Löb's Theorem and Curry's Paradox

Graham Priest

August 9, 2021

Departments of Philosophy, CUNY Graduate Center, University of Melbourne, and the Ruhr University of Bochum

The technique of self-reference is one that is very familiar to logicians. How to characterise it precisely, is not a simple question, but essentially it is when one takes some procedure and uses it, itself, as its own input.

The technique is something of the Jekyll and Hyde of logic. It is deployed to prove some of the most profound results in mathematical logic, such as Gödel's incompleteness theorems and Turing's halting theorem. However it is also deployed to prove paradoxical contradictions, such as the Liar paradox and Russell's paradox. Moreover, results of these two kinds are as intimately related as Jekyll and Hyde themselves. The difference between the profound results and paradox is often razor-thin.

One place in which this is the case concerns Löb's theorem and Curry's paradox. This pair is perhaps not very familiar to people who are not mathematical logicians. The point of this note is to make it more so, and ponder the significance.

A word on notation for non-logicians:

- $\neg A$ means: it is not the case that A
- $A \vee B$ means: A or B
- $A \to B$ means: if A then B
- $A \leftrightarrow B$ means: A if and only if B
- $X \vdash A$ means: A follows from the premises in X

1 Löb's Theorem: Background

Löb's theorem was proved in 1955 by the German mathematician Martin Löb.[1] The context of the theorem was Gödel's proof of his incompleteness theorems—we now know others as well—and their wake. The theorems concern axiomatic theories of arithmetic, that is theories which concern the natural numbers $(0, 1, 2, ...)$. Gödel showed how the symbols, statements, and proofs of such a theory can be coded as natural numbers. In an era when we are very familiar with the fact that computer programs—which, like proofs, are just sequences of statements of a formal language—can be stored as a binary number (sequence of 0s and 1s) in a computer's memory, this is not now an unfamiliar fact.

This coding allowed statements about formulas and proofs themselves to be expressed as purely number-theoretic statements. What Gödel showed next was that if an axiom system, T, is sufficiently strong (technically, that it can represent all primitive recursive functions), then there is a formula of arithmetic, $Prov(x, y)$ such that if π is a proof of A, $T \vdash Prov(\langle\pi\rangle, \langle A\rangle)$, and if it is not, $T \vdash \neg Prov(\langle\pi\rangle, \langle A\rangle)$. Here, $\langle\pi\rangle$ is the numeral of the code number of π, and $\langle A\rangle$ is the numeral of the code of A. That A is provable in T can then be expressed by the sentence $\exists x Prov(x,\langle A\rangle)$. Let us write this as $\mathcal{P}(\langle A\rangle)$—or just $\mathcal{P}\langle A\rangle$: I will omit double brackets to avoid clutter.

Third, and this is the really cunning part, Gödel showed that there was a formula, G, such that $T \vdash G \leftrightarrow \neg\mathcal{P}\langle G\rangle$. Effectively, G says: 'G' is not provable. So we have self-reference.

Finally, Gödel showed that if T proves G, that is, $T \vdash G$, then T is inconsistent. So if T is consistent, G cannot be proved. At this point, it is not hard to show that if T is consistent, and so G is not provable, G is true. (When I talk of truth here and in what follows, I mean what logicians call truth in the standard model. That is, the interpretation of the language of arithmetic in which symbols get their standard meaning.) Hence there are true sentences that cannot be proved.[2]

Gödel's self-referential construction is quite general, and can be used to show that for any formula, $C(x)$, there is a sentence, A, such that $T \vdash A \leftrightarrow$

[1] Löb (1955). For a modern treatment of the proof, see Boolos, Burgess, and Jeffrey (2007), ch. 8.

[2] For a more detailed discussion of the whole matter, see Priest (2017), chs. 14, 15. For a more technical presentation, see Boolos, Burgess, and Jeffrey (2007), chs. 15–18. See also Berlinski (2019).

$C\langle A\rangle$. Hence arises the question, proposed by the US logician Leon Henkin, of the provability or otherwise of the sentence which says of itself that it *is* provable—that is, a sentence, H, such that $T \vdash H \leftrightarrow \mathcal{P}\langle H\rangle$. It was to answer Henkin's question that Löb proved his theorem, which is as follows:

- For any sentence A, if $T \vdash \mathcal{P}\langle A\rangle \rightarrow A$ then $T \vdash A$

This is a surprising result. After all, if the axioms of the arithmetic are sound, anything provable is true, so every instance of $\mathcal{P}\langle A\rangle \rightarrow A$ is true. But Löb's theorem shows that you can prove it only for those instances for which you can prove A itself.

Anyway, given the theorem, a solution to Henkin's problem follows. Since $T \vdash P\langle H\rangle \rightarrow H$, $T \vdash H$.

