Kenny's Overview of Hofstadter's Explanation of Gödel's Theorem

Copyright (c) 1996 by Kenny Felder

In the nineteenth and early twentieth centuries, one of the big mathematical goals was to reduce all of number theory to a formal axiomatic system. Like Euclid's Geometry, such a system would start off with a few simple axioms that are almost indisputable, and would provide a mechanical way of deriving theorems from those axioms.

It was a very lofty goal. The idea was that this system would represent *every statement you could possibly make* about natural numbers. So if you made the statement "every even number greater than two is the sum of two primes," you would be able to prove strictly and mechanically, from the axioms, that it is either true or false. For real, die-hard mathematicians, the words "true" and "false" would become shorthand for "provable" or "disprovable" within the system. Russell and Whitehead's *Principia Mathematica* was the most famous attempt to find such a system, and seemed for a while to be the pinnacle of mathematical rigor.

Gödel's theorem dashed this hope completely. It didn't just find a hole in Russell and Whitehead's reasoning, which would presumably have been patched: it showed that the entire goal is unachievable. More specifically, Gödel showed that for any formal axiomatic system, *there is always a statement about natural numbers which is true, but which cannot be proven in the system.* In other words, mathematics will always have a little fuzziness around the edges: it will never be the rigorous unshakable system that mathematicians dreamed of for millennia.

In *Gödel, Escher, Bach,* Douglas Hofstadter presents his own version of Gödel's proof. This paper is my summary of Hofstadter's version of Gödel's theorem.

A Road Map of Where We're About to Go

Before I jump into the proof, I want to give an outline of where we're headed, and why.

A common trick in mathematics is to prove something by assuming the opposite, and then finding a contradiction. For instance, suppose I want to prove that there is an infinite number of prime numbers. So I assert the opposite: there is a finite number of primes, and `n` is the biggest one. Then I start reasoning logically from that premise. At the end of my train of logic, I find a paradox: I prove that 1=2, or find a prime number bigger than `n`, or some such. This means my original assertion must be false, so there is an infinite number of primes.

This is the form of proof we're going to follow. I'm going to start by explaining a system called Typographical Number Theory, or TNT, which is one attempt to represent all of math in an axiomatic way. I'm going to *assume that TNT works*—that is, assume that it really does encapsulate all of mathematics perfectly. (You will see me making reference to this assumption on a number of occasions.) And eventually, I'm going to be led to a contradiction.

The whole key to the proof is that it doesn't have anything to do with TNT *per se.* It simply assumes that TNT is perfect, and proceeds from there to a paradox. In doing so, it crushes any system which makes similar claims of perfection.

Typographical Number Theory (TNT)

Rather than using Russell and Whitehead's *Principia Mathematica* as his axiomatic straw man, Hofstadter invents his own, called "Typographical Number Theory" or TNT for short.

TNT expresses everything in terms of a few simple symbols. There are standard mathematical symbols such as +(plus), `*` (times), and =(equals). There are *variables,* represented by the letter `a` followed by primes: `a`, `a'`, `a''`, *etc.* There are standard logical symbols such as ~ (not), `V` (or), `E` (there exists) and `A` (for all). Finally, there are numbers, which are represented by the two symbols `0` (meaning zero) and `S` (meaning "the successor of"); so we count `0, S0, SS0, SSS0,` and so on. Note that we cannot express negative numbers or fractions, but that's okay, because we only care about natural numbers. (The "natural numbers" are all the integers above zero: TNT only attempts to address them.)

In TNT, what we would normally call "statements" are written as "strings"; that is, simple combinations of our allowed symbols. For instance,

```
~Ea:a*a=SS0
```

is a TNT string. It translates as "there does not exist any number `a`, such that `a` times `a` is two"; or, more concisely, "there is no square root of two." Since we are limiting our scope to natural numbers, this happens to be a true statement.

If we replace the `SS0` with `SSSS0,` we still have a perfectly fine TNT string, although it now happens to stand for a false statement.

