PY4601 Paradoxes: recent work on the Liar Paradox

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

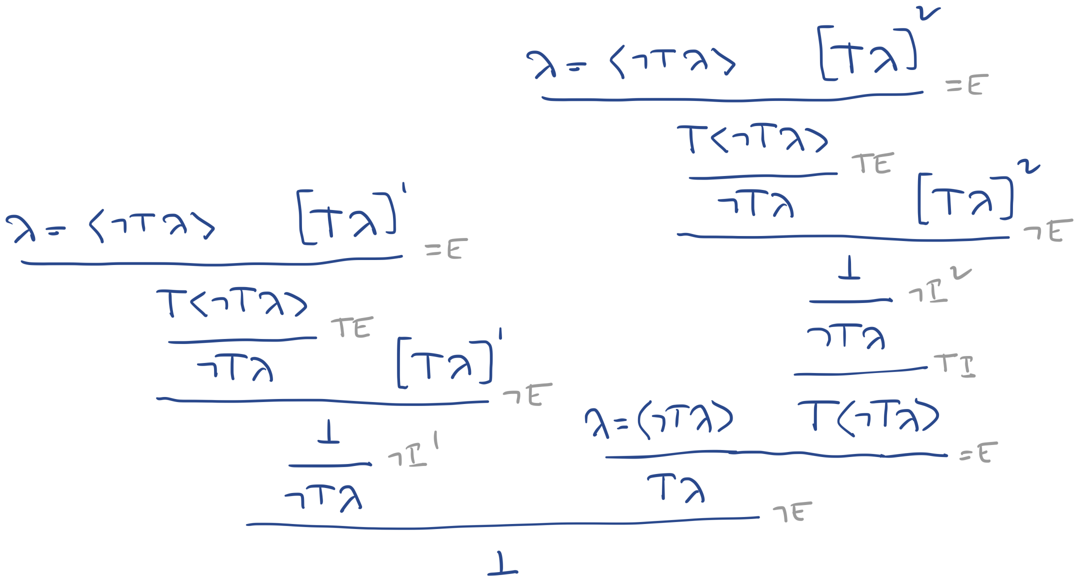
Lecture 11

In my first class (Week 10) I introduced Kripke's fixed point construction and said a little bit about its significance.

In this second class (Week 11), we look at a variety of perspectives on what on earth this construction might *mean*, more generally, noting *other* (non-Kripkean) perspectives which use broadly Kripkean insights for what turns out to be a very different kind of end result.

Then I 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.

As a reminder: here is the liar paradoxical proof.



Three ways to interpret $\{0, n, 1\}$

The fixed point construction is a technique to generate three-valued models for our language. These models assign sentences the values 0, n, and 1. There is more than one way to use these models to *interpret* a language, and so, to give an account of the liar paradox. For that, you need to explain the significance of these three "values" and, in particular, we need to explain the connection between our models and *validity*, if we are to gain any insight into *where* the paradoxical argument breaks down. It turns out that there are a number of different ways to interpret these three-valued models.

Truth-value Gaps (Kripke)

On this view, sentences assigned 1 are taken to be *true*, and sentences assigned 0 are taken to be *false*, and we interpret sentences assigned n as *neither true nor false*. This has been the standard way to understand three-valued models.

That is enough to interpret models as a way of representing what sentences in our language hold. It is *not* enough, by itself, to determine the validity of arguments. For that, we need to say more.

On the standard *two-valued* picture, we might say that an argument is valid if and only if it must be that whenever the premises are *true* so is the conclusion. So, using models, we can say this: an argument is *formally valid* if and only if, any model that assigns 1 to each of the premises of the argument must also assign 1 to its conclusion.

So, let's understand validity in the same way, as preservation of *truth*. Then, of all the steps in the liar paradoxical argument, $\neg I$ is not valid, and the argument breaks down at this point. In particular, the argument from $\lambda = \langle \neg T\lambda \rangle$ and $T\lambda$ to $\bot$ is *valid*(that is, given that $\lambda = \langle \neg T\lambda \rangle$, it is inconsistent to take λ to be *true*), but this does not mean, however, that, for this λ, we can take λ *false*, which is what is required for $\neg T\lambda$ to hold.

This is Kleene's three-valued logic, and it is one way to interpret these models. The story can be extended in different ways.

