC axioms. But he didn't succeed, and no one else did either.

Since we know that $|\mathbf{R}|$, whatever it is, is equal to $|\mathcal{P}(\boldsymbol{\omega})|$, and that (by definition), this = $2^{|\omega|}$ or $2^{\aleph_0}$, we can put the hypothesis very simply:

Cantor's Continuum Hypothesis (CH)

$2^{\aleph_0} = \aleph_1$

Clearly, the truth or falsity of the Continuum Hypothesis bears fundamentally on the question of what sets there "actually" are, or at any rate which sets there are according to the ZFC axioms. For instance, if the CH is true, then there is no set of real numbers that is larger than (of greater cardinality than) the set of naturals but smaller than set of *all* real numbers; if it is false, then there is such a set. In fact, we can

[5] More rigorously, just as we defined the ordinals in terms of pure sets, we'll want a definition of the transfinite cardinals in terms of pure sets as well. What, then, actually are the cardinals? Well, since we know that, for each cardinality, there is *at least one* ordinal that has that cardinality (although of course there are usually *many* ordinals with the same cardinality), and also that the for any set of ordinals, the set has a least member, we can just identify each cardinality with the least ordinal that has that cardinality. Thus, the cardinals really are (some) ordinals (for instance, $\aleph_o$ "actually" equals not only $|\boldsymbol{\omega}|$ but also **ω** itself, since **ω** is the smallest ordinal that has that cardinality) and we just use the two notations to talk about the same things under different aspects or guises.

generalize the CH in such a way that it bears on the shape of the *whole* hierarchy of finite and transfinite sets, the whole "set-theoretical universe" V:

Cantor's Continuum Hypothesis – Generalized Form (GCH)

For every ordinal *α*,

$2^{\aleph_\alpha} = \aleph_{\alpha+1}$

Seen another way, the issue of the truth or falsity of the GCH is really just the question of the "power" of the "power set" operation. By just how much does the power set of a given set exceed that set itself? A set of size $\aleph_1$ has, remember, the very next size that it is possible for any set to be, beyond $\aleph_0$. The question is whether this size is just as big as the power set itself; whether, in other words, the hierarchy of alephs (or possible sizes of sets) "measures" the excess of the power sets themselves. If so – that is, if the GCH holds – then the total universe of sets is relatively ordered and well-behaved. We can (as it were) use what we "know" from each level of set creation to "predict" what will happen when we form the power set. By contrast, if this is not so and the GCH is false, then the power set operation, we might say, introduces a kind of "immeasurable" excess in the hierarchy or series of sets; there's literally no telling how much "bigger" (in terms of the series of alephs) a power set will be than its original set.

What, then, is the status of the CH? We now know (perhaps sadly or perhaps suggestively) that it can't be either *proven* or *disproven* from the ZFC axioms. In other words, it is literally *independent* of the axioms: the axioms, and the system they define, just don't say whether it's true or false. This result is the combination of two very important results in set theory (and its meta-theory). First, in 1938, Kurt Gödel showed that the GCH can't be *disproven* in ZFC (this was Gödel's second most famous result, after the incompleteness theorem, which we'll study later). That is, it's impossible to prove the *negation* of the GCH in ZFC; we might put this as saying that it's *possible* (according to ZFC) that the GCH is true. However, in 1963, P. J. Cohen, by means of a profound technique called "forcing," showed that the GCH can't be *proven* in ZFC either. Thus (as we might say), it's equally possible (according to ZFC) that the GCH is false. This has sometimes been taken to suggest that the axioms actually under-describe the real universe of sets, that they don't really settle the question of what sets actually exist. We could take this to mean that the set theoretical universe is just inherently vague and ill-defined. Or we might think (as Gödel himself did) that the GCH has a determinate truth value (Gödel thought it was "actually" false) but we just haven't come up with the full (or correct) set of axioms to settle it.

Sets, Models, and "Reality"

When we raise questions such as the question of the truth of the GCH, or of the axioms themselves, we are asking about the truth or falsity of claims. Claims about what? We might just answer "sets" or "the set-theoretical universe." But also, given that we are identifying some of our sets with numbers (including transfinite ones), and we want our set-theoretical claims to underwrite actual arithmetic, we probably want to say something like "about sets, including those that 'are' numbers". Either way, when we take this attitude toward sets and/or numbers, we are essentially assuming that our set-theoretical

axioms describe some actual, well-defined "mathematical" reality, and that the conclusions we draw logically using our axioms can also be expected to describe this reality accurately and (we hope!) completely. But of course we cannot observe this "mathematical" reality using any of our five senses, so there is an important question to be raised about what this assumption actually means.

We already have met a couple of the notions that we need to talk about this assumption, namely the notions of **soundness** and **completeness**. Remember that soundness is the idea that a formal system tells "nothing but the truth" – that everything we can derive using the system can be assumed to be true in reality. And completeness is the idea that it tells "the whole truth" – that there is nothing true in reality that it doesn't tell us. Soundness and completeness are both "**semantic**" notions. That is, they have to do not only with the rules of the system, but their relation to some external reality that the system is supposed to be "about".

If we consider *just* first-order sentential logic, *without* any objects such as sets or numbers, we can show that this system by itself is both sound and complete. That is, for first-order sentential logic,

Soundness: If A ⊢B, then A ⊨B

This says, roughly, that if B follows from A *in the system* then any situation *in reality* in which A is true is *also* one in which B is true.

And we can also show

Completeness: If A ⊨B, then A ⊢B

Which says (again roughly) that if, in *any* situation in reality in which A is true, B is also true, then the system can derive B from A.[6]

That first-order logic is, by itself, sound and complete, is a useful and important result. On the other hand, though, first-order logic by itself is relatively impoverished. We can't use it to capture even basic mathematical and arithmetic operations, since we don't have any specified "objects" to use as numbers or any vocabulary to express their relationships. To move, now, to set theory and to "mathematical reality" in the specific sense that we are hoping to use sets to capture, we need to specify in more detail this notion of a "situation" that we are using to capture the ideas of soundness and completeness. The standard way to do this is to use the notion of a *model*. A model for a theory is a domain of objects that the theory is considered to be "about", and with respect to which we can think about the truth or falsity of its claims. Intuitively, to determine a model for a given theory, we have to specify objects to which each of its constant terms correlate, and ranges of objects for its variables to range over. Also, if there are any primitive predicates or relations (such as "element of") in set theory, we need to "interpret" these predicates by specifying the ranges of objects or pairs (etc.) that they hold of. The study of the

[6] This last result is also due to Gödel. Though this "Gödel completeness theorem" is much less famous than Gödel's *incompleteness* theorem, which we'll study in detail later in the class, in some ways it is just as important!

relationship between theories and their models is called *model theory*, or sometimes (since it lets us talk about semantic notions such as truth, soundness, and completeness) *formal semantics*. When we introduce the ZFC axioms, then, we want to see if there is a model for which all of the axioms hold. And ideally, we'd like this model to be the "set theoretical universe" or "mathematical reality" itself.

There is a corollary of Godel's *completeness*[7] theorem that tells us that any theory that is based on a *consistent* set of axioms has at least one model. So if the ZFC axioms are *consistent* (i.e. they cannot be used to prove both P and ~P, for any P), then there is some model, some "universe," of which they hold true.[8] Now, it would be nice if we could just identify this model with the "set-theoretical universe" or "mathematical reality". Then we could just assume that our ZFC axioms are true of that reality, that they really hold of the numbers that we've been familiar with since we started counting, and that all of their consequences are simply true of that reality itself. Can we do that? Unfortunately, not without further ado. For we can use model-theoretic reasoning to show that the ZFC axioms – and, in fact, *any* theory formulated in propositional logic – are insufficient to pick out the "intended model" of mathematical reality uniquely.

In particular, when we are dealing with theories that are supposed to cover the transfinite, things get tricky. There is a model-theoretical theorem proven by Löwenheim and Skolem that shows that it is impossible to make sure that our ZFC axioms pick out uniquely the "intended" model, since any axiomatic theory such as ZFC will pick out many (actually infinitely many) "unintended" models as well. This means that no axiomatized theory such as ZFC (or, more or less equivalently, Peano arithmetic) can *uniquely* specify the structure of the "actual" domain of natural numbers. Given any model that such a theory is true of, there will be many more "nonstandard" models, including lots of things that (intuitively) aren't natural numbers, in which everything proven by ZFC will still hold true.

In fact, things are even worse, as is shown by this corollary of the L-S theorem:

Löwenheim-Skolem theorem (downward version): Every theory that has an infinite (countable or uncountable) model has (infinitely many) models containing only *countably* many objects.

This implies that everything we can prove to be true using ZFC, including all the truths about the vast range of transfinite sets, which we took to be "about" the whole vast set-theoretical universe V (which, we assume, includes many uncountably big sets), is also true of universes that have only countably many sets! So in a certain sense, we cannot be "sure" that we are talking about anything bigger than a countable universe at all. And this, in turn, leads to a situation that seems, at least, paradoxical.

Skolem's paradox: All of our results about transfinite cardinality – including, for instance, that $|\mathbf{R}|>|\boldsymbol{\omega}|$, also hold true in a universe with only $\boldsymbol{\omega}$ sets.

[7] NB – not "incompleteness"

[8] As we shall see in great detail in the second half of the course, though, this is not a trivial assumption! In fact, we'll see that in a very real sense it is actually *impossible* to prove that ZFC is consistent.

What is paradoxical about this comes out if we consider a particular countable model, M. In that model, there are only **ω** sets, so even **R** (as defined in this model) has only, at most, **ω** elements! But our result that $|\mathbf{R}|>|\boldsymbol{\omega}|$ nevertheless still holds, even there.

What is going on, then? Though the situation seems paradoxical, there is actually a way to defuse it from the perspective of our general set-theoretical reasoning. To see this, consider what the claim that $|\mathbf{R}|>|\boldsymbol{\omega}|$ actually says. It says that, within the relevant universe, there are two sets, **R** and **ω**, and there is no one-to-one correspondence between them. Now, within the model M, R and **ω** both exist, and both are countable (since everything in M is). And both are defined exactly as they usually are. The trouble comes with the idea of one-to-one correspondence. What, after all, is a one-to-one correspondence? It is actually a function, or a set of ordered pairs. And although R and **ω** both exist, and there is (in an absolute sense) a one-to-one correspondence between them (since they are both countable), ***this one-to-one correspondence is not itself in M***. From a perspective where we can "stand outside" and look down on M, we can see that it exists (in the *general* set-theoretic universe, as we might say), but for an "inhabitant" of M, it doesn't exist, and so $|\mathbf{R}|>|\boldsymbol{\omega}|$ holds true.

We can resolve the paradox, in this way, from the perspective of model theory, where we "stand outside" the particular model M, and think about how it is embedded in a "bigger" universe of sets, V. That is, *to the extent that we can think of ourselves as "inhabiting" and "having access to" V rather than just M*, we can resolve the paradox. In fact, being able to "stand outside" in this sort of way is what we need to do in order to do model theory at all. Also, we can resolve the paradox, in this case, just because we have access (or assume we do) to sets in the broader hierarchy V, which goes far beyond M itself. Thus, by seeing (as it were from the perspective of V itself) that there is a one-one correspondence that is NOT visible to an "inhabitant" of M, we can resolve the apparently paradoxical situation.

However, though THIS paradox is thus resolved, there is, as Moore says, nevertheless still a deeper problem. The problem is: *in standing outside M to comment on it, how do we know we are taking the perspective of the real V (i.e., the 'real universe of sets') itself*? We could, after all, be essentially in the same situation as that of the inhabitants of M, trying to describe the whole of set-theoretical reality from the perspective of a limited, impoverished model of it. When we talk this sort of way, or do model theory at all, we essentially assume that when we talk about "sets" we are talking about *all* sets, or when we talk about elements we're talking about *all* elements. *If* we can assume that we can be sure we have access to *all* of the sets, then we can avoid the paradoxical implications of the L-S theorem. But, it seems, we cannot assume this, for at least two interlinked reasons. The first is that, as we have seen, according to the theorem itself, any model we can present as a model of our theory might actually be (from another perspective) something less than the "full" universe of sets. But the second is that, as we have seen, it is a consequence of the axiomatization developed in response to Russell's paradox that the "full" set-theoretical universe, V, doesn't even exist! In particular, there is no such thing as the set of all sets; so it is puzzling how we could claim that we can have any kind of access to it. There is, it seems, a way in which this "reality" (if we can understand it as such) irrevocably escapes, no matter how hard we try to capture it with formal theory.

Putnam and the "Skolemization of absolutely everything"

The problem that Skolem points out with respect to set theory is a problem that is specified in terms of formal semantics, where we are trying to think in formal terms about how our theories (in particular ZF or ZFC) are "true of" a reality, under the assumption that there is a reality that we are trying to describe. We can solve the problem – as above – *if* we can appeal to our "intended" interpretation of our set-theoretical theory – that is, that it is "really" about V, the "whole" universe of sets – or that it is about "the" set concept itself, on its "intended" interpretation. But it also looks like there is no *theoretical* way to support this, without appealing to something like a mysterious power of our minds to grasp **that** universe – and that interpretation's bearing on it – in a way which not only isn't but (in principle) *couldn't* be captured by any theory of this relationship itself. For, of course, any theoretical description of what we are grasping and how we are grasping it will itself be open to interpretation in terms of alternative models, as the L-S theorem points out. (Later on, we'll encounter a similar situation with respect even to "the" natural numbers themselves: since we won't be able to rule out alternative models by means of *any kind of* theoretical constraint, we'll have to wonder how it is that we can *know* that we are talking about *those* objects at all rather than some non-standard things: and this worry will extend, in principle, to *all* of arithmetic.)

As Hilary Putnam points out in his 1980 article, the problem we've run up against also has implications for the more general problem of the relationship of *all* of our scientific language to the "world" we take it to describe (and hence to the complex issue of "realism" vs. "anti-realism" about our scientific theories). A lot of twentieth-century philosophy of science was concerned with the question of how we can be sure that even our best scientific theories are *actually* describing reality, especially given that (in the past) a lot of successful scientific theories (e.g. the theory of Newtonian physics) have turned out to be false. More generally, we can ask how it is that we can guarantee that *any* theory is really "about" the world at all in the sense of being true of it, rather than just about itself. Here, as Putnam points out, there are two extreme positions: the strongly *realist* position is the "Platonist" one where the mind can just by itself grasp true reality (by means of non-natural powers) and see that our good theories are about *it*, and the strongly *anti-realist* position is the "verificationist" one where, instead of talking about theories being true or not, we should just talk about them as being verified or proven (so that we are always just talking "internal" to the theory). In between these positions, a lot of people have tried to come up with a middle way, where there are naturalistically understandable or physicalist ways we can understand that our theories are about reality itself.

But as Putnam points out, if we can't appeal to a non-natural and mysterious power of grasping what our theories are about, any amount of (even naturalistic) theory about *how* our theories connect to the world will *also* be open to a(n infinite) number of non-standard interpretations. Here, another corollary of the Löwenheim-Skolem theorem is relevant:

Löwenheim-Skolem theorem (second corollary): Any theory with a countable model has infinitely many *different* countable models.[9]

Think of our best scientific theories as containing a number of terms that are meant to refer immediately to the data of observation (e.g. "red", "hot", etc.) as well as a number of theoretical terms that are based on and defined in terms of these. It will also, of course, include mathematical vocabulary, including vocabulary for real-valued quantities. However, since all of our possible observations will only be (at most) approximations to real-valued quantities, it is plausible that our theory will only ever capture – at most – countably many observations (or data points). Given this, as Putnam argues, for *every possible* interpretation of the theoretical and observational language of the theory as referring to objects and relations, there will be infinitely many alternative possible interpretations which, though they preserve the truth-values of all of our theory's sentences, assign *different* objects and different relations to those same terms. For example, even our terms which are supposed to refer to ordinary material objects – for example the terms "dog" and "cat" – will get interpreted differently under the different interpretations (for example, the term "dog" may get an interpretation on which it refers to some dogs and some apples, or just to some pieces of furniture, etc.) This will *also* be true even for the terms that are supposed to fix the "basic observational data", e.g. the terms "red" and "hot" etc. And even if we try to "fix" this by means of our theory – for example by bringing into the theory an *explicit* "definition" of "dog" as (e.g.) "animal with four legs that barks" (etc.), then *all* of the terms in the definition will have (infinitely many) "non-standard" interpretations as well. Any such (theory-internal) "interpretation" will be, in other words, "just more theory," and (by the L-S corollary directly above) will fail once more to "fix" the identity of the objects and relations "up to isomorphism".

So given all of this, it becomes hard to say what it can mean that even our best scientific theories are "true" of reality. We could, of course, again appeal to the claim or idea that our minds just DO have access to the truth of our theory on the "intended" interpretation and that we can know by those means WHAT that interpretation is. But we then have to face that there is, in principle, NO way we could ever possibly describe how our minds have that access – or indeed verify that they "actually do". Or, we could just drop the idea that our theories describe a reality external to them at all. Maybe, what we are doing in theorizing – whether in mathematics or in science – is just moving around our own ideas and discursive rules, with no "external" constraint or reality. But this kind of anti-realism, which is captured in positions like verificationism (and – as we'll see in a couple of weeks – intuitionism in philosophy of mathematics), seems to suggest or imply that we cannot *do* science as access to a mind-independent reality at all, and raises big problems about what is meant by (or what we can understand as) truth at all.

[9] As model theorists say, these distinct models are distinct in that they are not *isomorphic* (so a way of putting the correlate is that no theory with a countable model determines its intended model "up to isomorphism")

Philosophy/Math 415/515: Fall 2025
History and Philosophy of Mathematics

Notes: Week 8

Priest and Inclosure

We saw a couple of weeks ago that there are some uncanny similarities between Russell's paradox, some of the paradoxes of absolute infinity (such as the Burali-Forti paradox and Cantor's paradox), and indeed the reasoning underlying Cantor's theorem itself. Although some of these results are paradoxical and others are not, it's worth exploring the underlying similarity to see what it teaches us about sets and language. Indeed, we saw that even with respect to the infinite, the standard, iterative conception (as expressed in Russell's theory of types and the axioms of ZFC) "buys" a kind of solution that saves consistency and allows us to talk about *certain* infinite sets, but only at the cost of prohibiting some others, and also leaves open questions about the scope of its *own* principles (for instance, about the size of the universe of all sets, V).

Graham Priest's remarkable contribution is to bring out this common structure, and thus to motivate the suggestion that there are indeed common issues underlying both the obvious paradoxes of absolute infinity and the derivation of the set-theoretical hierarchy itself, issues that also affect in a fundamental way what we should think about the status of language, thought, and their objects. In fact, Priest's claim in *Beyond the Limits of Thought* is that we are involved in this common structure of paradoxical self-reference whenever we think systematically about the *limits* of thought or language. This is indeed something that we do – or try to, at any rate! – constantly while doing philosophy. In the first part of the book, which we'll skip over, Priest documents how a certain kind of "contradiction at the limits of thought" occurs in the thinking of philosophers ranging from Plato and Aristotle, through Anselm and Aquinas, up to Kant and Hegel. Since all of these philosophers sought to understand the totality of what can be thought as existing, they all have taken positions, implicitly or explicitly, on thought's limits. But in so doing, they have all been involved in a certain kind of paradoxical structure. What is the paradox? Most generally, it is that in thinking about the limits of thought, we think something that (if we are successful) can indeed be thought. But then we are, seemingly, already beyond the very limits that we are thinking! Thus, assuming that everything thinkable stands within this limit, we seemingly must hold, contradictorily, both that the limits of thought are thinkable (we just did!) and that they are not (since to be thinkable they would have to stand within themselves).

Priest sees this exact structure of contradiction, which philosophy has long encountered implicitly, as standing directly behind Russell's paradox and the related paradoxes of set theory. In fact, it was Russell who first made this structure more or less explicit. For Priest, though, the essential structural commonality between what set theory treats as paradoxes to be avoided and positive results about infinity (like Cantor's theorem) to be embraced is actually a sign that both should submit to a unified treatment. In fact, according to Priest, *both* kinds of results – as well as many others in the history of philosophy – depend upon a kind of formal contradiction or paradox at the limits of thought and

expression, which we can now formalize and think about explicitly. This does not mean that we should dismiss Cantor's thought about the infinite as simply paradoxical and hence useless, however. Rather, it means we should recognize and attempt to formalize the paradoxes and contradictions that seem to run systematically through all the results of thinking about the totality of thought.

Diagonalization and Inclosure

At the heart of the common structure shared by the official paradoxes and other (apparently non-paradoxical) results about infinity is (as we've perhaps come to suspect) *diagonalization*. Roughly put, in the case of Russell's paradox, the Burali-Forti paradox, Cantor's paradox, *as well as* Cantor's theorem, we have an operation that allows us to take an initial set and use it to generate another set that has the *right form* to be a member of the first, but nevertheless *is not* a member of the first.

Thus we have a contradiction, whereby the newly generated element both is and is not a member of the totality; and Priest suggests that this kind of contradiction occurs very often, in fact whenever we want to think about a totality of a certain sort.

Priest formalizes this with a structure that he derives from Russell, and which Priest calls Inclosure:

First, suppose we have a set $\Omega = \{y | \phi(y)\}$, and $\psi(\Omega)$. [or in English: omega is the set of all phi's, and omega itself has the property psi] (This condition is called "existence")

Then suppose there is a function δ (intuitively, a "diagonalizer") such that:

For all x, if x is a subset of Ω such that $\psi(x)$, then:

(a) $\delta(x)$ is not an element of x (This condition is called "transcendence")

(b) $\delta(x)$ is an element of Ω (This condition is called "closure")

Once again, the intuitive idea is that the "diagonalizer" takes a set x and generates an element that isn't in that set, but nevertheless still has a certain form that makes it an element of a bigger set, Ω, of which x is a subset. Given the schema, however, we can always generate a contradiction. All we need to do is let $x = \Omega$. (We know we can do this, since Ω is a subset of Ω (every set is a subset of itself), and we know by assumption that Ω has the property ψ). Intuitively, what we're doing here is just applying the diagonalizer to the "maximal" or "total" set Ω, which is supposed to include everything having the property ϕ.

Then, by (a), we have: $\delta(\Omega)$ is not an element of Ω. However, by (b), we have: $\delta(\Omega)$ is an element of Ω. This is, of course, a contradiction, and Priest suggests it is the general contradiction underlying all paradoxes of this type.

For instance, we can get Russell's paradox from the schema as follows: Just let ϕ be the property of being a set; so that Ω is the set of all sets, and let ψ be the 'universal property' that everything has. (In some cases, like this one, we don't really have to do anything with ψ, so we can just let it be the 'universal property' and it drops out). Now let δ(x) be the function that takes x to the set of all elements of x that are not self-membered. Given this definition, it's easy to verify that conditions (a) and (b) hold; that is, δ(x) is not an element of x (Argument: δ (x) is not self-membered; since if it were it self-membered it would be a member of the set of elements in x that are not self-membered, which would produce a contradiction. Therefore δ (x) is not a member of x, since if it were it would be a member of x that is not self-membered, and hence would be a member of δ (x), which we just ruled out). But δ (x) is still a set, so condition (b) holds too. Thus, if we consider δ (V), we get an inclosure contradiction: δ (V) both is and is not an element of V.

Let's do one more case: the Burali-Forti paradox that is usually taken to show that there is no set of all ordinals. Here, we just let both ϕ and ψ be the property of being an ordinal (i.e. being a transitive set consisting only of transitive sets), and let δ(x) be the function that takes an ordinal to its successor (= S ∪ {S}). Then we just take Ω to be the set of all ordinals. By the definition of ordinals, the set of all ordinals is an ordinal (condition 1). And by the definition of successor, the successor of any ordinal is not an element of that ordinal (2a), and the successor of any ordinal is also an ordinal, and so is an element of Ω. Thus we have the contradiction: the successor of Ω both is and is not an element of Ω.

Diagonalization and the Domain Principle

This schema is very general, and seems to work as well to capture the structure of the more seemingly "semantic" paradoxes such as Berry's paradox of the least number not definable in 19 syllables or less. (Exercise for home: show how this fits into the inclosure schema, taking δ (x) as the least number not in x, ψ to be the property of being definable in 18 syllables or less, ϕ to be the property of being definable in 19 syllables or less, and thus Ω as the set of all numbers definable in 19 syllables or less.) But it also seems to work just as well for the (ostensibly non-paradoxical) results that 'generate' infinities out of other infinites (the underlying basis of which is Cantor's theorem). Why should this be so? The reason is that these 'positive' results, just as much as the paradoxes, rely on diagonalization. In each case, we can use a certain operation to "diagonalize out" of a certain set that should be total an element that both "is" and "is not" a member of the totality. This gives us, in general, an element that is (because it is the product of diagonalization) demonstrably not equal to any member of the totality, but is nevertheless of the "right form" to be.

Recall how this worked with the diagonalization that "demonstrated" that the cardinality of the reals is greater than that of the naturals (the demonstration we did on the first day). Basically, what we did was to suppose we had a definite totality of real numbers, in one-to-one correspondence with the naturals. Then we "diagonalized out" by means of a function to give us a "diagonal number" that demonstrably isn't equal to any of the numbers in our totality, but nevertheless still is (formally) a real number. This isn't an outright contradiction, as long as we assume that it makes sense to suppose that different infinite sets of reals are bigger or smaller than one another. However (as is often the case, it seems),

here it seems as if we've just traded in a contradiction involved in the very idea of a certain totality for the idea of a bigger totality.

In fact, the same strategy seems to underlie Cantor's theorem itself, in its more general form, and thus to underlie the whole basis of the "transfinite hierarchy" of sets. Cantor's theorem shows that the power set is essentially a diagonalization operator: it allows us to go from any set to a bigger set that's not an element of the first. Repeating this operation, we can generate the whole transfinite hierarchy of sets; but of course we get back into contradictions when we consider the set of all of these, the set of all sets. Then, as we saw, according to the standard formulations of set theory, we have to resolve these contradictions by other means, for instance by type theory, or by axioms that tell us that we can't form the set of all sets or that it's not really a set at all. The persistence of these issues might be taken to suggest that, in using diagonalization to "generate" the whole infinite hierarchy of sets, we haven't really solved the underlying contradictions involved in the very idea of diagonalization itself. At best we've deferred these paradoxes of totality, but they will reappear. In fact, if Priest is right, we have never really left them, since the underlying structure is the same in both cases.

If this is right, then it suggests that there's something unsatisfying about allowing this structure in some cases (for instance to 'generate' the transfinite hierarchy) and disallowing it in others (for instance when we get to the set of all sets and other such totalities). In fact, according to Priest, the standard set-theoretical 'resolutions' of the paradoxes, which yield conceptions such as Russell's theory of types and the standard, iterative conception of sets, are essentially just ways of papering over the paradoxical structure which is at the heart of all thinking about infinite totalities, and don't really provide an adequate motivation for the solutions they introduce.

To see this, consider the following principle, which Priest calls **The Domain Principle**:

Every "potential infinity" presupposes an actual infinity.

