PY4601 Paradoxes: Recent work on the Liar Paradox

Greg Restall, gr69@st-andrews.ac.uk

Lecture 10

In this class, I will introduce some ideas an constructions from *logic* that have been developed with an aim to give insight into the behaviour of the truth predicate in the light of the liar paradox. In particular, I will explain Kripke's fixed point construction to give *models* for a theory of self-referential truth. This should help you come to terms with the technical parts of Kripke's "An Outline of a Theory of Truth", (hereafter, OTT) so we can discuss the underlying *philosophy* of Kripke's position in our tutorial this week.

Then, in my second class (Week 11), we'll look at a variety of perspectives on what on this might *mean*, more generally, noting *other* (non-Kripkean) perspectives which use broadly Kripkean insights for a very different end. We will wrap up with a consideration of what scope there might be for a broadly *neutralist* perspective on truth and other concepts prone to self-referential paradox.

Formalising one paradoxical argument

Here is one version of the Liar Paradox:

> The liar paradox runs as follows. Consider a sentence that says of itself that it is not true:
>
> (λ): (λ) is not true.
>
> Suppose first that this sentence is not true. Then, since this is what it says, it is true after all. So supposing that (λ) is not true, we can conclude that (λ) is true. But if (λ) is true, then what (λ) says is the case. And what (λ) says is that it is not true. So (λ) is not true. But this contradicts what we had already concluded, namely that (λ) is true.
>
> Matti Eklund "Deep Inconsistency", *Australasian Journal of Philosophy* **80** (2002) 321-331 (from page 321)

One delightful (and infuriating) feature of the Liar paradox is that it is so *simple*. There are very few inference steps involved in the argument to the contradictory conclusion. I spell them out in detail here, and show how they combine to produce the argument to the contradictory conclusion.

The inference principles are

- **Negation elimination** $(\neg E)$, according to which, a statement A and its negation $\neg A$ are *contradictory*.
- **Negation introduction** $(\neg I)$, according to which, you can prove a negation $\neg A$ by first assuming the negand A an deriving a contradiction from this assumption.
- **Introduction for the truth predicate** (TI), according to which, you can infer $T\langle A\rangle$ from A.
- **Elimination for the truth predicate** (TE), according to which, you can infer A from $T\langle A\rangle$.
- **Identity elimination** $(=E)$, the principle of "indiscernibility" of identicals. We appeal to this when it is applied to the truth predicate. I.e., if $a = b$ and a is *true*, then so is b.

$$\frac{\neg A \quad A}{\bot}\,\neg E \qquad \frac{\begin{matrix}[A]^i \\ \vdots \\ \bot\end{matrix}}{\neg A}\,\neg I^i \qquad \frac{A}{T\langle A\rangle}\,TI \qquad \frac{T\langle A\rangle}{A}\,TE \qquad \frac{a=b \quad Fa}{Fb}\,{=}E$$

Each of these rules has the virtue of isolating a particular concept and describing some facet of what we can do with that concept, either as an *elimination* rule (what follows *from* the use of the concept) or an *introduction* rule (what we need to prove to be in a position to *conclude* a claim using that concept).

This is starkest in the case of the identity rule: it is about *identity*, and not any other concept at all. (The F in the rule stands in for any predicate. It is not restricted to any predicate in particular. You could press the case that Fa not only uses the concepts involved in the expression a and the expression F but also the *copula* that connects F to a, and if that is a concept involved in the rule for identity.)

The negation rules feature this strange marker: $\bot$. It can simply read as "that's a contradiction". I prefer thinking of it as not another statement which could be true or false, but something we say when we have reached a dead end in our reasoning. In the $\neg E$ rule, if we manage to prove $\neg A$ and also prove A (from various assumptions), we can say: well, *that can't happen*, and we can treat what we have got to as a *refutation* of one or other of our assumption we have appealed to on the way to this dead end. This makes sense of the $\neg I$ rule, which says: if I do reach a dead end (this kind of contradiction) in my reasoning, I can look among the assumptions and choose one and blame *that*. Since A cannot hold (along with the other assumptions), then given the other assumptions, we have $\neg A$.

