Self-reference

The following result is a cornerstone of modern logic:

> **Self-reference Lemma.** For any formula $\psi(x)$, there is a sentence ϕ such that $(\phi \leftrightarrow \psi([\ulcorner\phi\urcorner]))$ is a consequence of Q.

Proof: The proof breaks down into two parts. The hard part is to see what sentence ϕ to use, and the easy part is to verify that it works. In approaching the hard part, I'll be following Gyorgy Sereny's presentation in "Godel, Tarski, Church, and the Liar,"[1] although the main idea is already there in Gödel's original paper.[2]

The key idea goes back to the 6th century BCE, when the Cretan Epimenides said that Cretans always lie. Assuming for argument that the other statements made by Cretans are all blatent falsehoods, we find ourselves inexorably driven to the unhappy conclusion that, if what Epimenides says is true, it is false, whereas if what he says is false, it is true. It is doubtful that Epimenides realized the paradoxical nature of what he said, but someone who was fully aware of the cognitive disturbance was Eubulides of Miletas, a contemporary of Aristotle, who asked us to assess what is said by someone who declares, "What I am now saying is false." Eubulides is credited with other notorious paradoxes, notably the *sorites* (The observation that taking a single straw from a heap of straw still leaves you with a heap of straw leads, by multiple applications,

[1] *Bulletin of Symbolic Logic* 9 (2003): 3-25.

[2] "Über formal unentscheidbare Sätze der *Principia mathematica* und verwandter Systeme I." *Monatshefte für Mathematik und Physik* 38 (1931): 173-198. English translations in Jean van Heijenoort, ed., *From Frege to Gödel* (Cambridge, Mass., and London: Harvard University Press, 1967), pp. 596-616, and in Martin Davis, ed., *The Undecidable* (Hewlett, N.Y.: Raven Press, 1965), pp. 4-38.

to the conclusion that there is a heap of straw that contains no straw at all) and the *hooded man* (You do not know who the hooded man is, but you do know who your father is, even though, unknown to you, the hooded man is your father; this contradicts the logical principle that names that denote the same thing can be exchanged). Eubulides' formulation is in one way sharper than Epimenides', since it doesn't depend on the mendacity of all one's neighbors. However, it introduces a new level of complexity, since it contains the indexicals "now" and "I." You can avoid these complexities by looking at page 65 of the June 1969 issue of *Scientific American*,[3] where you will find the sentence, "The sentence printed in red on page 65 of the June 1969 issue of *Scientific American* is false," printed in red. This still isn't what we need for present purposes. For present purposes, we would like to reproduce a version of the liar paradox (with ψ is place of "is false") within the language of arithmetic, and contingent facts about who said what when and where aren't expressible within the language of arithmetic. What is expressible within the language of arithmetic is syntax, which we can express by means of the Gödel coding. So we would like to examine a version of the lair paradox that identifies the offending sentence in purely syntactic terms.

A purely syntactic version of the liar paradox is given by Quine:[4]

> "Yields a falsehood when appended to its own quotation" yields a falsehood when appended to its own quotation.

[3] This is Tarski's article, "Truth and Proof."

[4] "The Ways of Paradox" from *The Ways of Paradox and Other Essays*, revised ed. (Cambrdige, Mass., and London: Harvard University Press), 1976.

Quine's construction generalizes nicely. Given a property P, consider the following sentence:

> "Yields a sentence with property P when appended to its own quotation"
>
> yields a sentence with property P when appended to its own quotation.

The sentence is true if and only if it has property P.

This still isn't quite what we want, however, because it relies on a syntactic feature that English doesn't share with the formal language, namely, that you can form a sentence by concatenating a (possibly complex) noun phrase with a (possibly complex) verb phrase. An alternative formulation that does generalize uses the operation of substituting a noun phrase for a variable.[5] *Grelling's paradox*[6] asks us to partition open sentences into those that satisfy themselves and those that do not. "x contains fewer than ten words" contains fewer than ten words, and so it satisfies itself, unlike "x contains fewer than five words." "x is an open sentence of English"is an open sentence of English, so it satisfies itself. "x is an open sentence of Portuguese" is not an open sentence of Portuguese, so it does not satisfy itself. Also, "x is a horse" is not a horse, and thus it does not satisfy itself. Now consider "x does not satisfy itself." It would appear that it satisfies itself if and only if it does not.

To generalize the Grelling paradox, note that an expression of English satisfies an open sentence just in case the sentence obtained by substituting the quotation name of the expression for the variable in the open sentence is true. Thus "x contains fewer than ten words is true"

[5] Explicit variables are part of the dialect of English employed in math and science. Everyday speech employs pronouns for the much the same purpose.