There is some evidence that Cantor believed this principle, and in fact it is a key to his radical step forward in developing the theory of infinite sets. As we've seen, Cantor adopted basically two principles in order to generate the whole universe of ordinal numbers. The first is simply that every ordinal has a successor; we can generate this successor, as we've seen, using just the axioms of pairing and union, as S U {S} for any ordinal S. Second, is the principle that is used to generate limit ordinals such as **ω**: the principle that "if any definite succession of …[ordinals] …exists, for which there is no largest, then a new number is created…which is thought of as the *limit* of those numbers…" (quoted in Priest, p. 115). (In ZFC set theory, the possibility of creating limit ordinals is ensured by the axiom of infinity). This second principle, in particular, made possible the whole positive mathematical treatment of the infinite. For this radical step involved, of course, what Aristotle and others had treated as the merely "potential" infinity of natural numbers as presupposing an actual, completed set, the whole set of natural numbers. And it's not an unreasonable principle, in itself. For how are we to think of the "potentiality" of, say, counting to the next number, unless that next number is in some sense already "there," along with all of its successors as well? It's not as if we just freely create each number as we go; if we did, this wouldn't

be a potential *anything*, but just a random and chaotic collection. What's more, our constant use, in both ordinary language and mathematics, of variables ranging over certain infinite domains (such as 'z' in the statement: "Let z be a root of the equation $a^2+bx+c=0$. Then z has at least one value."), effectively demands that we know what domain these range over; and so this very use seems to presuppose not only the truth of the domain principle, but also that we can meaningfully know of and talk about these infinite (actual) totalities.[1]

On the other hand, though, the domain principle obviously also allows us to create certain sets (or totalities) that Cantor did *not* believe we could have any positive conception of. For example, since the principles of generating *all* ordinals are well-defined, it ought to allow us to generate the set of *all* ordinals, ON. We know from Burali-Forti (Cantor himself had earlier discovered the same result) that the assumption that this set exists generates a contradiction about the size of ON. But does this by itself establish that we cannot (and in fact do not) apply the domain principle here? In fact, in talking about ordinals we are able to make certain claims about them (in fact, there is a perfectly adequate definition of ordinals, as we've seen). These claims, on their face, are claims about *all* ordinals. Someone who holds that the set of all ordinals doesn't exist must hold either that we can't really make these claims, or that we don't really understand what they mean, or that they are in some way ambiguous or ill-defined. But none of these suggestions seem motivated, outside simply the desire to avoid paradox (or contradiction) at all costs.

Moreover, as Priest points out, the standard, iterative conception of sets (as formulated in ZFC) violates the Domain Principle. As we've seen, it's standard to picture this as an inherently "open" universe of stratified levels of "bigger and bigger" sets, taking the form of a large open "V". This standard picture implies that there is simply an "open" potential universe of sets, intrinsically ordered into levels, without any possibility of grouping them all together into a single, completed totality that includes all sets. The reason for excluding this, remember, was simply to avoid the Russell paradox and the related paradoxes of the Absolute. According to Priest, if we want to see what's really going on here, we may need to just face up to these paradoxes, and recognize that they will arise whenever we talk about certain kinds of totalities. This involves discerning a common structure which, though it leads inevitably to contradictions, is plausibly present in all of these (officially paradoxical and non-paradoxical) cases.

Parameterization and its Failures

If there is indeed reason to think that a common structure of contradiction (or inclosure) is at work both in the "generation" of the infinite hierarchy of sets and in the motivation of the claim that *certain* sets "can't" be formed, then there may also be reason to think that there's something wrong with the latter claim itself. In fact, Priest argues that the standard axiomatic "solutions" to the problem of Russell's paradox (and the set of all sets) are no solutions at all, since in fact they just amount to *ad hoc* denials of a problematic structure which seems definitely to exist (the very structure of inclosure itself). We can

[1] This is analogous to the examples of use of, e.g. "language" (=propositions as a whole) and "the universe" that we considered last week.

see this by noting that these standard solutions not only deny the (plausible) Domain Principle, but also have (as we have already noted) some extremely counterintuitive implications with respect to our ability to refer to, or think of, certain infinite totalities.

To see this, recall the basic shape of the solutions that both Russell and Zermelo-Fraenkel offer to the paradoxes. Russell used the theory of types to block them, whereas ZF set theory includes the axioms of foundation and separation. But the "intuitive" basis of both theories is the same: the idea is that sets are inherently "stratified" into levels, so that each set can be constructed or "formed" only at a definite level and no set can include itself as an element. It's also suggested that we can continue the "formation" process indefinitely, and there is no end to it; thus it isn't possible ("at any level") to group together *all* the sets, or to form the set of all sets not containing themselves. (This is a violation of the Domain Principle, for it tells us that there's a potential infinity of set formation without any corresponding actual, completed infinity of all sets).

Priest's general term for strategies like this for denying the possibility of referring to or forming totalities by indexing them is *parameterization*. The idea of using parameterization to avoid contradiction is simple: for example, if someone told me it is now both 1:00 and 3:00, this would be a contradiction, unless I parameterize and realize that it is 1:00 in Albuquerque and 3:00 in New York. In connection with the theory of types, Russell sometimes suggested that claims about all sets, all propositions, or all functions are "systematically ambiguous." That is, the idea is that a statement that seems to be about "all sets" is really a number of statements (actually an infinite number of statements) indexed to different parameterized levels or types: a statement about sets of type 1, a statement about sets of type 2, etc. However, as Priest notes, such a statement is not actually "ambiguous" in ordinary language, but is well expressed by the use of a universal quantifier. The fact that this – perfectly logically coherent – thought can't be expressed in a system adopting the theory of types doesn't show that the thought is incoherent or that it can only be expressed ambiguously, but just suggests that the theory of types may not be capable of expressing all thoughts. What's more, as Priest points out, there are perfectly coherent statements referring to totalities that can't be seen as "systematically ambiguous" at all. For instance, the statement "Propositions …must be a set having no total" cannot be seen as systematically ambiguous between the sentences: "propositions at stage 1 are a set having no total"; "propositions at stage 2 are a set having no total", etc. For on Russell's theory, there *is* a total set of propositions at *each* of these stages!

In fact, this points as well to an even deeper problem with parameterization and associated strategies. As Priest notes, in terms of the inclosure schema itself, the strategy of parameterization just amounts to denying, in each case, the existence of the totality Ω. But in many cases, it's highly plausible that some totality Ω exists, in the sense that it *can* be defined, even though its definition would violate the VCP. In fact, in many cases we have to refer to Ω itself to argue that it doesn't exist, thereby committing a kind of meta-inclosure contradiction that shows the implausibility of the original denial! For instance, consider the totality of propositions (or propositional functions). If such a totality exists, it must be defined in terms of propositions, and so violates Russell's VCP. Russell's solution was to deny that this totality exists and to parameterize. In particular, he holds that a statement that *apparently* refers to this

totality, such as "all propositions are either true or false" doesn't actually do so, since there is no totality of all propositions to be referred to. Rather, Russell takes statements like this to be "systematically ambiguous": this means that they really assert an *indefinite* (not infinite!) number of particular statements. In this case, the real meaning of "all propositions are either true or false" is supposed to be just that *given* any A that's a proposition, I can say of this A that it's either true or false. For Russell, however, there is no possible proposition actually referring to the totality of all propositions.

As Priest notes, however, this solution is rather disingenuous. If Russell is right, for instance, although we can *apply* the principle of bivalence ("A is either true or false") in each case, we can't say anything at all about *why* we can apply it in each case or even assert that it's true *in general*, even if it is and we know that it is. For to do that we'd have to refer to the totality of all propositions, which we can't do, according to Russell. In fact, things get even worse. For if Russell is right, he can't even say the very things he *does* say in laying out his own solutions. For instance, the very statement of the VCP: "Propositions … must be a set having no total" makes reference to the totality of propositions! And there's no way of replacing this reference with a parameterized or 'systematically ambiguous' version, for what it says is irreducibly *about* the totality of propositions (of which it denies the existence) and not about any individual proposition or any sub-collection thereof. It thus seems evident that any *statement* of the motivation behind parameterization winds up violating parameterization itself. If we think that this issue is a general one applying to the underlying issue of what is (in any way) definable, then this makes parameterization, in general, look like a rather bad option.

Of course, as Russell himself suggested, if we are just dealing with mathematical set theory, and if we think this is a domain that's separate from ordinary language, we might not worry about this too much: we might just lay down the axioms in such a way as to ensure parameterization, while refusing (for the purposes of "doing" set theory at least) to talk about why the axioms exist and are this way at all. Still, this would effectively mean just refusing to discuss why the axioms are the way they are, or where they come from (or perhaps censoring ourselves when we do). And this again would be disingenuous, since the choice of axioms obviously has *some* kind of motivation (in fact, in this case, it's precisely the *ad hoc* motivation of trying to avoid paradox) which we should presumably be able to discuss. In fact, even being able to do set theory at all seems to imply that we must be able to have the kind of knowledge which the VCP and parameterization deny that we can have. For when we do set theory (and decide which axioms to use, etc.) we're effectively reasoning "about" the set theoretical universe, V; we're trying, in other words, to consider what principles hold of this universe *as a whole*, or what is essential to anything that belongs to it. But the VCP and parameterization (as well as the standard, iterative conception of the set) deny, of course, that V exists as a definable whole. Nevertheless, we seem precisely to define it, at least implicitly, whenever we talk about the nature of sets or decide on axioms at all.

Another strategy that mathematicians and philosophical logicians have used to try to contain the paradoxes is to rigorously separate natural language from mathematics. If we could do this, then we might hold that the paradoxes of set theory can indeed be handled 'within mathematics' by means of the axioms, whereas the *statement* of the axioms is essentially done in natural language, which is

probably (so the strategy argues) "paradoxical anyway". An early proponent of this kind of strategy was F.P. Ramsey, who accordingly separated the paradoxes discussed by Russell into two separate groups. The first group, which includes Russell's paradox itself and the Burali-Forti paradox, he called the "mathematical" paradoxes. He separated these, though, from paradoxes like the Liar paradox and Berry's paradox, which he called "semantic" paradoxes. The idea was that while you could solve the first group of paradoxes "within" mathematics simply by laying down axioms that govern the formation of sets (in the manner of Russell or Zermelo-Fraenkel), whereas paradoxes in the second group were just a manifestation of the inherent inconsistencies of natural language, and so don't need to be solved in any rigorous manner. For decades, this division between the 'mathematical' and 'semantic' paradoxes governed the space of discussion about the paradoxes, and led mathematicians to come up with ever better descriptions of the "universe of sets" while denying that there was any deep problem involved in talking about this universe itself.

However, Priest argues by means of the inclosure schema that Ramsey was wrong to see the two supposed families of paradoxes as distinct at all. This is because *both* the supposedly 'mathematical' *and* the 'semantic' paradoxes have the same structure, the structure of inclosure. As we've seen, we can very naturally fit Russell's paradox and the Burali-Forti contradiction into this structure; but paradoxes like Berry's paradox and the Liar paradox itself (not to mention a variety of paradoxes about limits encountered in the history of philosophy) also fit just as naturally into the same logical structure of inclosure. It's plausible that if two kinds of problems have the same structure, they have the same kind of solution; and so Priest argues that Ramsey was in fact wrong to think that we can solve them by different means. Whatever we are "doing" in set theory, it's continuous with what we do in natural language when we talk about totalities and their elements, and we run into the same problems in both cases. That makes it plausible that the solutions, if such there be, are also going to be the same in both cases; this is what Priest calls the "principle of uniform solutions."

There's another reason to think that the 'mathematical' paradoxes can't so easily be separated – and handled separately from – the 'semantic' paradoxes of defining totalities, as Ramsey and others thought. Recall the intuition that originally motivated Cantor himself, as well as many of the early theorists of set theory. The idea was that *set formation* is basically the same as linguistic *predication*: that is, to predicate some property of an individual is essential the same thing as to include that individual in a set (so that if I judge something to be red, e.g., I effectively include it in the set of all and only the red things). If this is right, then there is certainly no point in separating the "mathematical" paradoxes from the "semantic" ones, since there is actually no real possibility of doing so. Of course, Russell's paradox itself might be taken as evidence against this intuition that set membership is the same as linguistic predication, in that it shows that we cannot, without contradiction, use a principle like Frege's basic law V in constructing sets. This *might* be taken to show that we were wrong, after all, to identify set membership with linguistic predication (this is, for instance, how Badiou takes it). But it might, just as well, serve as evidence that linguistic predication is just more complicated than we thought: that we're going to have to deal with contradictions wherever we refer to certain self-membered totalities, which is indeed practically everywhere we use language at all.

Finally, a third device that has been used to block contradictions from set theory is the one suggested by von Neumann in his own formalization of set theory: the idea of "proper classes." According to von Neumann, sets that are "too big" to exist without contradiction as sets (essentially, Cantor's "contradictory" infinities) are not sets, but rather "proper classes." Sets can be elements of classes, but proper classes can't be elements of sets. This is a little bit different from Russell's solution or the solution that's standard in ZFC, since it actually allows the "too-big" entities to exist (after a fashion) rather than simply denying their existence. But as Priest points out, it doesn't solve the underlying problem. For von Neumann gives us no compelling reason why (what he calls) proper classes can't be formed into sets, or even how to define "how big" something would have to be to be a proper class (other than telling us that anything that could be put into one-to-one correspondence with the universe would be "too big" in this sense). Also, von Neumann's theory *does* wind up denying the existence of certain sets anyway; for example, though it holds that the totality of ordinal numbers, ON, does exist as a class, it does not allow us to form (even as a class) the successor, ON U {ON}; so in this sense, it is really little different from Russell's or Zermelo's solutions.

Contradiction and Dialetheism

As we've seen, Priest gives some very strong arguments to show that we can't separate the "mathematical" paradoxes from the "semantic ones," that we're dealing with a unified structure in both cases, and indeed that if this is right, the strategy of parameterization is basically a non-starter for dealing with either (supposed) 'kind' of paradoxes. Indeed, what's most suggestive about Priest's arguments is the way that it lets us see the "set-theoretical" paradoxes, such as Russell's paradox, as exemplary of what happens in general, whenever we refer to self-membered totalities, for instance whenever we try to talk about the limits of thought and language. If this is right, then the inclosure schema (and structure) is indeed very illuminating and suggestive, and seems to be a profound, general structure of thought and language that philosophers have repeatedly run up against in their own attempts to think about totalities.

However, there is one bitter pill we're going to have to swallow, if Priest is right. If he is right, and the inclosure schema applies not only within set theory but to all thinking about the limits of thought and language, then thought about these limits *inherently* involves contradictions, and if we are going to think about these structures generally, we're going to have to think about the structure of contradictions which appear actually to exist. Thus, if we allow the actual possibility of thinking about the problematic totalities, it seems we'll have to admit that there are some *true contradictions*: that is, that there are at least some statements *P* such that *P* & ~*P* is actually true. This position – that there are some true contradictions – is called (following Priest himself) "dialetheism."

Philosophers have historically had a variety of attitudes toward contradictions. Leibniz, for instance, would certainly have denied the very possibility of such a thesis; but for a dialectical philosopher such as Hegel, contradictions are not only profoundly real but are the very engine of historical change and

development. But for *logicians* at least, contradictions have always been anathema: it's been presupposed (throughout the traditional history of logic as well as its twentieth-century versions) that a system that has any contradictions just isn't any good at all, and so that evidence of any kind of contradiction at all in a system is sufficient ground to throw it out. One reason for this is that most systems of logic imply the principle of *ex contradictione quodlibet*: "from a contradiction, anything follows." For instance, in the standard predicate logic of Frege, etc. (which you learned if you took first-semester logic) this principle is true. Given any contradiction, any statement of the form P& ~P, I can use the rules to derive any totally unrelated statement (for instance: "Pigs can fly to the moon.") Given a contradiction, I can prove anything; and a logic that can prove anything is clearly of little use.

Priest, however, made his name by being one of the first logicians to dispute this principle. He did so in an earlier book called *In Contradiction* by working out a *dialetheic* logic (actually, a set of such logics): that is, a logic that tolerates contradictions. In a dialetheic logic, we can indeed assert certain contradictions, and it's not the case that just anything can follow from them. To make sure of this, we have to modify the rules a bit (one way to do this, for instance, is to redefine the "conditional" so that A -> B isn't just the same as ~B v A, as it is in the "standard" logic). But we'll gain a logic where we can indeed assert certain contradictions, and draw out what are plausibly consequences of them without these contradictions "exploding" to imply just anything. If we can indeed form such dialetheic logics and think about them, there's no evident reason why we might not use them to think about the kinds of contradictions that Priest is concerned with in *Beyond the Limits of Thought* and that seem clearly to exist if the inclosure schema is as general as he suggests. Most prominently, these are contradictions of the form "X is thinkable and X is not thinkable," or "X is sayable and X is not sayable," where X is a thought or proposition involving the totality (of the thinkable or sayable) or its limits. If dialetheism is right, then these contradictions won't "explode" to imply just anything, but they will characterize all thinking of a certain type, the very type that philosophers, in thinking about the totality of whatever can be thought, said, or the totality of what exists, repeatedly seem to encounter.

Contradiction and Criticism

As Priest suggests, "totalizing is part of our conceptual machinery – like it or not." (p. 162). In fact, as Priest suggests elsewhere, the situation where we're involved in the thought of a totality, of which we ourselves (or that thought itself) is a member, is very general, and not at all limited to set theory or mathematics. In fact, Priest argues that the general situation of inclosure occurs just about whenever philosophers think about the limits of thought, language, definability, or expression. Let's see if we can put this all into a bigger picture that tells us something about philosophical thought about limits and boundaries, and what this kind of thought is doing and is capable of. Recall that, if Priest is right, and there is a common structure to the contradictions we encounter in set theory and the contradictions we encounter more generally in thinking about totalities, there are just two ways to deal with these contradictions. We can either acknowledge their existence and think dialetheically, or we can try to deny them by parameterizing. If we parameterize, we essentially hold that it's *only* possible to think or talk about the totality in question from a position *outside* of it. This is the essential idea behind Russell's vicious circle principle, as well as the prohibition of set self-membership that we see in Russell's theory

of types as well as the axioms of ZFC. In each case, we simply lay down rules to establish that we can't talk about or refer to a totality unless we can stand outside it, say at the next (higher) level of the type hierarchy, or (thinking constructively) by having the totality "on the table" in order to put it together with other sets. But if this principle is (somewhat) plausible (on some views of set formation) for sets, it's strikingly implausible when we get to many of the totalities that we (seem to) define in natural language. Here, it seems, we constantly define totalities that we ourselves stand within in defining them; and so parameterization just isn't a plausible option. To see this, and to see some of the implications of Priest's dialetheism and the inclosure structure itself, let's consider a few of these totalities:

1) **The English language.** As we've seen, if it's possible to refer to the totality of propositions in a language (or even to define the language itself), we're going to have to do so by means of members of that totality (propositions or words in English). If parameterization were right, then we'd only be able to do so by means of some statement or set of words that isn't part of the totality thereby referred to or defined. For instance, we'd have to use some *other* language (such as Chinese) to refer to English. It would follow that it's impossible for any monolingual English speaker to define, or even think about, the English language as a whole; in order to do so, we'd have to call in someone who speaks another language that's outside English itself. But clearly, we don't have to do this: it's clearly possible to use English to refer to itself.

2) **The totality of truths (in a language).** Thinking about the Liar paradox and related paradoxes which we'll consider in more detail later, Tarski gave a famous argument in the 1930s that no language can include a definition of what it is for a statement in that very language to be true. The reason is that if there is such a definition of truth, it's easily possible to construct a contradiction similar to the liar paradox, a statement that is true if it's false, or false if it's true. This led Tarski and others to hold that if we want to define truth in a language without contradiction we have to, essentially, parameterize: we have to "jump out" of the language and define it in a "meta-language" that is distinct from the language under consideration. Even if this is right, if we want to now consider truth in the "metalanguage," we're going to have to jump up yet another level, and define it in a "meta-meta-language." So simply in trying to talk about truth at all, we're involved in a whole infinite hierarchy of languages, without end. Of course, this is highly implausible with respect to English itself, since we do readily talk about truth-in-English in English, and we don't generally have available even a single meta-language to do so. The only plausible way to make sense of this is not to parameterize, but to hold that there are inherent limit-contradictions involved in thinking about truth itself.

3) **The passage of laws.** These general issues apply not only to questions of language and truth, but to more concrete issues as well, for instance in the theory of laws and politics. Suppose there is a country (like most developed countries today) in which laws are passed by some assembly such as a parliament or senate. This body, the parliament or senate, must decide ultimately what's legal and what's not. But in order to do so, this body *itself* must be legitimate; that is, it must be *legally* designated (or elected, or whatever) as the actually legitimate assembly. (Otherwise just anyone could get together and say that what they want is the law). Thus, the totality of the legal must be defined in terms of a member of

itself, and it seems we will inevitably face inclosure contradictions. The only way to avoid this problem is to parameterize, and to declare that the legal status of the ultimate maker of law depends upon a "higher" legal judgment. But then this legal judgment would have to depend on another legal judgment, and so forth; we might imagine an ultimate judge who is himself judged by a higher ultimate judge, and so forth, but where would it end? Clearly, here too, parameterization is implausible.

4) Sovereignty. There's a closely related issue involving the power capable of instituting a state or government (what's sometimes called the sovereign) itself. Clearly, in order to institute a government legitimately, a body or person must itself be legitimate; but it will generally only be legitimate in terms of the government that it, itself, institutes. (Thus, for instance, the constitutional convention that instituted the US was legitimate – in terms of that very constitution – but not in terms of, say, the British constitution). Thus we again have a structure of total self-inclusion, whereby the totality is defined in terms of a single member, and the possibility of inclosure contradictions arises. To insist on parameterization here would be to hold that a government can only be legitimately instituted by means of another government. Although this *sometimes* happens (e.g. in the case of "post-colonial" states that were formed by the decision of the previous, colonial occupiers), it clearly can't happen all the time, and in fact doesn't happen in most cases.

If Priest is right and the structure of total self-inclusion in these cases is indeed something like the inclosure structure, then it seems as if we'll have to deal with these paradoxes as the contradictions they are, without being able to rely on parameterization as a solution. This means that in thinking about the general structure of language or truth, or laws or state power, we'll have to think through these contradictions on their own terms, and consider their practical reality and actual effects, and that we can't necessarily rely on the possibility of "jumping out of the system" in the manner of parameterization to solve them.

In fact, there's a more general and perhaps deeper moral here for the nature of critical thought itself. For a long time, as Priest emphasizes, philosophers have thought about the limits of what can be known, thought or said. A historical high point of this thought is Kant, whose critical system attempts to draw a line around all that can be known by means of illuminating the basic principles and categories of knowledge and thereby delimit, or "criticize", the claims of thought to exceed this totality. Later on, this idea of critique as delimitation was applied to a huge range of other cases, running the gamut from problems of knowledge to problems of political and social action, thought, and language. But as philosophers also sometimes noted, the project of criticism as delimitation in the Kantian form leads to directly to the famous problem of "knowledge of the transcendent" (or, as it's sometimes called, the problem of "meta-critique"): that is, how is it possible to *know* anything about the boundaries of knowledge themselves? Kant sometimes seems to attempt to solve this problem by suggesting that we somehow (perhaps as "transcendental" rather than "empirical" subjects) can stand outside the boundaries of knowledge and survey them from without. This is essentially a version of parameterization, for it agrees on attempting to draw the boundaries from outside. But if, as we've suggested, this is an implausible and problematic solution, then it seems that the whole philosophy of criticism, and the whole attempt to take a critical stand on the knowable, thinkable, or sayable, must

perhaps be done from *within*, and that, as such, it will lead inevitably to inclosure contradictions. This suggests that, following the results of Russell, Cantor, and Priest, all plausible critical thought about these totalities must indeed work with the schema of inclosure, and with the kinds of inherently necessary limit-contradictions it suggests.

Philosophy 415: Fall 2025
History and Philosophy of Mathematics

Notes: Week 9

Now that we have learned the basic elements of transfinite set theory and understood the way it can be used to ground all the concepts of mathematics, we're in a position to consider how the structure of set theory – and in particular its appeal to the infinite – changed philosophical thinking about the nature of mathematics itself. In the early decades of the twentieth century, after Cantor's transformative discoveries, there emerged an interesting debate among different positions in the philosophy of mathematics. The positions that were articulated then are still with us today, though further developments, especially Gödel's incompleteness theorem, have complicated the picture somewhat, as we'll see. Still, along with some other views that we'll explore later (such as Platonism, naturalism, and fictionalism), the three positions that came to the fore in the "foundations debate" of the 1920s are still very important to philosophers' thinking about the traditional questions of the philosophy of mathematics. The biggest one of these questions is probably the question of what mathematics actually is: is it a process of creation, or one of discovery, or something else? If it is a process of discovery, what are we discovering? How do numbers, sets, functions and other "mathematical objects" actually exist (if they do)? And if they do not exist in space or time, how do we think about them and come to know things about them? All of these questions get trickier, in an obvious set of ways, when we think about the infinite and about the kinds of infinite sets that Cantor's mathematics suggests. Even if we could solve the problem about how we construct or discover relatively small *finite* numbers, for instance, it's far from clear that any solution that portrays us as doing this, over time, could also explain how we grasp or otherwise understand an *infinite* structure such as **ω** or (perhaps even worse in some ways) **R**. Yet Cantor's domain of transfinite sets – what Hilbert called Cantor's "paradise" – gives us the challenge of understanding how this kind of access might be possible, and how we might understand it concretely as supporting all of (the rest of) mathematics.

Pre-history of the "foundations" debate

Though the ancient Greeks knew about various domains of mathematics, they tended to favor geometry as the best example of a mathematical theory. Euclid worked out a system of five axioms (or "postulates") for geometry and gave many proofs of particular theorems using these axioms. However, Euclidean proofs typically look a bit different than modern ones. In particular, in Euclidean proofs there is typically a step of "construction". For example, if one is proving a theorem about isosceles triangles (i.e. triangles with two equal sides), one must actually "construct" or draw such a triangle. There are also further constraints upon the figure that is actually drawn or constructed. For example, it has to be constructible using a straight edge and a compass; one can't appeal to units of measure or introduce shapes that can't be "made" using just these tools. For the Greeks, this step of construction is essential in that it establishes that our geometric proofs are grounded in the real character of something that actually exists. Since we can actually construct the figures we're discussing, we can know that our geometric proofs really capture the character of actual space.

In the eighteenth and nineteenth centuries, new developments in mathematics raised questions about methods of proof, the "reality" of mathematical objects, and the whole idea of appealing to an outside "reality" in developing mathematical theories. The first of these developments was the discovery of calculus by Newton and Leibniz. Calculus was, from the beginning, seen as an extremely useful theory of actual physical phenomena, especially motion and change. However, as you know if you've taken a calculus course, one absolutely central concept of calculus is that of the differential, usually written dy/dx. This looks like a fraction or ratio of numbers, but it really isn't, since the "quantities" dy and dx are not numbers in an ordinary sense. Rather, they are sometimes called "infinitesimals," and Newton and Leibniz themselves thought of them as being quantities that are infinitely small, but still greater than 0. This raises, though, the question of what quantities like this could actually be, and many philosophers and mathematicians accordingly felt that there were deep problems in the "foundations" of calculus, despite the fact that the theory worked so well. Later, through the efforts of Weierstrass and others, the differential was given a different kind of "definition" in terms of the concept of a *limit*, which does not involve any idea of infinitely small but nonzero quantities (instead, the idea is that one approaches as close as one likes to a limit point, without ever reaching it). However, there was still the question of what limit points – and indeed, any point on the real line – actually *are*. It took Cantor's identification of points with sets, and Cauchy and Dedekind's definitions of real numbers in terms of sets of rationals, to clear this up in a set-theoretically motivated way. But this involves not only actually infinite sets, but even infinitely many sets of infinite sets, raising all the problems of the relationship between transfinite numbers that we've seen, including (decisively) the continuum problem itself.