The truth rules are simple, too, except they use the notion of *quotation*. For every sentence A we presume we have a singular term $\langle A\rangle$, which we can think of as A surrounded by quote marks. While A is the kind of thing we can assume in a proof, and we can prove or disprove, conjoin with something else, etc., $\langle A\rangle$ is something we can predicate properties of. A is a sentence which we can use to assert something. $\langle A\rangle$ is a singular term which refers to an object, namely, the sentence A.

With that squared away, the truth rules allow us to introduce quotation names and the truth predicate in one inference, and to reverse this inference. If $2+2=4$ then the sentence $\langle 2+2=4\rangle$ is *true*. Conversely, if $\langle 2+2=4\rangle$ is *true*, then $2+2=4$.

Using these five rules alone, we can start from the assumption that we have a sentence λ which says of itself that it is not true (i.e., a λ where we have $\lambda=\langle\neg T\lambda\rangle$), and we can prove a contradiction:

$$
\begin{prooftree}
\AxiomC{$\lambda=\langle\neg T\lambda\rangle$}
\AxiomC{$[T\lambda]^1$}
\RightLabel{$=E$}
\BinaryInfC{$T\langle\neg T\lambda\rangle$}
\RightLabel{TE}
\UnaryInfC{$\neg T\lambda$}
\AxiomC{$[T\lambda]^1$}
\RightLabel{$\neg E$}
\BinaryInfC{$\bot$}
\RightLabel{$\neg I^1$}
\UnaryInfC{$\neg T\lambda$}
\AxiomC{$\lambda=\langle\neg T\lambda\rangle$}
\AxiomC{$\lambda=\langle\neg T\lambda\rangle$}
\AxiomC{$[T\lambda]^2$}
\RightLabel{$=E$}
\BinaryInfC{$T\langle\neg T\lambda\rangle$}
\RightLabel{TE}
\UnaryInfC{$\neg T\lambda$}
\AxiomC{$[T\lambda]^2$}
\RightLabel{$\neg E$}
\BinaryInfC{$\bot$}
\RightLabel{$\neg I^2$}
\UnaryInfC{$\neg T\lambda$}
\RightLabel{TI}
\UnaryInfC{$T\langle\neg T\lambda\rangle$}
\RightLabel{$=E$}
\BinaryInfC{$T\lambda$}
\RightLabel{$\neg E$}
\BinaryInfC{$\bot$}
\end{prooftree}
$$

Locating different positions on this map

With this proof spelled out in such detail, we can categorise various positions on the paradox by way of distinguishing where they locate an error in the proof:

- ***Level* solutions** object to the assumption $\lambda=\langle\neg T\lambda\rangle$ that there is a sentence which says of itself that it is not true.
- ***No-proposition* views** (according to which any liar sentence does not express a proposition, and so, does *not* actually say of itself that it is not true) *also* object to this assumption, because on this view,

$\neg T\lambda$ can be truly said. It is just not truly said *by* the sentence λ.

- **Truth-value *gap* views** reject the inference rule $\neg I$. They agree that $\lambda = \langle \neg T\lambda \rangle$ and $T\lambda$ are inconsistent, but you cannot infer from this that $\neg T\lambda$ follows from $\lambda = \langle \neg T\lambda \rangle$.

- **Truth-value *glut* views** reject the inference rule $\neg E$. For friends of truth-value gluts, A and $\neg A$ need not be incompatible.

- **Revisionary projects** about truth reject either TI or TE, perhaps by way of restricting one or the other inference,[1] or by rejecting the use of the truth predicate altogether in favour of a pair of predicates, one of which satisfies the introduction rule and the other of which satisfies the elimination rule.[2]

- I know of no view in the literature that objects to the proof at the $=E$ inference. (Surely there is something to be explored there, though I can't quite see what it would be.)

Some of these diagnoses (e.g. the *gap* and *glut* views) analyse the liar paradox in terms of the distinctive behaviour of *negation*, saying that our traditional understanding of negation is shown to be incorrect by the paradoxical argument. Other diagnoses (e.g. the *level* and *revisionary* views, and perhaps no-proposition views, too) locate the culprit in the distinctive behaviour of the concept of *truth*. Perhaps one or other option seems plausible (after all, negation and the concept of truth seem to play an essential role in the paradoxical argument), but there are strikingly similar paradoxes that employ a different family of concepts. One is *Curry's paradox*, which Patrick mentioned last week.

