SELF-REFERENCE AND THE LANGUAGES OF ARITHMETIC

RICHARD G. HECK, JR.

1. A PUZZLE ABOUT SELF-REFERENTIAL REASONING

Consider the following sentence:[1]

(1) (1) is true iff the right-hand side of (1) is false.

The so-called T-scheme delivers:

(2) (1) is true iff [(1) is true iff the right-hand side of (1) is false].

By the associativity of the biconditional,

(3) [(1) is true iff (1) is true] iff the right-hand side of (1) is false,

and obviously, the left-hand side of (3) is a logical truth. Hence

(4) the right-hand side of (1) is false.

But

(5) the right-hand side of (1) is "the right-hand side of (1) is false",

and so:

(6) "the right-hand side of (1) is false" is false.

But now the T-scheme tells us that

(7) "the right-hand side of (1) is false" is true iff the right-hand side of (1) is false.

Hence,

(8) "the right-hand side of (1) is false" is true.

[1]See **?** for discussion of how similar examples arise in the discussion of deflationism.

But that contradicts 6.

That the T-scheme threatens to lead to paradox is not news. But it is surprising—at least, it was very surprising to me—how hard it is to replicate the argument just given in some of the formal systems in which the T-scheme is frequently studied. One such system is Peano Arithmetic formulated in the language $\{0, S, +, \times\}$ and augmented by a one-place predicate T. Call that system PA_S—'S' for *standard* language.[2]

The obvious way to proceed is as follows. Fix some standard Gödel numbering and consider the formula:

(9) $$T(x) \equiv \exists y(rhs(x,y) \wedge \neg T(y)),$$

where $rhs(x,y)$ is a formula representing the relation: y is (Gödel number of) the right-hand side of the biconditional (whose Gödel number is) x. Diagonalization then yields a formula G such that PA_S proves:

(10) $$G \equiv [T(\ulcorner G \urcorner) \equiv \exists y(rhs(\ulcorner G \urcorner, y) \wedge \neg T(y))],$$

where, as usual, $\ulcorner G \urcorner$ abbreviates the numeral denoting the Gödel number of the formula G. The surprise is that one cannot then replicate the argument rehearsed above for the inconsistency of (1)—not, at least, if one follows Gödel's proof of the diagonal lemma (**?**) or that in the standard textbook *Computability and Logic* (**??**). The problem is that the formula G that is delivered by these 'standard' proofs of the diagonal lemma is not a biconditional but an existentially quantified formula. It has no right-hand side, so we can get no further.

As it happens, in this particular instance, inconsistency with the T-scheme can be reached in another way. Not only is

(11) $$\exists y(rhs(\ulcorner G \urcorner, y) \wedge \neg T(y))$$

false, it is refutable: We can prove that G has no right-hand side. By the associativity of the biconditional, however, (10) leads to

(12) $$[G \equiv T(\ulcorner G \urcorner)] \equiv \exists y(rhs(\ulcorner G \urcorner, y) \wedge \neg T(y))],$$

whence PA_S proves:

(13) $$\neg[G \equiv T(\ulcorner G \urcorner)],$$

which of course contradicts the T-scheme. But that is not the intuitive argument given above, and it is just a happy accident that such an argument is available in this case.

A similar problem arises with the following example:

[2]It will not matter for our purposes whether induction is extended to include formulae in which T appears, except in section 3.4, where the issue arises tangentially.

(14) the right-hand side of (14) is true iff the left-hand side of (14) is false.

This sentence replicates the postcard paradox in a single sentence.[3] There is an obvious argument that (14) is paradoxical. Again, however, that argument cannot be formalized in PA_S, at least not in the most obvious way. The obvious way to proceed this time is to consider the formula

(15) $\exists y(rhs(x,y) \wedge T(y)) \equiv \exists y(lhs(x,y) \wedge \neg T(y))$.

Diagonalization, as standardly proven, then delivers a formula P such that PA_S proves

(16) $P \equiv \{\exists y(rhs(\ulcorner P \urcorner, y) \wedge T(y)) \equiv \exists y(lhs(\ulcorner P \urcorner, y) \wedge \neg T(y))\}$.

But again, the formula in question is not a biconditional. It therefore has neither a left- nor a right-hand side. Hence both halves of the right-hand side of (16) are false and, indeed, refutable, so PA_S actually proves the right-hand side of (16) and so proves P. Far from being inconsistent with the T-scheme, then, P is actually a theorem of PA_S (and, indeed, of far weaker systems).

The most common statement of the diagonal lemma for PA_S is as follows:

Lemma. *(Diagonal Lemma) Let $B(x)$ be a formula in the language of PA_S, with just x free. Then there is a sentence G such that PA_S proves: $G \equiv B(\ulcorner G \urcorner)$.*

And what I am calling the standard proof of it goes roughly as follows.[4]

Proof. Given a formula $A(x)$ with just x free, let the diagonalization of $A(x)$ be the sentence: $\exists x(x = \ulcorner A(x) \urcorner \wedge A(x))$. Diagonalization is a (primitive) recursive syntactic function and, as such, is representable in PA_S by a formula $\mathrm{diag}(x,y)$ meaning: y is the Gödel number of the diagonalization of the formula whose Gödel number is x. We want to show that, for every formula $B(x)$ with just x free, there is a sentence G such that PA_S proves: $G \equiv B(\ulcorner G \urcorner)$. So consider the formula: $\exists y(\mathrm{diag}(x,y) \wedge B(y))$. Its diagonalization is the sentence:

(17) $\exists x[x = \ulcorner \exists y(\mathrm{diag}(x,y) \wedge B(y)) \urcorner \wedge \exists y(\mathrm{diag}(x,y) \wedge B(y))]$,

which will prove to be the wanted G. Since $\mathrm{diag}(x,y)$ represents diagonalization and (17) is, as said, the diagonalization of $\exists y(\mathrm{diag}(x,y) \wedge B(y))$, PA_S proves:

[3]The postcard paradox is so-called because it can be formulated as follows: Imagine a postcard on one side of which is written "The sentence on the other side of this card is false"; the sentence on the other side is "The sentence on the other side of this card is true".

[4]This version is adapted from **?**, Ch. 15 The proof in the new edition (**?**, Ch. 17) is somewhat different, though not in ways that matter here.

Note that the proof does not actually use the fact that only x is free in $B(x)$: One can just as easily prove that, for each formula $B(x,y_1,\ldots,y_n)$, there is a formula $G(y_1,\ldots,y_n)$ such that PA_S proves: $G(y_1,\ldots,y_n) \equiv B(\ulcorner G(y_1,\ldots,y_n) \urcorner, y_1,\ldots,y_n)$.

The proof can also be extended to the case of multiple formulae (**?**, pp. 53-4): Given formulae $A_1(x_1,\ldots,x_n,\mathbf{y}),\ldots A_n(x_1,\ldots,x_n,\mathbf{y})$, where $\mathbf{y}$ represents additional parameters, there are formulae $G_1,\ldots,G_n$ such that, for each i, PA_S proves: $G_i(\mathbf{y}) \equiv A_i(\ulcorner G_1 \urcorner,\ldots,\ulcorner G_n \urcorner,\mathbf{y})$.