The other development of nineteenth-century mathematics that posed a deep question to the traditional idea of foundations was the development of non-Euclidean geometry. Euclid's fifth postulate, the so-called "parallel postulate," holds that two parallel lines never intersect. However, the fifth postulate always seemed "different" than the others – it does not seem as obviously definitional or logical – and it also could not be derived from the other four. In the early nineteenth century, several mathematicians discovered independently that it is possible to create a system of geometry that does not adopt this last postulate, but is nevertheless entirely consistent. Both *spherical* and *hyperbolic* geometry are non-Euclidean in that they do not adopt the fifth axiom but replace it with different assumptions. Later on, in the early twentieth century, it was discovered by Einstein that the geometry best suited for general relativity is non-Euclidean. So it appears that the actual geometry of large-scale space is non-Euclidean. All of this raises deep questions about the possibility of appealing to our natural assumptions or to the kinds of constructions we can make on (two-dimensional) paper to support claims about the actual nature of geometry and the structure of geometrical space.

Although he was actually one of the first to suggest[1] that Euclidean geometry might not be absolute – i.e. that it might not be absolutely necessary for all creatures – Kant thought that both geometry and arithmetic were grounded in the *form* or *structure* of our *intuitions*, or of how we experience the world. Here, "intuition" is just a fancy term for whatever we can know through perception, or whatever we can know directly and without thinking about it. According to Kant, space and time do not exist absolutely; rather, these are simply the forms of our intuition of objects in the world, the necessary structures of

[1] In his "inaugural dissertation".

how we must *know* them as humans. For Kant, moreover, all mathematical judgments are *synthetic a priori* – that is, though they are knowable in advance of particular experiences, they are not simply reducible to concepts or conceptual relationships. This is because, according to Kant, both arithmetical (e.g. '7+5=12') and geometric judgments ('The shortest distance between two points is a straight line') depend upon the structure of our intuition, i.e., the forms of space and time. In particular, arithmetic judgments are grounded ultimately in an intuition of *time* – since they are grounded in the possibility of *counting*, which for Kant essentially takes place over time – and geometric judgments are grounded in the structure of (our intuition of) space. In the *Critique of Pure Reason*, Kant further suggested that the geometry of physical space – at least as *we* can intuit and understand it – is necessarily Euclidean. The development of non-Euclidean geometry, as well as new ways of thinking about the possible analyticity of mathematical concepts and claims, made this look less plausible over the course of the nineteenth and early twentieth centuries, setting the stage for the new views of the basis of mathematics in the wake of Frege and Cantor.

Logicism/Platonism

In *The Foundations of Arithmetic*, Frege famously suggested that arithmetic (and perhaps other areas of mathematics) might be *reduced to pure logic*. Given the new quantificational logic that he had discovered, this, in fact, looked like it might for the first time be possible, and if it were possible it would provide a radical alternative to the Kantian picture of numbers and arithmetical judgments. In particular, the idea is this: if we can actually *define* numbers and all the basic operations on numbers (e.g. addition, multiplication, etc.) using *only* logic or logical vocabulary, and also use the logical laws we know of to govern legitimate inference, then we can reduce all actual proofs to purely logical structures. In this way we could provide mathematics with an absolute and rigorous method for knowing the truth and distinguishing it from falsehood. Mathematical truths would turn out to be analytic rather than synthetic – that is, they would turn out to follow completely logically from the structure of the basic axioms. And we would never actually have to appeal to intuition, or anything else about our actual structure of thought as humans; so in this way mathematics could be seen as rigorously *objective*, rather than as based in subjective structures of experience (as it is still, in a way, in Kant).

Frege's thinking about the basis and truth of mathematics is thus logicist, but his intuitions about *what* logic is are also strongly Platonist. "Platonism" has different definitions, and it's not clear that Plato himself would have subscribed to much or many of the modern interpretations of that term. But at its basis, the attitude of Platonism in mathematics is the attitude that mathematical objects *do* exist although they do not exist in space or in time (and are not just psychological or mental existents). Frege tended to think of "the" laws of logic as given absolutely and objectively, and sometimes even talks about a "third realm" beyond the physical/spatial and the mental/subjective where these laws exist or reside (along with all objective contents of thought, including all mathematical truths). Although it's (somewhat) obscure how we are supposed to have "access" to this realm, Frege thought that the objectivity of these truths required something like this, and that the logicist reduction could make it clear how the truths of mathematics were also grounded in the basic logical laws themselves.

Peano's axioms for arithmetic, which we considered a couple of weeks ago, encouraged Frege (and, slightly later, Russell as well) in thinking that a uniform reduction of mathematics to (something like) logic might actually be possible. The five Peano axioms, after all, allow us to define both numbers and all the basic arithmetical operations in general, and they do not in any way seem to depend on intuition or human psychology. All that they add to logic itself is one privileged element (the "zero"), one relation (of "successor"), the term "natural number," and some claims about the structure of this element, relation, and what is covered by this term. None of this seems to go very far (if at all) beyond basic logic itself, and it gives us a workable basis for all actual claims of arithmetic and number theory.

Frege's work was almost completely obscure until it was discovered and publicized around the turn of the century by Russell. At almost the same time, as we've discussed, Russell discovered his paradox, and this posed a deep problem for the assumptions that Frege had made (along with Cantor) about the coherence of supposing that each concept picks out, all by itself, a well-defined set. Frege himself was deeply troubled by this, and sometimes Russell's paradox is cited as putting an end to logicism, at least in its classical form. But Russell, as we've seen, thought he had a solution that would allow him to preserve the underlying idea of logicism, that mathematics, through set theory, could be reduced to what is essentially logic. This solution was the theory of types, which (as we've seen) imposes a strict hierarchy of "levels" on the sets that can be thought of as really existing or formed. Russell emphasized that, if we adopt set theory to define numbers, there is no sharp line between "mathematics" and "logic"; all we're doing in giving a foundation for all of mathematics is adding a couple of new symbols to logic (essentially, the sign $\in$ and the symbols for sets, { and }), and giving some rules that govern their use.[2] Along with the theory of types, Russell modified Frege's definition of number slightly, so that a number is actually (according to Russell) the set or class of *all* equinumerous sets of a particular kind. For instance, for Russell the number 2 is literally the set of all *pairs* of things.

The greatest and most dramatic outcome of Russell's logicism was the massive 3-volume *Principia Mathematica*, co-authored with Alfred North Whitehead and published in the first edition between 1910 and 1913 (the second edition was published in 1927). PM attempted to provide a complete logically based axiomatization, using set theory and Russell's theory of types, for all of arithmetic, and actually showed how to construct numbers and operations. (Notoriously, PM took about 200 pages to prove that 1+1=2). In *PM,* Russell and Whitehead employed an extended ("ramified") version of Russell's theory of types to make sure that set-theoretical paradoxes don't arise: but in relation to the problem of formulating claims about the totality of propositions or of functions (which seem to be needed even in formulating axioms such as the Peano axioms), Russell proposed a special "axiom of reducibility" which held that *every* propositional functions is equivalent to a *predicative* one (i.e. one which obeys the type-restrictions and thus avoids impredicativity). There was a lot of subsequent debate about the justification – if any – for the axiom of reducibility, and even Russell eventually admitted that there was no real (logical) reason it *has* to hold.

[2] At some points in his career, in fact, Russell even held that sets don't exist "strictly speaking" and are something like "logical fictions;" he thus tried to show how claims involving sets could be reduced to claims that are just about individuals.

The massive project of PM can also be seen as defining a *formal system* in the sense of Hilbert (see below), though it actually didn't provide a fully defined *syntax* for all of its primitive notions. Nevertheless, it was influential not only for the project of logicism, but equally for that of formalism, and Godel's incompleteness theorem was itself first formulated as a demonstration of the essential incompleteness of PM ("and related systems").

Intuitionism

Twentieth-century intuitionism has important antecedents in Kant and even ancient Greek mathematics, but it was first proposed in its modern form, partly in response to Cantor's theory of the transfinite, by the Dutch mathematician L. E. J. Brouwer. For the intuitionist, mathematical truth is not reducible to logic or to formalism; rather, it draws and depends *essentially* on the structure of human experience or intuition. Even if the Kantian appeal to *spatial* intuition must be rejected (in light of the possibility of non-Euclidean geometry and other issues), Brouwer insisted on the need for a basic intuition of *time* in providing the basis of all knowledge of number and arithmetic judgments.[3] This temporal dependence makes a crucial difference with respect to the infinite. In particular, for the intuitionist the infinite cannot generally mean the "actually completed infinite" (as it does for Cantor, e.g.). Rather, it means simply the "unlimited" in the sense of a process that *can* always be carried forward in time. Thus, it is reasonable to call the natural numbers "infinite" in the sense that it is always possible to count one more. But it is not reasonable to treat the totality of them as an *actually existing* set, or to make claims that are construed as actually holding for *all of them*, whether or not we have actually counted that high. This conception hearkens back to Aristotle's view of the infinite as always only potential, and of a completed infinite as actually impossible. However, for many types of intuitionists (those who are not "strict finitists"), there are certain kinds of infinities that are admissible. For instance, we can talk of the existence of **ω**, *given that we have a well-defined rule for "constructing" or producing each number from its predecessor*. But what we are really talking about is just this rule and its endless *possibility* of continuation; we are not talking about any actually existing set or totality with more than finitely many members. In general, for this kind of view, we can talk about an infinity when – and only when! – we have a particular well-defined rule which defines its unlimited possibility of continuation. This possibility is itself thought of as the possible continuation of a series indefinitely far, but not as a totality that ever exists "all together".

This has important consequences not only for such questions as those of the nature of the infinite and the nature of real numbers, but even for logic itself. One of the suggestions that Brouwer made early on was that, because of the necessary basis of mathematics in finite processes of construction or generation, it may be necessary to abandon the traditional logical "law of the excluded middle" (LEM) (or '*tertium non datur*') The LEM says that, for every judgment (or sentence) P, (P v ~P) holds. A closely related claim – the claim of *bivalence* – says that each claim P is either true or false, with no third or "in-between" value possible. To see why we might want to deny this, consider the development of what

[3] For Brouwer, in particular, the idea here is that our intuition of time gives us access to a basic "two-ity" or "two-oneness" that is instantiated in the experience of "before and after", and that this is the original basis for the "splitting" of 1 into 2, which then makes possible the progression to ever-higher numbers.

Brouwer calls a *free choice sequence*. For instance, suppose I have a ten-sided die; I roll the die repeatedly and write down the series of numbers. I might even consider this to be the "construction" of a real number between 0 and 1, taking each of the numbers as a digit in an indefinitely continued decimal expression of such a number. Now suppose I roll four times and obtain the sequence 5, 8, 3, 2 (or, equivalently, .5832…) I plan to keep going indefinitely, but I have only developed the first four elements so far. Now I ask myself, is there a 7 anywhere in the series? Here I am not just asking about the part of the series that has actually been developed, but about whether there is a 7 *anywhere* in the series as it *will be* w*hen it is finished*. In an obvious sense, this question has no answer. I can't say yet that it's either true or false that 7 appears in the complete series (this is different from a question such as "does a 5 appear?" which I *can* already answer affirmatively). Since we can't say that the claim that there is a 7 is determinately true *or* false, we apparently have to say that it lacks a truth-value, or that it is *indeterminate*.

Now consider a mathematical question, such as the question: is there a string of 7 occurrences in a row of the digit '7' anywhere in the decimal expansion of pi? For the non-intuitionist, this is presumably a question that *already* has a "yes" or "no" answer, even before we have discovered the answer or have any way of doing so. Pi is, after all, a perfectly well-defined and particular number; there is even a rule by which we can determine successively as many of its digits as we like. Of course, if there is somewhere in fact a string of seven 7's, we may find it one day through successively working out the digits. However, if there is not such a string, we'll NEVER find it this way; and even if there is, it may be farther "out there" in the decimal expansion of pi than we ever actually calculate out. Accordingly, for the intuitionist, pi is (in a very real sense) like the free choice sequence. We can't and shouldn't say that the claim "there is a string of seven 7's" is *either* determinately true or false. If, one day, such a string were found, then we could say the claim is true. But for now, we cannot legitimately give it either truth value.

Something similar holds, according to the intuitionist, with respect to proofs. In certain cases, we are entitled to make claims about "all numbers" or "all numbers of a certain sort." But in order to do so, we must have a proof that demonstrates the general claim in a finite number of steps, for instance a proof by mathematical induction that starts with the base case (n=0 or n=1) and shows that if a property holds of any particular n, it holds for n+1. Once a statement is proven in this kind of way, it can be held true for all numbers or all cases. By contrast, though, if we consider a statement like Goldbach's conjecture:

Every even number is the sum of two primes.

which has not yet been proven or refuted, we must reject the claim that the conjecture is either true or false. Relatedly, intuitionists and constructivists sometimes distinguish between "constructive" and "non-constructive" proofs. A constructive proof is one that actually exhibits or shows how to build an entity that is shown to exist, or shows *how* effectively to do something about which a general claim is made. By contrast with this, a proof by contradiction (for instance) is not constructive, and often

actually depends on the validity of the law of excluded middle in order to go through.[4] Because of this, intuitionists are often suspicious of supposed proofs by contradiction. Following Brouwer, Heyting developed a general framework for intuitionist logic that embodies both suspicions. Essentially, an intuitionist logic is derived from a classical one by: i) dropping the LEM (P v ~P for all propositions P) and ii) dropping the law of double negation elimination, (which classically tells us that we can move from ~~P to P for any P). Because of the second modification, intuitionists typically credit *reductio* proofs (i.e., so-called "non-constructive" proofs) only if they establish something *negative*: i.e., if I can derive a contradiction from A, I can reason that ~A is true (but interpreted intuitionistically, this just says that there cannot be a (constructive) proof of A); but I cannot reason from a reductio premise ~A to the claim A (since doing so would involve double negation elimination).

Because of this suspicion about nonconstructive proofs, as well as the general doubts about the completed infinite, the intuitionist is not prepared to accept Cantor's arguments for a hierarchy of transfinite sets. In particular, as Brouwer argues in "Intuitionism and Formalism," even if intuitionists can speak of the set ω (due to the obvious coherence of the law +1 which allows us to generate this series indefinitely), they are not prepared to accept Cantor's "diagonalization" argument in any general form. The intuitionist can say that, given any list connecting a *finite* number of natural numbers one-to-one to real numbers, it is always possible by means of diagonalization to come up with one more real number that is not on the existing (finite) list. But this clearly has no significance for the general question of "cardinality". Additionally, he can agree that it is not possible – so far as we know – to construct any law or rule that would take natural numbers one-to-one to real numbers. However, this is far from saying that the "set" of real numbers is "larger" than the set of natural numbers, or even that such a thing as the "set" of all real numbers even makes sense. In fact, since there is no effective law that generates *all* the real numbers in an ordered way, the intuitionist is prepared to doubt that such a set as that of all the real numbers even exists. For the same reason, instead of construing real numbers as infinite sets of rationals (in the manner of Dedekind), the intuitionist defines them as what Brouwer calls "spreads." In general, a "spread" is simply a rule for successively determining digits; if there is such a rule, the real number is considered legitimate, but if there is not, it is not.

More recently, Michael Dummett has adopted the intuitionist's characteristic denial of bivalence and applied it to other domains (besides mathematics) to characterize the underlying structure of "anti-realist" thought. For Dummett, anti-realism is the attitude that refuses to grant a truth value to claims which it is beyond our powers to know; thus, the anti-realist, by contrast with the realist does not think that the truth or falsity of any arbitrary claim is fixed, whether or not we can come to know it. Accordingly, the anti-realist does not consider claims that are beyond our capacity of verification to be (determinately) either true or false. This difference in attitudes, Dummett has argued, characterizes a large number of disputes in various domains between those who have the attitude that the meaning of claims is somehow given independently of our epistemological capacities (the realists) and those who

[4] For instance, if one assumes ~A in order to derive a contradiction (with the goal of showing A), and actually does derive a contradiction, for the intuitionist this only allows us to conclude that ~~A, which is *not* intuitionistically equivalent to A itself.

deny this, thinking that we can understand claims only insofar as we have the capacity to verify them, and thus that the idea of a truth completely independent of possible verification is nonsense.

Formalism

The last position that came out of the foundations debate in the 1920s – and, as we'll see, the one that is in many ways most influential for twentieth century thought and technology more broadly – is David Hilbert's project of *formalism*. Hilbert was a great mathematician who achieved many important results, including a formal axiomatization of all types of geometry (Euclidean and non-). In 1900, Hilbert announced a famous list of 23 problems, which in many ways set the agenda for twentieth century mathematics (the first of these was the problem of the truth of the continuum hypothesis). The key idea of the formalist program he announced, though, was the idea of a *formal system*.

Hilbert's innovation was to conceive of a radically new idea for thinking about mathematical reasoning. The idea is that it ought to be possible to reduce mathematical reasoning to formal reasoning involving *only* symbols and the rules for their combination and transformation, *without ever mentioning or having recourse to their "meaning*". More specifically, to specify a formal system, all that we have to do is specify a vocabulary of basic symbols, give rules for forming them into well-defined "sentences", and then give axioms and purely *mechanical* rules for deductive inference in the system. This idea, for obvious reasons, came to be called "formalism" and inspired what was called the "formalist program" (which is again just the program we talked about earlier). The formalist program has been *massively* influential in twentieth-century thought. As we'll see in more detail, the basic idea of formalism – that we can treat proof and calculation procedures simply as the manipulation of pure symbols that are in themselves meaningless – is the basis for the very idea and structure of the (electronic) computer, which Alan Turing, following out the consequences of Hilbert's program, would later formalize and develop. So it's possible to say that without Hilbert, we wouldn't today have electronic computers at all; and then, of course, we wouldn't then have any of the informational and communicational technologies (internet, cell phones, etc. etc.) that depend on them, and the world would be a very, very different place.

To get a feel for what a formal system is, recall the language and rules you learned in first-semester symbolic logic. There was a primitive vocabulary of a fixed number of symbols, or symbol types (such as the symbols for the truth-functions, the quantifiers, variables, and individuals). There were rules for forming from these symbols legitimate (or "well-formed") *formulas* or sentences, sentences that "say something" in the language of symbolic logic. Finally, there were rules for *deriving* one sentence from another (or from a set of others). The rules for forming sentences and for deriving sentences from one another were relatively simple, and they were also purely *syntactical*. This means that the rules themselves never concern the *meaning* of the various symbols (in the sense of what objects they stand for or represent). Instead, they just tell us how to intercombine symbols themselves, and how we can move from one symbol string to another. Because they are purely syntactic in this sense, this means that they are also purely *mechanical*. A computer can follow these rules as easily as people do (and in fact more quickly and accurately), and it's easy to design a computer program that does just this, taking

arbitrary premises written in the language of first-order logic and "grinding out" consequences in accordance with the rules.

Actually, we ourselves very often follow this kind of mechanical procedure in mathematical reasoning. For instance, consider the procedure for long multiplication that you learned in elementary school. There are just a few rules to learn, and given these simple rules, the procedure gives us an answer to the problem of multiplying any two integers, of whatever length. Also, the rules involve only symbols: what symbols we write where, and when to write a new symbol (and which one to write). This is thus essentially an (elementary) example of a mechanical "proof" procedure. Later on, when we have a formal system that can express all of arithmetic in logical terms, we can put this very same procedure (if we want to) in the logical language and use purely mechanical rules to get the result (these are in fact exactly the same rules that your pocket calculator follows to get the result, as well). The calculation of the product (say that 34 X 22 = 748) is then just the same as a purely mechanical and syntactic *proof* of the result (i.e. that 34 X 22 = 748).

Although this position thus draws on the possibility of capturing logical rules formally, Hilbert's motivation is a bit different from that of logicism. Unlike the logicist, who thinks that the truths of mathematics rest on substantive *logical truths* that are facts about the world or about a special domain of "logical reality", Hilbert's project emphasizes that the rules for a formal system are *completely arbitrary*. We can come up with any symbolism, any set of rules, that we like, and we can use them as long as we can be sure that they will not lead to a contradiction. On the other hand, Hilbert shared with the intuitionist the concern that there is something problematic about the infinite. Like many mathematicians, Hilbert was deeply impressed with the revolutionary new possibility of treating infinity in the way that Cantor made possible, and with the promise of this treatment for giving a new foundation to mathematics. He was so impressed, in fact, that he saw Cantor's hierarchy of transfinite cardinals as a kind of "paradise" and vowed that mathematicians should never be "expelled" from this paradisiacal realm. Nevertheless, like many others Hilbert has some "finitist" intuitions as well: he thought that there was something "fishy" about reasoning about infinite wholes, and indeed that the paradoxes already discovered confirmed this. After all, he reasoned, whatever we do with infinities and infinite sets, we certainly can't do anything that would take an infinite amount of time, for instance actually go through a procedure that requires completing an infinite number of distinct steps.

How, then, *is* it possible for us (finite beings that we are) to reason about infinite sets and totalities? According to Hilbert, it is indeed possible for us to reason about infinite totalities *if, and only if, we can do so by finite means*. For instance, suppose we want to reach a result which "quantifies over" or requires reference to an infinite domain: take, for example, the result that for every prime number there is some greater one. Even stating this result (which was proved already in ancient Greek times by Euclid) requires reference to an infinite domain: the infinite domain of numbers greater than any particular, given one. How, then, are we able to establish this result, which holds of an infinite number of numbers, and so can't possibly be treated as the "logical sum" (or disjunction) of each of them (i.e. "either 8 is a prime, or 9 is a prime, or 10 is a prime, … etc)? We do so by coming up with a *proof* in symbols. The proof uses particular symbols – variable letters – to *stand for* any arbitrary number within the mathematical domain, but importantly we *don't* have to consider each of these numbers

individually. Rather, we can reason using variables, and establish the requisite conclusion as holding for the *whole* infinite domain.

Hilbert's key insight is that when we do so, we give a *proof* that is itself – considered simply as a series of symbols – *finite* in length. Whatever the proof establishes, the proof itself (considered, again, *purely* as a series of symbols) is itself of finite length (it had better be, if we're ever going to give it!) and draws on a language with a fixed, finite number of symbols (or types of symbols). This is the kind of language that human beings can be expected to learn and master, and there's no obvious problem with supposing that we can also learn and master the kind of mechanical, syntactic rules that (Hilbert hoped) could define all such possibilities of proof. Thus, if we can make it clear how proof procedures can be reduced to such a language and such systems of simple, mechanical rules, we can show how such systems allow us to think about multiple infinities and whatever is true in mathematics, even though we remain finite beings ourselves.

In a certain way, this position is like intuitionism, in that it requires that a proof bearing on infinite structures must be given in finitely many steps, and that it be accessible to finite cognition through the "arbitrary" symbols and rules themselves. However, Hilbert disagrees with the intuitionist's idea that mathematics must be grounded in an experience of time, or in any "intuitive" provision of the object. In many cases in the history of mathematics, Hilbert reasons, the extension of mathematical reasoning to new domains has not depended on there being an actual "intuitive" demonstration of the existence of an object. For instance, Hilbert gives the example of the discovery of imaginary numbers (square roots of negatives). Even if we don't have any particular intuition of what the square root of -1 is or is like, we can still use the mathematics of imaginary numbers provided we can invent a symbol for it (*i*) and then go on to use this symbol in our system, without ever producing a contradiction. The fact that we can do so, alone, justifies the claim that *i* "exists" – and this is the only justification we need. Something similar, Hilbert reasoned, should go for our formal systems for dealing with the infinite. Provided we can symbolize infinite quantities and reason about them using this symbolism without contradiction, all our reasoning and proofs should be legitimate, and does not need to be limited or restricted in the way the intuitionist thinks.

Of course, this will only be helpful (and possible) *if* we can indeed come up with such formal systems: that is, if we can reduce truth in mathematics to provability in such a system. After all, wherever our systems take us, we can only be confident in their results if we can be confident that these systems capture *actual mathematical truth*: that is, that they are sound and complete. In 1925, Hilbert was aware of this, and he was also aware that there was as yet no successful proof that a formal system capable of capturing mathematics could be both sound and complete. Accordingly, he posed the first part of this question as an open and important problem. In particular, Hilbert was aware that we can be confident in our formal systems only insofar as we can be confident that they are *consistent*. If a system turns out to be inconsistent – that is, it is capable of leading to an inconsistency when we follow the rules – then (at least using the logic that Frege and Hilbert employed) it will prove anything at all, since it's possible (using classical logic) to derive anything at all from a contradiction. For instance, an inconsistent system will prove such absurdities as 1=0; and a system like this will hardly be useful for mathematics. Moreover, a system capable of proving such an inconsistency will also fail to be sound:

since there is some P for which it proves both P and ~P, and only one of these can be true, it will prove some falsehoods.[5] So we certainly want to be in a position to show that our systems are consistent.

In fact, what we really hope is that a formal system (capable of capturing mathematics) can also prove *its own* consistency. Failing this, we might look for a proof of the consistency of a system by means of *another* system; but then the problem would be that we'd have to find a consistency proof for this system, and we'd have to look for yet another system to do this in, and so forth without end. So Hilbert posed as the second on a famous list of unsolved mathematical problems the question: i) *is it possible for a formal system capable of capturing the truths of mathematics to prove its own consistency*? For reasons we'll go into later, this turns out to be the equivalent to the following two questions: ii) is it possible for a system capturing arithmetic to be both *sound and complete*? And: iii) is it possible to formulate a well-defined mathematical question which *no formal system* can answer? We'll turn to these questions in the second half of the class (after midterm break).

[5] Unless, that is, there are *true contradictions*, as Priest suggests (much later).

Philosophy 415: Spring 2023
History and Philosophy of Mathematics

Notes: Week 9

The Idea of a Formal System

In the first half of the course, we explored the mathematics of the infinite, and the development of set theory (including the theory of multiple infinities) as a basis for all of mathematics. We saw that Cantor makes a tremendous breakthrough over any previously existing theory of infinity through his willingness to theorize actual (rather than only potential) infinities as totalities, and that the set theory he created, once axiomatized, can in principle be used as the basis for *all* of mathematics. Yet we also saw that the "naïve" or intuitive idea of a set – the idea that a set is just any grouping whatsoever, and that there is a set that corresponds to any grouping we can name – leads to interesting and profound paradoxes, especially when the issues of totality and reflexivity are concerned. These paradoxes can be sidestepped, apparently, by coming up with the right axiom system – but in so doing, we restrict our intuitive idea of a set and limit its ability to deal with some of the widespread phenomena (in particular phenomena of self-reference and reference to totalities) that we might otherwise wish to use sets for.

In the second half of the class, we'll explore a deeply related set of results, which raise questions about the very nature of mathematics and indeed about reason itself. What is usually called "Gödel's incompleteness theorem " – actually, there are two closely interrelated incompleteness theorems – concerns the possibility of deriving *all* mathematical truths from a symbolic system based entirely on fixed and specifiable "mechanical" rules, what is sometimes called a "formal" system. To get a feel for what a formal system is, recall the language and rules you learned in first-semester symbolic logic. There was a primitive vocabulary of a fixed number of symbols, or symbol types (such as the symbols for the truth-functions, the quantifiers, variables, and individuals). There were rules for forming from these symbols legitimate (or "well-formed") *formulas* or sentences, sentences that "say something" in the language of symbolic logic. Finally, there were rules for *deriving* one sentence from another (or from a set of others). The rules for forming sentences and for deriving sentences from one another were relatively simple, and they were also purely *syntactical*. This means that the rules themselves never concern the *meaning* of the various symbols (in the sense of what objects they stand for or represent). Instead, they just tell us how to intercombine symbols themselves, and how we can move from one symbol string to another. Because they are purely syntactic in this sense, this means that they are also purely *mechanical*. A computer can follow these rules as easily as people do (and in fact more quickly and accurately!), and it's easy to design a computer program that does just this, taking arbitrary premises written in the language of first-order logic and "grinding out" consequences in accordance with the rules.