So much for the language. Now, as a formal system, it needs two things: axioms, and ways of deriving theorems from axioms. An axiom in TNT is simply a string; since there are only five, I will reproduce them all here.

```
Axiom 1: Aa:~Sa=0
Axiom 2: Aa:(a+0)=a
Axiom 3: Aa:Aa':(a+Sa')=S(a+a')
Axiom 4: Aa:(a*0)=0
Axiom 5: Aa:Aa':(a*Sa')=((a*a')+a)
```

All TNT derivations must start with those five strings, and only those five strings! If you want to test your understanding of TNT, you can attempt to figure out what mathematical statements these five strings represent: click here to verify your answers.

The proof methods are rules that allow you to transform one string into another. Hofstadter gives just over a dozen rules, such as the following.

Rule: The string `~~` can be deleted wherever it appears in any string.

Rule: For any variable `u`, the strings `Au:~` and `~Eu:` are interchangeable anywhere inside a string.

The way you use TNT is very well defined. You start off with the axiom strings. You then create new strings, by applying the allowed string manipulation rules. Any string which you produce in this manner is called a *theorem* of TNT. So a sample derivation might look like:

`Aa:~Sa=0` *(Axiom 1)*
`~Ea:Sa=0` *(By the second rule I printed above)*

This is a valid derivation of the string `~Ea:Sa=0`. Of course, it isn't the most tricky or interesting thing in the world. But I could presumably find another rule to apply to that second string, and yet another rule to apply to the resulting string, deriving strings which are further and further from axiom 1. In fact, Hofstadter shows some relatively long derivations that produce strings such as:

```
Theorem: S0+S0=SS0
Theorem: Aa:Aa':(a+a')=(a'+a)
```

It is important to understand that TNT, in and of itself, is not "about" numbers: it is simply a game involving symbol strings. You could program into a computer the set of *axiom strings,* and the set of *string manipulation rules,* and it could generate new theorem strings all day.

However, at a higher level, we do interpret the symbols. We recognize axiom 1 as "there are no negative numbers," and the first theorem above as "1+1=2." So in that all-important sense, TNT *is* about numbers. Each axiom string, and each rule, makes a statement which most people would agree is obviously true. So whenever we derive a theorem, we are inclined to believe that the statement it makes about numbers is also true. In that sense, we can legitimately say that deriving the string `S0+S0=SS0` "proves" that 1+1 is 2.

The Awesome Power of TNT

TNT has a very small set of mathematical and logical symbols. It has addition and multiplication: but it has no exponents, no roots, not even subtraction or division! Nonetheless, a "TNT proponent" would claim that you can translate *any numerical statement* into a TNT string. The reason is that advanced mathematical constructs are always built out of the basic operations of addition and multiplication.

For instance, there is no TNT symbol for "perfect square" or "prime number." However, we can translate the statements "4 is a perfect square" and "5 is a prime number" into TNT, as follows.

```
Ea:(a*a) = SSSS0
~Ea:Ea':(SSa*SSa') = SSSSS0
```

The first statement says, literally, that there is some number that multiplies by itself to create four. (It does not specify that the number is 2; just that *some* number exists with this property.) The second says that there are no *two* numbers, such that if you add two to each of them and then multiply them, you get five. (Recall that `S` means "the successor of," so

`SSa` means "the successor of the successor of `a`," or `a+2`.) Since the smallest number is 0, this is equivalent to saying "there are no two numbers greater than 1, which can be multiplied to give the answer 5."

Now, those are relatively easy: it can get a lot trickier. If you think about it for a while, you may be able to write "8 is a power of 2" in TNT, even though TNT has no exponents. If you think about it for a *really long* while, and know a lot more about number theory than I do, you may be able to write "100 is a power of 10" in TNT. But whether or not you can personally translate those statements, you can see that it might be possible. In fact, for the purposes of this paper, we are going to become "TNT proponents" ourselves, and assume without any proof that the following two statements are true.

1. Any statement you can make about natural numbers—no matter how complex, no matter how long, no matter how bizarre—can be written in a TNT string.