(18) $\quad \mathrm{diag}(\ulcorner \exists y(\mathrm{diag}(x,y) \wedge B(y))\urcorner, \ulcorner (17)\urcorner)$

By logic, then, PA_S proves that (17) is equivalent to $B(\ulcorner (17)\urcorner)$, as wanted. □

As is clear, however, the relevant sentence G—that is, (17)—is always an existentially quantified formula and is never a biconditional, a disjunction, or what have you.

What is the significance of these examples? In "Outline of a Theory of Truth", Saul Kripke remarks that the standard proof of the diagonal lemma "show[s] that [self-referential sentences] are as incontestably legitimate as arithmetic itself" (**?**, p. 692).[5] Now, in a sense, that is surely true. The formula G produced by this proof of the diagonal lemma certainly is self-referential—in a way. But although G refers to itself, it does not do so by name but rather by description. What G says is that there is one and only one formula that is a diagonalization of $\exists y(\mathrm{diag}(x,y) \wedge B(y))$ and that this formula satisfies $B(x)$. The formula in question is, as it happens, G itself, and provably so. It is therefore natural to compare G to Tarski's examples of 'empirical' liars, that is, formulae such as:

> ***There is one and only one sentence displayed in bold italics on page 4 of this paper, and it is not true.***

This sentence does not refer to itself by name, but it does describe itself uniquely, and so in that sense does refer to itself.

The point here is easy to overlook—or, at least, it was easy for me to overlook. Consider the standard example of a liar sentence:

(The Liar) The Liar is not true.

This sentence refers to itself by name. Now consider the liar sentence produced by diagonalization. We consider the formula $\neg T(x)$ and diagonalize, thus getting a formula Λ such that PA_S proves:

(19) $\quad \Lambda \equiv \neg T(\ulcorner \Lambda \urcorner)$

There is supposed to be a sentence here that 'says of itself' that it is not true. Comparing (19) with The Liar, one might naturally suppose that the self-referential sentence was $\neg T(\ulcorner \Lambda \urcorner)$. But it is not. The sentence $\neg T(\ulcorner \Lambda \urcorner)$ refers to a sentence all right, but it simply refers to Λ, and Λ is most certainly not the sentence $\neg T(\ulcorner \Lambda \urcorner)$. Rather, Λ is constructed in accord with the standard proof of the diagonal lemma and so is the sentence:

(Λ) $\quad \exists x[x = \ulcorner \exists y(\mathrm{diag}(x,y) \wedge \neg T(y))\urcorner \wedge \exists y(\mathrm{diag}(x,y) \wedge \neg T(y))]$

[5]It is clear that Kripke has this sort of proof in mind, for he says: "Gödel…showed that, for each predicate $Q(x)$, a syntactic predicate $P(x)$ can be produced such that the sentence $(x)(P(x) \supset Q(x))$ is demonstrably the only object satisfying $P(x)$. Thus, *in a sense*, $(x)(P(x) \supset Q(x))$ 'says of itself' that it satisfies $Q(x)$." (**?**, p. 692) Note the remark I have italicized: Perhaps Kripke was uncomfortable with the idea that the mentioned sentence is *truly* self-referential.

As noted above, this formula says, roughly: There is a sentence that is a diagonalization of $\exists y(\mathrm{diag}(x,y) \wedge \neg T(y))$, and it is not true. As it happens, the sentence in question is Λ itself. So, Λ does in some sense refer to itself, namely, by description. But it does seem noteworthy that the 'self-referential' sentence diagonalization produces is quite different from the sentence that was our inspiration. And it is worth emphasizing again that $\neg T(\ulcorner \Lambda \urcorner)$ is not self-referential in any sense: The only sentence to which it refers is Λ.

Of course, in many cases, perhaps even in most cases, this difference makes no difference. Since Λ is provably equivalent to $\neg T(\ulcorner \Lambda \urcorner)$, even if it is not *actually* $\neg T(\ulcorner \Lambda \urcorner)$—whereas The Liar really is just "The Liar is not true"—we can work back and forth across the biconditional when we seek to replicate the liar paradox in PA_S. But what the examples presented earlier show is that the difference does make a difference in some cases.

The foregoing thus suggests two questions. First, the preceding shows that there is natural, informal reasoning involving self-reference that cannot be replicated in PA_S *via* the diagonal lemma.[6] In what kinds of theories can the preceding reasoning be formalized, then? Second, the limiations of PA_S and the usual form of the diagonal lemma make for some amusing and perhaps instructive examples. But are they of any real significance?

2. Two Solutions to the Puzzle

The difficulties we encountered above with these examples

(1) (1) is true iff the right-hand side of (1) is false.

(14) The right-hand side of (14) is true iff the left-hand side of (14) is false.

derive from the fact that these sentences make reference to features of their own syntax.[7] The formulae that are produced by the standard proofs of the diagonal lemma, however, do not preserve the syntactic structures of the formulae from which they are produced: If we diagonalize on $T(x) \equiv \exists y(rhs(x,y) \wedge T(y))$, then we get a formula G that is provably equivalent to $T(\ulcorner G \urcorner) \equiv \exists y(rhs(\ulcorner G \urcorner, y) \wedge T(y))$, but G does not have the same syntactic structure as this formula and so (for some purposes) is not an appropriate formalization of the self-referential sentence (1). What we need, then, is a strengthened form of the diagonal lemma that produces a sentence that has the same syntactic structure as the sentence on which we are diagonalizing.

[6] As a couple readers pointed out, this is, strictly speaking, not surprising, since no one would have expected reasoning involving reference to the orthography of English sentences—consider "This sentence is true if, and only if, the clause contained in it that begins with the word 'the' is false"—to be replicable in PA. But the examples we are discussing exploit nothing but resources that are anyway needed in theories of truth for the language of arithmetic: We need, for example, to be able to say that a biconditional is true if, and only if, its left- and right-hand sides have the same truth-value. It therefore is, or so it seems to me, unexpected that the foregoing reasoning cannot easily be formalized.

[7] As a result, it is reasonably easy to formulate examples that illustrate the problem with which we're concerned by using purely syntactic predicates. (Consider, for example: "This sentence is a biconditional".) Such examples do not, however, suggest—as, I take it, the above examples do—that this phenomenon might be any more than a curiosity.