Formalising ***a different*** self-referentially paradoxical argument

Curry's paradox is formulated not using negation, but using the *conditional*.

I will give an example of Curry's paradox here, but instead of using the concept of *truth*, I will use the concept of *property possession* (or, if you prefer, *class membership*[3]).

The idea is simple. For any sentence $A(x)$ in which the variable x is free, we can think of the property of *being a thing that satisfies* $\langle A(x) \rangle$. For example, given the sentence "x is green" the property of *being a thing that satisfies* $\langle x$ *is green*$\rangle$ is simply the property of being green. We introduce a shorthand notation $\{x : A(x)\}$ to talk about such a property, and we say $b \in \{x : A(x)\}$ to say that b is one of the things that has that property.

We can talk *about* properties in the same way that we talk about anything else. In particular, properties can *have* properties, like other things can have properties. Every property has the property of *being a property*, as one example.

Some very plausible rules of property exemplification go like this: if $A(t)$ holds, then t exemplifies the property $\{x : A(x)\}$. And conversely, if t exemplifies the property $\{x : A(x)\}$ then we have $A(t)$. The green things are all and only the things that exemplify the property of being green.

That motivates the property exemplification rules below, which, together with the very straightforward rules for the conditional, suffice to derive a version of Curry's paradox.

$$\frac{A \to B \quad A}{B}\,{\to}E \qquad \frac{\begin{matrix}[A]^i \\ \vdots \\ B\end{matrix}}{A \to B}\,{\to}I^i \qquad \frac{A(t)}{t \in \{x : A(x)\}}\,{\in}I \qquad \frac{t \in \{x : A(x)\}}{A(t)}\,{\in}E$$

We work with a very simple (but, admittedly, bizarre) property. We choose an arbitrary proposition p and consider this property: being a thing such that *if* it exemplifies itself, then p. For short, we call this property c. The $\in E$ and $\in I$ rules when applied to $c \in c$ have a very interesting structure:

Let 'c' be shorthand for '$\{x : x \in x \to p\}$', for some given p.

$$\dfrac{c \in c}{c \in c \to p}\,{\in}E \qquad\qquad \dfrac{c \in c \to p}{c \in c}\,{\in}I$$

With these instances of $\in E$ and $\in I$ we can reason like this:

$$\dfrac{\dfrac{\dfrac{\dfrac{[c \in c]^1}{c \in c \to p}\,{\in}E \quad [c \in c]^1}{p}\,{\to}E}{c \in c \to p}\,{\to}I^1 \qquad \dfrac{\dfrac{\dfrac{\dfrac{[c \in c]^2}{c \in c \to p}\,{\in}E \quad [c \in c]^2}{p}\,{\to}E}{c \in c \to p}\,{\to}I^2}{c \in c}\,{\in}I}{p}\,{\to}E$$

The result is a proof of the proposition p, which was whatever proposition we cared to choose in the first place. We can prove *anything* we like, from no premises at all. That means our rules have gone too far. But where?

Noticing the parallels

This proof has a remarkably similar structure to the paradoxical liar proof. Notice, though, that it uses *none* of the same inference rules. If this is the "same" kind of problem under a different guise, then we have an opportunity to refine our diagnosis, to give a more general account that can help in a wider range of settings. Thankfully, some of diagnoses of the liar paradox given above *do* generalise in a relatively natural way:[4]

- ★ ***Level* solutions** object to the assumption that the statement $c \in c$ is well-formed. Properties may have properties, but properties at level n have higher-order properties at level $n+1$.

- ★ ***No-proposition* views**: perhaps to generalise this view to the *property* case you would have to reject the coherence of the definition of the putative property $\{x : x \in x \to p\}$ on grounds other than *levels*.

- ★ **Truth-value *gap* views**: to generalise this to the Curry-paradox setting, you would reject $\to I$ in parallel with $\neg I$. The motivation for this seems, as yet, unclear… But notice that this doesn't seem to be about *truth values* any more. What could motivate the rejection of $\to I$?

- ★ **Truth-value *glut* views**: to generalise these it would seem that the natural place to object here would be $\to E$ in parallel with the inference rule $\neg E$. But $\to E$ is *modus ponens*, and can we do without *that*? (This is the crux of the Restall/Priest disagreement, mentioned in footnote 4.) Again, the connection with truth value gluts is not so clear.