2. If such a statement is *true,* its TNT string can be derived as a theorem from the axioms. If the statement is *false,* we can derive its negation from the axioms. (Meaning, the same string with a ~ symbol in front of it.)

Hofstadter calls these two claims "TNT *expresses* number theory" and "TNT *represents* number theory," respectively. Assuming these two statements are true is equivalent to assuming that TNT succeeds in its lofty goal of symbolically representing all of number theory in a cohesive system. And based on this assumption, we are going to lead ourselves to an inevitable contradiction, and thus prove we were wrong to assume that any system could make these claims.

Gödel's Theorem: The Very End of the Proof

At this point, I'm going to jump all the way to the *end* of Gödel's proof, and fill in the middle later. (Thought all math was linear, didn't you?) The critical step is to take the following statement, which Hofstadter calls "sentence G," and translate it into a TNT-string.

`Sentence G: This statement is not a theorem of TNT.`

Now, ask yourself this question: is sentence G true or false?

If sentence G is false, then it is a theorem of TNT. Then we have a valid theorem which is false, and the whole system falls apart.

So it must be true. But if it is true, then it is not a theorem of TNT. Which means that *sentence G is true, but it is not provable within TNT.* That is Gödel's "incompleteness." He showed that TNT, although it may be perfectly consistent and always correct, cannot possibly prove *every* true statement about number theory; there is always something which is true, which the system cannot prove. So we're done!

Except that, as you may have noticed, this is totally ludicrous. After all, TNT makes statements about *numbers,* and sentence G is a statement about a *statement* (itself). So while writing a TNT-string for "100 is a power of 10" might be very difficult, it seems reasonable to grant that it's *possible;* but translating sentence G into TNT seems about as likely as yodeling in sign language.

Take a moment to make sure that you buy both of the following facts. First, *if* we could translate sentence G into TNT, we would have defeated TNT. Second, there is *no way* to translate sentence G into TNT as we have formulated it. Once both of those seem obvious, you will be ready for what is probably the most brilliant part of Gödel's proof, which is that he found a way to talk about statements in a language that was only meant to discuss numbers. And what enabled him to do this was nothing more profound than a slight change in notation.

A Slight Change in Notation

The change we're going to make to TNT is the simplest change possible, in a lot of ways. We aren't going to change the axioms. We aren't going to change the rules. We're just going to change the symbols.

Specifically, we're going to replace each symbol with a *three-digit number.* So instead of `0`, we will write 666; instead of `S`, 123; instead of =, 111; and so on. The numbers are chosen completely arbitrarily, following only two rules: every number has three digits, and no two numbers are the same. So TNT now starts to look much uglier. For instance, here are a few of the statements I gave in TNT earlier, translated into the Gödelized version.

TNT statement: `~Ea:a*a=SS0`

Gödelized: 223333262636262236262111123123666

TNT rule: The string ~~ can be deleted wherever it appears in any string.

Gödelized: The string 223223 can be deleted wherever it appears in any string.

You may or may not see why I'm bothering to change the notation in this way, but concentrate for the moment on convincing yourself of the following fact: *nothing has changed.* If TNT was valid, the Gödelized version is exactly as valid. The two systems are completely interchangeable.

If you've taken that small leap with me, I will now ask you to make one that is only slightly larger. Note that in our new notation, every string is a number. Because of this, we can change the form of our *rules,* from typographical manipulations to mathematical functions. I'll give you an example, using a rule which is not actually a real rule, just to keep things simple.

Fake rule: Whenever a string ends in the symbol "000", you can replace that symbol with "005".

The same fake rule, written differently: Whenever a number is a multiple of 1000, you can add 5 to it.

I chose this "fake rule" because it translates very simply. The rule about removing 223223 from a string would be trickier to write as a function, but it could certainly be done. And so, in this manner, we are going to change all our rules from *typographical,* to *mathematical,* manipulations of number-strings.

Once again, I urge you to take a moment to convince yourself that we haven't changed anything: our original TNT string-manipulation rules have been turned into mathematical functions, but our new system is still *exactly the same as the old,* just written differently.