2 Löb's Proof

Against this background, let us now look at Löb's proof. This uses three additional statements concerning $\mathcal{P}$. Again, if T is sufficiently strong, these can be proved, as was already clear to Gödel. These are:

[1] If $T \vdash A$ then $T \vdash \mathcal{P}\langle A\rangle$

Roughly: if you can prove A, you can prove that you can prove it.

[2] $T \vdash \mathcal{P}\langle A \rightarrow B\rangle \rightarrow (\mathcal{P}\langle A\rangle \rightarrow \mathcal{P}\langle B\rangle)$

Roughly: if you can prove $A \rightarrow B$ and A then you can prove B.

[3] $T \vdash \mathcal{P}\langle A\rangle \rightarrow \mathcal{P}\langle\mathcal{P}\langle A\rangle\rangle$

Roughly: you can prove that if A is provable then it is provable that it is provable.

Löb's proof can be put it a few different ways, but they all come to much the same thing. One way of putting it goes as follows. The proof uses a few assumptions about the underlying logic of T. I will note these as we go along. *Modus ponens* is the inference $\{A, A \rightarrow B\} \vdash B$. Contraction is the $\{A \rightarrow (A \rightarrow B)\} \vdash A \rightarrow B$.

Suppose that $T \vdash \mathcal{P}\langle A\rangle \rightarrow A$. By Gödel's self-referential construction, we can find a sentence, L, such that:

- $T \vdash L \leftrightarrow (\mathcal{P}\langle L\rangle \rightarrow A)$

(Take $C(x)$ to be $\mathcal{P}(x) \to A$.) So:

- $T \vdash L \to (\mathcal{P}\langle L\rangle \to A)$

By [1]:

- $T \vdash \mathcal{P}\langle L \to (\mathcal{P}\langle L\rangle \to A)\rangle$

and so by [2] and *modus ponens*:

- $T \vdash \mathcal{P}\langle L\rangle \to \mathcal{P}\langle(\mathcal{P}\langle L\rangle \to A)\rangle$

Hence by [2] again and the transitivity of $\to$:

- $T \vdash \mathcal{P}\langle L\rangle \to (\mathcal{P}\langle(\mathcal{P}\langle L\rangle)\rangle \to \mathcal{P}\langle A\rangle)$

But $\{B \to (C \to D)\} \vdash C \to (B \to D)$. Hence:

- $T \vdash \mathcal{P}\langle\mathcal{P}\langle L\rangle\rangle \to (\mathcal{P}\langle L\rangle \to \mathcal{P}\langle A\rangle)$

So by [3] and the transitivity of $\to$ it follows that:

- $T \vdash \mathcal{P}\langle L\rangle \to (\mathcal{P}\langle L\rangle \to \mathcal{P}\langle A\rangle)$

By Contraction:

- $T \vdash \mathcal{P}\langle L\rangle \to \mathcal{P}\langle A\rangle$

and by our supposition about A and transitivity again:

- $T \vdash \mathcal{P}\langle L\rangle \to A$

By the original characterisation of L:

- $T \vdash (\mathcal{P}\langle L\rangle \to A) \to L$

So by *modus ponens*:

- $T \vdash L$

By [1] again:

- $T \vdash \mathcal{P}\langle L\rangle$

So by a final application of *modus ponens*:

- $T \vdash A$

as required.

3 Curry's Paradox

Curry's paradox was published in 1942 by the US logician Haskell Curry[3]—though paradoxes in the same family were known to some medieval logicians.[4] It can be formulated using the notions of set, property, or truth. Here, let me give the version that uses truth. Like the Liar paradox, it appeals to the apparently obvious principle that a sentence is true just if things are the way it says. (Logicians, following Tarski, often call this the T-Schema.) So, for any sentence, A (provided that it does not use context-dependent words, like 'you' and 'now'):

- $\mathcal{T}\langle A\rangle \leftrightarrow A$

Here, $\mathcal{T}x$ is the predicate 'x is true', and $\langle A\rangle$ is a name for A. (It need not be obtained by arithmetic coding, but it could be.)

The Liar paradox concerns the sentence, S, such that $S \leftrightarrow \neg\mathcal{T}\langle S\rangle$; that is, which says of itself that it is not true, and deduces both S and $\neg S$. If one appeals to the principle of Explosion (or to give it its medieval name, *ex contradictione quodlibet sequitur*) ($\{B, \neg B\} \vdash A$, for arbitrary A and B), then an arbitrary conclusion follows. Curry's paradox establishes an arbitrary conclusion A, but without explicit mention of negation or the use of Explosion. Again, it can be formulated in a number of different ways, though these all come to much the same thing. Here is one standard way.