- ★ **Revisionary projects** can take a stricter view of property formation and application, rejecting one or other of $\in I$ and $\in E$ in just the same way that they reject either TI or TE.

Notice that in *each* of these responses, we need to do surgery on different concepts, whether *properties* and *exemplification* or *conditionality*, in addition to the concepts of *truth* or *negation*. It would be good to have some kind of guiding principles to help us in such surgery, and to formulate those principles in terms that are independent of any particular guise in which this kind of malady arises.

Kripke's Model: why it's needed

Kripke's insight in OTT[5] is one way to attempt to articulate the idea that a key diagnostic tool in all of these paradoxes of self-reference is the role of *groundedness* and *ungroundedness*. Paradoxical sentences (whether the liar paradox or Curry's paradox or the truth-teller) send us into a kind of ungrounded process of evaluation in which our everyday rules spin their wheels and never come in contact with something "outside" which might ground the evaluation. The heavy-handed way to deal with this is to impose levels and never permit even the *possibility* of ungrounded claims. But Kripke's construction shows that this might be a more aggressive treatment than necessary. The aim is to show that there is some way to interpret sentences in a language involving the truth predicate in such a way as to make sure that each sentence A and $T\langle A\rangle$ always the same truth value, and which allows for sentences like λ which are liar paradoxical.

You should first think about why the naïve first thought for how to interpret the truth predicate...

> ? Assign $T\langle A\rangle$ whatever truth value you assigned A: that is, for any given model m, simply make sure that $m(T\langle A\rangle) = m(A)$.

...*cannot* work as a guide to assigning truth values.

Here is why: Consider the truth-teller sentence τ for which we have $\tau = \langle T\tau\rangle$. What advice does this rule give us about the value of $T\tau$? Since $\tau = \langle T\tau\rangle$, we assign $T\tau$ the same value as $T\langle T\tau\rangle$, and the rule says this should have the same value as $T\tau$. In other words, the advice has simply gone around in a very tight circle and does not actually tell us the value of $T\tau$! The case is worse for the liar sentence: What do we assign $T\lambda$? Since $\lambda = \langle\neg T\lambda\rangle$, we assign $T\lambda$ the same value as $T\langle\neg T\lambda\rangle$, and the guideline above tells us that you should assign this the same value as $\neg T\lambda$. The guideline, in other word, tells us to go around in a circle, and to assign $T\lambda$ the same value as $\neg T\lambda$. In traditional two-valued logic we can't do this, but maybe we *can* do this if we have more values to apply. But the guideline *doesn't* tell us to assign $T\lambda$ a truth value on the basis of something *simpler* (which is *already* determined). It tells us to assign the truth value of $T\lambda$ in terms of the value of something *more complex* (in this case, $\neg T\lambda$). This *might* be possible, but there is, as of yet, no guarantee that assigning truth values in this way is at all possible.

Kripke's model: $\{0, n, 1\}$ and refinement

Let's step back and think about how we might do assign an interpretation to the truth predicate.[6]

First, let's be clear about models. Each *model* assigns values to sentences in our language. In Kripke's construction, in particular, each sentence is assigned one of the *three* values 0 (for falsity) 1 (for truth) and n (for a third status, which you can think of as "neither true nor false" for the moment). The crucial idea in Kripke's construction is that we think of n as "less specific" than either 0 or 1, and 0 and 1 as both being *more defined* than n.

So, there is a kind of "ordering" among the values where $n \sqsubset 0$ and $n \sqsubset 1$ but no other ordering obtains between the values (so 0 and 1 are *incomparable* by $\sqsubset$.) The strict order $\sqsubset$ induces the order $\sqsubseteq$ (which stands to $\sqsubset$ as $\leq$ stands to $<$ for numbers), where we define $x \sqsubseteq y$ to hold if and only if $x \sqsubset y$ or $x = y$. If $x \sqsubseteq y$ then we think of y as being at least as *defined* as x, or y as being a possible update to x which might (or might not) resolve the less specific value n into a more specific 0 or 1, but a 0 or 1 cannot be resolved to anything other than the value it is.