2.1. **Enriching the Language.** One way to get a stronger form of the diagonal lemma is to enrich the language in which we are working. The problem we are discussing arises because PA_S is term-poor. If one works with a language that has more terms than PA_S does, then this problem does not arise. Suppose, in particular, that we expand the language of arithmetic so that it contains function symbols for all primitive recursive functions and add as axioms the equations that primitively recursively define the relevant functions: Call the resulting theory PA^+. Then we can prove the diagonal lemma in the following strong form:[8]

Lemma. *(Strong Diagonal Lemma) Let $B(x)$ be a formula containing just x free. Then there is a term τ_B such that PA$^+$proves: $\tau_B = \ulcorner B(\tau_B)\urcorner$.*

Then, of course, PA^+ will also prove $B(\tau_B) \equiv B(\ulcorner B(\tau_B)\urcorner)$, and the formula $B(\tau_B)$—which is now playing the role of G—really is self-referential: The formula denoted by τ_B really is $B(\tau_B)$. And obviously $B(\tau_B)$ will have the same syntactic structure as the formula $B(x)$ on which we were diagonalizing, except for the fact that the former contains a complex term where the latter contains only a variable. Thus, for example, in the case of (1), the strong diagonal lemma produces a term τ such that PA^+ proves:

$$(1') \qquad \tau = \ulcorner T(\tau) \equiv \exists y(rhs(\tau, y) \wedge \neg T(y))\urcorner.$$

By the laws of identity, then, PA^+ proves:

$$(20) \qquad T(\tau) \equiv T(\ulcorner T(\tau) \equiv \exists y(rhs(\tau, y) \wedge \neg T(y))\urcorner).$$

One instance of the T-scheme is then:

$$(21) \qquad T(\ulcorner T(\tau) \equiv \exists y(rhs(\tau, y) \wedge \neg T(y))\urcorner) \equiv [T(\tau) \equiv \exists y(rhs(\tau, y) \wedge \neg T(y))]$$

So, if (21), then:

$$(2') \qquad T(\tau) \equiv [T(\tau) \equiv \exists y(rhs(\tau, y) \wedge \neg T(y))],$$

and so by the associativity and reflexivity of the biconditional:

$$(4') \qquad \exists y(rhs(\tau, y) \wedge \neg T(y))$$

But now $T(\tau) \equiv \exists y(rhs(\tau, y) \wedge \neg T(y))$ really does have a right-hand side, namely, $\exists y(rhs(\tau, y) \wedge \neg T(y))$, and PA^+ proves this fact. That is, PA^+ proves:

$$(5') \qquad rhs(\tau, y) \equiv [y = \ulcorner \exists y(rhs(\tau, y) \wedge \neg T(y))\urcorner]$$

So PA^+ proves that, if (21), then:

[8]This version of the diagonal lemma seems to have made its first appearance in **?**. It figures importantly in discussions of some non-classical theories of truth, such as Kripke's (**?**), since $\Lambda \equiv \neg T(\ulcorner \Lambda \urcorner)$ is not true (it is, in fact, paradoxical) and so had better not be a theorem. See **?** for some related observations.

(6′) $\neg T(\ulcorner \exists y(rhs(\tau, y) \wedge \neg T(y))\urcorner)$.

But now the T-scheme delivers:

(7′) $T(\ulcorner \exists y(rhs(\tau, y) \wedge \neg T(y))\urcorner) \equiv \exists y(rhs(\tau, y) \wedge \neg T(y))$,

and so PA^+ proves that, if (21) and (7′), then:

(8′) $T(\ulcorner \exists y(rhs(\tau, y) \wedge \neg T(y))\urcorner)$,

since (4′) is the right-hand side of (7′). But (8′) contradicts (6′). So PA^+ proves that (21) and (7) lead to contradiction, and the formal argument is simply a transcription of the informal argument considered earlier.

Consideration of other such examples—I'll leave the formal derivation of a contradiction from (14) as an exercise—makes it plausible that a great deal of informal reasoning involving self-reference, though it cannot be replicated in PA_S using the diagonal lemma, can be replicated in PA^+ using the strong diagonal lemma. There is, of course, such reasoning that cannot be formalized in PA^+, since it appeals to mathematical principles stronger than any available in PA^+. But the contrast between PA_S and PA^+ in which we are interested concerns the form of the diagonal lemma that is provable in the two theories, and that contrast is a function not of their strength—PA_S and PA^+ have the same proof-theoretic strength, since PA^+ can be interpreted in PA_S *via* the usual definition of primitive recursive functions in the latter—but rather of the expressive power of the languages in which they are formulated. We could therefore replace PA_S in the discussion to follow with ZFC, formulated in the usual language $\{\in\}$, since this language, like that of PA_S, is term-poor: The diagonal lemma as it is typically proven in ZFC is of the weaker rather than the stronger form, and so the informal arguments we have been discussing cannot be formalized in ZFC either, at least not straightforwardly.

So let us ask this question: Can informal reasoning involving self-reference that can be replicated in PA^+ using the strong diagonal lemma also be replicated in PA_S, somehow or other?

2.2. **The Structural Diagonal Lemma.** The informal reasoning that establishes the inconsistency of (1) and (14) can be replicated in PA_S *via* a form of the diagonal lemma that is stronger than the standard form mentioned earlier. We may call it the "structural" form of the diagonal lemma:[9]

Lemma 1. *(Structural Diagonal Lemma) Let P be a truth-functional schema in (distinct) sentence-letters $p_1, \ldots, p_n$. Let $A(x)$ be the substitution instance of P in which each p_i has been replaced by a corresponding formula $A_i(x)$ containing just x free. Then there is a formula G such that:*

[9] I do not know to what extent this result can be extended so that *G* reflects the *quantificational* structure of $A(x)$.

(i) G is the substitution instance of P in which each p_i has been replaced by a corresponding formula G_i;
(ii) $\mathrm{PA}_S \vdash G_i \equiv A_i(\ulcorner G \urcorner)$*;*
(iii) $\mathrm{PA}_S \vdash G \equiv A(\ulcorner G \urcorner)$.

In fact, (iii) follows from (i) and (ii): If G just is the result of substituting G_i for p_i in P, and $A(x)$ is the result of substituting $A_i(x)$ for p_i in P, then $\mathrm{A}(\ulcorner G \urcorner)$ is the result of substituting $A_i(\ulcorner G \urcorner)$ for p_i in P; and so, since PA_S proves: $G_i \equiv A_i(\ulcorner G \urcorner)$, it certainly proves $G \equiv A(\ulcorner G \urcorner)$. I should perhaps emphasize that there is no misprint in condition (ii): The condition is *not* that PA_S should prove: $G_i \equiv A_i(\ulcorner G_i \urcorner)$. This weaker condition would not serve our purposes.[10]

Let me give a simple example to illustrate what the structural diagonal lemma says. Let $A(x)$ be: $A_1(x) \vee (A_2(x) \wedge \neg A_3(x))$, x being the sole free variable. Then the structural diagonal lemma says that there is a sentence $G = G_1 \vee (G_2 \wedge \neg G_3)$ such that PA_S proves that each subsentence G_i is equivalent to $A_i(\ulcorner G \urcorner)$ and that G itself is equivalent to $A(G)$.