Because, although nothing fundamental has changed, something very interesting has happened. Our original TNT strings were interpreted as being "strings about numbers." That was a very familiar idea: after all, we all learned in first grade to treat "1+1=2" as a string about numbers! But in the Gödelized version, we have something slightly different: *numbers about numbers.* In this system, the number 123666112666111123666 means "1+0=1"; and the number 123666111666 means "1=0".

Now, you may recall that Gödel's proof required us to write a TNT string about a TNT string; and we said that was impossible, because TNT strings are only about numbers. Although we've still got a way to go, you might suspect that the Gödel notation is a step in the right direction, since it blurs the distinction between numbers and statements. And because this is the step where things blur, it is the most important time for you to keep your understanding as *clear and unblurred* as possible, and to make sure that the idea of Gödelized TNT is as straightforward and incontrovertible as you can make it.

A Closer Look at Gödelized TNT

One good clarification technique is to note that we have three different ways now of talking about exactly the same information: number theory, TNT, and Gödel-numbered TNT. All three systems have *axioms,* and *rules of production* for turning those axioms into *theorems;* but these elements look very different in the different systems. I've tried to summarize the entire thing in the following chart.

Mathematical Logic	TNT	Gödel-numbered TNT
An axiom is an "obvious" statement about natural numbers.	An axiom is a statement string	An axiom is a number
A rule of production is a logical way to work with axioms.	A rule of production is an allowed string-manipulation mechanism.	A rule of production is an allowed mathematical function.
The theorems you produce are new statements about natural numbers.	The theorems you produce are new strings.	The theorems you produce are new numbers.

Now, I made the point above that in the Gödelized version of TNT, we actually have numbers that are *about* numbers. Of course, all such numbers are not "true." I gave the examples earlier of 123666112666111123666, which means "1+0=1", and 123666111666, which means "1=0". We recognize immediately, on the mathematical level, that the first statement is true and the second is not. In Gödelized TNT, we would say the same thing by saying that *the first number is a valid theorem, and the second is not.*

This means we have defined a new property of natural numbers. It is a property which we can call "theoremhood," and every number either has it, or doesn't. The property can be defined as follows.

Definition: A number has *theoremhood* if it corresponds to a valid theorem of TNT—or, in other words, to a true statement about numbers.

However, I prefer the following definition.

Alternative definition: A number has *theoremhood* if it is possible to create that number from our small set of axiom-numbers, by the application of our small set of function-rules.

The first definition goes up to the levels of TNT and mathematical logic. I prefer the second because it is a completely mathematical statement, which does not have to refer to "TNT" or "statements" at all. In other words, saying "this number has theoremhood" is kind of like saying "this number is a perfect square" or "this number is prime" that we discussed earlier: it's a complicated mathematical construct, which can be built out of simpler operations. This fact is going to allow me to pull the most important sleight of hand in this entire paper, so watch me closely during the next couple of paragraphs.

I'm going to start off with a true fact, and write it three different ways: in terms of math, in terms of TNT, and in terms of Gödelized TNT.

1. "Zero equals zero" is true.
2. The string `0=0` is a valid TNT theorem (*ie* can be derived from axioms).
3. The number 666111666 has the theoremhood property.

Remember, we're assuming that TNT represents every possible mathematical statement: so if sentence 1 is true, sentence 2 must also be true! Sentence 3, of course, says exactly the same thing as sentence 2, since Gödelized TNT is just string TNT with a different notation.

Now, I want to focus on sentence 3. I mentioned earlier that "theoremhood" is just a mathematical concept, like "primeness" or "squareness." In fact, we could rewrite this sentence as "It is possible to take certain numbers, apply certain functions to them, and arrive at 666111666."

Now, once again, remember that we are assuming TNT can express *any mathematical statement, no matter how complex.* So presumably it can handle this one: we could take the sentence "666111666 has theoremhood," and translate it into TNT!

The resulting string would presumably be true, leading us to the following claim, which I will again write in three equivalent ways.

1. "666111666 has theoremhood" is true.
2. The TNT string for "666111666 has theoremhood" is a valid TNT theorem.
3. The Gödel number for the TNT string for "666111666 has theoremhood," has theoremhood.