We can interpret the logical concepts in our language by way of "truth tables" with respect to these three values. One way to specify them is by way of the following rules, which specify the conditions under which a sentence gets the value 0 or 1 (so a sentence gets the value n if the other clauses do not give it a value):

- $m(\bot) = 0$ always.
- $m(\neg A) = 1$ iff $m(A) = 0$; $m(\neg A) = 0$ iff $m(A) = 1$.

- $m(A \wedge B) = 1$ iff $m(A) = 1$ and $m(B) = 1$; $m(A \wedge B) = 0$ iff $m(A) = 0$ or $m(B) = 0$.
- $m(A \vee B) = 1$ iff $m(A) = 1$ or $m(B) = 1$; $m(A \vee B) = 0$ iff $m(A) = 0$ and $m(B) = 0$.
- $m(A \to B) = 1$ iff $m(A) = 0$ or $m(B) = 1$; $m(A \to B) = 0$ iff $m(A) = 1$ and $m(B) = 0$.

These values can be summarised in a table like this:

A	B	$\bot$	$\neg A$	$A \vee B$	$A \wedge B$	$A \to B$
0	0	0	1	0	0	1
0	n	0	1	n	0	1
0	1	0	1	1	0	1
n	0	0	n	n	0	n
n	n	0	n	n	n	n
n	1	0	n	1	n	1
1	0	0	0	1	0	0
1	n	0	0	1	n	n
1	1	0	0	1	1	1

The idea is simple: a complex formula gets a traditional value (0 or 1) just when its components are determined *enough* (by having values 0 or 1) to assign the value required. So, e.g., a conjunction is *true* when both conjuncts are true. It is false when one conjunct (at least) is false. And in the remaining cases, it is left n.

Think of a *model* of the language as assigning values 0, n and 1 to every sentence of the language, in such a way that these rules are respected.

Each connective in the language respects the order $\sqsubseteq$ in the following way: Think of models as ordered by refinement: $m_1 \sqsubseteq m_2$ iff $m_1(p) \sqsubseteq m_2(p)$ for each atom p (so m_2 differs from m_1 only by resolving some atoms that were valued n by m_1 to be valued 1 or 0 by m_2, but it never changes a 1 to a 0 or vice versa), then this fact is extends to the entire language: $m_1(A) \sqsubseteq m_2(A)$ for every formula A. (I leave verifying this to you as an exercise if you want to work through the details, or consult my *Proofs and Models in Philosophical Logic*, p. 44, 45 to see how this is done.)

Kripke's model: stages and the fixed point

Now, we'll see how we can define a model m_* that will, as a matter of fact, always assign A and $T\langle A\rangle$ the same truth values. The process of doing this will respect the intuition that if the process of evaluating a sentence of the form Tx is *grounded*, it will be assigned one of the truth values 0 or 1, but *ungrounded* sentences might be assigned the value n.

Start with some model m_0 that interprets the *entire language* in whatever way we please, except that each sentence of shape Tx is assigned the value n. At stage 0, the truth predicate is totally *undefined*. We will be constructing m_0, m_1, m_2, etc… for very many stages of evaluation.

Now, given any model m_i that we have defined so far, we will define the *next* model in the series, m_{i+1}. We assign its atomic sentences in just the same way as m_i does, except for the T-sentences. Here, we assign $T\langle A\rangle$ the same value that the model m_i assigns to the sentence A. The idea is that if A has been assigned a value 0 or 1 in the process of evaluation, then at the *next* stage (at least), we can assign $T\langle A\rangle$ that value, too.

The crucial feature of this process is that for every sentence at all, $m_i(A) \sqsubseteq m_{i+1}(A)$. That is, the values assigned to sentences as we go along this series are *more and more refined*. If A has is assigned some value n at stage i and then 1 (say) at stage $i+1$, then at stage $i+2$, $T\langle A\rangle$ gets value 1 too, and from then on, its truth value is settled and does not change any more. Sentences flip from n to 0 or to 1, but once a sentence has the value 0 or 1 it is fixed from then on.