If we had the structural diagonal lemma, then we could resolve our puzzlement about (1) and (14). Consider (1), again. We start as before with the formula:

(9) $\qquad T(x) \equiv \exists y(rhs(x,y) \wedge \neg T(y))$

and now apply the structural diagonal lemma to get a formula G of the form $G_1 \equiv G_2$ such that PA_S proves:

(22) $\qquad G_1 \equiv T(\ulcorner G \urcorner)$
(23) $\qquad G_2 \equiv \exists y(rhs(\ulcorner G \urcorner, y) \wedge \neg T(y))$
(24) $\qquad G \equiv [T(\ulcorner G \urcorner) \equiv \exists y(rhs(\ulcorner G \urcorner, y) \wedge \neg T(y))]$

Working in PA_S, we can now replicate the informal reasoning that led us to a contradiction. By (24) and the associativity of the biconditional, $[G \equiv T(\ulcorner G \urcorner)] \equiv \exists y(rhs(\ulcorner G \urcorner, y) \wedge \neg T(y))]$. So, by the T-scheme, $\exists y(rhs(\ulcorner G \urcorner, y) \wedge \neg T(y))$. And now G really is a biconditional and it really does have a right-hand side, namely, G_2, and this fact is provable in PA_S: $rhs(\ulcorner G \urcorner, y) \equiv y = \ulcorner G_2 \urcorner$. So $\neg T(\ulcorner G_2 \urcorner)$. But then by the T-scheme again, $\neg G_2$ and so, by (23), $\neg \exists y(rhs(\ulcorner G \urcorner, y) \wedge \neg T(y))$. Contradiction.

2.3. **Proof of the Structural Diagonal Lemma.** The structural diagonal lemma would thus solve the problem with which I opened this note. So let's prove it. We can derive it from the diagonal lemma as Gödel originally proved it.[11]

[10] With this weaker condition replacing the stronger one, the lemma would follow from the generalized diagonal lemma—the version that applies to several formulae simultaneously. See note 4.

[11] Note again that nothing in the proof actually requires the assumption that $A(x)$ contains *only* x free, so the proof about to be given easily adapts to a proof of the form that allows parameters.

Proof. We need some definitions. Given a formula $A(x)$ containing just x free, the standard proof of the diagonal lemma to which I referred earlier yields a sentence $G_{A(x)}$ such that PA_S proves: $G_{A(x)} \equiv A(\ulcorner G_{A(x)} \urcorner)$. We can think of the proof as defining a syntactic function, whose value for a formula $A(x)$ is the formula $G_{A(x)}$. This function is obviously recursive, so it is representable in PA_S. Let $spdl(x,y)$ represent it. So $spdl(x,y)$ means: y is the Gödel number of the formula that the *s*tandard *p*roof of the *d*iagonal *l*emma yields when applied to the formula whose Gödel number is x. We may, of course, also think of $spdl(x,y)$ the other way around, as it were: Since it defines a one-one relation, we may think of it as determining *from* a formula $G_{A(x)}$ the formula $A(x)$ from which it was constructed—if, of course, $G_{A(x)}$ is a formula of the appropriate kind, that itself being decidable.

The function that we actually need is definable from this one: Given a formula $G_{A(x)}$, recover the formula $A(x)$ from which it was constructed; then substitute $\ulcorner G_{A(x)} \urcorner$ for x in $A(x)$, thus getting our old friend $A(\ulcorner G_{A(x)} \urcorner)$, which we shall call the *Gödel equivalent* of $G_{A(x)}$—so called because it is the formula the standard proof of the diagonal lemma shows to be equivalent to $G_{A(x)}$. Clearly, Gödel equivalence is also recursive and so is representable in PA_S by a formula $gdleq(x,y)$, which means: y is the Gödel number of the the Gödel equivalent of the formula whose Gödel number is x. And so in general, then, PA_S proves:

$$gdleq(\ulcorner G_{A(x)} \urcorner, y) \equiv y = \ulcorner A(\ulcorner G_{A(x)} \urcorner) \urcorner$$

Now, by the supposition of the structural diagonal lemma, $A(x)$ is a substitution instance of a formula P containing the sentence-letters $p_1, ..., p_n$, where each p_i is replaced by $A_i(x)$. Let $Gdleq_{A(x)}(x)$ be the formula that results from replacing each p_i by: $\exists y(gdleq(x,y) \wedge A_i(y))$. Then the *standard* proof of the diagonal lemma gives us a formula S such that PA_S proves:

$$S \equiv Gdleq_{A(x)}(\ulcorner S \urcorner).$$

Claim. $\mathrm{PA}_S \vdash gdleq(\ulcorner S \urcorner, y) \equiv y = \ulcorner Gdleq_{A(x)}(\ulcorner S \urcorner) \urcorner$.

Proof. Since $gdleq(x,y)$ represents Gödel equivalence, it is enough to show that $Gdleq_{A(x)}(\ulcorner S \urcorner)$ is indeed the Gödel equivalent of S. But S is the formula you get by diagonalizing on $Gdleq_{A(x)}(x)$: That is, S is $G_{Gdleq_{A(x)}(x)}$. So the Gödel equivalent of S is the formula you get by substituting the numeral for the Gödel number of S into $Gdleq_{A(x)}(x)$, that is: $Gdleq_{A(x)}(\ulcorner S \urcorner)$, as wanted. □

As it happens, the proof of the diagonal lemma in **?**, pp. 53–4 also delivers the structural diagonal lemma. The point is that what Boolos calls 'pseudo-terms' are eliminated in a way that gives the newly introduced quantifiers narrow scope: It is, in a sense, 'compositional', so it does not alter the truth-functional structure of the sentence to which it is applied. (Thanks to Albert Visser for explaining this point to me.) We might therefore say that what the examples with which we began show is that the treatment in **?** is superior to the treatment in **?**, because the former does not give rise to anomalies to which the latter does give rise.

Consider now $Gdleq_{A(x)}(\ulcorner S\urcorner)$: It will turn out to be our wanted formula G. By construction, it is itself a substitution instance of P, where each p_i is replaced by: $\exists y(gdleq(\ulcorner S\urcorner, y) \wedge A_i(y))$. These are the G_i. What we must show is that, for each i, PA_S proves: $G_i \equiv A_i(\ulcorner G\urcorner)$. That is, we must show that PA_S proves:

$$\exists y(gdleq(\ulcorner S\urcorner, y) \wedge A_i(y)) \equiv A_i(\ulcorner Gdleq_{A(x)}(\ulcorner S\urcorner)\urcorner).$$

But this follows immediately from the claim.