Actually, we ourselves very often follow this kind of mechanical procedure in mathematical reasoning. For instance, consider the procedure for long multiplication that you learned in elementary school. There are just a few rules to learn, and given these simple rules, the procedure gives us an answer to the

problem of multiplying any two integers, of whatever length. Also, the rules involve only symbols: what symbols we write where, and when to write a new symbol (and which one to write). This is thus essentially an (elementary) example of a mechanical "proof" procedure. Later on, when we have a formal system that can express all of arithmetic in logical terms, we can put this very same procedure (if we want to) in the logical language and use purely mechanical rules to get the result (these are in fact exactly the same rules that your pocket calculator follows to get the result, as well). The calculation of the product (say that 34 X 22 = 748) is then just the same as a purely mechanical and syntactic *proof* of the result (i.e. that 34 X 22 = 748).

For logicians and mathematicians working in the first decades of the twentieth century, it was a widely shared hope that we could capture as much of mathematics as possible – perhaps all of it – by means of such formal systems and mechanical rules. As we saw earlier, logicians like Frege saw this as the great hope of logic in relation to mathematics. By working out such a set of rules, it would finally be possible to gain clear, rigorous criteria to distinguish actual mathematical proofs from false ones. No longer would it be necessary to rely on "intuition" or guesswork to tell when a mathematical result had really been proven. Many philosophers also shared the confidence expressed in Hilbert's motto that in mathematics, "there is no *ignorabimus*." That is, in mathematics, though there may be things we don't know at any given time, there shouldn't be anything that's absolutely unknow*able*. When faced with a well-defined but unsolved problem, a mathematician should never simply throw up her hands and declare the problem unsolv*able*: rather, she should go to work, using current methods or trying to find a new method for finding the answer. And the method ultimately reached should itself be maximally clear, rigorous, and capable of being followed by anyone. This is nothing other than the ideal of a rigorous methodology of proof, which finds expression in the actual development of formal systems.

For mathematicians who though this way, the development of set theory up to about 1920 or so was a powerful inspiration. Despite the paradoxes that had already been discovered, it seemed it would be possible to put all of set theory on a uniform, solid axiomatic basis, and as we saw a few weeks ago, if this is possible it is also possible to translate *all* of mathematics into set-theoretical (plus logical) terms. However, even despite the apparent success of these attempts, the question remained open whether a formal system like Principia Mathematica (or ZFC) can indeed capture *all* mathematical truths. It might indeed be possible to *translate* all mathematical statements into the notation of logic and set theory; but can we indeed get a formal system that does what PM was supposed to do in terms of proof: that is, that *proves* all of mathematical statements that are actually true, and doesn't prove any that are actually false? Recall our distinction (discussed in the second week) between two (desirable) aspects of a theory: soundness and completeness. That a theory is sound means that it proves *only* truths: it never "goes wrong" and proves something that's false. That a theory is complete means that it proves *all* truths: everything that is actually true in the domain can be proven by the theory (in this case, using only its mechanical and syntactic rules of inference). The whole project of finding a logicist reduction of mathematics – of using set theory and logic to formalize mathematics completely by giving it a determinate, mechanical formalization – is just the project of finding a formalism for mathematics that is both sound and complete. The question whether it's possible to think of mathematics as a formal

system, indeed, appears to depend on just this question: whether there can be a system (such as PM attempted to be) for formalizing mathematical inference that is indeed both sound and complete.

Given this, it's now possible to state the dramatic and completely surprising result of Gödel's theorems. The result is that this question has to be answered *in the negative*: there is *no* formal system for mathematics that is both sound (proves only truths) and complete (proves all the truths). And in fact, we don't even need to apply to very abstract or advanced domains of mathematics to show this. One way of putting Gödel's result is as follows: for any system that is capable of capturing (even so much as) whole-number *arithmetic*, there is a relatively straightforward, true, and purely *arithmetical* statement that such a system can't prove. As we'll see, there are lots of ways of understanding the significance of this result, and it's not straightforward how we *should* think of it in terms of truth, meaning, and the powers of systems. But one (perhaps loaded) way of putting what this shows is that there is something "about" mathematics that essentially exceeds the "powers" of any formal system, something about truth that goes beyond anything that "mere rules" and symbols can capture. We'll explore these consequences in more detail later. For now, it's sufficient to note that Gödel's theorem shows, of any formal system that has enough power to capture even basic arithmetic, that such a system cannot be both sound *and* complete: if it is sound, and only proves truths, it won't prove all of the (even straightforward) truths; and if it's complete, and proves all the truths, it will also prove some falsehoods.

Recall that we discussed some of the motivations for Hilbert's project when we discussed formalism, a few weeks ago. The idea was that we can make sense of the infinite if, and only as long as, we can prove claims about it using only *finite* reasoning, i.e. proofs following mechanical procedures that we can carry out in a finite number of steps. The idea of "mechanical procedures" and rigorous proof, here, is captured by the idea of a formal system. But of course, there are many possible formal systems, so we need to say more about what makes one "good" or "usable" for our purposes. Clearly, the minimal requirement for the system to be usable is that it be consistent (since, using classical reasoning about consistency, at least, a contradictory system proves everything).

In fact, what we really hope is that a formal system (capable of capturing mathematics) can also prove *its own* consistency. Failing this, we might look for a proof of the consistency of a system by means of *another* system; but then the problem would be that we'd have to find a consistency proof for this system, and we'd have to look for yet another system to do this in, and so forth without end. So Hilbert posed as the second on a famous list of unsolved mathematical problems the question: i) *is it possible for a formal system capable of capturing the truths of mathematics to prove its own consistency*? For reasons we'll go into later, this turns out to be closely related to the following two questions: ii) is it possible for a system capturing arithmetic to be both *sound and complete*? And: iii) can any *formal system* correctly answer all well-defined mathematical questions (without also giving incorrect answers)? Very surprisingly, as we'll see, Gödel resolves all of these questions by showing that the answer to all of them is "NO".

Tarski and Defining Truth

Before we take our first look at Gödel's theorem, let's look again at something closely related that we've essentially already seen before. One goal that we might have for a formal system (or indeed, for a natural language like English) would be to define "truth" in it. That is, we would like to be able to define a predicate that applies to sentences, in such a way that it holds of just the sentences that are true and not of the sentences that are false. It's certainly not unreasonable to think that we should be able to do this: after all, we use the English word "true" to mean just about exactly this, and we understand its meaning (even if we don't always know WHICH sentences are true and which are false!) In fact, intuitively, however we define "true", it seems as if we should be able to say, of a sentence G, that it is true, just when we are able to say G itself. (For instance, if I am prepared to say, "It's raining outside," I should also be prepared to say, "It's true that 'it's raining outside'" (and vice versa). In some sense, the two sentences are "equivalent").

Tarski suggested that we can capture this intuition, for a formal language, with the following schema for the truth-predicate TRU, which we are trying to define:

(Schema T) For any sentence G, TRU(G) $\leftrightarrow$ G

Note that this isn't itself the definition of TRU. The idea is rather that, whatever definition of TRU we come up with, it should accord with this schema. This is how we will know that we have actually defined a truth-predicate, rather than some other property or feature.

All this is going fine, so far, but we start to run into trouble as soon as we observe that in English and other natural languages it's possible to come up with "liar" sentences, i.e. sentences of the form:

This sentence is not true

Or

P: P is not true

Actually, it'll turn out (though it will take us a few weeks to show this) that we can come up with sentences of the same form in formal systems as well. In particular, all we need in order to come up with sentences of this form is a device for giving sentences of the system themselves "names" (such as P) in the system; then we can let sentences "talk" about themselves (this will be key, also, to proving Gödel's theorem). We'll also show that any system with enough complexity to "talk about" (i.e. prove truths of) ordinary arithmetic is sufficiently complex to do this – i.e. to "talk about" itself in a way allows us to construct these kinds of "self-referential" sentences.

But now we're in a position to see why it's actually impossible to define truth within a (noncontradictory) system. For suppose we COULD define truth somehow, in a way that fits with Tarski's schema T above. Then there is a well-defined and usable truth predicate. But then (given the

possibility of constructing "self-referential" sentences) we could construct a sentence L that "says of itself" that it is not true, i.e.

L ←→ ~TRU(L)

But then, from Tarski's convention T (which, we are assuming, characterizes our truth predicate, if there is one), we would also have L ←→TRU(L). Thus, combining the two, we have TRU(L) ←→ ~TRU(L) – a contradiction.

We thus must conclude, according to Tarski, that there is NO definable truth predicate for a system like this. That is, it is not possible to define the property of truth within the system itself, unless the system is inconsistent.

Tarski himself thought that the actual definition of truth for a language must be possible – i.e. there must actually BE some property of truth that works in accordance with schema T – but that this result showed that the definition must be given in *another* language, a "metalanguage" to be used for discussing the language in question. If we can always use a metalanguage that is not L to define truth-in-L, then we can define it in a way that doesn't allow the Liar sentence to come up. If we're in the business of creating formal languages, it's also not unreasonable to think we could use a natural language, such as English (for instance) to do this work: we wouldn't get a definition of truth-for-Principia Mathematica, e.g., within Principia Mathematica, but maybe we could do this in English.

However, it's not clear Tarski's solution can work if thought of as applying to natural languages such as English itself. For example, even if we DO define truth-for-PM in English, for example, in so doing we'll appeal to an intuitive understanding that we already have of truth itself (as we discuss it ordinarily in English). How could we define this? Obviously, it's not the case that in order to understand truth as we ordinarily use it, we have to be able to speak ANOTHER natural language. So clearly, there's some understanding of truth that we already have and that we presumably should be able to define without leaving the language we ordinarily speak. Tarski thought that this just showed that natural languages are irreducibly inconsistent, and so there was no point in attempting formal definitions of properties like truth for them. However, even if they ARE inconsistent in this way, might there be ways of thinking about the actual possibility of these definitions within them?

Gödel's Theorem: Preview

Let's get a first look at what Gödel's first incompleteness theorem says and how it's established. We'll prove it in more detail – actually using two different, but ultimately equivalent, methods – later on. It's often said that Gödel's theorem shows that there are mathematical *truths* that are not captured by any formal system – and thus that there is a "dimension" to truth, the kind of truth that *we*, at any rate, can "see", that no formal system can itself see or prove. As we'll see, this is indeed *one* way of putting the significance of Gödel's result, but it depends on a specific interpretation of a lot of what goes on in proving it, and there are other ways of understanding what it's saying. So to really see what's going on, we need to get clearer about what we mean by truth, proof, and reason.

At a very rough, general level, the strategy will be this. We're going to show how to generate a sentence in the system of PM (or any similar system) that "says", of *itself*, that it (that very sentence) is *not provable* in that very system. Thus, we'll come up with a sentence, GS, that says (intuitively):

GS: I am not provable in PM.

or, equivalently:

GS: GS is not provable in PM.

Note the similarity, but also the difference, to the "Liar" sentences above. Like the Liar sentences, the sentence GS is "self-referential"; i.e. it (intuitively) "talks about itself." But what it says about itself is different. Rather than "saying" of itself that it's not *true*, it "says" of itself that it's not *provable*. These are importantly different notions. In particular, there is a difference between something being true, and its being provable in a specific system – although of course we *hope* that the two notions will be able to converge. (In fact, Gödel will effectively show that they cannot).

We can also put this more formally, using some logical notation. We're going to show that there is some sentence GS such that:

$\vdash_{PM}$ GS $\leftrightarrow$ "GS" is not provable in PM

That is, we're going to be able to derive from the axioms of PM that GS holds if and only if it is, itself, not provable.

If we can indeed come up with such a sentence and formulate it in the language of PM, then we'll have a system that's intuitively "true" but can't be proven. Why? Well, either GS is true or it's false. First, suppose that GS were false. Then what it says (that GS can't be proven in PM) would be false, and this would mean that it *can* be proven in PM. Then there would be a proof of it, and so – *assuming* that PM is sound, and hence only proves truths – it must be true. But we started out by assuming it was false, so this is a contradiction. Therefore it can't be false, and must be true. We can therefore conclude that GS is true.

However, what does GS 'say'? It 'says' that it is not provable. As we have just seen, this is true: it really is not provable! We therefore conclude that GS is indeed true, but unprovable in PM. The situation might be frustrating, indeed, but note that there's no contradiction here. Note again that this is again different from the Liar paradox, where even assuming that the sentence is TRUE leads to a contradiction – here, if we assume it's true, there's no contradiction, since it may very well be true but unprovable. Therefore, we conclude, there are definitely truths that *aren't* provable by PM (or any system with an equivalent demonstrative power). In this sense, we say following Gödel, that PM and all related systems are *incomplete*.

Thus, though GS is true, we can never use our system to prove it. We can also show that we'll never come up with a proof of its negation, ~GS, in the system either. For if we did have such a proof, then we'd have a proof of the opposite of what GS says: that is, we'd have a proof that GS IS provable. Again assuming soundness, that would mean that both GS and ~GS are provable, which would mean (at the very least) that our system, PM is inconsistent, hence unsound. (We can take a quicker route to the same conclusion by noting that we have already observed that GS is true; thus if a system were to prove ~GS, it would prove a falsehood, and hence be unsound).

Thus we can conclude that, assuming PM is sound, the GS is (at least) "undecidable" in PM: PM can't prove either GS or its negation.

Notice a few things about this informal meta-proof, though. First, notice that it *is* a meta-proof about PM, not a proof *in* PM itself. In order to argue about what's true and what's provable, we have to effectively stand outside the system, and talk about what's involved in the system itself, which we do in English or in some other "meta-language." So when we do this, we're essentially presupposing the consistency of English (or whatever meta-language we're using), and we haven't yet proved *this*. Second, notice that in stating the meta-proof we've been sort of cavalier about the meaning of "truth." We haven't defined it, though we have presupposed that it's a *different* idea from provability-in-a-system. Third, notice that in showing there'd be a contradiction if the GS were false, we had to assume that PM is in fact *sound*. In particular, we had to assume that it is *consistent* and that it never proves a contradiction.[1] Of course, if PM isn't consistent, then (assuming standard logic) it's of no use to us whatsoever and we might as well just give up. So it may that we're obviously just justified in assuming soundness, or at any rate consistency. But it's important to note that this is a substantive assumption, and one we haven't demonstrated yet.

What do we actually mean when we talk about a sentence in a formal system being "true" or "false"? There are lots of different ways of thinking about this, but on the standard account, truth is a *semantic* rather than a syntactic notion. That is, unlike provability, it's not something that we can just ascribe to systems and their purely rule-based properties. To say that a sentence is true, then, is *not* just to say that it is provable in some system, but that it stands for something which is "actually the case" somewhere outside the system. This gives us the distinction between syntax, which concerns purely formal properties of systems, and "semantics," which concerns also properties of what these systems are *about* (although it took a long time to get clear on this, and some of the reasons for making the distinction actually come from Gödel's theorem itself).

Usually, these days, if we want to talk semantics, we talk about *models*. A model is just a system of objects and relationships (we can also think of it as a kind of "possible world"). In order to "interpret" any system in terms of a model, we have to assign the primitive names of the system to objects in the

[1] Assuming classical bivalent logic, soundness implies consistency. For if the system is inconsistent then there is some P such that it proves both P and ~P; but one of these is false, so the system is unsound. Since inconsistency implies unsoundness, soundness implies consistency by contraposition.

model, and assign the variables ranges among the objects in the model, and assign the relations in the system to relations holding in the model. To say that some sentences is true-in-a-model, then, is to say that the objects in that model are actually related as *that* sentence says they are (under *that* assignment of meanings). If we want to talk about something being necessarily true (for instance tautologies such as A v ~A) what we mean is that it's true in *all possible models* and under *all possible* assignments of the names and relations. We can then also talk about a sentence being *entailed* by another one: A ⊨B. What this means is that in any model and assignment where A is true (where the objects actually stand in the relations it asserts them to stand in), B is true also. Again, we hope this relation will match the syntactic relation of implication (A ⊢B) for a particular system, but we do *not*, so far, know that this is the case. In fact, it will only be the case if the system is sound and complete. But at any rate, we can now say what we mean by talking about the *truth* of GS. To say that it is true is just to say that in *every* model of PM (and, in particular, in the 'intended model' which is the 'actual numbers' standing in the actual relationships they do stand in), the sentence we're trying to derive (which says that it itself is unprovable) is true.

So how are we going to get this remarkable sentence and show that it actually is derivable in PM? Remember that what we're going for is:

$\vdash_{PM}$ GS $\leftrightarrow$ "GS" is not provable in PM

If we could just write this down and show that it holds, we'd be done with it. The trouble, however, is twofold: first of all, what follows the turnstyle isn't yet a sentence in PM (rather, it's still in English, using words like "provable" and "PM"); and second, we haven't shown that such a sentence *can* even be written in PM, let alone shown to follow from its axioms. Most of the hard work of the proof is going to be in showing how we can in fact translate this English sentence into PM, or any other formal system that's also capable of capturing arithmetic.

Notice, though, that this is going to be tricky, not only because we have to effectively "translate" provability into purely arithmetical notions, but because we'll have to show that such systems can include devices of *self-reference*, which is what we need in order to write the sentence following the turnstyle. Even though natural languages such as English obviously include such devices of self-reference (we can just write "this very sentence" or, as we've done here, use quotation marks to name the sentence itself), it's quite surprising that we can actually do the same with purely formal systems as well. By means of a truly ingenious device, Gödel will show that we can in fact capture the *entire proof-logic* of a system *within itself* , and thus make it possible to treat the provability (or not) of a sentence – *considered simply as a syntactic string* – as a straightforward numerical (in fact, arithmetical) property *of that sentence itself* (again, *considered simply as a syntactic string*). This is, in a way, just exploiting the underlying idea of Hilbert's program – that proof and provability is just a matter of simple, mechanical rules for manipulating finite strings symbols. In fact, it's *because* the rules are mechanical and easily definable and operate only on finite strings that we can formulate them determinately as mechanical and ultimately arithmetical properties of the symbol strings themselves. This will eventually make it

possible rigorously to formulate - *within* the system itself – a "provability" predicate that holds of *just* those sentences of a certain form, those that are provable, according to the rules, within the system (and we know that this *is* a certain form because the rules themselves are determinate and mechanical). And this is what will make it possible to write, and to prove, the sentence above.

Before we prove the result in the way Gödel himself did, though, we'll prove an equivalent result that is in some ways easier to motivate, and shows many of the same features. This is the result proven by Turing in 1936 (five years after Gödel's proof). It shows that there are perfectly well-defined questions in mathematics that cannot be resolved by an "effective procedure": that is, by a procedure using the purely mechanical and finitist means that Hilbert suggests. To show this, Turing had to first formalize the ideas underlying Hilbert's program in an absolutely rigorous fashion, and his particular formalization is in fact the one that laid the basic groundwork for the architecture of the digital computer (and in fact for all computers we use today). Given this formulation, Turing was able to show that there are indeed well-defined problems which *cannot* be solved by any such procedure (and so, intuitively, can't be solved by any computer). This result turns out to be equivalent to Gödel's result, since the rules of PM or any similar system can in fact be encoded in terms of Turing's architecture. But Turing's result also demonstrates the implications of this result in a somewhat different light, showing both how profound Hilbert's conception itself is and, surprisingly, that (perhaps *because* it is so powerful) it can be brought to a remarkable and dramatic point of failure that throws into question the very ideas of truth, method, and reason on the level of their own formal determination.

Philosophy/Math 415/515: Fall 2025
History and Philosophy of Mathematics

Notes: Week 11

Having considered last week the general idea of a formal system, and having gotten a first look at Gödel's incompleteness theorems, now it's time to get down to details. We'll prove a version of Gödel's incompleteness theorem this week, one that is due to Alan Turing and is in some ways easier to understand than Gödel's own version. Turing's version is also fascinating and useful because, in coming up with the proof, Turing *also* formalized the idea of a computer, which gives us the basic abstract architecture for ALL computers that we have today. We'll learn this basic architecture, and then see how it leads to the result that there are some well-defined and even fairly straightforward procedures and problems that CANNOT be solved by anything WITH that (kind of) architecture – that is, that cannot be resolved by any determinate and finitely stateable algorithm at all. It's highly interesting, and perhaps surprising, that the very argument that thus can be taken as showing the inherent limitations of algorithmic computation is the very same argument in which Turing came up with the structure that is followed by all algorithmic computers today. As we'll see, though, this is really no accident, and the conjunction of the results promises to show us something interesting and important about the very nature of human thought in relation to rules, methods, and procedures.

What is a Computer?

When Turing wrote the groundbreaking article "On Computable Numbers, with an Application to the Entscheidungsproblem" in 1936, there were no electronic or digital computers. What was first called a "computer," in fact, was a *human* whose job was to calculate large products, sums, and the like (perhaps with the aid of a slide rule or mechanical adding machine). Nevertheless, as we saw last week, Hilbert had posed the problem of what such a human following simple, mechanical rules – or an equivalent machine – could actually accomplish in terms of computing and proving mathematical truths. His formalist ideas, along with the development of formal systems of logic, in fact led Hilbert to think that there must be an algorithmic procedure for determining *any* mathematical truth whatsoever. This is just the idea that there must be a sound and complete formal system for mathematics – a system that proves ONLY truths (soundness) and that proves ALL truths (completeness). In 1928, in fact, Hilbert put this idea as a famous challenge to mathematicians, what he called the "Decision Problem" (or "Entsheidungsproblem," in German).

Hilbert's Decision Problem

We can distinguish between two forms of Hilbert's Decision Problem, a "wider" and a "narrower" form. First, the "wider form":

Is there an *algorithmic procedure* for deciding the truth or falsity of any statement of mathematics expressed in first-order logic plus arithmetical symbolism? That is, could there possibly be a machine

which, when given any arbitrary statement in this symbolism, eventually determines correctly whether it is true or false?

This version of Hilbert's decision problem captures, of course, what we would ideally want a mathematical "truth decider" to do. If we could have such a machine, in a real sense there would be no need for human mathematicians any more: any problem could just be fed into the machine, and after some finite amount of time we would get the right answer.

However, the "wide" formulation employs the notions of truth and falsehood, which are philosophically loaded and tricky. Thus it is useful to work with the decision problem in a "narrower" form that doesn't mention the semantic notions of truth and falsehood:

<u>"Narrow" form:</u>

Given a finite system of (presumably sound) axioms in a language capable of representing the notions of arithmetic, is there an *algorithmic procedure* for deciding, of a given statement, whether it follows from those axioms or not?

If we could get a positive answer to the "narrow" question, and *we also knew that our axioms were complete*, i.e. that they are sufficient to prove either P or ~P for every P, then we would have a positive answer to the "wide" question as well. On the other hand, a negative answer to the "narrow" form will also imply a negative answer to the "wide" form as well.

Philosophers and mathematicians quickly got to work trying to show either that there could or could not be such a machine. In order to do so, however, it was necessary first to formalize the idea of an *algorithmic procedure*. Church came up with one such formalization (called the lambda-calculus) and Turing came up with another at the same time; later on it was shown that the two formulations (along with others) are equivalent. But Turing's is much more intuitive and easy to work with, and also yielded the formal architecture of everything that we today call a "computer," so we'll go with his formalization.

The idea, as, we saw last week, is to capture everything that a human operator or a machine can do by means of rigorous, determinately specified rules. So the important considerations in coming up with the system are that:

i) The rules can be written down in a finite amount of space

ii) What the rules tell us to do in any particular case is always determinate; i.e. there is no room for disagreement or opinion about whether a given rule has been followed correctly or not in any particular case.

These are features of the familiar computational procedures that we learn in elementary school, for instance the procedures for computing long division and multiplication. The whole procedure can be formalized in a finite amount of space, and it's always clear how to follow the rules in each case. It's not

hard, as well, to get a computer or a calculator to follow just those rules (or equivalent ones) to come up with the answers to math problems.

In addition to these requirements, we would *like* it to be the case that our system of rules is *sound*: that is, that it never leads us to a result that is false. In practice, we often assume soundness, and since soundness implies consistency (assuming non-dialetheic logic), this involves assuming that our systems are consistent as well. However, that our system of rules for arithmetic is sound (or even consistent) is not something we have proven yet (indeed, as we will see, it's not in general something we *can* prove using the system itself).

Turing Machines

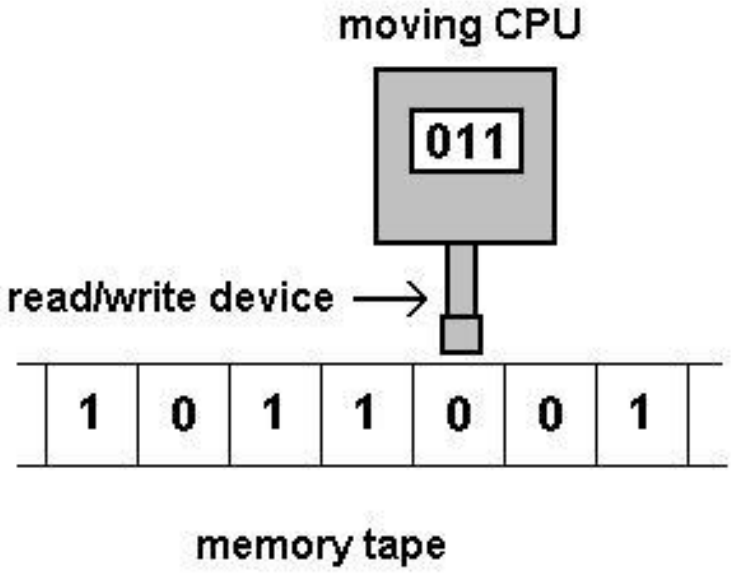
Turing therefore thought about what we would need to provide an abstract characterization of a system – any system – that satisfies i) and ii). He came up with the idea of a "computing machine," or as it has since been called in his honor, a Turing Machine.

Basically, a Turing Machine consists of 3 abstract parts. First, there's going to be a space for *input* and *memory*. In today's computers, this space is implemented as the RAM or hard disc memory that the computer uses to store what you type in and work on that input. In Turing's architecture, though, we can just consider this to be an arbitrarily long tape with symbols inscribed on it. For simplicity, we'll assume that only the symbols "**0**" and "**1**" are used and that the tape is divided into small squares or cells: one symbol per cell. Second, there is a *processing unit* that's going to work on the tape to read and transform what's written there. This is the CPU in today's computers; in Turing's architecture, we can assume that it's a kind of "read/write" head that can move right and left along the tape and sometimes transform a "**0**" into a "**1**" or vice versa. Finally, there is a *set of instructions* that tell the processor what to do in each case. This corresponds to what we would today call a *program* – it captures the *algorithm* by which the computer does whatever processing work it does. We suppose that the processor can be in a number of different internal "states" (call these A, B, C, D, etc.) and that what it does when it finds that a certain square has a "0" or a "1" depends only on what is in that square and what state it's in. Thus, for instance, a typical instruction might be:

<A, 0 -> 1, L, B**>**

This instruction says: IF you are in state A and you find a "0" on the tape, THEN change this to a "1", move one square LEFT, and go into state B.