Now, this process can go on infinitely far. (For any number i, you can formulate a sentence with i 'T's prefixing some sentence, which will only get its value at stage m at the earliest.) In fact, given the sequence $m_0, m_1, m_2, \ldots$ going on for every natural number, we can define a new model m_ω that collects together all the models we've made before! ($m_\omega(p) = 1$ iff $m_i(p) = 1$ for some finite i; $m_\omega(p) = 0$ iff $m_n(p) = 0$ for some finite i, and $m_\omega(p) = n$ otherwise.) In fact, if the language we are evaluating is expressive enough, we might not stop at level ω, but have to continue. After all, we can define $m_{\omega+1}$ by applying the same rule as before, and assign a settled value to some T-sentence $T\langle A\rangle$ where A was first settled at level ω.[7]

Will this *ever* stop, or will it go on for ever and ever with no place to stop at all? This is the point at which things get even more technically challenging, and you can skip these details unless they interest you. The process we have defined so far for every natural number, and beyond to ω, $\omega+1$ and so on, can be extended to every *ordinal* number, way off into the transfinite. (The idea is that you can keep on adding one, and do that indefinitely, and at each infinite stack of adding ones, you sum everything up into a *limit*, so we have $\omega+\omega$, $\omega+\omega+1$,... and onto $\omega+\omega+\omega$, so following the pattern, we have $2\times\omega$, $3\times\omega$, $4\times\omega$,..., and onto $\omega\times\omega$,... and way beyond *that*, too.) How far off into the transfinite? Well, there are *so many* of these ordinals that there is an important sense in which there are *more* ordinals than sentences in any given language we are interpreting in our models.

This is important, because as we define this humongously large "sequence" of models m_0, m_1, $\ldots$, m_κ, $\ldots$ at each and every stage along the way the models are getting more and more refined, at least in the sense that we have

$$m_0 \sqsubseteq m_1 \sqsubseteq \cdots \sqsubseteq m_\kappa \sqsubseteq m_{\kappa+1} \sqsubseteq \cdots$$

and in such a humongous sequence of models once a sentence is given the value 0 (or the value 1) it keeps it, forever more. Given that there are more steps along the way in this "sequence" than there are formulas in the language, *eventually* we have to run out of formulas to be assigning these settled values. We will eventually have a pair of models m_κ and $m_{\kappa+1}$ which *agree* in their values, and once that is done, the sequence stops: $m_{\kappa+2}$ will assign to the T-sentences $T\langle A\rangle$ the values that A was assigned in $m_{\kappa+1}$ and these are, by hypothesis, just the same as the values that A was assigned in m_κ, which determines the values that $T\langle A\rangle$ has in $m_{\kappa+1}$. So, every T-sentence is fixed in value from $m_{\kappa+1}$ to $m_{\kappa+2}$, and the *other* atomic sentences were unchanging, so *every* sentence has the same value from $\kappa+1$ to $\kappa+2$, and so, things are fixed from then on.

Such a model is called a *fixed-point* of the sequence, and fixed points are special. They assign *exactly* the same value to A and to $T\langle A\rangle$ for every sentence A, ensuring that in a very tight sense, the TI and TE rules hold in full generality: the sentences are assigned the same semantic values. We did this by allowing *all* grounded sentences (any sentence for which there is any conceivable process of evaluation that can interpret $T\langle A\rangle$ by first assigning a value, 0 or 1 to A) to eventually be assigned a settled truth value—no matter how long that process might take—leaving only the remainder to have the unsettled value n. This is a *model* of some way that truth can be understood without resorting to levels or otherwise restricting the truth rules in their full generality.

The Upshot

What can we learn from this exercise?

Whatever we've done, it's both *less* and *more* than a "solution" to the original liar paradox. It's much less than a "solution" because there is so much more to say than just accepting that sentences are assigned truth values as they are in a fixed point of a scheme like Kripke has constructed. If we think of the values 0 and 1 as "false" and "true", then I suppose something assigned as n is neither true nor false, and if we say that in our preferred model of the liar sentence $m(T\lambda) = n$ and so, the liar sentence is neither true nor false, which seems to commit us to saying that the liar sentence is not true, which is just what the liar sentence says. We have a simple revenge paradox, which itself needs an answer.

So, we need to do more to use this to comprehensively diagnose the liar paradox and its cousins.

But there is another sense in which its much *more* than just an account of the liar paradox. Nothing in the reasoning we looked at said anything special about *negation*. The key idea was that *all* the logical concepts in the construction preserved the refinement order. This constrains the interpretation of negation, but it *also* equally constrains the interpretation of the conditional, so Curry-paradoxical sentences are interpreted in just the same sort of way as liar paradox sentences.