Given A, we form the sentence which says of itself that if it is true then A. In other words, by some form of self-reference, we form a sentence, C, of the form $\mathcal{T}\langle C\rangle \rightarrow A$. The T-Schema for C then gives us:

- $\mathcal{T}\langle C\rangle \leftrightarrow (\mathcal{T}\langle C\rangle \rightarrow A)$

and so:

- $\mathcal{T}\langle C\rangle \rightarrow (\mathcal{T}\langle C\rangle \rightarrow A)$

By Contraction, we get:

- $\mathcal{T}\langle C\rangle \rightarrow A$

From the T-Schema for C in the other direction, we get:

[3]Curry (1942). For a general discussion of the paradox, see Shapiro and Beal (2018).

[4]See, e.g., Priest and Routley (1982).

- $(\mathcal{T}\langle C\rangle \to A) \to \mathcal{T}\langle C\rangle$

So by *modus ponens*:

- $\mathcal{T}\langle C\rangle$

And by *modus ponens* again:

- A

Since A may be obviously absurd (e.g., 'Donald Trump is a frog'), this cannot be the case—and even if it is true (e.g., 'Donald Trump is corrupt'), one ought not to be able to prove it like this.

4 And So?

It is not hard to see that structurally the same thing is going on in the proof of Löb's theorem and Curry's paradox. (At least where the conditional used in Curry's paradox is that same as that in the formal arithmetic. Arguably, there are different kinds of conditionals; and each will generate its own Curry paradox.) Let us write $\mathcal{Q}$ for either $\mathcal{P}$ or $\mathcal{T}$. Then by self-reference we form a sentence, S, of the form $\mathcal{Q}\langle S\rangle \to A$. Using the properties of $\mathcal{Q}$, we then show that $\mathcal{Q}\langle S\rangle \to (\mathcal{Q}\langle S\rangle \to A)$, and so by Contraction, $\mathcal{Q}\langle S\rangle \to A$. A couple of applications of *modus ponens* then deliver A. However, in one case the reasoning delivers a theorem. In the other it delivers something unacceptable. What is going on here?

The standard story is something like this. The proof of Löb's theorem is fine. The result is perhaps surprising; but then many mathematical results are such. The proof of Curry's paradox is fallacious. That, at least, is impossible to gainsay. The Liar paradox delivers a contradiction. If one is a dialetheist one can simply accept the argument and its conclusion. The use of paraconsistent logic, not validating Explosion, prevents the contradiction spreading where it should not go.[5] With Curry's paradox it is impossible to accept the argument: we have a direct proof of everything. Given that all the logical moves in the paradoxical argument are in the proof of Löb's theorem, and so correct, the only other possibility is that the T-Schema must be rejected in full generality. This view about truth is buttressed by appeal to

[5]See, e.g., Priest (2006).

a construction of Tarski, according to which there is hierarchy of languages. There is a T-Schema at every level of the hierarchy (except the first), but the Schema at any level is guaranteed to hold only for sentences of the level below. (Of late, it is worth nothing, logicians have been more sympathetic to the view that the T-Schema is correct. Various logical principles are problematised instead.[6])

Anyway, given this, the T-Schema (for all A : $\mathcal{T}\langle A\rangle \leftrightarrow A$) is not a property of truth. It holds only for a certain class of As. (It is well known that there are consistent theories of arithmetic plus a truth predicate where the Schema holds for certain syntactically defined classes of sentences.[7]) In the same way, the provability of what we might call the Löb Schema (for all A: $\mathcal{P}\langle A\rangle \rightarrow A$) is not a property of provability. It holds only for a certain class of As, namely, those that are themselves provable.

This analysis is problematic, however. The analogy between $\mathcal{T}$ and $\mathcal{P}$ is not as straightforward as it might appear. Those who hold the view concerning truth in question hold that not all instances of the T-Schema are true. By contrast, all instances of the Löb Schema *are* true. They are just not provable. In one sense, this is just a version of Gödel's incompleteness theorem; but matters are more significant than that.

As is clear to anyone familiar with Gödel's proof, the heuristic it uses is a paradox for provability analogous to the Liar paradox for truth. Specifically, let G be a sentence of the form $\neg\mathcal{P}\langle G\rangle$. Since $\mathcal{P}\langle G\rangle \rightarrow G$, $\mathcal{P}\langle G\rangle \rightarrow \neg\mathcal{P}\langle G\rangle$. Hence $\neg\mathcal{P}\langle G\rangle$, that is, G; and we have just proved this, i.e., $\mathcal{P}\langle G\rangle$. So we have a paradox. Clearly, this uses the Löb Schema, and in particular, its instance for G, at the first step. If this is not provable in a formal arithmetic, then this argument cannot be reproduced in the formal system to show that it is inconsistent. What is left of the argument just shows that the arithmetic is incomplete.