- For Kripke, sentences assigned n are *meaningful*, but do not express propositions, since they do not have determinate truth conditions. The logic of *sentences* is Kleene's three-valued logic, while propositions behave in the classical manner.
- It is possible, instead, to think of liar paradoxical sentences as not only meaningful, but as expressing *propositions*. In that case, the logic of propositions will allow for truth-value gaps.

There is much to like in this interpretation. However, a large problem looms. The key interpretive claim, the that λ is neither true nor false, cannot be modelled in the theory as something that is itself *true*. The theory seems to undercut itself: the claim that λ is not true—since, according to the theory itself, this *is* the claim λ itself—is to be rejected since it has value n. But this claim, that λ is neither true nor false—and hence, that it is not true—is the central claim of the theory of the paradox itself. The theory undercuts itself at this crucial point. This is the revenge paradox.

So, it seems reasonable to explore alternative *interpretations* of the $\{0, n, 1\}$ structure, and it turns out that there are two approaches that avoid this kind of revenge, by giving an alternative interpretation of the intermediate value n.

Truth-value Gluts (Priest)

We interpret sentences assigned n as *both* true and false. If validity is understood as preservation of truth (that is, an argument is *valid* if whenever the premises are 1 or n then so is the conclusion), then the paradoxical argument indeed breaks down at a different point. Now it is the inference $\neg E$ that is not valid, at least given that we understand $\bot$ as never true. (In particular, the argument from $\lambda = \langle \neg T\lambda \rangle$ and $T\lambda$ to $\bot$ is *invalid*, because as the Kripke construction shows, we can assign $T\lambda$ and $\neg T\lambda$ both the value n (which is enough to count as *true*, on this view), and so, the inference from here to $\bot$ is invalid, since this is, indeed, possible.[1])

The resulting logic is Graham Priest's *logic of paradox* (LP), a well-known *paraconsistent* logic.[2] A logic is said to be paraconsistent (with respect to a given negation concept $\neg$) when a contradictory pair of sentences $A, \neg A$ need not entail every sentence whatsoever. That is, $A, \neg A \nvdash B$.

It is worth pausing at this point and verifying that the evaluation clauses for the logical connectives that are used in the model construction—

- $m(\bot) = 0$ 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 $m(B) = 0$.

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

—still make sense when the value n is interpreted as "both true and false", so 0 can be read as "false only", and 1 as "true only", and it seems, indeed, that these clauses can be seen as appropriate ways to understand the logical connectives, though now the material conditional $A \rightarrow B$ has the unfortunate feature that it can be true (that is, both true and false, i.e., has value n) when the antecedent is true (that is, both true and false, i.e., n) and the consequent fails to be true (that is, has the value 0). It follows that indeed, *modus ponens* fails for this "material conditional" interpreted in this way.

In Priest's account of the liar paradox, the liar sentence λ is both *true* and *false*. He grants both $T\lambda$ and $\neg T\lambda$, and rejects the claim that this pair is *inconsistent* in the strong sense of being not jointly possible.

This LP approach does *not* seem susceptible to revenge problems in quite the same way that the gap approach. Now, the central claim, that the liar is true and it is false, is expressible in the framework, and is indeed expressed in the framework in a claim that is at the very least *true* (or both true and false): by Priest's lights, $T\lambda \wedge \neg T\lambda$ turns out to be (at least), true.

However, the approach does seem to be subject to *expressibility* worries: in particular, it is impossible, in the current framework, to express *exclusion*, given that *negation* no longer has the function of expressing exclusion. Here, $\neg A$ no longer excludes A, at the level of truth, and it is *very* hard to extend the framework with a device expressing exclusion without either (a) undercutting the general motivation for truth-value gluts or (b) falling prey to revenge problems featuring the new notion of exclusion.

For the latter half of the 20th Century, the *glut* and the *gap* approaches mentioned here were the two major interpretations of these three-valued models.

However, we can use these three-valued models in another way, which remains, in an important sense, *neutral* between the gap and glut approach.

Strict-Tolerant Logic (Cobreros, Égré, Ripley, van Rooij)

Instead of making a categorical choice about sentences interpreted as n, we interpret sentences assigned n as *tolerantly* true but *strictly* false. There are (at least) *two* standards for assertion, the *strict* standard—if you like, think of this as avoiding all falsehood, similarly to supervaluationist interpretations of vagueness—and the *tolerant* standard—if you like, think of this as including all truth, similarly to subvaluationist interpretations of vagueness.