Similarly, the groundedness interpretation for *truth* applies equally well to *properties* (we say $t \in \{x : A(x)\}$ is assigned the same value at stage $i+1$ that $A(t)$ is assigned at stage i), and exactly the same story can be told here. The crucial insight of the construction is the stages of construction, the refinement ordering and a fixed point. The particular details of how these are applied, to any given logical connective, or any concept that threatens ungroundedness, like truth or property abstraction, is mere detail. The point is not a specific one about the semantics of this or that connective, or of this or that concept like truth or property instantiation.

More importantly, though, the *model* gives us reassurance that at the very least of a minimal kind of coherence of a set of rules. The fixed point construction shows that any rules which hold *in that model* do not, in and of themselves, lead us into paradox. A model or a system of models provides a proof of concept that any system of rules that apply in models like that cannot give rise (by themselves) to any conclusion that *does not* also hold in those models. That does *not* mean that no paradoxes can arise, of course, but it is a kind of safety net, saying that if we are happy with limiting ourselves to principles like *these*, our position is at least, in this minimal sense, coherent. And *that*, in this area, is a kind of reassurance that is surprisingly rare.

In the *next* class I will explore what kind of insight we might draw from this when it comes to thinking more generally about the liar paradox and other paradoxes of self-reference.

---

1. See, Stephen Read's "The Liar Paradox from John Buridan back to Thomas Bradwardine" *Vivarium* **40** (2002), 189–218 for a good introduction to these issues. ↩
2. See Kevin Scharp, *Replacing Truth*, OUP, 2013 for an extended presentation of a view of this form.↩
3. I describe all of this in terms of *properties* and *exemplification* rather than *classes* and *membership*, but the reasoning is exactly the same in either case. Classes satisfy an extra condition, *extensionality*, not satisfied by properties. If C and D are classes with exactly the same members, then they are identical; while P and Q might be *distinct* properties exemplified by exactly the same things. The class of equiangular triangles in the Euclidean plane is exactly the same as the class of equilateral triangles in the Euclidean plane. However, the property of being an equilateral triangle need not be identical to the property of being an equiangular triangle. Since extensionality is not involved in Curry's paradox, we need not worry about whether we are reasoning about classes or properties.↩
4. To track down some of the considerations here around the costs and benefits of uniformity in diagnoses of the paradoxes of self-reference, see the critical discussion between Graham Priest "The Structure of the Paradoxes of Self-Reference" *Mind* **103** (1994), 25–34, and Greg Restall "Deviant Logic and the Paradoxes of Self Reference" *Philosophical Studies* **70** (1993), 279–303. (Although published earlier than Priest's 1994 paper, Restall's paper was written partly in response to an earlier version Priest's paper, which was presented at the 1991 *Australasian Association for Logic* conference.)↩
5. This insight was not solely had by Kripke. Amazingly, the same idea was formulated by Brady in 1971, Gilmore in 1974 and Martin and Woodruff in 1975, all apparently independently. See Greg Restall, *Proofs and Models in Philosophical Logic*, Cambridge University Press, 2022 (page 44) for references.↩
6. It will turn out that what we do with the truth predicate could *equally* apply to properties and exemplification, as I will mention below. It is a general technique for how to deal with possibly *circular definitions*. For more on this general theme, and for an exploration of a related approach that treats circular definitions in their generality, I recommend Anil Gupta and Nuel Belnap's *A Revision Theory of Truth*, MIT Press, 1993.↩
7. If the language has quantifiers, and if we have a primitive predicate B that (in the ground model m_0) is true of every T-free sentence, and false otherwise, and a primitive relation R that (in the ground model) holds between ϕ and ψ just when ψ is the sentence $T \cdots T\phi$, that is, ϕ prefixed by some number of T-predicates, then the sentence $(\forall x)(\forall y)((Bx \land Tx \land xRy) \to Ty)$ will only be assigned 1 at level ω, and so, $T\langle(\forall x)(\forall y)((Bx \land Tx \land xRy) \to Ty)\rangle$ will only be assigned 1 at $\omega+1$, $T\langle T\langle(\forall x)(\forall y)((Bx \land Tx \land xRy) \to Ty)\rangle\rangle$ will get its value at $\omega+2$, and so on...↩