The Diagonal Lemma:
An Informal Exposition

Richard G Heck, Jr

At the heart of Gödel's incompleteness theorem is the so-called 'diagonal lemma' whose purpose is to allow us to construct self-referential sentences, that is, sentences that refer to themselves and say something about themselves. In most of the usual expositions, the diagonal lemma is presented as an arithmetical fact, and its proof is complex and difficult to follow. Students are often left with the sense that it is all some kind of dark magic. In fact, however, the basic idea behind the diagonal lemma is easy enough to understand.

We are going to give an informal, but still perfectly mathematical, exposition of the fixed point theorem, not in terms of arithmetic but in terms of syntax.[1] The diagonal lemma is not fundamentally about arithmetic. It is about syntax, about linguistic expressions. The arithmetical form is a consequence of the fact that "syntax can be coded in arithmetic" through Gödel numbering. That is no doubt important to the application of the diagonal lemma to arithmetic, but the syntactic facts are what are fundamental. So, we're talking about syntax, about expressions.

What we are going to prove, then, is the following:

> Let $A(x)$ be a sentence-frame of English containing the variable x. Then there is an English expression t_A that is itself a name of the English sentence $A(t_A)$ that one gets by substituting the very term t_A for the variable x in $A(x)$.

The sentence $A(t_A)$ is then a sentence that says of itself that it has whatever property $A(x)$ expresses.

Clearly, we have a notion of substituting one expression for parts of some other expression. We'll be especially interested here in expressions

[1] This sort of approach originates with Quine. See the chapter on Protosyntax in his *Mathematical Logic*.

containing the nonsense word 'blurg'. We could use variables, as we just did, or spaces, or whatever. It doesn't matter.

Suppose we have two expressions, then, E_1 and E_2. We know what it means to substitute E_2 for all occurrences of the word 'blurg' in E_1. For example, if E_1 is:

> Some expressions are blurg

and E_2 is:

> amusing

Then the result of substituting E_2 for all occurrences of 'blurg' in E_1 is:

> Some expressions are amusing

If E_2 is:

> are are are

then the result is:

> Some expressions are are are are

As said, are talking about expressions. There's no assumption they make any sense.

Of course, there's no need for E_1 and E_2 to be different. So if E_2 is the same as E_1, then the result is:

> Some expressions are some expressions are blurg

That's nonsense again, but that's fine.

A special case of this sort of substitution arises if E_2 is not just an expression but a *name* of an expression. So let E_1 be the expression:

> blurg contains four words

and let E_2 be the expression:

> Bob's favorite English sentence

Then the result of substituting E_2 for all occurrences of 'blurg' in E_1 is:

> Bob's favorite English sentence contains four words.

Whether this is true will depend upon what Bob's favorite English sentence is. To take another example, let E_2 be:

> “There are pigs in the yard”

Then the result is:

> “There are pigs in the yard” contains four words

which is not true. Then again, E_2 could be:

> “Bob’s favorite English sentence”

and now the result is:

> “Bob’s favorite English sentence” contains four words

which is true. It’s critical to distinguish the quotes case (mention) from the non-quotes case (use). Confusion ensues otherwise.

We could have E_2 be E_1 again, in which case we get nonsense:

> blurg contains four words contains four words

More interestingly, we can have E_2 be a quote-name of E_1 itself. I.e., E_2 could be:

> “blurg contains four words”

and then the result is:

> “blurg contains four words” contains four words

which is true.

Focus attention now on this special kind of self-substitution, that is, on the operation:

> The result of replacing all occurrences of ‘blurg’ in E_1 with its own quote-name

(that is, with a quote-name of E_1). The example we just did is one case of this kind of self-substitution, with E_1 being:

> blurg contains four words

since E_2, in that example, was just a quote-name of E_1. Other examples are easy to construct. E.g, if E_1 is:

> Bob is blurg

then the result is:

> Bob is “Bob is blurg”

which isn’t true, since Bob is a cat (my cat), not a sentence.

Consider now the expression:

> The result of replacing all occurrences of ‘blurg’ in blurg with its own quote-name

What is the result of replacing all occurrences of ‘blurg’ in that very expression with its own quote name? It is:

> The result of replacing all occurrences of ‘blurg’ in “The result of replacing all occurrences of ‘blurg’ in blurg with its own quote-name” with its own quote-name

So this expression actually names itself. We just proved that it does through simple calculation.

We are now ready to prove the diagonal lemma. Let A(blurg) be any expression you wish containing the word ‘blurg’. What we want to show is that there is an expression that names the result of substituting a quote-name for that very expression into A(blurg). That is: There is a term t that denotes the very sentence: $A(t)$. We will construct t explicitly. We’ll just do an example, but it will be clear that it generalizes.

Let A(blurg) be:

> blurg is weird

We will construct an expression that names the result of substituting that very expression for ‘blurg’ in “blurg is weird”. That is: We will have a term t such that:

> $t =$ “t is weird”

Consider the expression:

(1) The result of replacing all occurrences of ‘blurg’ in blurg with its own quote-name is weird.

What is the result of replacing all occurrences of ‘blurg’ in (1) with its own quote name? That is, which expression does the expression:

(2) The result of replacing all occurrences of ‘blurg’ in "The result of replacing all occurrences of ‘blurg’ in blurg with its own quote-name is weird." with its own quote-name

name? Well, just take (1) and replace the (unquoted) occurrence of 'blurg' in it with the result of putting (1) in quotes. So we get:

(3) The result of replacing all occurrences of 'blurg' in "The result of replacing all occurrences of 'blurg' in blurg with its own quote-name is weird." with its own quote-name is weird.

So (2) names (3). But (3) just is (2) followed by the words "is weird", i.e., it is the result of replacing 'blurg' in "blurg is weird" with (2). Roughly:

(2) = "(2) is weird"

So (3) says of itself that it is weird.

Abbreviations may help. Abbreviate:

The result of replacing all occurrences of 'blurg' in blurg with
its own quote-name

as: The diagonalization of blurg. Then we consider:

(2′) The diagonalization of "The diagonalization of blurg is weird."

and note that it is a name of:

(3′) The diagonalization of "The diagonalization of blurg is weird." is weird.

Which is to say:

The diagonalization of "The diagonalization of blurg is weird." *is* the expression "The diagonalization of 'The diagonalization of blurg is weird.' is weird.".

As wanted, again.

The informal character of the foregoing should not mislead. It is perfectly good mathematics, and we can easily enough formalize it all in a formal theory of syntax. Translating it all into arithmetic is what we need Gödel numbering to do, but, as I've said, self-reference is primarily a syntactic idea, not an arithmetical one.