I will call this the *ST* approach to truth, and to the liar paradox, where "ST" stands for **Strict/Tolerant.**[3] On this picture, the three values $\{0, n, 1\}$ are interpreted as follows: to assign a sentence the value 0 is to say it is false (at either the strict or tolerant standard), to assign the value 1 is to say that it is false (again, at either standard), and an assertion of a sentence assigned n is taken to be (according to our model), *true* when evaluated to the *tolerant* standard, but *false* when evaluated to the *strict* standard.

Immediately, we can see connections with the gap approach (the gap approach exclusively attends to the *strict* standard), and the glut approach (which exclusively attends to the *tolerant* standard). We will see, though, that this hybrid approach to the paradoxes allows for a new account of the liar paradox, which combines features of the glut picture and the gap picture, but which allows for a new uynderstanding.

Once we have *two* standards for assertion, this raises the question: how are we to understand logical validity? If we ignore the S/T distinction, and focus merely on evaluating assertions strictly, the preservation of truth gives you exactly the same result as truth-value gaps. It's preservation of the value $\{1\}$. Understood in this way, you have the logic of truth-value gaps. If you focus merely on *tolerant* truth, the preservation of tolerant truth gives you exactly the same result as truth-value gluts, the preservation of $\{1, n\}$. We can incorporate either the gap approach or the glut approach with no modifications.

However, another notion of logical validity makes sense, given strict and tolerant standards.

To introduce this alternative, it helps to think first about what it is for an argument to be *invalid*. An argument is invalid if it is (in some sense) OK to assert the premises and deny the conclusion; or equivalently, if it is (in some sense) possible for the premises to be true the conclusion to be false. With the distinction between tolerant and strict assertion in mind, we can now ask the question: is that to be understood *strictly* or *tolerantly*?

If we are to have a *strict* counterexample to an argument, this would be some interpretation that makes the premises *true* (when understood strictly) and the conclusion *false* (also understood strictly). In terms of our models, we say an argument is *invalid*, on this picture, On this view, an argument is *ST*-valid if it is never the case that the premises are all 1 and the conclusion 0) then, believe it or not,[4] *all* of the inference rules in our argument count as valid! (On this interpretation, as with others, $\bot$ always has the value 0, and so, it is neither strictly nor tolerantly true.) In fact on this interpretation of validity, any argument that is logically valid according to classical logic is also valid in ST-logic.[5]

We call this notion of logical validity the ST (for **S**trict/**T**olerant) account of validity. The result is a different way to interpret the same three-valued models. The intermediate value is not simply a *gap* (it is a gap when sentences are evaluated *strictly*) and it is not simply a *glut* (it is a glut when sentences are evaluated tolerantly). It has aspects of *both* interpretations, but it is unlike either.[6]

So, we can keep the truth predicate rules, and keep every classically valid logical inference. That is, frankly, *incredible*. In the ST sense, every *logical* step of the paradoxical proof is *valid*.

However, there is a catch.

So, every classically valid argument is ST-valid, and the truth predicate rules can be also taken to be valid (since they are ST-valid on every model for the Kripke construction). Call the resulting theory, ST$^{\mathsf{T}}$, the ST theory of truth. ST$^{\mathsf{T}}$-validity is not *transitive* in the following sense. The argument from $\lambda = \langle \neg T\lambda \rangle$ to $T\lambda$ is ST$^{\mathsf{T}}$-valid (you can't make the premise 1 and conclusion 0). The argument from $\lambda = \langle \neg T\lambda \rangle$ to $\neg T\lambda$ is *also* ST$^{\mathsf{T}}$-valid (you can't make the premise 1 and the conclusion 0 here, either). Also, the argument from $T\lambda, \neg T\lambda$ to $\bot$ is also ST$^{\mathsf{T}}$-valid, since it's *classically* valid. But you cannot *chain these together* to construct an argument from $\lambda = \langle \neg T\lambda \rangle$ to $\bot$. We have a model (any Kripke fixed point will do) where $\lambda = \langle \neg T\lambda \rangle$ is assigned 1 and $\bot$ is assigned 0, so we have a *counterexample* to the argument. Logical validity, in this ST sense, is not *transitive*.

So, on this view, $T\lambda$ and $\neg T\lambda$ are both *tolerantly* true, and are both *strictly* false. Every inference step in the argument to the contradictory conclusion is at least ST-valid, in that they never have strictly true premises and strictly false conclusions. However, the strict truth of the hypothesis $\lambda = \langle \neg T\lambda \rangle$ is not enough to ensure (thank goodness!) the strict truth of the contradiction $\bot$, which is impossible. Along with the *gap* account, the the step $\neg I$ is where the proof fails to preserve the strict truth of the premises to the strict truth of the conclusion: at the step in which $\neg T\lambda$ is inferred, discharging the assumption $T\lambda$, we can only conclude that $\neg T\lambda$ is at least *tolerantly* true.