If we want, we can start spiralling up forever at this point. This sentence 3 is a new mathematical statement, and could be the "1" of another triad: we could translate it into TNT, Gödelize that string, claim theoremhood for our new number, and so on. What we have is a series of ever-longer statements, each of which says "the statement before me is a theorem." It's enough to make your head spin.

Fortunately, we don't have to take it that far. In fact, the only point I'm making at all is that "666111666 has theoremhood" is a mathematical statement, which can be translated into TNT. More generally, *it is possible to write a TNT string which says that another TNT string is (or is not) a valid theorem.*

Why is that so important? Well, if you remember Gödel's proof, it hinged on writing a TNT statement that talks about itself. We said at the time that it couldn't be done, because TNT strings are only about numbers. But now, we have suddenly found a way to write TNT strings that claim theoremhood for other TNT strings!

Actually, in all fairness, I should mention that we haven't actually "found a way." I never took a string like "666111666 has theoremhood" and translated it into TNT. I simply asserted that it is *possible* to translate that string, because that string is a mathematical statement, and *we are assuming that TNT can express all mathematical statements.* This is one of the most fascinating parts of this proof. The only thing we need to know about TNT, is that it *claims* to be all-powerful; we can use that fact to lead us all the way to the end, without going through the hard work of figuring out how to use the power.

But G is Still Harder to Write Than it Looks

At this point, you may well think we're done. We have shown that sentence G destroys TNT's claim to completely represent natural numbers. All we have to do is show that sentence G can, in fact, be translated into TNT. And we've done the hard part there, which is showing that "TNT Theoremhood" is a mathematical concept that can be expressed in TNT. So we know that the following sentences could all be written in TNT.

5 is not a theorem of TNT.

10 is not a theorem of TNT.

123666111666 is not a theorem of TNT.

...and so on. All we need to do is find such a sentence with the following peculiarity: the number that it talks about, happens to be the Gödel number of the sentence itself. Is that so hard?

Actually, it's impossible, as you can readily convince yourself with the following observation. The number 1 is "S0" in TNT, or 123666 in Gödelized TNT. Similarly, the Gödel number for 10 is 123123123123123123123123123123666 (ten "S"s and a 0). In general, the Gödel number for *any* number, is much bigger than the number itself! So if you wrote down the sentence "10 is not a theorem of TNT," its Gödel number would definitely not be 10, it would be much bigger...and so for any number. A TNT string cannot possibly be big enough to "contain" its own Gödel number.

So does that mean that it's still impossible to write sentence G? No, it just means that some indirectness will be required. Sentence G will not contain its own Gödel number explicitly, but will instead refer to it indirectly. For instance, we can refer to "10" as "5*2", which has a much lower Gödel number. Sentence G has to have some very clever, indirect reference to its own Gödel number. And the rest of this paper—and of Gödel's proof—is just such a clever method.

"Arithmoquining" to Get TNT Sentences About TNT Sentences

We start with a simple mechanism for writing a TNT sentence about another TNT sentence. The mechanism is called *arithmoquining,* and here's how it works. You start with any sentence that has a free variable, which we'll call `a`. To arithmoquine the sentence, you take the Gödel number of the entire sentence, and replace all occurrences of the variable `a` with that number.

For instance, start with the sentence `a=S0`. The Gödel number of this expression is 262111123666; so arithmoquining, we get "262111123666=1". That's all there is to arithmoquining: it's so simple, you might wonder why we're bothering with it.

The reason is that arithmoquining gives us a generalized way to write one TNT statement about another TNT statement. Just to keep things visually clear, I like to write the two statements in pairs; the first sentence has a variable, and the second sentence is the arithmoquine of the first.

T: `a = S0`
A: *The Gödel number of Sentence T* is 1.

The first sentence is about an arbitrary Thing (hence "T") represented by a free variable. The second sentence is the Arithmoquine ("A") of the first. We can write sentence A a bit more concisely, as follows.

A: *Sentence T* is 1.