[6] Grelling, Kurt, and Leonard Nelson. "Bemerkungen zu den Paradoxien von Russell und Burali-Forti." *Abhandlungen der Fries'schen Schule neue Folge* 2 (1908): 301-334.

satisfies itself because "'x has fewer than ten words' has fewer than ten words" is true. If an expression doesn't satisfy an open sentence, then the result of substituting the quotation name of the expression for the variable in the open sentence is false. Thus "x is a horse" doesn't satisfy itself, and so "'x is a horse' is a horse" is false. An open sentence S does not satisfy itself if and only if the sentence obtained by substituting the quotation name of S for its variable is false. Thus Grelling's paradox consists in asking whether a false sentence is obtained from "A false sentence is obtained from the open sentence x when its quotation name is substituted for its variable" when its quotation name is substituted for its variable. This gives us a version of Eubulides' paradox in which the paradoxical sentence is identified entirely by its syntactic features.

The sentence "This sentence is false" is a sentence that asserts its own falsity, but it does so by making use of the demonstrative "this," and demonstratives aren't available in the formal language of arithmetic. "A false sentence is obtained from 'A false sentence is obtained from the open sentence x when its quotation name is substituted for its variable' when its quotation name is substituted for its variable" likewise asserts its own falsity, and it does so without relying on demonstratives. Right at the moment, our focus isn't on the liar paradox. We'll come back to talk about the paradox later on, although what we'll have to say won't be even remotely satisfying. Right now, however, our interest in self-reference, and we're in luck, for the same construction works generally. Specifically, given a property P, the sentence "A sentence with property P is obtained from the open sentence 'A sentence with property P is obtained from the open sentence x when its quotation name is substituted for its variable' when its quotation name is substituted for its variable" is true if and only if it has property P.

Transferring this construction from English to the formal language, we use Gödel numbers in place of quotation names. Specifically, we define a function Z taking a number to the Gödel number for the numeral for that number, thus:

$$Z(0) = \ulcorner 0 \urcorner = 4$$

$$Z(n+1) = \ulcorner [n=1] \urcorner = \text{Pair}(4,Z(n))$$

The usual technique for converting recursive definitions to explicit definitions shows that Z is Δ.

The partial function that takes the code number of a formula and the code number of a term to the code of the formula obtained by substituting the term for free occurrences of the variable "x"in the formula is Δ.[7] We explicitly wrote out the formula for substituting a term into a term, and the formula for substituting a term into a term is completely analogous. I won't write it out, but I could if I wanted to. We just noted that the formula Z taking a number n to [n] is Δ. Composing the two, we see that the function g given by:

g(n) = the code of the sentence obtained from the formula coded by n by
substituting [n] for free occurrenes of "x," if n is the code of a formula
whose only free variable is "x";
= 0, otherwise;

is Δ, so there is a formula γ(x,y) that functionally represents it.

Continuing with our formalization of the generalized version of Grelling paradox, with "ψ(x)" taking the place of "property P," let θ(x) be the formula

[7] "x" isn't really a variable of the formal language; the official variables are "x_0," "x_1," "x_2," and so on, but I'll pretend "x" is a variable, because too many subscripts are annoying.

$$(\exists y)(y \text{ is a sentence} \wedge \gamma(x,y) \wedge \psi(y)).$$

Let k be the Gödel number of k, and let ϕ be the sentence

$$(\exists y)(y \text{ is a sentence} \wedge (\gamma([k],y) \wedge \psi(y)).$$

Then $\ulcorner\phi\urcorner = g(k)$, and so

$$Q \vdash (\forall y)(\gamma([k],y) \leftrightarrow y = [\ulcorner\phi\urcorner]).$$

Also,

$$Q \vdash [\ulcorner\phi\urcorner] \text{ is a sentence.}$$

Consequently,

$$Q \vdash ((\exists y)([\ulcorner\phi\urcorner] \text{ is a sentence} \wedge \gamma([k],y) \wedge \psi(y)) \leftrightarrow \psi([\ulcorner\phi\urcorner]),$$

that is,

$$Q \vdash (\phi \leftrightarrow \psi([\ulcorner\phi\urcorner]).☒$$

Generalized Self-Referential Lemma. For any formula $\psi(x,z_1,z_2,\ldots,z_n)$, there is a formula $\phi(z_1,z_2,\ldots,z_n)$ such that:

$$Q \vdash (\forall z_1)(\forall z_2)\ldots(\forall z_n)(\phi(z_1,z_2,\ldots,z_n) \leftrightarrow \psi(\ulcorner\phi\urcorner,z_1,z_2,\ldots,z_n)).$$

Proof: In the proof of the Self-Referential Lemma, the extra variables quietly go along for the ride.☒