So, according to the ST framework, the claim that the liar sentence is neither true nor false is acceptable in the sense of being at least *tolerantly* true ($\neg(T\lambda \vee \neg T\lambda)$ is indeed assigned n), while negation expresses exclusion at the very least in the *strict* sense. It is impossible for A and $\neg A$ to both be *strictly* true—in fact, it is impossible for A to be strictly true and $\neg A$ to be even merely tolerantly true. We must be tolerant in both sides, if we are to grant both A and $\neg A$, in exactly the same sort of way that we would be shifting in our standards if we are to grant both “that’s red” and “that’s not red” in vagueness cases where we classify the same object as red and as not red.

There is much more to do to work out the details of this kind of approach to the paradoxes. In particular, more must be said about these two standards for assertion, and how these relate to the underlying notion of truth. A distinctive feature of this approach is that there is a single truth predicate, rather than two distinct truth predicates (for strict truth and tolerant truth), since it is designed to satisfy the constraint that A and $T\langle A \rangle$ have the same semantic value. If there were *strict truth* and *tolerant truth* predicates, then wherever those two predicates differ in judgement concerning a sentence A, at least one of their semantic values must differ from the semantic value of A. So, this is not a view that distinguishes two kinds of truth. So, the connection between truth and successful assertion must be carefully articulated. There is a plausible sense in which to assert A under the conditions that A is true is to *succeed* (in some sense) in the assertion, even if in a lucky fashion (in the case that I simply guessed or asserted without knowledge). But if assertion has two different standards, one more strict, and one more tolerant, does this mean that there are two different criteria for this kind of success? What *are* these standards, how do they arise, and how, exactly, do they apply? It seems that shifting standards are a fruitful way to understand the paradoxical nature of the truth predicate (and property ascription, and more . . .), but there is much more to be done to spell out the view.

[1] (If, on the other hand, we take $\bot$ to simply mean “some contradiction or other is true” (that is, $\bot$ has value n), then on *this* interpretation, $\neg E$ is valid while $\neg I$ is invalid, just as it is in Kripke's interpretation.)↩

[2] See Graham Priest, “The Logic of Paradox”, *Journal of Philosophical Logic*, **8** (1979), 219–241, for the canonical introduction to LP, or his *In Contradiction: A Study of the Transconsistent* (Edition 2), Clarendon Press, 2006, for a book-length treatment.↩

[3] See Pablo Cobreros, Paul Égré, David Ripley and Robert van Rooij, “Reaching Transparent Truth”, *Mind* **122** (2013), 841–866.↩

[4] That is, each of these rules are valid if (a) identity is constrained so that, at the very least, if $a = b$ is assigned the value 1 then Fa and Fb are never assigned 1 and 0 (or 0 and 1) respectively (or you can impose the much stronger constraint, to the effect that if $a = b$ is assigned the value 1 then Fa and Fb have the same value, and this weaker condition is automatically satisfied), and (b) the truth predicate is constrained so that $T\langle A\rangle$ and A are never assigned 1 and 0 (or 0 and 1) respectively (and similarly, you can impose the much stronger constraint to the effect that A and $T\langle A\rangle$ are assigned the same value, as happens in models generated by the Kripke construction.)↩

[5] Here is why: If we had some three-valued model m which is an ST-counterexample to the argument from X to A, then we have $m(X) = 1$ and $m(A) = 0$. Simply refine m into a two-valued evaluation m' that assigns 1 or 0 to each atom, by picking an arbitrary value for anything previously assigned the value n. Since, by our hypothesis, all the logical concepts are preserved under refinement, in this new model we still have $m'(X) = 1$ and $m'(A) = 0$, and this is a two-valued counterexample to our argument.↩

[6] We can also define a different notion of validity, given the interpretation of the third value. This notion of logical consequence is broadly *tolerant*, in that a *tolerant* counterexample—a way to make the premises tolerantly true and the conclusion tolerantly false—counts as a counterexample to the argument's validity. On this view, *any* argument form has a counterexample. If the atomic sentences are all n, then the complex formulas all have value n, so *every* formula is tolerantly true and tolerantly false. The notion of TS-validity seems less useful than ST-validity.↩