Here's a picture of a Turing machine to help you get a sense of it:


moving CPU
011
read/write device
1
0
1
1
0
0
1
memory tape

The whole procedure for the machine is going to be specified completely by a finite set of such instructions, what Turing calls a "machine table." With this in mind, we can easily come up with some simple "programs" or "machine tables" for simple computations. Let's start with a machine for *adding one* to any number input. We can assume that a number is input as a string of "1"s: for instance 7 is 1111111. Let's also assume that this input is written on the tape alone, with just an unending string of 0s preceding it and another string of 0s following it, and that when the machine starts the processor head is somewhere to the left of the input string, and that the machine starts in state A. Intuitively, we're just going to make sure that we find the beginning of the string of 1's, move along it to the end, and then add one more 1, then stop. Then the following algorithm will accomplish this:

<A, 0 -> 0, R, A> [If in state A and you find a 0, leave it the same, move right, and stay in state A]

<A, 1 -> 1, R, B> [If in state A and you find a 1, leave it the same, move right, but go into state B]

<B,1 -> 1, R, B> [If in state B and you find a 1, leave it the same, move right, and stay in state B]

<B, 0 ->1, R, C> [If in state B and you find a 0, change it to a 1, move right, and go to state C]

Intuitively, the first two instructions lets us "find" the first 1 in the string and tell us to change to state B when we find it. Then the last two instructions let us move along the string of 1s until the end and change the first 0 after the string of 1s to a 1. State C, which we go into at the end, is essentially a "STOP" state – since there's no further instruction as to what to do in state C, the machine just stops here, having performed its duty.

In a similar way, we can easily come up with machine tables for subtracting one from a number, for adding together two numbers, for multiplying two numbers, etc. In fact, we can very easily convince ourselves that *any* reasonable algorithmic procedure conforming to the requirements of being finitely stateable and determinate can indeed be put in the form of a finite Turing machine table. The rules for any such procedure can be finitely stated – we always work with only finitely many instructions – and are completely determinate – for the instructions always determine what should be done in each case. The "programs" might get to be very complicated, but they still conform to these two basic principles. In fact, this is the basic idea behind all computer programming. The programs that we run on today's computers (such as Microsoft Word or Mathematica) might run to millions of lines of code. But the underlying idea is exactly the same as with Turing's machine tables.

Universal Turing Machines

So far, we've worked with what might be called "fixed-program" Turing machines. This means that each machine was defined by a single machine table, and this machine table gave a single algorithm that performs just one procedure. This is already quite a useful formulation, since it lets us capture Hilbert's idea of an algorithmic procedure and think about what can, and what can't, be done by any algorithmic procedure. However, the machines we've worked with so far are limited in that each has only a *single*

program. In this respect they are like a pocket calculator (or better yet, a calculator that only performs *one* function): they carry out a fixed procedure which is "encoded" in the construction of the device, but there's no way (without altering the hardware) to get them to do any other procedure. However, as you know, most computers (such as your laptop or even, these days, your cell phone) *aren't* like this: they aren't limited to just one procedure or algorithm, but can implement *any number* of different programs or procedures. All you have to do is make sure the right program is "entered" into the computer, either by being typed in by hand or (these days) read from a CD-ROM or downloaded from the web. In fact, this idea of a *programmable* computer was also arrived at by Turing, and can easily be derived from the basic idea of the Turing Machine itself. Such a programmable computer is what is called a "Universal Computing Machine" or "Universal Turing Machine" since it can in principle compute the result of *any* algorithm performed on *any* input.

To go from our fixed-program Turing Machines to a Universal Turing Machine, all we have to do is notice that the various algorithms, expressed by the various machine tables, can each *themselves* be considered simply as finite sets of symbols. Thus the machine table for our "add one" machine, above, is itself just a list of 25 or so symbols, including letters, numbers, commas, and brackets. Since it can be captured this way, why not just write these symbols down in the memory space of our Turing Machine itself? (This is equivalent to "loading in" a program into the computer's memory). In fact, we can do this easily. All that we have to do first is convert the series of symbols that makes up the machine table into a unique number (Turing calls this its "description number.") There are various schemes for doing so, but the essential thing is that since the machine table itself is a finite symbolic expression drawing on a finite "alphabet" of symbols, we can always *enumerate* the tables, assigning a unique integer to each distinct table. Our four-instruction "add one" table, for instance, might turn out to convert to the number 4,293,484 (although the actual number we'll get will of course depend on what specific enumeration scheme we use).

Now all we have to do is inscribe *this* number at the beginning of the tape and come up with a machine that, when it reads *this* number, will "simulate" the "add-one" machine and perform its action on whatever comes next on the tape. That is, we'll first write" 4,293,484" on the tape; and then we'll write, say, "7", and we want our machine to read both the first number and the second number and replace the second number with an "8." We also want it to be the case that if we had put a different "program" in as the first number – say the "program" for the "subtract-one" machine – then our Universal Machine would have done what this program demands, in this case replacing the "7" with a "6". In fact, it can take any well-defined algorithmic procedure at all, and given its numerical encoding will do with the input number *exactly* what that algorithmic procedure demands.

Can we in fact come up with a Turing Machine (table) that thus "simulates" any other Turing Machine whatsoever when given its description number? In fact we can, and although the construction is a bit more complicated than that of the Turing Machines we've considered, such a Universal Turing Machine can indeed be given in a description of a few pages in length. This will be a machine, remember, that is in principle capable of computing any algorithm that can be computed at all. In fact, all modern programmable computers have the basic architecture of the Universal Turing Machine: although there

may be practical limitations to storage space or time available to perform the relevant computations, they all have the same structural ability to implement, in principle, *any* algorithmically specifiable procedure whatsoever.

The Halting Problem

As we've seen, with the formalization of a Universal Turing Machine (UTM), we have formalized the idea of computability itself. The idea here is that *any* procedure that is computable at all can be computed by a UTM; equivalently, *any* problem that is solveable by regular means is solveable by a UTM. Since procedures that can be implemented by a UTM are sometimes called "effective procedures", this idea is sometimes expressed as what's called the "Church-Turing Thesis"

Church-Turing Thesis:

Anything that is computable at all is computable by an effective procedure, i.e., by a UTM.

Since this is just an attempt to formalize the "intuitive" notion of being computable at all, we can't really give a proof of this thesis. The claim is just that Turing machines, and in particular a UTM, indeed captures whatever we might mean by their being a way to "compute" any number or the answer to any given problem.

In any case, if we accept the Church-Turing Thesis, we now have at our disposal a formal way of confronting the question of whether there are intrinsic limits to what is computable or decidable by formal means. We can now return to the question of Hilbert's decision problem:

1) Is it possible by means of effective procedures (i.e., Turing Machines) to settle the truth or falsehood of any arbitrary mathematical claim?

As we saw last week, this is in fact equivalent to two closely related questions:

2) Can a formal system (such as Principia Mathematica) prove all and only mathematical truths?

3) Can a Universal Turing Machine solve all well-defined mathematical problems?

Turing will show that the answer to all of these questions is, surprisingly, "NO." Interestingly, this is equivalent, as well, to giving a positive answer to this question:

4) Are there real numbers that are not "computable," that is, cannot be arrived at by *any* formal procedure stateable in terms of a finite set of definite rules?

As we've known since the beginning of the class, there are many real numbers that are not rational, that is, are not expressible by any ratio of two natural numbers. Pi and the square root of 2 are examples of such non-rational reals. However, for both pi and root 2, although they are not rational, there *are* simple procedures for computing them to any desired degree of accuracy; although these procedures don't give us the exact number in any finite amount of time, we can get arbitrarily close, and we can say

that these procedures nevertheless indeed "determine" the number, in the sense that if we carried them infinitely far, we would get pi (or root 2) itself. These numbers are thus *computable* in that there is a finitely stateable procedure that computes them; in a certain sense, by stating this procedure we capture exactly what pi *is*. Turing's surprising result entails that there are real numbers that are *not* computable in this sense: that is, that there are real numbers for which there is *no* finitely stateable procedure that would yield them, even if we are allowed to apply the procedures an infinite number of times.

To begin to see how Turing used his idea of the Turing Machine to give a negative answer to the first question, we need to consider the issue of *halting*. If we design our fixed-program Turing Machines the right way, they will eventually halt given any input: this means that they come to the final, STOP state and don't do anything else. But it's also possible that a machine, given a certain input, might never halt: that is, it might get into a loop where it keeps performing a set of instructions again and again and never ends. (This is a familiar problem for programmers of modern computers as well!). In this case, we say that the machine (given that input) *fails to halt*. We can also think about the question of halting as applied to different intuitively stated problems (remember that what we're going for is an algorithmic means for solving any such problem). For instance, consider the problem:

P1. Given *m* and *n*, find *m+n*.

There is, of course, a well-defined procedure for computing this, and given any m and n, no matter how big, the Turing Machine that carries out this procedure will eventually halt. (Of course, our input might be given in the wrong format or not at all: in that case we don't worry about whether it halts or not). However, there are also problems that are not like this, such as:

P2. Find an odd number that is the sum of two even numbers.

We can easily come up with a procedure for *testing* each odd number, in succession (1, 3, 5, etc.) to find out if it's the sum of any two evens. This procedure is also easily realized as a machine table, and programmed into a UTM. However, the problem is that, as WE humans can easily see, given this program, the machine is never going to halt: that is, it'll never reach a STOP state where it gives us an answer.

In the first case, it's easy to see that the machine will always halt, and in the second case it's easy to see that it'll never halt. But some cases aren't so obvious. For instance, consider:

P3. Given *n*, find a number that is not the sum of *n* square numbers.

As a matter of fact, this procedure will halt *only* if *n* is 0, 1, 2, or 3. If n is 4 or greater, it turns out that it's impossible to find a number that is not the sum of n squares, so our machine won't ever stop. But to show THIS, we need to prove a high-level theorem in number theory called Lagrange's Theorem. If we didn't know Lagrange's Theorem, we might put in a value, say 5, and wait for the program to come up with a number that's not the sum of any 5 squares. The trouble is that we wouldn't know how long to

wait before giving up and concluding that there is no such number. At any finite moment, it might be, for all we know, that we just haven't waited long enough, and at the very next moment the computer will come up with one.

In other cases, it's not even known whether a given machine will halt or not. This is the case, for instance, for the following:

P4. Find an even number greater than 2 that is not the sum of two primes.

The claim that this does not halt – that there is no such number – is equivalent to Goldbach's conjecture, which has still not been proven today. So if we implemented a procedure for successively testing each even number in succession, it would run through all the even numbers that have so far been tested, finding that each of these is the sum of two primes. It MIGHT indeed run forever – this is the content of Goldbach's conjecture – but we won't KNOW that it's going to run forever unless we're prepared to wait an infinite amount of time.

Given that many of these procedures (for instance P2) are effectively "looking" for something that doesn't exist at all, we can't expect all of the procedures to halt. But what we might reasonably expect is to be able to determine algorithmically *which* procedures will eventually halt and which will not. Being able to determine this is, in fact, equivalent to being able to determine the truth or falsity of the various conjectures associated with these problems. For instance, *if* we could determine that P4 does NOT halt, this would be equivalent to proving Goldbach's conjecture ("Every even number >2 is the sum of two primes). On the other hand, if we could determine that it DOES halt, this would be equivalent to a refutation of Goldbach's conjecture. Either way, we would have solved the problem – settled the truth of the conjecture – by algorithmic means.

We can call this general problem – the problem of determining *whether* a given machine with description number *P* halts when given an input *n* – the "halting problem". As we have already come to suspect (and will argue more rigorously later), a general algorithmic solution to the halting problem would be equivalent to a positive answer to Hilbert's Decision question (question 1 above). For given an algorithmic means for answering the halting problem for any given procedure, we can (assuming our procedures are sound) use this answer to determine the truth or falsity of the associated conjecture. Conversely, if the halting problem CAN'T be solved in general, then there is no algorithmic means for determining whether certain procedures are ever going to halt or not.

This – that the Halting problem can't be solved in general – is in fact just what Turing proves. We'll go through the proof quickly, since we'll prove something equivalent (Gödel's incompleteness theorem) in more detail later. Since the proof is complicated, don't worry if you don't understand everything the first time – just try to get a sense for the "big picture" and we'll fill in all of the details later.

We're trying to settle the question of whether there is an algorithmic solution to the halting problem: that is, whether there is an algorithm (or Turing Machine table) H, such that H answers the question whether or not machine # *m* halts given input *n*, for *every m* and *n*. Like pretty much all of our

interesting proofs so far, this is going to be a *reductio ad absurdum*: we'll start out by assuming that there IS such a machine H, in order to derive a contradiction, which shows that there is NOT such a machine H. So let's start by assuming H does exist and it can settle the question whether any given machine (with description # *m*) halts when given any input (*n*). That is, when given inputs m and n, H will yield a "1" if machine #m halts when given input n, and will yield a "0" if it does not. The important thing to notice here is just that if H is going to give us either answer, it must *itself* halt: if H just runs on forever, it won't be able to do what it's supposed to do and solve the halting problem.

Let's symbolize the action of machine #m on input n as $M_m(n)$, and the action of H on inputs m, n as H(m,n). (Remember that the all of the machines M_x can be enumerated in order).

(1) Supposing that H(m,n) exists (and always halts), we can define another machine H' as such:

a. If H(m,n) yields 0, H'(m,n) yields 0

b. If H(m,n) yields 1 or anything else, H'(m,n) goes into an infinite loop and does not halt

(It's easy to define such a machine if H itself is definable).

Thus, assuming that H and H' are doing what they're supposed to, we have:

(2) H(m,n) yields 0 iff $M_m(n)$ *does not* halt.

This, after all, is just what H is supposed to do. And also, given how we've defined H',

(3) H'(m,n) halts (and yields 0) iff $M_m(n)$ *does not* halt.

Now we can, if we like, consider just the values for m and n where m=n. This might seem unmotivated, but there's no reason we can't do it: in effect, we're just considering the cases where the description number of a machine, and the input that that machine is working on, turn out to be the same. In these cases, we have

(4) H'(n,n) halts iff $M_n(n)$ *does not* halt.

Now, consider what such an H'(n,n) actually is. Since it takes the same value in both places, we can also consider this to be a computation on just *one* input, n. But remember also that all of the computations on one input can be enumerated: each possible computation just *is* some M_x. So H'(n,n) is going to be equivalent to some one of these M_x's, say $M_k(n)$. Now all we need to do is consider what happens when we let n=k. Then we have, substituting into (4),

(5) H'(k,k) halts iff $M_k(k)$ *does not* halt.

But H'(k,k) is just $M_k(k)$! So we have:

(6) $M_k(k)$ halts iff $M_k(k)$ *does not* halt.

This is a contradiction. This means that H(k,k) is in fact *incapable* of ascertaining the halting status of $M_k(k)$. It follows that there cannot be a procedure H which is capable of determining the halting status of every other procedure. Thus there is no algorithmic solution to the halting problem; and thus there is no algorithmic method for determining the truth or falsity of any arbitrary mathematical conjecture. Hilbert's decision question must therefore be answered in the negative, and so must be (as we'll soon see) the question of whether a formal system such as Principia Mathematica can ever be both sound and complete. On the other hand, there *are* procedures -- such as $M_k(k)$ itself – that can never be shown by a formal procedure not to halt, even though (as "we" can see) they do not halt. Thus, there are mathematical "truths" that we can see, but no formal procedure can verify. Such, at least, is the interpretation of Turing's result that brings it closest to Gödel's own interpretation of his own incompleteness theorems.

We'll delve into the details of what went on with this proof in more detail later, but for now just notice some interesting analogies to things we've seen before. For instance, one essential part of the proof was showing that we can effectively *enumerate* all of the machine tables or specifiable procedures. In other words, we can make a numbered list of all of the procedures (or programs), M_1, M_2, M_3 etc. We did this by correlating the actual symbols that expressed the procedures in the machine tables with numbers, and since there are (only) countably many combinations of symbols, we know there are only countably many possible procedures. When we then went on to hypothesize the universal Turing machine H and consider the application of one of the machines, M_k, to its "own" description number, what we were doing was, in a certain sense, diagonalizing. In other words, we were using the convertibility of programs into numbers to allow one of the machines to achieve a kind of "self-reference," or effectively to try to consider or compute some of its own properties (in this case, its own halting status). This gave us a contradiction, and so demonstrated the existence of something that *can't* be computed in this way at all. As we'll see, exactly the same device will be at the heart of Gödel's own incompleteness theorems. We'll number the symbols of the language that we're using in order to produce a kind of "self-reference," and this itself will allow us to diagonalize and "demonstrate" a truth that our system cannot itself prove. Thus the device of diagonalization that Cantor originally discovered turns out to have a profound bearing on the very nature of understanding and truth! But to see just what this bearing amounts to, we'll have to wait for the details of Gödel's result.

In the meantime we can quickly argue that Turing's proof of the unsolvability of the halting problem is sufficient to prove a version of Gödel's first theorem.

Gödel's first theorem (version): No system sufficient to represent arithmetic can be both sound and complete.

To show what we want to, we need only see that if there could be such a system, this would imply the *decidability* of all well-defined arithmetic decision problems stateable in its vocabulary. In particular, the decision function h(m,n) which determines the halting status of machine #m when given input n (yielding 0 if it does not halt and 1 if it does) is a perfectly well-defined mathematical function. But as we have seen, it is not computable: there is no algorithm that computes it. Thus, given that soundness

and completeness implies decidability (in this sense), the undecidability of h(m,n) implies that no system (capable of expressing it) can be both sound and complete.

But why does soundness and completeness imply decidability (for a system capable of expressing arithmetic)? If such a system is sound and complete, note that it proves either A or ~A (and never both) for *all* mathematical statements A (this is just the definition of "syntactical" completeness). In particular, given that such a system would prove either A or ~A, we could specify an algorithmic procedure that would *decide* whether A holds or ~A holds. Here is the procedure:

1) Using the axioms and derivation rules for the system (of which there are finitely many), prove all the consequences of the axioms that can be proven by just one application of a rule.

2) Prove all consequences of the existing theorems that can be proven by one more application of a rule.

3) Repeat (2).

This procedure gives us an *enumeration* of everything provable by the system (i.e., we could use it to make a list, in order, of everything it will prove). But then, *if the system is complete*, it proves either h(m,n) or ~h(m,n) for every m,n. So to effectively decide h(m,n) for a given m, n, we just have to follow the procedure above, and wait for it to yield one or the other as a theorem (if h(m,n) shows up, write down 1; if ~h(m,n) shows up, write down 0). If the system is complete, this will always happen in finite time and thus h(m,n) will be decided by it. But since (as we've shown), there's no possible procedure that decides h(m,n) for all m,n, it follows that no formal system (capable of expressing functions like h(m,n)) can be both sound and complete.

Philosophy/Math 415/515
History and Philosophy of Mathematics

Notes: Week 12

Preliminaries to Gödel 's Theorem

Having explored Turing's way of proving a theorem equivalent to Gödel's first incompleteness theorem, we can now get to work on Gödel 's own proof. As we saw a couple of weeks ago, the key idea is going to be to prove that there is a sentence, GS, in PM (or any other sufficiently strong system) that is provably equivalent to the assertion of its own unprovability. Thus, we're looking to show that there is a sentence GS such that:

$\vdash_{PM}$ GS $\leftrightarrow$ "GS" is not provable in PM

Or, more formally,

$\vdash_{PM}$ GS $\leftrightarrow$ ~Prov("GS")

Here, Prov("x") is going to be a "provability predicate" that says that "x" can actually be proven by PM. That is, it says that the sentence (or syntactic string) "x" comes at the end of a sequence of sentences that begins with one or more of the axioms and actually *is* a valid proof in PM, according to its rules for well-formed formulas and derivations.

This is as far as we've gotten so far, but there's still a lot to do. In particular, we have to show how to write the Prov("x") predicate in PM and make sure it does what we want it to do – that is, pick out just those sentences that are actually provable in PM. And, we need to make sense of the quote marks as well, which aren't symbols in PM. And, of course, we need to prove the whole thing.

Gödel Numbering

We've seen that if we're going to have any hope of proving the theorem, we need to show how systems such as PM (or even ZFC) can incorporate a kind of "self-reference". That is, since the Gödel sentence GS is going to be one that, intuitively, says "I am not provable!" we need to figure out how to get it to say this in PM. English and other natural languages come equipped with devices of self-reference, such as "I" and "this very sentence…", but it's not at all evident that a formal system like PM is capable of such devices. (In fact, it might seem quite surprising that ZFC is capable of such devices, since part of the intent of the axioms, as we saw earlier, was precisely to rule out self-reference in the form of set self-membership. It turns out that even *without* set self-membership, we *can* achieve self-reference in ZFC by means of Gödel's methods). How can we show that a system like PM can achieve self-membership, allowing its sentences in a certain way to "talk about themselves?"

The answer lies in an ingenious "trick" of coding that Gödel discovered. Intuitively, the idea is simple. Remember that, for a formal system, we have to have a limited (finite) vocabulary of basic symbols, and we have precise, rigorous rules for combining these symbols into well-formed formulas as well as for

deriving formulas from one another. Gödel 's idea is called " Gödel numbering" or "arithmetization of syntax", and it effectively gives us a way of talking *about* the system and the properties of its formulas *within* the system itself. We do this by code-*numbering* the symbols and the formulas composed of them. Then we can show that each formula corresponds to exactly one (whole, natural) number. Then – and here's the key innovation – *we can talk about all the properties of formulas – including whether they are provable or not within the system – by talking about the* ***arithmetical*** *properties of their code numbers*. Because there are only a finite number of symbols, and because the rules that determine what can be formed and derived from the strings are completely mechanical (remember that these are the two original requirements of Hilbert's program), we can talk about all these properties and relations – including the regular relations of derivation – as regular, arithmetical properties and relations between numbers. This will allow us to come up with a predicate (which is really just a predicate holding of certain numbers) that holds of a symbolic string (actually, of the number for a symbolic string) if and only if it is provable in the system.

Let's see how to do it in more detail. For simplicity, we'll work with a fragment of PM that has unlimited variable signs ('x_1', 'x_2', etc.) but only one basic term sign, '0'. We'll also have a sign 'S' which lets us write natural numbers: thus we'll express 1 as 'S0', 3 as 'SSS0', etc. Our axioms for the system will just be the Peano axioms (which we discussed a few weeks ago):

1. Every number has a unique successor

2. If the successor of *n* = the successor of *m*, then *n*=*m*

3. There's a distinguished element *z* (written '0') such that no *n* has *z* as a successor.

4. If this distinguished element has property P, and the successor of *n* has property P whenever n does, then all numbers have the property P.

In addition to the signs for 0 and successor, our vocabulary of basic signs to work with comprises also variable signs (such as 'x' and 'y'), quantifier signs ('∀ ' and '∃ '), signs for the logical connectives (for instance '&' and '~'), simple mathematical signs ('=', '+', and 'X') and punctuation marks (for instance parentheses). To begin with, let's just "assign" a unique odd number to each of these. We'll use '#(x)' to refer to the number assigned to x.

Sign	#
0	3
S	5
+	7
X	9
=	11
~	13
&	15
V	17

→ 19
↔ 21
etc.

Now that we've got the basic symbols coded, we need a way to code *strings* of symbols as well. To do this, we'll take advantage of a result from elementary number theory: *every number has a unique prime factorization*. That is, every number can be expressed uniquely as

$2^a \times 3^b \times 5^c \times 7^d \times \ldots$

For some *a, b, c, d*, etc. Accordingly, suppose I want to encode the following string of three symbols:

0=0

According to our chart above, '0' has the Gödel number 3, and '=' has the Gödel number 11. So transformed into numbers, this gives us the string 3,11,3. Now, using the prime factorization trick, we can just encode this as:

$2^3 \times 3^{11} \times 5^3$

Which, as it turns out, is equal to 117,147,000. The Gödel numbers for strings get very big very fast, but the important thing is just that each string is transformable into exactly one number, and each number is transformable (by means of its unique prime factorization) into exactly one string. In this way, we can "code-number" *all* of the (finitely long) strings possible using the symbols of PM, giving a unique number to each one. Of course, not all of these strings are going to be WFFs; many are just "garbage" such as '→x=&'. But the important thing is that at least *some* of the strings are going to be legitimate sentences, and some of *these* are going to be provable sentences, and now, by way of their coding into numbers (since PM can talk about all the finite numbers) we have a way of talking about these in the system itself.

Given the device of "self-reference" that Gödel numbering allows, we can now alter slightly, as well, the statement we're trying to prove:

$\vdash_{PM}$ GS ↔ ~Prov(#(GS))

That is, the "provability predicate", Prov, is just going to be a predicate that holds of certain *numbers*: in particular (if we build it right) the Gödel numbers of sentences that are provable from the axioms in PM.

Recursive Functions

Now that we have a way of talking about strings of symbols in PM with PM, we'll be interested in the properties of *certain* of these strings in particular. We're interested, in fact, only in the WFFs, and among these we're particularly interested in those that are *theorems*, or in other words those that are provable from the axioms. We'll ultimately try to show that we can distinguish all of these by means of the Prov predicate (which we'll build up) and that we can actually use this predicate in PM to pick out the right Gödel numbers (the Gödel numbers of the theorems of PM).

To begin to show this, though, we need one more idea: the idea of a *recursive function*.[1] Intuitively, a recursive function is just a function that can be defined by finitely many rules. In fact, we'll see that the recursive functions correspond exactly to the ones that are Turing-machine computable (which is an idea we discussed last week). That is, the recursive functions are going to be just those that are computable by Turing machines.

What is more, we can use these notions to help settle the question of which *sets* are decidable. A set (for instance, the set of all well-formed formulas of PM) is called *decidable* if and only if there is a computable way of deciding its membership (which things are elements and which are not). To connect this with the notion of computable functions (recursive functions) we can identify each set with its *characteristic function*; this is a function whose range is just (0,1), and the idea is that, if CF_H is the characteristic function of set H, then:

$CF_H(x) = 1$ iff $x \in H$

$CF_H(x) = 0$ otherwise.

That is, the characteristic function returns 1 if we have a member, and 0 if we do not. Given this, we can say that *a set is decidable if and only if its characteristic function is recursive*.

Let's go ahead and define the recursive functions; again, the idea is that these are going to be just the functions that we can expect our system (given that it has a finite set of mechanically applicable rules) to be able to compute (or, in other words, that we can expect our system to prove hold of the numbers that they do). To begin with, let's start with two very simple functions:

The *zero function z(x)*, which gives 0 for any value, is recursive.

The *successor function suc(x)*, which gives n+1 for any n, is recursive.

[1] Note, though, that this use of the term 'recursive' in connection with functions is very different from the use in connection with sets that we discussed near the beginning of the course).

Intuitively, it's obvious that these should be recursive, as they're very easily computed (as an exercise, you can come up with a Turing Machine for each one).

What about the operation of finding the *n*th number in a given list of *k* numbers (k greater than or equal to n)? This is called a "projection" function and it seems clear that it, too, should be recursive. After all, this is easily done with a Turing machine, and is clearly always computable in a regular way and in finite time. So we'll add all of these "projection functions" (actually there are infinitely many (countably many) of them, corresponding to the different *n*s and *k*s) as recursive as well.

Now that we have these basic functions, we want to see how to "build up" other functions from these. The idea is going to be that if the basic functions we use are computable, then we want the functions that we "build up" from them to be computable as well. We'll use two operations that let us build a new function from functions already defined. First, we can compose functions:

1) If functions *f* and *g* are recursive, then the function *h* which is obtained by performing first *g* and then *f* is also recursive.

For instance, if we have shown that the doubling function (which gives 2n for n) is recursive (and this is easily shown), then we can compose it with the successor function (which we know is recursive) to show that the function *2n +1* is recursive.