As noted earlier, it now follows that PA_S proves: $G \equiv A(\ulcorner G\urcorner)$. In this case, what follows is that $Gdleq_{A(x)}(\ulcorner S\urcorner)$ is provably equivalent to the formula that results from replacing each p_i in P by $A_i(\ulcorner Gdleq_{A(x)}(\ulcorner S\urcorner)\urcorner)$. □

The proof may make more sense as applied to an example. Suppose as earlier that $A(x)$ is $A_1(x) \vee (A_2(x) \wedge \neg A_3(x))$. Then $Gdleq_{A(x)}(x)$ is:

$$\exists y(gdleq(x,y) \wedge A_1(y)) \vee [\exists y(gdleq(x,y) \wedge A_2(y)) \wedge \exists y(gdleq(x,y) \wedge A_3(y))]$$

Diagonalization then yields a formula S such that PA_S proves:

$$S \equiv \exists y(gdleq(\ulcorner S\urcorner, y) \wedge A_1(y)) \vee [\exists y(gdleq(\ulcorner S\urcorner, y) \wedge A_2(y)) \wedge \exists y(gdleq(\ulcorner S\urcorner, y) \wedge A_3(y))].$$

The right-hand side of this formula is our formula G. Each G_i is the corresponding part of G. That is, G_i is: $\exists y(gdleq(\ulcorner S\urcorner, y) \wedge A_i(y))$. Since G is the Gödel equivalent of S, PA_S proves: $gdleq(\ulcorner S\urcorner, y) \equiv y = \ulcorner G\urcorner$, and so each G_i is provably equivalent to $A_i(\ulcorner G\urcorner)$. Hence G itself is provably equivalent to: $A_1(\ulcorner G\urcorner) \vee (A_2(\ulcorner G\urcorner) \wedge \neg A_3(\ulcorner G\urcorner))$.

3. Self-Reference and the Languages of Arithmetic

The informal reasoning that demonstrates the inconsistency of (1) and (14) can thus be replicated in PA_S *via* the structural diagonal lemma. The examples with which we began thus serve only to highlight the significance of that form of the diagonal lemma. But there are other examples that pose quite a different problem.

Consider the following two principles:[12]

(Not) The negation of a sentence A is true iff the sentence A itself is not true.

(Disq) A sentence of the form $\ulcorner t$ is true$\urcorner$ is true iff t denotes a sentence that is itself true.

These two principles—the usual clause for negation and the disquotation principle—taken together, are intuitively inconsistent. Consider again the liar sentence:

(The Liar) The Liar is not true.

[12] I have discussed the philosophical significance of (Disq), and of the results to be discussed here, elsewhere (**?**). Note that the term t in (Disq) may be an arbitrary closed term.

Since The Liar is the sentence "The Liar is not true", The Liar is true if, and only if, "The Liar is not true" is true. By the first of the two principles above, however, "The Liar is not true" is true if, and only if, "The Liar is true" is not true. By the second, "The Liar is true" is not true if, and only if, The Liar is not true. So The Liar is true if, and only if, The Liar is not true. Contradiction.

This argument may be formalized straightforwardly in PA^+. We may formalize (Not) and (Disq) as follows:[13]

(Not_F) $\quad T(\ulcorner\neg A\urcorner) \equiv \neg T(\ulcorner A\urcorner)$

(Disq_F) $\quad T(t) \equiv T(\ulcorner T(t)\urcorner)$

where t may be any closed term. This theory is easily seen to be inconsistent *via* the argument just discussed. For, by the strong form of the diagonal lemma, there is a term λ such that PA^+ proves $\lambda = \ulcorner\neg T(\lambda)\urcorner$ and therefore proves:

$$\begin{aligned} T(\lambda) \equiv & T(\ulcorner\neg T(\lambda)\urcorner), \text{ by identity} \\ \equiv & \neg T(\ulcorner T(\lambda)\urcorner), \text{ by } (\mathrm{Not}_F) \\ \equiv & \neg T(\lambda), \text{ by } (\mathrm{Disq}_F) \end{aligned}$$

As we shall see, however, there are models of $\mathrm{PA}_S + (\mathrm{Not}_F) + (\mathrm{Disq}_F)$, which is therefore consistent.

A great deal of care is needed here regarding what theory is being said to be consistent. As noted above, the expression $\ulcorner A\urcorner$ is an abbreviation for the numeral that denotes the Gödel number of the expression A. So, for example, if the Gödel number of A is 3084, then $\ulcorner A\urcorner$ abbreviates: $\underbrace{S\cdots S}_{3084}0$. The important point for present purposes is that what $\ulcorner A\urcorner$ abbreviates depends upon what Gödel numbering we are using. For this reason, there is no such thing as *the* theory $\mathrm{PA}_S + (\mathrm{Not}_F) + (\mathrm{Disq}_F)$. There are as many such theories as there are Gödel numberings.

---

[13] These axioms should be understood as restricted to sentences of the language in question. Strictly speaking, that is to say, "sentence$(t) \rightarrow$" should precede both axioms. I will ignore this fact below to simplify the exposition. I will also stick here with the case of truth rather than deal with the more general notion of satisfaction. What follows should smoothly extend to that case.

As Albert Visser pointed out to me, one needs to take (Disq_F) in the form given in the text: The weaker principle

(Not^*_F) $\quad T(\ulcorner A\urcorner) \equiv T(\ulcorner T(\ulcorner A\urcorner)\urcorner)$

will not suffice, since it then will not follow that $T(\lambda) \equiv T(\ulcorner T(\lambda)\urcorner)$, λ being a term rather than a numeral. Indeed, the resulting theory is then consistent.

Visser also raised the question what the relation is, in general, between the weaker principles, such as (Not^*_F), and the corresponding stronger priniciples, such as (Not_F) , and he speculated that the stronger principle can be derived from the weaker one if we have the following rule of inference, which acts as a principle of extensionality: From $\vdash A \equiv B$, infer $\vdash T(A) \equiv T(B)$. Visser was right: The proof is by induction on the depth of the embedding and is fairly straightforward. For example, suppose $\vdash \lambda = \ulcorner A\urcorner$. Then $\vdash T(\lambda) \equiv T(\ulcorner A\urcorner)$, by identity, so also $\vdash T(\ulcorner T(\lambda)\urcorner) \equiv T(\ulcorner T(\ulcorner A\urcorner)\urcorner)$. So, given (Not^*_F), we thus have $\vdash T(\lambda) \equiv T(\ulcorner T(\lambda)\urcorner)$.

In many contexts, this point is not particularly relevant and so is not explicitly noted.[14] When we were discussing different forms of the diagonal lemma above, for example, we did not specify any particular Gödel numbering that we were using, for we did not really need to do so: No matter what Gödel numbering g we are using, there will be, for each formula $A(x)$, a formula G such that PA_S proves: $G \equiv A(\ulcorner G \urcorner^g)$, where $\ulcorner G \urcorner^g$ means: the numeral denoting $g(G)$, that is, the Gödel number of G, where g is the Gödel numbering we are using. Something similar is true of the argument just given: No matter what Gödel numbering g we are using, there will be a term λ_g such that PA^+ proves: $\lambda_g = \ulcorner \neg T(\lambda_g) \urcorner^g$. Which term λ_g is will depend upon which Gödel numbering we are using, but its existence does not so depend, and so the proof of the inconsistency of $\mathrm{PA}^+ + (\mathrm{Not}_F) + (\mathrm{Disq}_F)$ therefore does not depend upon which Gödel numbering we are using, either. For that reason, we may pretend that there is such a thing as *the* theory $\mathrm{PA}^+ + (\mathrm{Not}_F) + (\mathrm{Disq}_F)$, although in fact there is not: There are many such theories, one for each Gödel numbering of the expressions of the language of PA^+. All of these theories are inconsistent, however, and parallel proofs demonstrate the inconsistency in each case.