When you write it that way, it's important to remember that by "Sentence T" you really mean the Gödel number of sentence T. But this way of writing it does have the advantage of showing, very clearly, that sentence A is always *about* sentence T. To give a more complicated example,

T: `a = SS0 * a - SSSS0`
A: *Sentence T* is 2 times *sentence T* minus 4.

Here, as before, sentence T is neither true nor false, since `a` is unspecified. Sentence A, the arithmoquine of Sentence T, is a blatantly false statement about a specific number.

I should mention that, although I like writing these sentences in pairs, you can always write the arithmoquine as its own complete sentence. In this case, we could say:

A: "a=SS0*a-SSSS0" is 2 times "a=SS0*a-SSSS0" minus 4.

Now, instead of a sentence about another sentence, we have a sentence about a phrase in itself. But all we've really done is rearrange things to make them less readable. We still definitely do *not* have a sentence about itself, which is what we need in order to write sentence G.

Finally, Writing Sentence G

Before you read any further and watch me write sentence G, I would encourage you to try to write it on your own. You have the basic problem, which is to write a sentence that says "I am not a theorem"; and you have the basic tool, which is arithmoquining to write a sentence about another sentence.

If you have tried it, I would guess that your first shot was something like this.

T: `a` is not a valid TNT theorem-number.
A: *Sentence T* is not a valid TNT theorem-number.

But that clearly doesn't do it. Sentence A is not the sentence G we're looking for, since it isn't about itself, it's about sentence T.

So you have to get a little trickier. And without any further ado, I'm going to present the answer, so this is your last chance to figure it out on your own...

T: The arithmoquine of `a` is not a valid TNT theorem-number.
A: The arithmoquine of *Sentence T* is not a valid TNT theorem-number.

Ta-da! This is a minor variation of the "first try" I presented a moment ago. But sentence T adds a wrinkle, in the sense of referring to the *arithmoquine* of our free variable `a`. So sentence A, which is the arithmoquine of sentence T, is also *about* the arithmoquine of sentence T. In other words, it's about itself!

Just to get that feeling of completion, I'm going to write sentence A as one big sentence without sentence T. This is the actual sentence G we've been looking for.

G: The arithmoquine of "The arithmoquine of `a` is not a valid TNT theorem-number" is not a valid TNT theorem-number.

You cannot, surprisingly, make it any shorter than that: that is sentence G, written out in English. Of course, to actually write it in TNT, you would have to figure out how to write "arithmoquine" and "valid theorem-number" in TNT, which would be an incredibly long, arduous, and (frankly) boring task. But fortunately, you don't have to bother. It is enough to know that both arithmoquining and theoremhood are mathematical properties, involving finite predictable manipulation of natural numbers—which, we have already said, both of them are. TNT claims that it can express all mathematical properties, so it should be able to express those two, so we should *in principle* be able to translate the above sentence into TNT; which is all it takes to undermine the system. Ironically, its very power of expression is what defeats TNT in the end.

What Now?

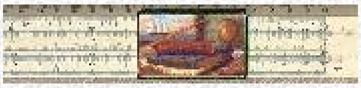
If you followed all that, you can now say with reasonable confidence that you "understand Gödel's theorem." That is, you understand why no formal mathematical system can ever hope to represent all statements about natural numbers.

As I see it, there are three directions you can go from here. The first direction is *down,* to a more mathematical level. The explanation I have given is very "high-level," and would not satisfy a real mathematician for an instant. By learning more about the math involved, you can work the proof to ever finer levels of detail, and make it ever more rigorous and bullet-proof.

The other way to go is *up,* to a more philosophical level. There are many people who believe that the human mind, based on neurons and physical principles, is just a very sophisticated formal system. Does Gödel's theorem imply the existence of *facts that must be true, but that our minds can never prove?* Or even stronger, that our minds can never believe—or strongest yet, ever conceive?

The third direction you can go is *sideways,* to lunch. Who wants to spend his whole life worrying about abstract mathematical theorems?

Gary and Kenny Felder's Math and Physics Help Home Page



www.felderbooks.com/papers

Send comments or questions to the author