Second, we can do a kind of "recursion" on functions themselves (this is what actually gives the "recursive" functions their name). Suppose we have recursive functions *f* and *g*, and *g* takes two arguments more than *f* (i.e., if *f* is an n-place function, *g* is an n+2-place function). Then we can define a function *h* (which has n+1 places) such that:

i) $h(x_1, \ldots x_n, 0) = f(x_1, \ldots x_n)$

ii) $h(x_1,\ldots x_n, \mathrm{suc}(x)) = g(x_1,\ldots x_n,x,h(x_1,\ldots,x_n,x))$

This looks a bit tricky but the intuitive idea is simple. We define *h* by defining a "0" value which is essentially the value of *f*, and then letting the h-value of any number be determined by *g* and the *h*-value for the predecessor. In this way we've fixed both *f* and *g* (since they're recursive by stipulation) and we've also fixed an h-value for any number whose predecessor's *h*-value is fixed (and since we've fixed an *h*-value for 0, this will fix all of them). If we define h in this way, we can say that *h* is also recursive, since it too can be expressed (and computed) by means of a finite set of rules.

Let's try a simple example, the definition of the multiplication function, X, by recursion. We assume that addition is recursive (this is easy to show: exercise for home). Then we can just define multiplication for any arbitrary *x*, *y*, by means of the following recursive definition:

i) $x \times 0 = 0$

ii) $x \times \mathrm{suc}(y) = x + (x \times y)$

In this kind of way, we can easily convince ourselves that most ordinary arithmetical functions are recursive. What's more, any recursive function is going to be *representable* in PM (or whatever system we want to use).[2] To see what this means, recall that an n-place function is just a big set of ordered n+1-tuples of numbers. Thus, for instance, the 2-place function a+b is just a big set of ordered triples <a,b,c> where c = a+b in each case. So the idea of *representability* is that for each such set, there's a statement in PM (or whatever our system is) that holds of only the members of the sets. For instance, in this case, since the function a+b is representable in PM, for any <a,b,c> where c=a+b we have:

$\vdash_{PM}$ PLUS (a,b,c)

And for any <a,b,c> where c does not equal a+b we have:

$\vdash_{PM}$ ~PLUS(a,b,c)

Because of this equivalence, we can say that the formula PLUS(x,y,z) in PM *represents* the (intuitive) function of addition. The fact that there is such a representation of this function in PM follows, in fact, from its being a recursive function.

In Gödel 's original paper, he started by defining 45 recursive functions, showing that some of them were recursive, and then used these to define the Prov(*x*) predicate. We won't do all of this work in detail (it's laborious and unrewarding), but let's motivate the work we will do next time by just noting that all of the following are recursive. In most cases, we should be able easily to give a quick argument, in fact, to that effect, simply by noting that the rules for WFFs as well as derivations in PM are perfectly regular and mechanical and treat only the possible configurations and transformations of strings of signs. We also need to note that we can treat *sets* – for instance the set of all numbers which are the Gödel numbers of WFFs – in terms of characteristic functions that decide them (the function f that, for any x, returns 1 if x is a member of the set in question, and 0 if it's not). Then the following functions are all recursive, and their associated sets decidable:

Prime(*x*): The characteristic function (c.f.) of the set of all primes.

Var(*x*): The c.f. of the set of numbers which are the Gödel numbers of variable signs.

At(x): The c.f. of the set of numbers which are the Gödel number of a well-formed ATOMIC formula of PM. (We can define this recursively using the rules for atomic WFFs).

Fl(x): The c.f. of the set of numbers which are the Gödel numbers of (any) WFFs of PM. (We can define this recursively using At(x) and the rules for forming WFFs from atomic WFFs).

Finally, one more that will prove very useful in coming up with the Gödel sentence itself. Suppose we have a formula ϕ in PM that has one free variable (for instance: '∃y (y X 2 = x)', which has x as a free variable). This formula has a Gödel number; suppose this number is *m*, and suppose that the Gödel

[2] We won't prove this because it takes some tedious work, but we'll just give it some intuitive motivation.

number of the one free variable (in this case, 'x') is *n*. Now, we can transform ϕ into another formula by substituting the value *p* for the free variable (for our example we obtain '∃y (y X 2 = *p*)'), and this transformed formula also has a Gödel number, which is computable given *m,n,* and *p*. This Gödel number of the transformed formula, in fact, is computable by means of the (recursive) function:

Sub(*m,n,p*).

This allows us, in general, to calculate the Gödel number of a sentence that "talks about" any particular number, which can then also be the Gödel number of a sentence. Since this is recursive, we can now use the apparatus of Gödel numbers to formulate any claim we like (that is also representable in PM) about any particular sentence. We're well on our way to formulating the claim of the Gödel sentence itself, that this sentence *itself* is not provable.

Philosophy 415: Fall 2025
History and Philosophy of Mathematics

Notes: Week 13

Proof of Gödel's Theorem.

With all the preliminaries in place, we're now finally ready to get going on proving Gödel's first incompleteness theorem. Recall that we're going to prove, formally, that there is a sentence, GS, in PM such that

$\vdash_{PM}$ GS $\leftrightarrow$ ~Prov(#(GS))[1]

To get this, we have to first finish "building" the "provability" predicate, Prov; and then we have to show that such a sentence actually exists.

Last week, we convinced ourselves that various functions are recursive. In general, any function that can always be computed in a finite number of steps by following definite, mechanical rules will be recursive, and we can convince ourselves of this by noting that recursivity is equivalent to Turing-machine computability. In particular, we convinced ourselves that various arithmetical functions (such as SUC(x) and PLUS(x,y,z)) are recursive, and also that functions that operate on Gödel numbers of various syntactic strings (such as the functions At(x) and Fl(x), which discern atomic formulas, and all formulas, respectively) are also recursive.

Now let's consider a new function, called PRF(x,y). PRF is going to be the characteristic function of the relation that holds between x and y when x is the Gödel number of a *proof* of the sentence of which y is the Gödel number. To understand this, recall that a *proof* in PM (or any other formal system) is a sequence of (finitely many) lines, each of which is a well-formed formula, and each of which follows from one or more of the axioms together with one or more other (previous) lines of the proof, according to the rules. We can give this whole sequence (excluding the last line) a (big) Gödel number x, and then give the last line itself a Gödel number y (using our scheme for Gödel numbering). Then we'll want the relation PRF(x,y) to hold just in case the proof with the number x really is a proof of the claim with the number y.

Assertion: PRF(x,y) is recursive.

Again, we won't actually prove this, but it's a straightforward result of the fact that all the proof rules are syntactical and purely "mechanical" (or, equivalently, Turing-computable). That is, given that we have the axioms (that the set of axioms are decidable) and know all the rules, it's easy to build a

[1] Recall that #(GS) is the Gödel number of the formula 'GS'. Sometimes we'll drop the parentheses and just call this '#GS'.

function that "checks" whether any given sequence of lines really is a proof of the last line. This is just what the PRF(x,y) function does.

But now that we've got the PRF predicate defined, it's an easy matter to define what we're looking for, the predicate Prov. Remember that we're just looking for a predicate that's the characteristic function of the set of provable statements (or theorems). The provable theorems are going to be just those that have a proof. So we can just let Prov(y) be equivalent to the assertion that y is the Gödel number of a statement for which there is some x that is the number of its proof. Equivalently,

$\text{Prov}(y) =_{\text{def}} \exists x(\text{PRF}(x,y))$

This just says that the formula with Gödel number y is provable if and only if there's an x such that x is the number of a sequence that proves it – just what we want. Note, however (and this is important) that although this is a definition of Prov(y), and it shows that Prov(y) is (weakly) *representable* in PM,[2] it is *not* a recursive definition. Since it is defined as involving an existential assertion, there's no way we can define this simply by using the rules for recursive definitions that we discussed last week. So we are not defining Prov(y) recursively, and it is probably not a recursive function at all.

In fact, the guess that it is not recursive corresponds directly to something we've already seen: that Prov(y) is not decidable, as Turing showed. Of course, Prov(y) is recursively *enumerable* (or, "one-way" decidable); that is, we can list *in order* all the provable statements. But what we can't expect to do is "test", for any arbitrary statement, whether it's provable or not. The question whether an arbitrary statement is provable corresponds, as we now see, to the question of whether *there is* a number with a certain property. We can search through the numbers as long as we like, and we might get lucky and find that a proof of our statement shows up, but if there is no proof, we will never find this out in this way. Because of this we will never know (if we haven't found it yet) whether we just haven't found it yet, or whether there is none.

"Self"-reference

At any rate, though, we know have a predicate that allows us to "say" that *something* is provable (by way of making a claim about its Gödel number). The next step is to use this as a device of *self*-reference. After all, we want a sentence that says something about the provability of (not just any sentence) but *it itself*. How are we going to do this?

Recall one of the functions that (we convinced ourselves) was recursive from last time, the function SUB(m,n,p). SUB(m,n,p) is the Gödel number of the sentence we get by substituting p for the free

[2] Recall that a property is *weakly* representable iff there is some predicate P(x) such that if a number, y, has the property P(y) is provable in the system. Strong representability requires, in addition, that if a number does *not* have the property, then ~P(y) is provable in the system; this condition does *not* hold in the case of our Prov() predicate. In the following I shall use "representable" to mean just "weakly representable".

variable (which has the Gödel number n) in the sentence with Gödel number m.[3] So, this lets us talk "about" the sentences we get by making different sorts of substitution (in general, by substituting a number for the free variable in any formula with one free variable). Now we're going to use this SUB operation to make a claim about the sentence we get by making a certain substitution in this way. The point is going to be that the sentence we get by making the precise substitution is going to be *just* the Gödel sentence itself. And we're going to say of *this* sentence – the one we get by making a certain substitution in another sentence – that it, itself, is not provable.

OK, now we're almost there (can you feel the excitement?) Let's consider first a formula, G, with one free variable, namely y:

G(y) : ~∃x (PRF(x, SUB(y, 33, y))).

To read this, it's helpful to recall that when we originally assigned the Gödel numbers, 33 was our number for the variable sign 'y'. So what this says, for each *number* y, is that there is no proof of a certain formula. Which one? The formula that we get *from* the formula with *number y*, by replacing *in that formula* every free occurrence of the *variable* '*y*' with the number for *that* formula (i.e., with the number *y*) itself. This might seem a little perverse (note that we don't yet have a single formula, just a whole bunch of formulas for different y's), but now we can make the final step to displaying the Gödel sentence in all of its glory. To do this, all we have to do is note that G(y) *itself* has a Gödel number, q. Now we simply can put:

GS: ~∃x (PRF(x, SUB(q, 33, q))

Notice that GS just *is* the formula that we get by substituting q (the number) for 'y' (the variable) in G(y). But what does it say? It says that there doesn't exist a proof of SUB(q, 33, q). But what is SUB(q, 33, q)? Well, it's the number of the formula that we get by a certain substitution operation, namely the operation for substituting for 'y' (the variable) in G(y), q itself (which is, remember, the number for that very formula, G(y)). So we can now take GS as a whole as asserting that THAT formula, which is obtained from G(y) by the relevant substitution operation, is unprovable. But the formula obtained from G(y) by means of the relevant substitution operation (viz., substituting q for 'y') just is GS itself. So GS says that GS is unprovable.

Great! We have our Gödel sentence. We've created a sentence that says of its own number that the formula with that number is unprovable. Given our other definitions, we can now just write:

GS ↔ ~Prov(#(GS))

[3] This really means that we are substituting in *the numeral* (in the system) for the number p. I will sometimes talk about "substituting the number" in the following, but we should remember that this always really means substituting a numeral that is supposed to represent that particular number.

Fixed Point Theorem

We've now got the Gödel sentence 'itself'. But have we proved Gödel's theorem? No, not yet. For in order to prove the theorem we have to not only manifest the sentence, but show that it is indeed derivable from PM. That is, we have to show that:

$\vdash_{PM}$ GS $\leftrightarrow$ ~Prov(#(GS))

The important part here is getting what the turnstyle says, namely that the whole thing is derivable in PM. To get this, we'll prove a general theorem called the "fixed point theorem" or the "diagonalization lemma" (we'll see why in a minute) that holds for *all* formulas with one free variable.

Fixed Point Theorem:

For *any* formula with one free variable, $\phi(y)$, there is a sentence such that:

$\vdash_{PM}$ $\sigma \leftrightarrow \phi(\#\sigma)$

σ is going to be the formula that "says" that its own Gödel number has the property ϕ; in other words, we can think of sigma as the sentence "I have the property ϕ".

Proof of the Theorem

Let's first use our SUB function to define a closely related function, SB. In particular,

1. SB(m) = SUB (m, 33, m)

Intuitively, SB(m) is the Gödel number of the formula that we get by substituting the Gödel number of a formula, P(y), for 'y' in P(y) itself. Thus, SB(m) "says" *of itself* that it has some property, P, where m is the Gödel number of P(y). Or equivalently,

2. SB(#P(y)) = #P(#P(y))

Let's also (recursively) define another formula, $\theta(m,z)$, such that:

3. $\vdash_{PM}$ $\theta(m,z) \leftrightarrow z = SB(m)$

$\theta(m,z)$ just holds, in other words, whenever z is the result of the SB operation performed on m, that is, whenever z is the Gödel number of the sentence that we get when we take a formula (with number m) and substitute into itself its own number.

We also define $\psi(x)$:

4. $\Psi(x) =_{def} \exists w\ (\theta(x,w)\ \&\ \phi(w))$

$\Psi(x)$ says that the number of the sentence that we get by substituting its own Gödel number into the sentence with Gödel number x has the property ϕ.

And finally, we let

5. $\sigma =_{def} \psi(\#\psi(y))$

Then σ says *that* the number of the sentence we get by substituting the number of ψ(y) for 'y' *in* ψ(y) has property ϕ. But since σ just *is* the sentence we get by substituting the number of ψ(y) for 'y' *in* ψ(y), this just means that σ says that *its own* number has property ϕ.

Now we're ready to prove our theorem. We'll first prove $\sigma \rightarrow \phi(\#\sigma)$, and then $\phi(\#\sigma) \rightarrow \sigma$.

First half:

6.	σ	Assumption
7.	ψ(#ψ(y))	By 5 (definition of σ)
8.	∃w (θ(#ψ(y),w) & ϕ(w))	By 4 (def. of ψ) and 7

Now we can suppose, without loss of generality, that w=c (existential instantiation). Then:

9.	θ(#ψ(y),c) & ϕ(c)	
10.	θ(#ψ(y),c)	From 9 (& - elimination)
11.	c = SB(#ψ(y))	By 10 and 3 (def. of θ)
12.	c = #ψ(#ψ(y))	By 11 and 2 (def. of SB)
13.	c = # σ	By 12 and 5 (def. of σ)
14.	ϕ(c)	From 9 (& -elmination)
15.	ϕ(#σ)	From 13, 14

We have now established ϕ(# σ) on the assumption of σ. Thus we have:

$\sigma \rightarrow \phi(\#\sigma)$.

Now let's do the other direction.

Second half:

6a.	ϕ (#σ)	Assumption
7a.	#σ =#ψ(#ψ(y))	By 5 (def. of σ) and taking the Gödel number of both sides.
8a.	#σ=SB(#ψ(y))	By 7a and 2 (definition of SB)

9a.	θ(#ψ(y), #σ)	By 8a and 3 (definition of θ)
10a.	θ(#ψ(y), #σ) & ϕ (#σ)	By 6a and 9a, &-introduction
11a.	∃w(θ(#ψ(y), w) & ϕ (w))	By 10a, ∃-introduction (putting w for #σ)
12a.	Ψ(#ψ(y))	By 11a and 4 (definition of Ψ)
13a	σ	By 12a and 5 (definition of σ)

So, we're done: we now have proven ϕ (#σ) on the assumption of σ, so we have ϕ (#σ) → σ. And since we also have σ → ϕ (#σ) from before, we now have what we were looking for: σ ↔ ϕ (#σ). QED.

Finishing the Proof of Gödel's First Theorem.

Now we're pretty much done and we just have to put on the finishing touches. In particular, since we know by the fixed point theorem that $\vdash_{PM}$ σ ↔ ϕ (#σ) for *any* one-place predicate ϕ representable in PM, and since we know that Prov(x) is representable in PM, we now have our Gödel sentence GS, and we know that:

$\vdash_{PM}$ GS ↔ ~Prov (#GS)

In *one* sense, this is all that Gödel proved: that there is a sentence that, through Gödel numbers, "asserts" "of itself" that it does not have the property Prov() (where *we* know, at any rate, that Prov() was "supposed" to hold of just those sentences that are indeed provable in PM). However, we recall that Gödel's first incompleteness theorem says something stronger, and in ordinary English, namely:

G1. A formal system capable of capturing arithmetic cannot be both *sound* and *complete*.

To get to this claim, however, we still have to do a little bit of work. Remember the "quick" proof of G1 from a couple of weeks ago? Then, we gave a simple argument that GS "must be" true and unprovable. However, we gave this argument in something of a hand-waving way, in English rather than in the formal system, and we assumed soundness in order to show incompleteness (and hence the truth of GS). It's time to formalize what we can, and leave aside what we can't, so that we can get maximally clear on what we have *really* proven and what we can *really* say on purely formal grounds.

To begin with, there's a problem with just assuming soundness (even if only to show incompleteness, on that condition). Remember that soundness is a *semantic* notion, involving the key idea of *truth*. We don't really have a grip on what we're showing if we just assume this without specifying what idea of truth we're talking about, but it would also be best if we could make the proof as completely formal as possible, and make do with as little as possible of a semantic nature. But recall also that soundness implies consistency (for any system that is sound must also be consistent), and that consistency can be formulated as a purely *syntactic* notion. Accordingly, we can improve the proof if we could show that

the (syntactic) assumption of the consistency of a system *alone* is sufficient to show that it's incomplete, viz., that there's some G such that it never proves either G or ~G. Actually, we can't quite do this yet, given the resources that we have, but we can prove a closely related result using the related notion of (not consistency but) **ω**-consistency (pronounced "omega-consistency"). That is, we can go ahead and prove:

Any formal system that is i) capable of expressing arithmetic and ii) **ω**-consistent is incomplete.

(As we'll see later on, after Gödel himself wrote, someone named Rosser showed how to "strengthen" the proof to show what we really want to show, namely that a formal system that is consistent (at all) is incomplete. We'll look at this proof shortly, but for now let's just see what Gödel himself did).

So, what does it mean to say that a system is **ω**-consistent or **ω**-inconsistent? Consider the following scenario, which is perfectly plausible. Suppose a formal system proves ∃x P(x) for some numerical property P. But suppose it also proves ~P(0), ~P(1), ~P(2),…, and indeed proves ~P(n) for *every* n in ω (the set of all natural numbers). This situation is bad, but it's perfectly coherent, and there's no outright contradiction (i.e., there's no two formulas, G and ~G, such that the system proves them both). If this is the situation, we call the system **ω**-inconsistent; otherwise we call it **ω**-consistent. Notice that a system can be **ω**-inconsistent and still be consistent (in the ordinary sense); thus **ω**-consistency is a stronger condition than simple consistency. That is, **ω**-consistency implies (ordinary) consistency, but not vice versa.[4]

Now we're ready to use what we've got to prove that if a system (of the right sort) is **ω**-consistent, then there is a sentence (in fact, our friend GS) such that it will never prove either GS or ~GS. In particular, given that we have

(0) $\vdash_{PM}$ GS ↔ ~Prov (#GS)

we'll argue:

1) If our system is consistent, then it does not prove GS.

2) If our system is **ω**-consistent, then it does not prove ~GS.

For 1): Assume, contrary to fact, that $\vdash_{PM}$ GS. Then, since provability is (weakly) represented in PM by the Prov() predicate, as we have shown, we would have $\vdash_{PM}$ Prov(#GS). Then, by (0), we'd have $\vdash_{PM}$ ~GS. But then we'd have $\vdash_{PM}$ GS and $\vdash_{PM}$ ~GS, so our system would be inconsistent.

[4] What would it be like for a system actually to be **ω** –inconsistent, though? To generate an example of **ω** –inconsistency, it suffices to add to our Peano axioms the claim that there is some object, x, such that S(x) (the successor of x) = x. This is not inconsistent with the Peano axioms themselves, and given this, some c will have a property (i.e. being its own successor) that is provably not a property of any natural number (i.e. is not a property of *n* for any *n* in **ω**).

For 2): Assume, contrary to fact, that $\vdash_{PM}$ ~GS. If this were the case, then by (0), we'd have $\vdash_{PM}$ Prov(GS). But, since Prov(y) is just the same as ∃x PRF(x,y), we'd then have $\vdash_{PM}$ ∃x (PRF(x, #(GS)). But, assuming PM is **ω**-consistent, it is consistent (since consistency is a weaker notion than **ω**-consistency) and we have just shown (under (1)) that if it's consistent, there is no proof of GS. That is (as we have seen) there is no number which is the number of a proof of GS, and since PRF is (fully) represented in PM, for each natural number n, $\vdash_{PM}$ ~PRF(n, GS). But if we have: $\vdash_{PM}$ ∃x (PRF(x, #(GS)) and: for each natural number n, $\vdash_{PM}$ ~PRF(n, GS), then PM is **ω**-inconsistent. Therefore if $\vdash_{PM}$ ~GS then PM is **ω**-inconsistent.

Putting it all together, then, we have the complete proof of Gödel's incompleteness theorem:

G1. For any system that is capable of expressing arithmetic, if the system is **ω**-consistent, there is a sentence GS such that the system proves neither GS nor ~GS.

What have we actually proven? – Model Theoretic issues and ω-consistency

Now that we have Gödel's first theorem, purely as a formal theorem, clearly in view, it's time to zero in on the more philosophical question of what it really shows. In particular, it's important that we don't just assume that Gödel's own favored philosophical interpretation of his theorems is the right one, just because it *was* Gödel's own interpretation. In fact, Gödel himself was always quite rigorous about signaling the difference, in his own work, between the purely formal results and the interpretation he favored, although commentators and expositors of his work haven't always been as careful.

To begin with, let's think a bit more about the claim that GS is shown to be a *truth* of mathematics that isn't provable by the system. We know that GS is just a "normal" arithmetic statement in the language of the system (one calling for the existence of a number with a certain complex arithmetic property), and we also know that from the proof itself we can "see" that GS is not provable, so we can "see" in a certain sense that what it "says" is true. We can also recall Tarski's theorem, and notice that, if there *were* a representable truth predicate, we could use the fixed point theorem to render the system inconsistent by constructing a provable "liar" sentence. So it appears that both results show that our system cannot "capture" "mathematical truth" in some sense. But in order to remove all the scare quotes, it's important to get clearer (as clear as we can) about the notion of "truth" itself and how it works.

Assuming (as mathematicians and logicians usually do) that it is meaningful to talk about mathematical truths in some way that is not simply *reducible* to truth-in-a-system (for instance by appeal to our "intuitions" about the numbers, or what we already know prior to any formal system about mathematical truths), Tarski's theorem shows that no formal system is ever going to capture syntactically the totality of these truths. The usual way to talk about truth formally after Gödel and

Tarski, then, is to develop a *semantic* account of truth, which is done by means of model theory.[5] Models are "possible worlds" composed of objects standing in certain relations, and roughly speaking, what "truth" means in this sense is truth-in-a-model. That is, to say that a sentence is true-in-a-model is to say that there is a model (or possible domain) in which the objects really do stand in the relations that the sentence (interpreted as standing for those objects) says that they stand in. And when we talk about "arithmetical truths", we're talking about the truths of how objects are related in what we think of as the "intended model" for our formal systems of mathematics: namely, the model consisting in the "actual" natural numbers, standing in the relations that they "actually" stand in.

Given all of this, we can now say that Gödel's and Tarski's theorems together show that there is no formal system that captures all and only the truths (what is "really" true) in this intended model, the universe (sometimes called '**N'**) of the "actual" natural numbers, standing in the "actual" arithmetical relationships in which they do actually stand. In fact, the Gödel sentence is just such an assertion of an arithmetical truth about certain numbers (in fact, the assertion that there does not exist a number with certain properties), and if we accept that the Gödel sentence will never be proven, it seems we must accept that this assertion is true. Thus it appears – and this is, of course, Gödel's own "Platonist" interpretation – that we can indeed "grasp" something that is really true, of the "actual" numbers, but that our system (any system) cannot prove. If so, this is truly a remarkable result, in that it establishes definitively that our mathematical insight goes beyond any finite set of rules (and hence that, in a very important sense, we are not computers). But before we leap to this conclusion, let's make sure we've understood all the different aspects of what's going on here.

Given the first incompleteness theorem, which appears to show that there is a "truth" that cannot be proven or refuted by our system, one of the first things to ask is: why not just add this "truth" to our system, say as a new axiom? Thus, it might seem we could make our "incomplete" system PM complete by just adding GS as an axiom; then we wouldn't be able to say that the new system (call it PM*) can't prove or disprove GS, since it would prove GS trivially. However, this is actually no help. For it's then possible to come up with another, slightly different Gödel sentence for the new system, and it will again be possible to show that the new system can neither prove nor refute this new sentence. So it seems like incompleteness is here to stay, and we're not going to remedy it even by successively adding the truth of each successive Gödel sentence.[6]

---

[5] It's worth noting, though, that model theory only really develops historically after Godel and Tarski's results, and as a response to them, so it's not evident that it captures what really "was" the meaning of truth before they did their work.

[6] What happens if we iterate this process up to (countable) infinity, adding a countably infinite number of distinct Godel sentences, so that each new system remedies the incompleteness of the former one? Interestingly, this totality then amounts to something that *is* complete: there's no longer any sentence such that neither it nor its negation can be proved. However, the price to pay is that what we have is *no longer finitely axiomatizable at all*; that is, there is no finite (or finitely specifiable) set of axioms for it. So in an important sense, it's no longer a "formal system" at all (in Hilbert's sense).

However, something that's interesting model-theoretically *does* happen if we add to a system such as PM the negation of its Gödel sentence, i.e. ~GS. We thereby obtain a new system, call it PM*; and since we know that PM doesn't prove GS, we know that this new system is consistent (provided that PM itself is). Since (as has been proven elsewhere) any consistent system has at least one model, we also know that this new system has a model: that is, there is a 'mathematical' structure of objects that behave exactly as its axioms and theorems claim they behave. However, there is something 'odd' about PM*; in particular, it proves ~GS, and so it also proves Prov(#GS), which itself, remember, amounts to the assertion that *there is an x such that x is the number of a proof of GS*. Note also that PM* includes all of the axioms of PM (plus one more), so whatever is provable in PM is also provable in PM* (and perhaps other things as well). In particular, as we showed in proving the second half of Gödel's first incompleteness theorem, for any *natural number n*, PM proves that n is not the number of a proof of GS (i.e, 1 is not the number of a proof of GS, 2 is not the number of a proof of GS, etc.) And since this is provable, for each natural number n, in PM, it's also provable, for each natural number n, in PM*.

Thus, PM* is in the odd position that it proves that *there is an x such that x is the number of a proof of GS*, even though *for each natural number n*: 0, 1, 2, … , it also proves that *that natural number* is *not* the number of a proof of GS. In fact, we've met this situation before: PM* is, precisely, **ω**-inconsistent. That is, it proves as an existential assertion that there exists a number with a certain property (in this case, being the number of a proof of GS) even though, for every natural number n, it also proves that that natural number doesn't have the property in question. Remember, though, that a **ω**-inconsistent system may still be consistent (in the ordinary sense), and in fact that, as we've just argued, PM* *is* consistent (in the ordinary sense) provided that PM itself is. Because it's (simply) consistent, moreover, it has a model. What should we say about this model, in which it's both true that there is a *number* that has the property in question and that none of the *natural numbers* 0, 1, 2, … etc. has this property? This is possible, in fact, on the assumption that there is at least one "number" that is not a natural number: that is, there is at least one object in the model that, though it behaves in certain respects *like* a natural number, it is not any of the natural numbers 0, 1, 2, etc. Following Hofstadter, we might call these non-standard denizens of the realm of mathematics "supernatural numbers," and let's call the model that makes all the claims of PM* true (including claims that turn out to be about "supernatural" numbers) **M**. This model is certainly different from the "standard" or "intended" model **N**, since it includes at least one "supernatural" number, but on the other hand everything that's provable from the standard axioms of PM (which was meant to, but failed to, capture N) also holds here.