However, the instances of the Löb Schema are true; indeed, since 'prove' means something like 'establish', they would seem to be true by the very meaning of 'prove'. And since the aim of an axiom system is to capture the truths of some subject, one should expect to be able to have an axiom system for arithmetic in which they are provable. If this system is not to be trivial (i.e., such that everything is provable) the proof of Löb's theorem must fail. Given the structural parallel between the proof of Löb's theorem and that of

[6]See, e.g., Kripke (1975), Priest (1987), Field (2008), Beall (2009).

[7]See, e.g., Boolos, Burgess, and Jeffrey (2007), p. 287, 23.1.

Curry's paradox, it would seem that same thing must account for the failure of the argument there; and this gives us only two suspects: the principles of *modus ponens* and Contraction (for the conditional involved).

The more dubious of the two would appear to be Contraction. There has been little investigation of arithmetics based on logics in which Contraction fails—so far. But given what we know about proofs in formal arithmetic, there is a definite suspicion that many of the standard number-theoretic results would not be provable in such theories.

Though initially less promising, the failure of *modus ponens* is actually more so. In paraconsistent logics Explosion fails. That is, there can be situations where for some A and B, B and $\neg B$ hold, but A does not. But then the Disjunctive Syllogism, $\{B, \neg B \vee A\} \vdash A$, also fails. ($\neg B \vee A$ holds since its first disjunct does.) In standard mathematics $A \to B$ is understood as $\neg B \vee A$. In other words *modus ponens* fails.

Moreover, we know that there are axiomatic theories of arithmetic, T, whose underlying logic is paraconsistent, in which everything true (in the standard model of arithmetic) is provable.[8] So, in particular, T is complete; that is, for any A, either $T \vdash A$ or $T \vdash \neg A$. The theories are inconsistent, but contradictions do not spread everywhere because of the failure of Explosion. (Gödel's first incompleteness theorem shows that every appropriately strong theory of arithmetic is either incomplete *or* inconsistent. The second disjunct is usually ignored since it is assumed that the theory is based on a logic where Explosion is valid.) *Modus ponens* seems such an integral part of reasoning that it would naturally be thought to be virtually impossible without it. What these results show is that this is not so. All arithmetic truths—and so all the standard results of number theory—are provable in such theories without it.

In these inconsistent arithmetics all instances of the Löb Schema, $\mathcal{P}\langle A\rangle \to A$ are provable; so, as might be expected, are the Gödel undecidable sentence, G—that is, $\neg\mathcal{P}\langle G\rangle$—and its negation. The proof of Löb's theorem fails, as it must, since the theory is non-trivial. Whether the Henkin sentence, H—that is $\mathcal{P}\langle H\rangle$—is provable in these arithmetics is currently unknown. In that sense, Henkin's original question[9] is still open.[10]

[8]For details of this and what follows, see the second edition of Priest (1987), ch. 17.

[9]Effectively raised in this context by Shapiro (2019).

[10]Many thanks go to Hartry Field for his helpful comments on an earlier draft of this piece.

References

[1] Beall, J. (2009), *Spandrels of Truth*, Oxford: Oxford University Press.

[2] Berlinski, D. (2019), 'The Director's Cut', *Inference* 5(1), https://inference-review.com/article/the-directors-cut.

[3] Boolos, G., Burgess, J., and Jeffrey, R. (2007), *Computability and Logic*, 5th edn, Cambridge: Cambridge University Press.

[4] Curry, H. B. (1942), 'The Inconsistency of Certain Formal Logics', *Journal of Symbolic Logic* 7:115–117.

[5] Field, H. (2008), *Saving Truth from Paradox*, Oxford: Oxford University Press.

[6] Kripke, S., (1975), 'Outline of a Theory of Truth', *Journal of Philosophy*, 72: 690–716.

[7] Löb, M. H. (1955), 'Solution of a Problem of Leon Henkin', *Journal of Symbolic Logic* 20: 115–118.

[8] Priest, G. (2017), *Logic: a Very Short Introductio*n, 2nd edn, Oxford: Oxford University Press.

[9] Priest, G., and Routley, R. (1982), 'Lessons from Pseudo-Scotus', *Philosophical Studies* 42: l89–199.

[10] Priest, G. (1987), *In Contradiction*, Dordrecht: Martinus Nijhoff; 2nd edn, Oxford: Oxford University Press, 2006.

[11] Shapiro, S. (2019), 'Inconsistency and Incompleteness Revisited', ch. 22 of C. Başkent and T. M. Ferguson (eds.), *Graham Priest on Dialetheism and Paraconsistency*, Berlin: Springer Nature.

[12] Shapiro, L., and Beall, J. (2018), 'Curry's Paradox', in E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/curry-paradox/.