We are about to see, however, that, in the investigation of axiomatic theories of truth formalized in PA_S, it can matter very much which Gödel numbering we are using and so which of the many different formalizations of (Not) and (Disq) we consider.[15] In particular, we shall see that, if we use certain Gödel numberings—the most natural ones—then the relevant formalization is consistent, but, if we use other Gödel numberings—quite unnatural ones—then the relevant formalization is inconsistent. To put the point differently, there are consistent theories of the 'form' $\mathrm{PA}_S + (\mathrm{Not}_F) + (\mathrm{Disq}_F)$ and there are inconsistent theories of the 'form' $\mathrm{PA}_S + (\mathrm{Not}_F) + (\mathrm{Disq}_F)$—where the form is given content by our choosing a particular Gödel numbering.

3.1. **The Consistency of Some Theories of the Form $\mathbf{PA}_S + (\mathbf{Not}_F) + (\mathbf{Disq}_F)$.**

Let me begin by proving the former claim.

For the moment, leave open which Gödel numbering of the language of PA_S we are using. I shall now define a set $\mathbf{E}$ and prove that the interpretation of the language of PA_S that interprets T by $\mathbf{E}$ and is otherwise standard verifies both (Not_F) and (Disq_F)—*if* the Gödel numbering satisfies a certain condition. This is a condition all the usual sorts of Gödel numberings do indeed satisfy.

[14] Thus, to illustrate with just one example (chosen only because we shall discuss this paper further below): In their investigations of axiomatic theories of truth **?**, Friedman and Sheard never specify a particular Gödel numbering, apparently assuming without comment that their results will not depend upon which Gödel numbering is used. As we shall see, this assumption appears to have been correct, but that is only because of other assumptions they make. Had they instead considered a different base theory formulated in a different base language—in particular, had they worked in PA_S rather than in PA^+—the assumption would have needed defense and would have failed but for their assumption of what I shall call below the 'basic truth-principles'.

[15] That there are contexts in which it matters what Gödel numbering we use is a central point of Feferman's classic (**?**). But those sorts of cases, connected with Gödel's second incompleteness theorem, are very different from the ones considered here.

Consider some closed sentence S. This sentence may contain various sub-formulae of the form $T(s_i)$, where s_i is a closed term (which need not be a numeral) that denotes the Gödel number of a closed sentence S_i of the form: $\neg^k T(t_i)$,[16] where k may be zero and t_i is also a closed term (which, similarly, need not be a numeral). Now consider the result of replacing each such term s_i by a left quote, followed by the sentence A_i, followed by a right quote. Of course, the result of this replacement is not a sentence of the language of arithmetic, but never mind that: It is a sentence of a different language, one that will serve our purposes here, and—importantly for later—it can be discussed in the language of PA_S using Gödel numbering. So, that said, this new expression may itself contain closed sentences of the form $T(s_i)$, where the s_i are again closed terms that are the Gödel numbers of sentences S_i that are also of the form: $\neg^k T(t_i)$. If so, replace s_i by: 'S_i', and then continue the process.[17] Call this the process of *quotational expansion* and, if the process terminates in a given case, call the end result the quotational expansion of the original sentence. Call a sentence *even* if it has a quotational expansion that contains an even number of negation symbols and *odd* if it has a quotational expansion that contains an odd number of negation symbols. The obvious induction shows that, if a sentence has a quotational expansion at all, it has a unique one. So no sentence is both even and odd. But we are not assuming, and have not proven, that every sentence has a quotational expansion, so we are not assuming, and have not proven, that every sentence is either even or odd.

Without addressing that issue, however, we can already use the notion of quotational expansion to show that all theories of this form

$$\begin{array}{ll} (\mathrm{Not}\rightarrow) & T(\ulcorner\neg A\urcorner) \rightarrow \neg T(\ulcorner A\urcorner) \\ (\mathrm{Disq}_F) & T(t) \equiv T(\ulcorner T(t)\urcorner) \end{array}$$

are consistent, no matter what Gödel numbering we might be using. Let $\mathbf{E}$ be the set of even sentences. Interpret the standard language of arithmetic in the usual way and take the extension of T to be the set $\mathbf{E}$. Then it is easy to see that (Disq_F) is true: The first step in the quotational expansion of $T(\ulcorner A\urcorner)$ is to replace it with: 'A', so if A has a quotational expansion A', the quotational expansion of $T(\ulcorner A\urcorner)$ is just: 'A''; hence either both are even or neither is. And $(\mathrm{Not}\rightarrow)$ is also true: If $\neg A$ is even, then it has a quotational expansion that contains evenly many negation symbols; but the quotational expansion of A is just the quotational expansion of $\neg A$ without the initial negation, so it is not even but odd.

It should be clear that a corresponding argument shows that the corresponding theory is consistent in PA^+. But we cannot prove that the full theory $\mathrm{PA}^+ + (\mathrm{Not}_F) + (\mathrm{Disq}_F)$ is consistent—fortunately, since we have already seen that it is *in*consistent—because we cannot show that

[16] Here, $\neg^k A$ means: $\underbrace{\neg \ldots \neg}_{k} A$.

[17] To be completely precise here, we may need to make use of the kind of machinery found in **?**.

$(\text{Not} \leftarrow) \qquad \neg T(\ulcorner A \urcorner) \rightarrow T(\ulcorner \neg A \urcorner)$

is true in the corresponding interpretation. That a sentence A is not even is, for all we have said so far, compatible with its not having a quotational expansion at all: It won't have one if the process of quotational expansion does not terminate. And if the process does not terminate when applied to A, then it will not terminate when applied to $\neg A$, either, in which case $\neg A$ will not be even—nor will it be odd—and $(\text{Not} \leftarrow)$ will be false. If, on the other hand, the process terminates for *every* sentence A, then every sentence is either even or odd. Hence, if A is not even, then it is odd, that is, it has a quotational expansion that contains an odd number of negation symbols. But then, the quotational expansion of $\neg A$ is just the quotational expansion of A with an extra negation in front and $\neg A$ is therefore even. Hence, $(\text{Not} \leftarrow)$ is true.

We may state this result as a lemma:

Lemma 2. *Given a particular Gödel numbering for the sentences of the language of a theory* Θ*, the corresponding theory of the form* $\Theta + (\text{Not}_F) + (\text{Disq}_F)$ *is consistent if the process of quotational expansion—applied to sentences of the language of* Θ *using that Gödel numbering—always terminates.*

Of course, what we proved above already implies that, if the theory in question is PA^+, then it is impossible to define a Gödel numbering with respect to which the process of quotational expansion will always terminate: $\text{PA}^+ + (\text{Not}_F) + (\text{Disq}_F)$ is inconsistent no matter what Gödel numbering is being used. Nonetheless, it is worth noting that we can prove this fact directly.