What interesting meta-conclusion can we draw, then, from the existence of **M**, a 'non-standard' model in which everything that **PM** proves is nevertheless true? Well, recall that our "semantic" interpretation of Gödel's theorems and (especially) Tarski's theorem as showing that there are "*arithmetical truths*" that PM can't capture, but "we" can, depended on our sense that *we* (at any rate) have access to the "intended model" **N**. This was supposed to be a model in which the only objects are the "actual" natural numbers, and there were no supernatural numbers. However, since the axioms of PM hold just as well for the nonstandard model **M** (and indeed for any number of *other* nonstandard models, **M**', **M**'', etc. obtained by adding iterations of ~GS and GS for each level), we might wonder what establishes that *any*

one of the models *is* the "actual" (intended, "standard") one. The situation might very well be, rather, that any set of axioms (the axioms of PM or any others) picks out a whole set of distinct models that are consistent with those axioms, and which of these is the "true" or "intended" model is more or less up to us, as long as we preserve the truth of the axioms. So this, again, would be reason to believe that things are more complicated than simply talking of "Gödel's *incompleteness* theorem" tends to suggest. In other words, the fact of *syntactic* undecidability – that for each consistent system, there is some sentence such that neither it nor its negation is provable – would, though certainly interesting and important, not really suffice to show that there is a "truth" that any given system cannot prove, or indeed that there is anything much to the gloss that is usually given to Gödel's first theorem, namely the notion that we "actually" have insight into "truths" that are essentially beyond the capacity of *any* system to prove.

What have we really proven? 'Philosophical' interpretations of the First Theorem

Let's go back to the first theorem as we proved it. Here's a perfectly rigorous statement of it:

G1) For any formal system capable of capturing arithmetic, if the system is **ω**-consistent, there is a sentence GS in the language of the system such that the system proves neither GS nor ~GS.

This is enough to establish the (syntactic) *undecidability* of the sentence GS. Since to say that GS is undecidable in this sense is just to say that neither GS nor ~GS is provable, in stating the theorem this way we are just saying that for every system (of the right kind) there is some sentence that's syntactically undecidable. So far, we haven't said *anything* about incompleteness.[7] Nor have we said anything of a *semantic* nature, for instance anything about the actual *truth* of any statement; everything we have said so far is purely syntactic in nature.

But we probably do want to say more than this; at any rate, we want to explore the implications of the various assumptions that we might have to make if we *do* want to say more. Let's consider a series of interpretations that we might make of the formal system, each one stronger than the last, and consider what we need to assume about the nature of mathematical truth in order to get out various stronger conclusions.

G1 Interp. 0) (Truth as Mere Provability) If we are skeptics about there actually being *any* sense of "mathematical truth" that outstrips provability-in-a-system, then we might just go for a kind of 'absolutely minimal' interpretation. On this interpretation, the Gödel sentence does indeed bear witness to the syntactic undecidability of any system that is **ω**-consistent (or any system that is consistent, on the Rosser modification). *However*, since there is, on this interpretation, no distinction between truth and provability-in-the-system, the first theorem itself (which, recall, says that $\vdash_{PM}$ GS ↔ ~Prov(#(GS))) just degenerates into a "liar" sentence for the system:

[7] It's notable, in connection with this, that Gödel's original paper was called "On Formally Undecidable [not: "incomplete"!] Propositions of Principia Mathematica and Related Systems I", and that the paper actually mentions incompleteness *only once*, in fact in a footnote.

$\vdash_{PM}$ GS ↔ ~Tru(#(GS))

But if we let this happen, then we immediately have a contradiction within the system, since GS is now true (in the sense of being provable in the system) if it's false, and false (in the sense of not being provable in the system) if it's true, and our system is inconsistent. So if we take this first, absolutely minimal interpretation, we must interpret Gödel's first theorem as showing that there is in fact *no* formal system (capable of representing arithmetical functions) that is so much as consistent. If this were right, it would be a depressing conclusion, since it would seem to show that there's no use at all to trying to come up with any formal system whatsoever, since they're all going to be inconsistent anyway. However, it's worth noting that, in a sense, we can't absolutely rule this out in the case of any particular formal system, since we know from Gödel's second theorem (about which more, next week) that no formal system (of the right type) can *prove* its own consistency anyway.[8]

G1 Interp. 1) (Minimal Incompleteness) If we want to say anything about (not just syntactic undecidability but) actual incompleteness, therefore, we're going to have to presuppose some notion of truth, however minimal, that outstrips simply provability in the system. The most minimal notion of such truth is probably ***bivalence.*** Bivalence is simply the claim that, for any pair of sentences P and ~P expressible in the system, just one of these (either P or ~P) is true, and the other of the pair is false. If we can assume bivalence in this sense, then we can say that Gödel's theorem does indeed yield a kind of incompleteness. For, since we know that every expressible sentence is either true or false, either GS or ~GS is true. Since we know that our system (provided it's **ω**-consistent) doesn't prove either one, there's at least one truth – either GS or ~GS – that it doesn't prove. So in this sense, it's incomplete.

This gets us a kind of minimal incompleteness, but notice that it's pretty minimal. In particular, it gives us no reason (so far, at any rate) to assume that we know *which* of GS or ~GS actually *is* true. What's more, we don't even really know what kind of truth we're talking about. All that we've assumed is that each sentence expressible in the language has a determinate truth value, either true or false, and its negation has the opposite truth value. We haven't yet said anything about what these sentences really "say," at all, though. In particular, even assuming that they're bivalent, we haven't yet established that they capture (anything like) *mathematical* truths: that is, truths that are in any real sense "about" the "actual" numbers, their actual relationships, and so forth. And so, even though we've established that every system of a certain sort must be incomplete in the sense that there's a true statement it doesn't

[8] Moreover, as we know from reading Priest, this might not be such a bad option after all. We could, after all, admit all of this and, going dialetheist, still have a very useful system (albeit one that contains some contradictions). Historically, of course, a big part of the reason this first interpretation was usually passed over was that people assumed that a system containing contradictions was just useless, since they didn't understand the possibility of dialetheic logic. But if, as is anyway very plausible, virtually all of the systems we're really concerned with in actual life are going to be inconsistent in some way, we might indeed conclude that we must go for an interpretation along these lines in interpreting the significance of the formal results for these 'everyday' systems.

prove, we haven't done anything to say what this true statement is, nor to establish that "we" can determine that it *is* true.

G1 Interp. 2) (Weak Platonism) Recall the "informal" argument that we initially gave for Gödel's incompleteness theorem when we first introduced it. For this argument, we *assumed* the soundness of the system we were working in, and then we gave an informal argument (in English) by contradiction that was supposed to establish that GS is *actually* true. What was it, though, that we assumed when we assumed soundness? Well, in order to assume that the system produces *only* truths, we have to give some kind of semantic (rather than purely syntactic) characterization of what truth amounts to. The usual model-theoretic way of doing this is to assume that there is a model (i.e. a set of objects in relations) that we are discussing, and that the truths we're looking for are just the statements that hold true of the objects and relations in that model. So if we assume *this,* we can say that Gödel's theorem indeed establishes that there is a truth that is not provable, and indeed that this truth is precisely GS, and indeed that we can – in a certain sense – 'see' that GS *is* a truth. However, there's still a catch. To say that there is *a* model for a system is not yet to say *what* that model is; and so in assuming that there is *some* model that makes it legitimate to talk about truth or falsity for sentences of PM, we haven't yet shown that this model consists in what we intuitively think of as the actual natural numbers, and nothing else. In view of the availability of alternative models that we explored, it may even turn out that there is *no* way to interpret the "truths" of PM as referring (only) to the familiar natural numbers. So some of our "truths" might not turn out to be *mathematical* truths at all (they might, for instance, turn out to be truths about exotic objects such as the supernatural numbers we explored). So even if we can establish, on these assumptions, that GS is a truth, we still don't have what we most likely want, the intuitable *truth* of a statement that our system cannot prove.

G1 Interpr. 3) (Strong Platonism). Suppose, however, we are *realists* about the finite natural numbers and the "universe," **N**, that contains all and only these numbers, standing in the actual relations that we (informally) use arithmetic to calculate and compute. That is, we think that these are real objects that really exist somewhere, and so there is always a determinate fact of the matter about the truth or falsity of any claim about them. Given this belief, we can then say that our system is sound if it delivers only truths in the sense of truths *about* this model, and that these are all "mathematical" truths in the ordinary sense. On this view, these are truths about the natural numbers that hold, or don't, completely independently of whether we come to know them or not, and completely independently of *any* formal system or *any* interpretation of one.

If we do presuppose *all* of this, we can then finally say that Gödel's first theorem establishes that there is, for each system, an ordinary *mathematical truth* that that system can't establish but that *we* can, again by assuming soundness and arguing by contradiction to truth, in some sense "see" is true, and hence that in this specific sense our own powers of reasoning about truth always outstrip the powers of the specific system with which we're presented. Notice that this still doesn't show that we have insight into a single truth that *no* formal system can prove, since each Gödel sentence for a particular system can indeed be proved by some *other* system. However, given that every time we are presented with a system we can "see" the "truth" of its Gödel sentence, whereas *it* cannot – and indeed that we can even

see that *this* will be the case for *each and every system* – we might still argue that we now have evidence that the human mind has a power of "seeing" truths that outstrips the power of *any* formal system.

Philosophy 415/515: Fall 2025
History and Philosophy of Mathematics

Notes: Week 14

Gödel's Theorem(s): Finishing up

With the main formal work of proving Gödel's first incompleteness theorem done, we're over the main peak. We've proven in PM (or whatever system we're working in) that there's a sentence that asserts that it itself is not provable, and hence that if the system is at least **ω**-consistent, that it does not prove either this sentence or its negation (and hence is syntactically incomplete). But there are still a few more formal things we want to do to firm up the proof, and we still have to prove a related theorem—Gödel's second incompleteness theorem – which, even more than the first, really puts an end to Hilbert's program and the hope of completely formalizing mathematical proof and reason.

Rosser's Theorem

Remember that last week we proved something that only approximates what we'd really like to prove. In particular, we showed that:

1) If our system is consistent, then it does not prove GS.

2) If our system is **ω**-consistent, then it does not prove ~GS.

In the second case, we had to weaken the conditional, since all that we've shown so far in this case is that if ~GS IS provable, then the system is **ω**-inconsistent (and NOT, so far, that it's *simply* inconsistent.) Since **ω**-consistency is a stronger condition than simple consistency – every **ω**-consistent system is consistent, but not every consistent system is **ω**-consistent – we'd really like to weaken this condition and prove

3) If our system is (simply) consistent, there is a sentence GS' such that our system does not prove either GS' or ~GS'

If we could do this, then we'd have proven what we really wanted – that a system that is (simply) consistent can't be complete. But to do so, we need to "firm up" the claim of (2) so that it doesn't rely on **ω**-consistency, but only simply consistency.

Gödel himself didn't do this, but a few years after Gödel wrote, someone named Rosser showed how to. The basic idea is that we can take advantage (once again) of the fact that every proof has a unique Gödel number. We can use this fact to make a list of all the proofs, in order of their numbers. Then we'll build a self-referential sentence called a Rosser sentence; it'll be similar to the Gödel sentence but will make use of this list of proofs. In particular, rather than just saying "I am not provable," (as the Gödel sentence does), the Rosser sentence, RS, will say

RS: RS is not provable by a proof that comes (in the list) before a proof of my negation, ~RS.

Given that we can construct the Gödel sentence, and the proofs can be ordered in a list, we can easily construct this Rosser sentence (though of course it'll be different from the Gödel sentence GS itself). But now we can go ahead and prove that if our system is (simply) consistent, neither RS nor ~RS is provable:

1) Assume RS is provable, that is, that there is a proof of RS. Assuming that the system is consistent, this means that there is no proof of ~RS, and hence no proof of ~RS before (in the list) the proof of RS. Thus the proof of RS also proves: RS is provable by a proof that comes (in the list) before any proof of ~RS. But this is the same as ~RS, the negation of the Rosser sentence. Thus if RS is provable, ~RS is provable, and the system is inconsistent. Therefore if the system is consistent, it does not prove RS.

2) Assume ~RS is provable, that is, that there is a proof of ~RS. Then, once again, assuming the system is consistent, this means that there is no proof of RS, and hence no proof of RS that comes before the proof of ~RS. Thus the proof of ~RS also proves: RS is not provable before any proof of ~RS. But this is the same as RS. Thus if ~RS is provable, RS is provable, and the system is inconsistent. Therefore if the system is consistent, it does not prove ~RS.

So, with this in mind, we have now proved an "improved" version of Gödel's first incompleteness theorem:

Gödel-Rosser Theorem: If a system strong enough to represent arithmetic is (simply) consistent, then there is a sentence RS, such that the system proves neither RS nor ~RS.

Gödel's Second Incompleteness Theorem

By proving Gödel's First Incompleteness Theorem, either in its original or its Rosser-improved version, we've proven that there is, for any sufficiently strong system that's (omega-) consistent, a sentence that is (at least) *syntactically undecidable* – that is, such that the system proves neither it nor its negation. This already deals a strong blow to Hilbert's formalist project – especially since it is equivalent to the unsolvability of the decision problem, which should be solvable if the formalist project is going to work. However, one might still hope to achieve another goal that Hilbert had set out for formal systems: that formal systems ought to be able to prove their own consistency. Remember that Hilbert had set this out as an essential condition for the success of his idea of formal systems, since a proof of consistency is what we need in order to make sure that we can trust a particular system at all. Moreover, what we'd really like to be able to get (if we want to complete Hilbert's program) is a proof of the consistency of a system within that system itself – otherwise, we have to rely on another system, which needs its own proof of consistency, and so on. Strikingly, however, Gödel's First theorem already gives us the resources we need to show that we can't do this – we can't prove the consistency of a (sufficiently strong) system within that system itself – and

this is the content of Gödel's (so-called) *Second* Incompleteness Theorem. Accordingly, we really have to abandon Hilbert's program and his hope for an internal consistency proof.

To begin with, we need a formula that expresses the claim that our system *is* consistent, a so-called "consistency statement" for our system. This is, in fact, easily done. Remember that (using the classical kind of logic we're using throughout the proof of Gödel's theorems), if the system is inconsistent, then it will prove anything at all. So we can take an (obviously) inconsistent statement, say 1=0, and say that our system is consistent just in case it doesn't prove *that*. This works, because if our system is consistent it'll never prove 1=0, and if it is inconsistent, it definitely will prove that (since it will prove anything). So let's just make the consistency statement for PM, Con(PM), equivalent to the claim that '1=0' is not provable:

$Con(PM) =_{def} \sim Prov(\#(1=0))$

Given this, what can we say about consistency given what we've already shown? Well, recall that the first part of the demonstration of the first incompleteness theorem showed that:

If PM is consistent, then GS is not provable.

With our consistency statement, we can now write this in the language of PM as:

$Con(PM) \rightarrow \sim Prov(\#GS)$

In fact (although we won't do this in detail here) we can do more than just write this; we can actually *prove* it in PM itself. That is, we can show that:

$\vdash_{PM} Con(PM) \rightarrow \sim Prov(\#GS)$

To do this, what we need to do is basically replicate the proof of the first half of Gödel's first theorem (which we originally proved in English) *within* PM itself. But we can in fact do this, since the proof really just depended on regular, arithmetical relations that are all representable in PM. In fact, we can show that we can do this, provided that our "provability predicate" Prov obeys certain relations within PM.[1] In particular, given our

0. $\vdash_{PM} GS \leftrightarrow \sim Prov(\#GS)$

And these properties of the provability predicate, we can argue as follows:

[1] More specifically, to show this we have to show that:

i) If $\vdash_{PM} A$, then $\vdash_{PM} Prov(\#A)$

ii) $\vdash_{PM} Prov(\#A) \rightarrow Prov(\#Prov(\#A))$

iii) $\vdash_{PM} Prov(\#(A \rightarrow B)) \rightarrow (Prov(\#A) \rightarrow Prov(\#B))$

1. Suppose $\vdash_{PM}$ Prov(#GS)
2. Then $\vdash_{PM}$ ~GS (by (0 and 1))
3. Then $\vdash_{PM}$ Prov(#~GS) (by condition i, below)
4. Then there is a proof of GS and a proof of ~GS (appealing to 1, 3, and the representability of Prov()).
5. Then $\vdash_{PM}$ 0=1 (since from a proof of GS and of ~GS we can generate a proof of 0=1)
6. Then $\vdash_{PM}$ Prov(#0=1) (by condition i, in the footnote above)
7. Thus if $\vdash_{PM}$ Prov(#GS) then $\vdash_{PM}$ ~Con(PM) (1-6)
8. Thus $\vdash_{PM}$ Prov(#GS) → ~Con(PM) (from 7)
9. Thus $\vdash_{PM}$ Con(PM) → ~Prov(#GS) (from 8, contraposition).

Now, the second incompleteness theorem follows easily. Recall that ~Prov(#GS) is just equivalent to GS, the sentence that asserts its own unprovability. Thus, given that we have (by replicating Gödel's first theorem within PM):

$\vdash_{PM}$ Con(PM) → ~Prov(#GS)

we also have:

$\vdash_{PM}$ Con(PM) → GS

But that means that if we had a proof of Con(PM), we would have a proof of GS. But we know, again from the first incompleteness theorem, that there is no proof of GS. Therefore, there is no proof of Con(PM).

Thus, we have Gödel's second incompleteness theorem:

G2. For any system S, capable of capturing arithmetic, if S is consistent then S cannot prove that it is consistent.

What have we really proven? 'Philosophical' interpretations of the First Theorem:

Let's go back to the Gödel-Rosser Theorem as we proved it. Here's a perfectly rigorous statement of it:

Gödel-Rosser Theorem: If a system strong enough to represent arithmetic is (simply) consistent, then there is a sentence RS, such that the system proves neither RS nor ~RS.

This is enough to establish the (syntactic) *undecidability* of the sentence RS. Since to say that RS is undecidable in this sense is just to say that neither RS nor ~RS is provable, in stating the theorem this way we are just saying that for every system (of the right kind) there is some sentence that's syntactically undecidable. So far, we haven't said *anything* about incompleteness.[2] Nor have we said anything of a *semantic* nature, for instance anything about the actual *truth* of any statement; everything we have said so far is purely syntactic in nature.

But we probably do want to say more than this; at any rate, we want to explore the implications of the various assumptions that we might have to make if we *do* want to say more. Let's consider a series of interpretations that we might make of the formal system, each one stronger than the last, and consider what we need to assume about the nature of mathematical truth in order to get out various stronger conclusions.

G1 Interp. 0) (Truth as Mere Provability) If we are skeptics about there actually being *any* sense of "mathematical truth" that outstrips provability-in-a-system, then we might just go for a kind of 'absolutely minimal' interpretation. On this interpretation, the Gödel sentence does indeed bear witness to the syntactic undecidability of any system that is **ω**-consistent (or any system that is consistent, on the Rosser interpretation). *However*, since there is, on this interpretation, no distinction between truth and provability-in-the-system, the first theorem itself (which, recall, says that $\vdash_{PM}$ GS ↔ ~Prov(#(GS))) just degenerates into a "liar" sentence for the system:

$\vdash_{PM}$ GS ↔ ~Tru(#(GS))

But if we let this happen, then we immediately have a contradiction within the system, since GS is now true (in the sense of being provable in the system) if it's false, and false (in the sense of not being provable in the system) if it's true, and our system is inconsistent. So if we take this first, absolutely minimal interpretation, we must interpret Gödel's first theorem as showing that there is in fact *no* formal system (capable of capturing arithmetic) that is so much as consistent. If this were right, it would be a depressing conclusion, since it would seem to show that there's no use at all to trying to come up with any formal system whatsoever, since they're all going to be inconsistent anyway. However, it's worth noting that, in a sense, we can't absolutely rule this out in the case of any particular formal system, since we know from Gödel's second theorem (about which more, next week) that no formal system (of the right type) can *prove* its own consistency anyway.[3]

---

[2] It's notable, in connection with this, that Gödel's original paper was called "On Formally Undecidable [not: "incomplete"!] Propositions of Principia Mathematica and Related Systems I", and that the paper actually mentions incompleteness *only once*, in fact in a footnote.

[3] Moreover, as we know from reading Priest, this might not be such a bad option after all. We could, after all, admit all of this and, going dialetheist, still have a very useful system (albeit one that contains some contradictions). Historically, of course, a big part of the reason this first interpretation was usually passed over was that people assumed that a system containing contradictions was just useless, since they didn't understand the possibility of dialetheic logic. But if, as is anyway very plausible, virtually all of the systems we're really

G1 Interp. 1) (Minimal Incompleteness) If we want to say anything about (not just syntactic undecidability but) actual incompleteness, therefore, we're going to have to presuppose some notion of truth, however minimal, that outstrips simply provability in the system. The most minimal notion of such truth is probably ***bivalence.*** Bivalence is simply the claim that, for any pair of sentences P and ~P expressible in the system, just one of these (either P or ~P) is true, and the other of the pair is false. If we can assume bivalence in this sense, then we can say that Gödel's theorem does indeed yield a kind of incompleteness. For, since we know that every expressible sentence is either true or false, either GS or ~GS is true. Since we know that our system (provided it's **ω**-consistent) doesn't prove either one, there's at least one truth – either GS or ~GS – that it doesn't prove. So in this sense, it's incomplete.

This gets us a kind of minimal incompleteness, but notice that it's pretty minimal. In particular, it gives us no reason (so far, at any rate) to assume that we know *which* of GS or ~GS actually *is* true. What's more, we don't even really know what kind of truth we're talking about. All that we've assumed is that each sentence expressible in the language has a determinate truth value, either true or false, and its negation has the opposite truth value. We haven't yet said anything about what these sentences really "say," at all, though. In particular, even assuming that they're bivalent, we haven't yet established that they capture (anything like) *mathematical* truths: that is, truths that are in any real sense "about" the "actual" numbers, their actual relationships, and so forth. And so, even though we've established that every system of a certain sort must be incomplete in the sense that there's a true statement it doesn't prove, we haven't done anything to say what this true statement is, nor to establish that "we" can determine that it *is* true.

G1 Interp. 2) (Weak Platonism) Recall the "informal" argument that we initially gave for Gödel's incompleteness theorem when we first introduced it. For this argument, we *assumed* the soundness of the system we were working in, and then we gave an informal argument (in English) by contradiction that was supposed to establish that GS is *actually* true. What was it, though, that we assumed when we assumed soundness? Well, in order to assume that the system produces *only* truths, we have to give some kind of semantic (rather than purely syntactic) characterization of what truth amounts to. The usual model-theoretic way of doing this is to assume that there is a model (i.e. a set of objects in relations) that we are discussing, and that the truths we're looking for are just the statements that hold true of the objects and relations in that model. So if we assume *this,* we can say that Gödel's theorem indeed establishes that there is a truth that is not provable, and indeed that this truth is precisely GS, and indeed that we can – in a certain sense – 'see' that GS *is* a truth. However, there's still a catch. To say that there is *a* model for a system is not yet to say *what* that model is; and so in assuming that there is *some* model that makes it legitimate to talk about truth or falsity for sentences of PM, we haven't yet shown that this model consists in what we intuitively think of as the actual natural numbers, and nothing else. In view of the availability of alternative models that we explored last week, it may even

---

concerned with in actual life are going to be inconsistent in some way, we might indeed conclude that we must go for an interpretation along these lines in interpreting the significance of the formal results for these 'everyday' systems.

turn out that there is *no* way to interpret the "truths" of PM as referring (only) to the familiar natural numbers. So some of our "truths" might not turn out to be *mathematical* truths at all (they might, for instance, turn out to be truths about exotic objects such as the supernatural numbers we explored last week). So even if we can establish, on these assumptions, that GS is a truth, we still don't have what we most likely want, the intuitable *truth* of a statement that our system cannot prove.

G1 Interpr. 3) (Strong Platonism). Suppose, however, we are *realists* about the finite natural numbers and the "universe," **N**, that contains all and only these numbers, standing in the actual relations that we (informally) use arithmetic to calculate and compute. That is, we think that these are real objects that really exist somewhere, and so there is always a determinate fact of the matter about the truth or falsity of any claim about them. This position is sometimes called "Platonism", since Plato may have held some version of it. Given this belief, we can then say that our system is sound if it delivers only truths in the sense of truths *about* this model, and that these are all "mathematical" truths in the ordinary sense. On this view, these are truths about the natural numbers that hold, or don't, completely independently of whether we come to know them or not, and completely independently of *any* formal system or *any* interpretation of one.

If we do presuppose *all* of this, we can then finally say that Gödel's first theorem establishes that there is, for each system, an ordinary *mathematical truth* that that system can't establish but that *we* can, again by assuming soundness and arguing by contradiction to truth, in some sense "see" is true, and hence that in this specific sense our own powers of reasoning about truth always outstrip the powers of the specific system with which we're presented. Notice that this still doesn't show that we have insight into a single truth that *no* formal system can prove, since each Gödel sentence for a particular system can indeed be proved by some *other* system. However, given that every time we are presented with a system we can "see" the "truth" of its Gödel sentence, whereas *it* cannot – and indeed that we can even see that *this* will be the case for *each and every system* – we might still argue that we now have evidence that the human mind has a power of "seeing" truths that outstrips the power of *any* formal system.

Formalism, Platonism, and Gödel

Along the lines of the last interpretation, Gödel himself thought that his first theorem provided a powerful argument *for* the Platonist view that the realm of numbers exists in itself, independently of the proof methods we use to establish truths about it, and indeed that his results show that we have a special kind of "insight" or ability to "see" truths about this realm that no formal system can capture. Of course, this position is to a certain extent circular, since as we've seen Gödel essentially had to assume Platonism in order to interpret his own result this way. What considerations are there, antecedently, to suggest that this kind of assumption is justified, or not, and what effect does Gödel's first theorem actually have on their plausibility?

For the first several decades of the twentieth century, as we've seen, the philosophy of mathematics was more or less divided among four positions: formalism, intuitionism, logicism, and Platonism. We've

already explored the formalist position through Hilbert: this is the position that the notion of "mathematical truth" is just equivalent to the notion of provability-in-a-system, and that there is no semantic "meaning" to the claims of mathematics beyond their place in a well-defined formal system. These ideas are what yielded the formalist program, which (recall) sought ultimately to find a formal theory that's both complete and which could prove its own consistency. Of course, Gödel's first and second incompleteness theorems dealt major blows to both of these commitments. Nevertheless, it's not clear that Gödel's results completely kill the formalist program, or at any rate that we might not still even in light of them pursue some version of it. For instance, the formalist can clearly interpret Gödel's first theorem in the sense of interpretation 0, above, and even in a way use Gödel's second theorem, which says that no formal system can prove its own consistency anyway, as evidence for this position. Moreover (see below) even if formal systems cannot prove their *own* consistency, there may still be useful ways of formally proving the consistency of various systems using other systems, and so saving at least some of the hopes of the formalist program.

The fourth position, Platonism, was of course Gödel's position (and there is evidence that Cantor and Frege essentially held it as well). There are certainly some problematic commitments internal to Platonism – for instance, holding that numbers are "real" in themselves, though not physical or spatial, raises the question of how we have any knowledge of or contact with them at all. However, most working mathematicians more or less implicitly accept some kind of Platonism in their quest for (what they see as) the discovery of the truths of mathematics, and its antecedently plausible that questions like the question of whether Fermat's last theorem is true or false have a determinate answer, and had such an answer even before they were settled by mathematicians one way or another. (Do we really believe, with the intuitionists, that the theorem only *became* true when Andrew Wiles proved it, a few years ago? Or that the question whether there is a string of seven '7s' anywhere in the decimal expansion of pi literally has *no* true or false answer, until and unless we discover that there is?)

Additionally, there's another kind of motivation that tends to make the antecedent case for Platonism attractive, and hence makes something like a Gödel-style Platonist interpretation of Gödel's first theorem itself more plausible. Recall when you first learned arithmetic, in grade school: you were taught about the familiar numbers: 1, 2, 3, and so forth, along with techniques for operating with them and relating them to each other. Presumably, you had a sense that you were learning about *something*, and that the things your teacher was teaching you were actually true *about* this something (or somethings). In one sense, Platonism is just the view that this grade-school intuition was right, that you really were learning about objects, and that they really do fit together as a totality into a coherent universe, the universe of natural numbers **N**. If this were *not* so, indeed, it would certainly be surprising that we can reason so easily about the (apparent) denizens of this universe and that we agree, universally, about what's true about them; we'd need at the very least to come up with an alternative explanation of these facts.

What's more, later on, when we come to formalize this knowledge and think about different systems of axioms for handling numbers, we have the sense that some of these possible axioms (for instance the Peano axiom that every number has a successor) are simply *true*; they simply seem to "force

themselves" on us in a way that's hard to explain if they're not actually "true of" the real domain of numbers. Moreover, once we have argued in this way that we have axioms and rules of inference that are "true" in the sense of matching up with the way numbers are in themselves, it's easy to argue that the whole point of our mathematical reasoning is simply to discover further of these truths. Gödel himself thus thought that the fact that we can apparently just "see" that certain axioms are true was a powerful consideration in favor of his kind of Platonism.

On the other hand, though, as we know from the first part of the course and discussed last week, things get trickier when we want to axiomatize the numbers themselves in terms of set theory, especially if we go along with Cantor in extending this axiomatization to infinite domains. Here there are, indeed, easily formulable claims which, as we saw, are not decided by any of the axioms of ZFC or by any other "natural"-seeming set of axioms: as Gödel himself helped to prove, Cantor's Continuum Hypothesis is a prime example of such an "undecidable." So it may not be the case that, as Gödel thought until his death, we really can always simply "discover" new axioms that will settle the truth of various undecidable claims, or that we can simply rely on our mathematical "intuition" to discover the truth (especially in cases involving infinite sets, such as the CH).

The Second 'Incompleteness' Theorem

What, then, about Gödel's second theorem, which (fairly uncontroversially) shows that no system that is consistent can prove its own consistency – that is, can prove a consistency statement for itself?

G2) No system (of sufficient strength to capture arithmetic) that *is* consistent can prove a statement of its own consistency.

This is the theorem that, much more than the first, really convinced people of the untenability of Hilbert's formalist program, for given Hilbert's assumptions and ideas, it seemed to show conclusively that, even if we could somehow come up with a trustworthy formal system, we could never really know that is *was* trustworthy. In fact, it seemed to demonstrate that systems that were already (implicitly or explicitly) 'in use', such as ZFC or PM, might not be consistent at all, that, in fact, there is nothing we can do or say internal to these systems to preclude the possibility that an inconsistency might show up tomorrow, and the whole system be ruined. (Of course, at this point, no one thought that we could do *anything* useful with an inconsistent system). And if these systems cannot even prove their own consistency, there is certainly no hope of getting them to prove their own soundness, so we really (it seems) have no reason to "trust" them at all, and we might as well just give up the whole game of formalizing mathematical reasoning at all.

Thus it seemed to many that the implication of Gödel's second theorem was the "skeptical" result that we can never "be sure" of the consistency of any system that's useful to us at all. Does it actually imply this? Well, in a sense, 'yes,' but also in a sense, 'no.' Of course, as Hilbert knew, the "best case" for guaranteeing the trustworthiness of our systems would be to find what he was looking for, a proof of

the consistency of a system within that system itself. But the fact that we cannot find this doesn't necessarily mean that we couldn't possibly become assured of the consistency of a particular system by *other* means; and here questions about mathematical truth again become relevant. In particular, if we were, for instance, strong Platonists like Gödel, completely committed to the existence of the "realm" of natural numbers **N** as well as the possibility of our having insight into this realm, we could simply argue as follows: the axioms of our systems are true (because we can 'see' that they are); and the rules for inference in our systems always preserve truth (because we can 'see' that they do). Therefore, since we're only ever moving from truths to truths, and all truths (assuming a non-dialetheic universe) are consistent with one another, we have good reason to believe that our systems are consistent (even if we cannot use them to *prove* this fact internally).

In fact, there's another important consideration here which tends to undermine the idea that Gödel's second theorem even gives us reason to *doubt* the consistency of our formal systems to begin with. The "skeptical" position is the one that says that, given that a certain system – say ZFC – cannot *prove* its own consistency, we have reason to doubt that it *is* consistent. But consider, for a moment, the alternative: suppose ZFC *could* prove its own consistency. Then, given Gödel's second theorem, we would be in a position to conclude that ZFC is actually inconsistent! Remember that the second theorem just says that *if* a system *is* consistent, it can't prove this; if it's inconsistent, on the other hand, then (using classical inference rules) it can 'prove' anything, including the statement of its own consistency. As Berto points out, a system that *could* prove its own consistency statement would therefore be something like a person who goes around loudly and constantly proclaiming to everyone: "I am not crazy!"; it wouldn't be long, of course, before we began to suspect that this behavior is *precisely* proof that the person *is* crazy. Thus, it appears that, far from giving us any positive reason to doubt whether a system like ZFC is consistent, the lack of a proof of consistency *within* the system, given Gödel's second theorem, actually provides a strong source of evidence (perhaps the best we can have) that the system *is* consistent.

Still, we might wonder what "justifies" positively our faith that the systems we are using are consistent; as long as we don't actually "have" a consistency proof it still seems like there may be some residual possibility that a system which has served us well so far might suddenly start producing contradictions at any moment, say tomorrow. As noted above, we can argue for the consistency of our systems by appealing to Platonistic intuitions about the truth of their axioms and rules of inference. But even if we don't make such strong, Platonistic assumptions about truth and models, we might pursue a related strategy to help "assure" ourselves of the consistency of various systems. Rather than proving a system consistent using itself (which we know we can't do), we might try to prove it consistent using *another* system. Of course, it will only be possible to prove the consistency of a particular system in another system that is stronger than it, itself (in the intuitive sense that the latter system proves "more"). And if such a proof is designed to give us confidence in the consistency of the first system, it will only be as good as our antecedent confidence in the consistency of the second one, in any case.

Nevertheless, there are proofs of this sort, and they are somewhat instructive. Most significantly, there is a proof due to Gentzen of the consistency of a fragment of PM (enough, in particular, to capture the

Peano axioms). This proof works by a form of induction on the length of formulas of this fragment of PM, essentially arguing that if there is no inconsistency among all the formulas of a certain length, n, or less, then there is no inconsistency among the larger set of formulas of length n+1 or less. However, in order to make the proof work, the induction has to be continued up to the level of a large transfinite (but still countable) ordinal; and so the proof can only be formulated in a portion of ZFC that contains all the axioms that allow for transfinite sets. Moreover, there is (so far at least) no comparable proof of the consistency of ZFC itself (or of this particular fragment of it) in *any* system that anyone has discovered; and the consistency of ZFC is, in any case, in many ways much less intuitively "obvious" than the consistency of the Peano axioms themselves.

Philosophy 415: Fall 2025
History and Philosophy of Mathematics

Notes: Week 15-16

Minds and Machines

Now that we have a pretty good "fix" on what Gödel's first and second incompleteness theorems actually are and say, we are ready to consider the bearing of the theorems on one of the most fascinating questions of 20th century (or any) philosophy: could we all really just be machines? In the 20th century, this old question takes the form of the question of artificial intelligence or AI: could our minds really be computers? That is, could it be that our minds and reasoning are really equivalent to a mechanistic computer? If we could answer "yes" to this, then this would show that computers could be minds, too, so we would vindicate the hopes of the research program that has sometimes been called "strong AI." Proponents of strong AI believe (to a first approximation) that the mind really is just a computer, and so, conversely, computers could also be minds: that is, there is no fundamentally reason to think that we couldn't eventually build a computer that can think in a human sense.

Interestingly, the first prominent 20th century proponent of something like strong AI was Alan Turing himself, and through his work and Gödel's we seem to have the background we need in order to turn the resources of metalogic and proof theory to addressing the question of AI. In particular, remember the equivalence between Hilbert and Russell's idea of a formal system, on one hand, and Turing machines, on the other; and also recall that every actually existing computer is, formally speaking, just a Turing machine. So it seems as if we can simply reduce the question: Is the mind a computer? To the question: Is the mind a Turing machine? And in fact, a lot of work in the twentieth century, especially in "cognitive science" has been more or less predicated on the idea that the mind is a computer in this sense. In particular, cognitive science analyses of various cognitive abilities and tasks treat the mind as a system that takes in information (say from the senses), "processes" it according to well-defined rules, and then yields "output" in the form of behavior. The task of cognitive science is then largely defined as just that of finding out these rules, determining (if you will) which formal system we are implementing.

…And Gödel

But if we conceive of things this way, it seems as if we're already well underway to an *anti*-AI argument that uses Gödel's theorems (or, equivalently, Turing's own proof of the insolubility of the halting problem) to show that human minds are NOT computers in this sense. In particular, if we can use Gödel's theorems to show that the human mind has the ability to grasp or otherwise KNOW some truth or truths that NO formal system could "know" (in the sense of being able to prove), then it seems we will have shown that the mind has capacities or abilities that essentially outstrip those of any formal system. And this would seem to be enough to show that the mind is not a computer (or Turing machine).

Although Gödel himself seems to have thought something *like* this (though not quite – see below!) was right, he was always very careful with the "philosophical" interpretation of his own results, and consequently was very guarded and cautious about expressing things just this way. The first person to make the "philosophical" argument against mechanism about the mind explicitly and prominently was J.R. Lucas in 1961, and more recently Roger Penrose has defended an updated version of Lucas's argument in a series of books.

To a first approximation at least, we can already see how the argument is supposed to go. Remember the first, "informal" argument that we gave several weeks ago to establish the conclusion that the Gödel sentence for a particular system, S, yields a mathematical truth that the system can't prove (but that we can see is true). The idea was, given that we have

$\vdash_S$ GS $\leftrightarrow$ ~Prov("GS")

we can consider the possibility that GS is false. If it's false, then "what it says," i.e. that it's not provable, is also false, so GS is provable. But then GS is provable, which is impossible given the soundness of our system. So (the argument goes) we can conclude that GS is in fact true, and that, since it (truly) asserts its own unprovability, it's also unprovable.

If all of this were OK, we could now say that we've, just by way of the preceding argument, "seen" the truth of a particular sentence, GS, which our system *cannot* prove. In this sense, our abilities to "see" or "establish" truth would go beyond those of the particular system. However, as we've already seen in detail, there are already several problems. First, for this all to go through, we have to assume the soundness (or at least the consistency) of the system. And we know from Gödel's second theorem that we can't use the system under consideration *itself* to establish this: no system can prove its own consistency (except an inconsistent one!) Second, we have to assume at least a weakly Platonistic view of truth: that the truths are "out there" waiting to be discovered, in a sense that's independent of the proof methods of any particular system. Third, even granting all this, there are questions about the sense in which we actually "establish" the truth of GS at all through this argument. Certainly, what we've said above doesn't amount to a "formal proof" (in the sense internal to the system) of GS, and so it's not so clear that we've actually "proven" the truth of GS at all. But if the demonstration is just an instance of plausible "intuition," or something like that, we might wonder whether we really want to agree that we "know" the truth of GS (in the mathematically relevant sense) at all.

In addition to these difficulties, there are some further problems with taking even the "informal" Gödelian argument to establish the anti-mechanistic thesis that the mind outstrips the abilities of *any* computer (or Turing machine). To begin with, the argument above *only* purports to establish that there is *for the system S* a sentence that that system cannot prove but whose truth we can see. So we can in this sense "outstrip" the abilities of the system S by "seeing" the truth of *its* GS, but remember that every sentence has *its own* GS, so there's no way that we can take this to show that we've "seen" a truth that *no* system can prove. In fact, there is no such truth; since for any claim at all, we can cook up a system that proves it (for instance just by adding that claim as an axiom). Lucas's response is that the

anti-mechanist claim is just of the following form: give me *any* (consistent) system you like, and I will show you a truth that *that* system can't prove (but is true). If this is right, it doesn't mean that there is some *one* truth that no system can prove, but it appears to mean that the mechanist will never be able to come up with "the true" mechanistic description of the mind. For every time the mechanist comes up with a description, saying that the mind is some Turing machine T, the anti-mechanist can come up with the Gödel sentence for T and argue that the mind can see that *this* Gödel sentence is true but unprovable. So it would seem that, if we can solve the other problems, then it might be possible to make the anti-mechanist argument this way.

Adding 'all' the GSs

Another part of Lucas's argument, though, is to establish something more than just that given any particular system, we can see the truth of the (unprovable) GS *for it*. He takes it that when we see the demonstration of Gödel's theorem itself, we see something more general, that in a very general way human reasoning outstrips not just some machines, but any machine whatsoever. We can capture this claim by considering what happens when we "add" to a system S its particular GS. Then, of course, we get a new system (call it S_1) that proves the GS for the original system; but this will, of course, have its own, different GS (call it GS_1) that *it* can't prove. And so on. Now, we could imagine iterating this process as many times as we want, and since the process of "generating" the Gödel sentence is in each case itself mechanical, we can integrate *this* process into a formal system, a kind of "super-system" that proves GS, GS_1, GS_2, GS_3, … G_n, for *every* effectively computable ordinal *n* (up to **ω** and in fact somewhat beyond, since given **ω** we can effectively compute **ω** +1, etc.) Now, Lucas's claim is that we could take *this* system, which appears to prove all the possible Gödel sentences, and still come up with a new Gödel sentence for *it*. That is, we could come up with a sentence that even the "super-system" (which proves '*all*' the GSs) can't prove. In this sense, we would have seen "something" that goes beyond any of the ability of any of the iterated systems to prove.

However, things are actually once again more complicated than this suggests. In particular, while it's right that given the postulated super-system we could indeed "Gödelize" it to generate a new Gödel sentence that *it* can't prove, the form of *this* "mega-Gödel" sentence will be different from that of all the GSs in the series we've already "included." In particular, even to write it down we will have to write down a very high transfinite ordinal. And so, in order to "see" that the new mega-Gödel sentence is true, we'll have to establish that *this* very high ordinal is effectively computable. It's not in fact clear that we can do this; in any case, our degree of confidence in the truth of the "mega-Gödel" sentence will have to be *much less* than that of our confidence in any of the lesser (especially the finite-level) ones. So once again, it seems less than certain that we've established, in this case, a "truth" that goes beyond anything mechanical systems can prove, and so the anti-mechanist argument seems less than decisive, once again.

What about consistency?

Let's go back, then, to the simpler version of the Lucas anti-mechanism argument: give me *any consistent* system S and I'll give you a sentence, GS_S that's true but S can't prove. Of course, this only works for consistent systems, since an inconsistent system proves *everything* (and hence proves GS_S too). And we need, of course, to assume that S is consistent in order to argue that S is true. But here we have to come back to the stumbling block we mentioned before: how can we be sure that the system S under consideration *is* consistent, given that (as Gödel's second theorem shows) no S can prove its own consistency?

Lucas argues that this isn't really a problem, since the mechanist *himself*, in proposing a particular formal system as *the* formal system that captures the workings of the mind, must *assume* that the system he's proposing is consistent. If it wasn't consistent, after all, it wouldn't be a viable contender for capturing the ideal formal principles behind human reasoning at all. Accordingly, Lucas argues, since all parties can assume consistency of the systems under discussion, we can just go ahead and presuppose this in arguing that each of the GSs is true.

Once again, though, this isn't quite right. The problem is that it's one thing to *assume* that a system is consistent, and quite another to *know* it; but if we are going to *know* that a GS for a particular system is true, we need to *know* that the system is consistent (and not merely assume it). If we want to know this with "mathematical certainty," moreover, we're going to have to prove it mathematically, and this is of course just what Gödel's second theorem shows us we *can't* do, at least with the resources of the system itself. So again, it looks like the anti-mechanist argument is in trouble.

The anti-mechanist now might respond that it *is* possible, after all, to know the consistency of various sytems; maybe we can't prove the consistency of a system *in that system itself*, but why not just use another one? If we can prove the consistency of a system S, for instance in *another* system which is itself one we can trust, then it would indeed seem that we are justified in assuming that S is consistent and that its Gödel sentence is true. In fact, since we showed in the course of proving Gödel's second theorem that the consistency statement for a sentence, Con(S), is *equivalent* to its Gödel sentence, this strategy just amounts to the strategy we considered before of "adding" the Gödel sentence at each level. And it's still true that, at each level, we can argue that the system at the level immediately below is consistent, and we can then convince ourselves of the truth of the GS sentence in the *new* system. But the problem comes, once again, in assuming that we can iterate this for *all* systems. The issue here is essentially the same as the issue we ran into when we considered the Gentzen "consistency" proof for a fragment of PM, which is conducted in ZFC and employs high transfinite ordinals. Once again, in order to assume the total iteration to be possible (and the consistency of each of the weaker systems to be shown) we have to work in a system involving very high transfinite ordinals, and our confidence in the consistency of *this* system will likely again be rather weak.

Weakening the conclusion

All things considered, then, it seems the anti-mechanist hasn't established the conclusion that the mind can't be a Turing machine, but perhaps he has established a weaker, conditional conclusion: IF we are given a formal system, and can KNOW it to be consistent (and hence sound), the mind is not THAT formal system. For knowing the consistency of the system is equivalent to being able to prove its Gödel sentence; but no consistent system can prove its own Gödel sentence. From this, we can draw an interesting disjunctive conclusion which is weaker than the original one, but interesting nonetheless:

EITHER: i) The mind is not a Turing machine; or ii) the mind *is* a Turing machine that cannot know its own consistency (and hence soundness).

So putting things all together, we can say that we *have* concluded that the mind, at any rate, is not following a "knowably sound" algorithm for grasping mathematical truths. Either it is not following an algorithm at all (Lucas's conclusion) or, if it *is* following a (sound) algorithm, it *cannot know* with mathematical certainty that this algorithm is consistent (and hence that it is sound).

In fact, this is exactly what Gödel thought too, and what he was careful to say when he expressed himself on the issue. For instance, in 1963 he said:

> Before my results had been obtained it was conjectured that any precisely formulated mathematical yes or no question can be decided by the mechanical rules of logical inference on the basis of a few mathematical axioms. In 1931 I proved that this is not so. i.e.: No matter what & how many axioms are chosen there always exist number theoretical yes or no questions which cannot be decided from these axioms. Combining the proof of this result with Turing's theory of computing machines one arrives at the following conclusion: Either there exist infinitely many number theoretical questions which the human mind is unable to answer or the human mind … contains an element totally different from a finite combinatorial mechanism [such as a Turing machine]…

Notice the "either…or" of the last sentence. Gödel himself certainly favored the second alternative, that the human mind is different from any Turing machine. He even went on to say that "I hope I shall be able to prove on mathematical, philosophical and psychological grounds that the second alternative … holds". But he knew that his own theorems don't suffice to formally show that the second alternative is right, and that the first is equally a possibility. In particular, if the mind IS following some formal algorithm, then there will always be some well-defined mathematical problems that it can't resolve one way or the other (in fact, we can show that there will be certain polynomial equations for which such an algorithm can't determine whether or not they have solutions). These problems will then be, in a sense, "absolutely unsolvable," since the human mind can't solve them, and no algorithm that we can see to be sound can solve them either.

So, summing up, it does seem that the Gödel theorems face us with an interesting dilemma. Either: i) all truths are, as such, knowable (with mathematical certainty) by the mind – and then the mind is not a

formal system and no formal system can do what it does in knowing some of these truths; or ii) there are truths that are "absolutely unknowable," not only by any computer but by us as well (or at any rate, if we are suspicious of this notion of truths, well-defined problems that are *absolutely* unsolvable). The first, which Gödel favored, is the "optimistic" conclusion: there is something quite unusual and remarkable about human mathematical cognition, some kind of innate "ability" or capacity of insight that exceeds any possible machine or mechanical procedure. The second, which Turing favored, is the "pessimistic" conclusion: on this conclusion, the only implication that Gödel's theorem has is that we *cannot* completely survey our own cognitive abilities, that there is an inherent limit to how much we can know about ourselves. But as to which of the two we should favor, that still seems to be anybody's guess.

Idealization and the Hypothesis of Mechanism

There is, though, one more important issue that we should consider before leaving the question of mechanism. This is the issue of whether "the mechanistic hypothesis" is even so much as *coherent*. Throughout the discussion so far, we've assumed that it is: that is, that the hypothesis that "the mind is a Turing machine" or that its workings are "completely captured by a formal system" is, whether true or false, at least meaningful. But it's actually not at all evident that this is the case; and if it isn't, then mechanism is wrong, not so much because it's *false* as because it's *incoherent*.

To begin with, let's consider what we *could* mean by the hypothesis that the mind of some particular thinker, P, is a Turing machine or is capturable by a formal system. Which one? Well, when we talk about a formal system (or a Turing machine) we're talking about something that has the capacity to produce an *infinite number* of results, and we're trying to model the structure of this capacity. But of course any human being will only ever come out with (or "discover") a finite number of truths. So to think about the mind of such a thinker as implementing a Turing machine at all, we're going to have to idealize and talk about all the truths he *could* prove if he were to live for an infinitely long time. Moreover, this is going to have to include truths that can only be proved, using that system, by means of proofs that would take longer than one billion years to state, or that would take up more atoms than there are in the universe to write down, etc. So we're going to have to abstract from *these* limitations as well. Additionally, in order to talk about the "formal system" that's behind my actual performance in mathematical reasoning, we'll have to be able to distinguish in a motivated way between those of my utterances that do, and those that don't, actually reflect the operation of this "formal system." Obviously, not everything I say is supposed to be the outcome of mathematical reasoning, and even if I say that something is known to me as the result of following a reasoning procedure, I might be mistaken (either about the procedure I followed or indeed about its truth). We also have to distinguish the utterances that are meant to express knowledge at all from those that are not, for instance those uttered when I talk in my sleep or random sounds that happen to form in my mouth. So we're going to have to presuppose that what we're talking about as resulting from the "formal system" is just the

"truths known through reasoning," though it's not clear for various reasons how to separate these out from everything else that I say.

What's more, even in terms of the (finite number of) things we actually *do* say or believe, it's almost certainly true for each of us that there's at least one inconsistency in there somewhere: as human beings, we all are at least a little bit inconsistent. So suppose (as is surely not implausible) that at some point I believe (or state, or whatever) both A and ~A for some A. If both statements are the result of my "mathematical reasoning procedure," then that procedure is inconsistent, and so (assuming that anything follows from a contradiction), it can't really be a "reasoning procedure" at all, since it's unsound! So we're going to have to make a choice as to which one (A or ~A) to treat as the "actual" result of my "formal system," presumably picking the one that is, upon reflection, the one that I *should* have come out with, i.e. the one that's true.

These considerations can be taken (and are sometimes taken) even further. If it's not possible, without some degree of idealization, to establish on the basis of the actual performances of any human thinker *which* formal system that thinker is "actually" implementing, then the same goes for actual machines as well! That is, even for actual computers, like PCs and MACs, there's no unique way to determine "which" Turing machine they're implementing. For after all, in order to determine this, we'd have to distinguish between performances that are "right" according to the procedure and those that are "wrong," and abstract from the possibility that there are power surges, machine parts breaking, etc. that yield "unintended" behaviors. In a certain sense, then, we have to idealize even in order to establish that mechanism is true of *machines*!

It might seem as if, in the case of actual computers at least, we can just "read off" the unique Turing machine they're implementing from their software: just look at the code that was programmed in to find out which Turing machine they're implementing. But even this isn't as straightforward as it seems. The problem is that, as Searle and other philosophers have pointed out, even for explicitly programmed machines, there's a sense in which every machine is implementing any program you like, on *some* interpretation of that program. For instance, if I suitably (and by a highly non-trivial interpretation scheme) interpreted all the behavior of my computer, including the whirring noises from the hard drive, the flickering of the screen, etc., in the right way, I could argue that it's "implementing" any abstract algorithm that I like. So even "reading off" the explicit program isn't going to work, since even *that* only determines a certain Turing machine under a certain interpretation.

The upshot is that we have to make all kinds of idealizations whenever we talk about *any* actually existing system "as" or as "implementing" a formal system or Turing machine. Lucas argues that all of these idealizations are legitimate, and indeed it's true that we make these kinds of abstractions and idealizations quite often in talking about mathematics, and indeed about human abilities. For instance, it's quite standard in doing cognitive science to presuppose a distinction between "actual performance" and "ideal competence"; the idea is that what we're really trying to find out when we analyze various human behaviors and capacities is the structure of the underlying rules and algorithms that determine what we *can* do, not what we actually *do* do. What's more, there's a sense in which we draw this

distinction, implicitly at least, *whenever* we try to establish mathematical truths by proof at all. For after all, whenever we do this, we're presenting these claims as the outcome of a system of reasoning, and we're presenting that system of reasoning as at least presumably sound. Whenever we criticize somebody else's conclusions as incorrect, we're presupposing a common (and sound) system of mathematical reasoning. In fact, it's not clear what it would mean to claim to have "established" a mathematical truth unless we have something like a proof of it in a system that we take (implicitly at least) to be sound. So one claim that the mechanist can make here is that even if there is a problematic "idealization" required even to make sense of the claim of mechanism, this is an "idealization" that we *must* make if we're going to talk about the structure of mathematical reasoning at all.

Now, however, there's another problem. For if the mechanist argues that we *must* be able to idealize in this way in order to talk about what we "can" rationally establish in mathematics at all, then the mechanist's thesis (i.e. that all mathematical truths that we can establish can be established by a formal procedure) turns out to be tautological. Now this is just the thesis that what can be established by a formal system can be established by a formal system; and to argue in this way would just be, obviously, to restrict the notion of mathematical ability in a way that would beg the question against the Gödel-Lucas-Penrose anti-mechanist position. What's more, if we *do* argue this way – and hence opt for the "pessimistic" interpretation of the implications of Gödel's theorems (above), we'll have to accept that *even though* in a sense we "must" be able to consider the ideal structure "actually underlying" our mathematical reasoning, even if we do so, we will never *know* what it "actually is." So even if we consent to all the idealization we must perform in order to make the mechanist hypothesis (so much as) coherent, we still will certainly never have what we wanted out of an "ideal competence" description of our reasoning, say of the kind we're looking for in cognitive science. So either way, it seems as if a certain kind of anti-mechanist (or at any rate, anti-cognitive-science) result is established by Gödel's theorems, after all: namely that *even if we are* (in some sense) "actually" following sound procedures to reach mathematical truths, there is no way we will ever come to know *what* these procedures "actually" are.