Fix a Gödel numbering for the language of PA^+. By the strong form of the diagonal lemma, there is a term τ such that PA^+ proves: $\tau = \ulcorner T(\tau) \urcorner$. Since everything PA^+ proves is true, the term τ really does denote the Gödel number of the sentence $T(\tau)$. Hence, the first step of the process of quotational expansion would have us replace τ in $T(\tau)$ by: '$T(\tau)$', thus arriving at: $T(\text{`}T(\tau)\text{'})$. This sentence now contains a sentence of the form $T(t)$ where the term $t$ denotes a sentence of the form: $\neg^k T(u)$: Unfortunately, the term in question is just $\tau$ again, and the next step in the process of expansion simply yields: $T(\text{`}T(\text{`}T(\tau)\text{'})\text{'})$. The process of expansion as applied $T(\tau)$ thus fails to terminate.

On the other hand, if the language is the standard language of arithmetic, then the process of quotational expansion will terminate if we use either of the Gödel numberings employed by **?** or **?**, ch 15, and indeed most of the ones usually employed.[18] Such Gödel numberings satisfy the following conditions:[19]

[18]There are so many, of course, that there is really no prospect of checking them all.

[19]Here t and u are arbitrary closed terms. The Gödel numberings mentioned also satisfy similar conditions with respect to the other truth-functional operators and even with respect to quantifiers, though we shall not need these facts here.

$$\ulcorner 0 \urcorner \geq 0$$
$$\ulcorner St \urcorner > S(\ulcorner t \urcorner)$$
$$\ulcorner t + u \urcorner > \ulcorner t \urcorner + \ulcorner u \urcorner$$
$$\ulcorner t \times u \urcorner > \ulcorner t \urcorner \times \ulcorner u \urcorner$$
$$\ulcorner T(t) \urcorner > \ulcorner t \urcorner$$
$$\ulcorner \neg A \urcorner > \ulcorner A \urcorner$$

Let us call such a Gödel numbering *regular*. By the obvious inductions:

$$\ulcorner t \urcorner \geq \mathrm{den}(t)$$
$$\ulcorner \neg^k T(t) \urcorner > \ulcorner t \urcorner$$

where $\mathrm{den}(t)$ is the denotation of the term t. Hence:

$$\ulcorner \neg^k T(t) \urcorner > \mathrm{den}(t).$$

That is to say, regular Gödel numberings are well-founded with respect to the process of quotational expansion: If one has a sentence of the form $T(s)$, where s is the Gödel number of a sentence of the form $\neg^k T(t)$ and t is a closed term, then the term t must denote a number that is strictly less than what s denoted. If so, then the process of quotational expansion must terminate: The denotations of these terms cannot decrease endlessly.

We thus have the following result:

Theorem 3. *Let g be a regular Gödel numbering. Then the result of adding the axioms*

(Not$_F^g$) $\quad T(\ulcorner \neg A \urcorner^g) \equiv \neg T(\ulcorner A \urcorner^g)$
(Disq$_F^g$) $\quad T(t) \equiv T(\ulcorner T(t) \urcorner^g)$

to PA$_S$ is consistent.[20]

Something even stronger is true: This theory is consistent if PA_S is, because it is actually interpretable in PA_S. We can define the notions of quotational expansion and of an even sentence in the language of PA_S and then, so long as our Gödel numbering is provably regular, prove in PA_S that every formula has a quotational expansion and so is either odd or even. Having done so, we can then add to PA_S the definition $T(x) \equiv \mathrm{Even}(x)$ and prove (Not_F) and (Disq_F) as above. And it can be proven in PA_S (and in far weaker theories, too) that we can do all of that. So PA_S itself proves that, if PA_S is consistent, then so is PA_S+(Not_F^g)+(Disq_F^g), so long as g is provably regular.

[20]Indeed, the result of adding these axioms to the set of all truths in $\{0, S, +, \times\}$ is consistent.

3.2. **A Result In Which the Notion of Self-Reference Occurs Essentially.** We can adapt the foregoing to state a result in which the notion of self-reference occurs essentially, that is, in which the assumption that there exists a self-referential sentence of a certain form cannot be replaced by an assumption to the effect that something of the form $G \equiv A(\ulcorner G \urcorner)$ is provable.

Let $\mathscr{O}$ be an interpreted language containing a predicate T, intended to be understood as a truth-predicate. Consider the principle:[21]

(25) $\ulcorner T(t) \urcorner$ is true iff the denotation of t is true.

This principle is of interest because it is the obvious principle to adopt as governing the truth-predicate contained in $\mathscr{O}$: What (25) says is that a sentence of $\mathscr{O}$ of the form $\ulcorner T(t) \urcorner$ is true if, and only if, the sentence denoted by t is itself a true sentence of $\mathscr{O}$. The principle (25) is thus perfectly analogous to such familiar semantic clauses as:

(26) $\ulcorner t$ is red$\urcorner$ is true iff the denotation of t is red

and

(27) $\ulcorner t$ is even$\urcorner$ is true iff the denotation of t is even.

It is important to note, moreover, that (25) fully respects the distinction between object-language and meta-language: There is no truth-predicate here that applies to sentences of the meta-language; in particular, both the predicate 'is true', which belongs to the meta-language, and the predicate T, which belongs to the object-language, can be supposed intelligibly to apply only to sentences of the *object*-language.[22] So one might hope that (25) would be immune to inconsistency. But it is not. If the object-language contains a self-referential liar sentence, then only very weak assumptions are needed to generate a contradiction within a classical meta-language.[23]

So suppose that $\mathscr{O}$ contains a truly self-referential liar sentence, that is, that $\mathscr{O}$ contains a term λ that denotes the sentence $\ulcorner \neg T(\lambda) \urcorner$. That is, suppose that:

(28) the denotation of λ is $\ulcorner \neg T(\lambda) \urcorner$.

One instance of (25) is then:

(29) $\ulcorner T(\lambda) \urcorner$ is true iff the denotation of λ is true.

[21] I am now using corner quotes in Quine's way, to denote, in this case, the result of putting t in the argument place of T. (Note that 'T' is a *name* of a predicate, not a predicate.)

[22] The language $\mathscr{O}$ does not respect this distinction.

[23] Those familiar with the literature on the liar paradox will recognize that the phrase "within a classical meta-language" hides several other assumptions that can be and have been denied in an effort to avoid inconsistency. But the point here is not to introduce a new form of the liar paradox—I'm sure this one can be accomodated by extant approaches—but to contrast forms of it that depend upon diagonalization with one that makes direct use of self-reference. In short: My intent here is to throw some light on self-reference, not on truth.

But then, by (28) and (29):

(30) $\ulcorner T(\lambda)\urcorner$ is true iff $\ulcorner \neg T(\lambda)\urcorner$ is true.

If, now, we have the usual clause for negation:

(31) $\ulcorner \neg A\urcorner$ is true iff A is not true,

that will deliver:

(32) $\ulcorner \neg T(\lambda)\urcorner$ is true iff $\ulcorner T(\lambda)\urcorner$ is not true.

But then, putting (30) and (32) together, we have:

(33) $\ulcorner T(\lambda)\urcorner$ is true iff $\ulcorner T(\lambda)\urcorner$ is not true.

And that is a classical contradiction, one derived in an informal meta-theory that is therefore inconsistent.[24]

Formalizing the foregoing, we thus have the following result.

Lemma 4. *Let $\mathscr{O}$ be a language (not necessarily a first-order language) containing a predicate T and a unary operator $\neg$; let Σ be a classical semantic theory for $\mathscr{O}$ formulated in a language $\mathscr{M}$ containing the following (primitive or defined) expressions: $term_{\mathscr{O}}(x)$, x is a term of $\mathscr{O}$; $sent_{\mathscr{O}}(x)$, x is a sentence of $\mathscr{O}$; $neg(x,y)$, x is the negation of y; $den_{\mathscr{O}}(x,y)$, x denotes y in $\mathscr{O}$; and $true_{\mathscr{O}}(x)$, x is true in $\mathscr{O}$. Suppose that Σ proves:*

(1) $term_{\mathscr{O}}(\lambda) \wedge den_{\mathscr{O}}(\lambda, \ulcorner \neg T(\lambda)\urcorner)]$, *for some term* λ
(2) $term_{\mathscr{O}}(t) \wedge sent_{\mathscr{O}}(s) \wedge den_{\mathscr{O}}(t,s) \rightarrow true_{\mathscr{O}}(\ulcorner T(t)\urcorner) \equiv true_{\mathscr{O}}(s)$, *for all s and* t
(3) $sent_{\mathscr{O}}(s) \wedge sent_{\mathscr{O}}(s') \wedge neg(s',s) \rightarrow true_{\mathscr{O}}(s') \equiv \neg true_{\mathscr{O}}(s)$, *for all s and* s'

Then Σ is inconsistent.

This result is closely related to one discussed in section 3.1: We can use theorem 3 to show that, if our Gödel numbering is regular, then PA_S plus (2) and (3) from the lemma is consistent, even if the object-language is the language of PA_S itself. Of course, if there is to be a possibility that this language contains a truly self-referential sentence, sentences must be numbers. So fix some standard Gödel numbering of the sentences of the language and identify sentences with their Gödel numbers: That turns (2) into ($Disq_F$) and (3) into (Not_F)—more or less—and we know that the resulting theory is consistent. What saves the theory from inconsistency is the fact that (1) fails: There is no *truly self-referential* liar sentence in this language.

[24] To reach a formal contradiction, of the form $p \wedge \neg p$, one has to use classical rules—in particular, classical structural rules—that can be denied (**?**). But, again, we are working in a purely classical meta-language here. See note 23.

We thus cannot simply replace (1) in the lemma with the assumption that, for some sentence Λ, $\Lambda \equiv \neg T(\ulcorner \Lambda \urcorner)$ is true in $\mathscr{O}$, that is, with:[25]

(1′) for some sentence Λ, $T(\ulcorner \Lambda \equiv \neg T(\ulcorner \Lambda \urcorner) \urcorner)$.

It simply isn't true that a (classical) theory satisfying (1′), (2), and (3) must be inconsistent: We have just seen a counter-example. To be sure, there are various ways inconsistency can be restored. We could, for example, add the usual clause for the biconditional:[26]

(Bicon) $T(\ulcorner A \equiv B \urcorner) \equiv [T(\ulcorner A \urcorner) \equiv T(\ulcorner B \urcorner)]$.

But that is an assumption we did not have to make for the proof of lemma 4 and whose addition would therefore weaken the result. We did not assume in the lemma that $\mathscr{O}$ so much as *contains* a biconditional, nor anything in terms of which one might be defined, let alone assume anything about how such an expression might behave.[27]

3.3. **The Strong Diagonal Lemma in the Standard Language of Arithmetic.** As noted earlier, most of the Gödel numberings actually used in practice are regular. What we have seen is thus that informal arguments involving self-reference cannot always be replicated in PA_S using the usual sorts of Gödel numberings. One such argument is the informal argument that (Not_F) and (Disq_F) are inconsistent: Since $\mathrm{PA}_S + (\mathrm{Not}_F) + (\mathrm{Disq}_F)$ is consistent with respect to regular Gödel numberings, there is not going to be a derivation in PA_S of a contradiction from (Not_F) and (Disq_F), so long as we are using a regular Gödel numbering.

The question is still open, however, whether this particular informal argument can be replicated in PA_S using some other Gödel numbering, one that is not regular: Is there a Gödel numbering with respect to which the corresponding theory of the form $\mathrm{PA}_S + (\mathrm{Not}_F) + (\mathrm{Disq}_F)$ is *in*consistent? The answer to this question is "yes". In fact, we can show something stronger: We can prove that there is a Gödel numbering of the formulae of the language of PA_S with respect to which we can prove the *strong* form of the diagonal lemma that was used to establish the inconsistency of $\mathrm{PA}^+ + (\mathrm{Not}_F) + (\mathrm{Disq}_F)$, and we can then use the same argument used to establish the inconsistency of $\mathrm{PA}^+ + (\mathrm{Not}_F) + (\mathrm{Disq}_F)$ to establish the inconsistency of $\mathrm{PA}_S + (\mathrm{Not}_F) + (\mathrm{Disq}_F)$, with respect

[25] One might have wanted to suggest we should replace (1) with the assumption that $\Lambda \equiv \neg T(\ulcorner \Lambda \urcorner)$ is provable. But in what theory? $\mathscr{O}$ is not a theory: It is a language. Certainly we can consider a theory $\mathscr{T}_{\mathscr{O}}$ formulated in $\mathscr{O}$ and assume that it proves $\Lambda \equiv \neg T(\ulcorner \Lambda \urcorner)$, but, as a little experimentation will show, we will need to make further assumptions, for example, that if $\mathscr{T}_{\mathscr{O}}$ proves $A \equiv B$, then A is true iff B is true.

[26] Given (Bicon) and that $T(\ulcorner \Lambda \equiv \neg T(\ulcorner \Lambda \urcorner) \urcorner)$, we then have $T(\ulcorner \Lambda \urcorner) \equiv T(\ulcorner \neg T(\ulcorner \Lambda \urcorner) \urcorner)$. By (Not_F), $T(\ulcorner \neg T(\ulcorner \Lambda \urcorner) \urcorner) \equiv \neg T(\ulcorner T(\ulcorner \Lambda \urcorner) \urcorner)$; by (Disq_F), $\neg T(\ulcorner T(\ulcorner \Lambda \urcorner) \urcorner) \equiv \neg T(\ulcorner \Lambda \urcorner)$; hence, $T(\ulcorner \Lambda \urcorner) \equiv \neg T(\ulcorner \Lambda \urcorner)$. See section 3.4 for more on this.

[27] For this same reason, we did not assume in lemma 4 (anything that implies) that all the T-sentences are truths of $\mathscr{O}$, since we did not assume that the T-sentences for $\mathscr{O}$ could even be expressed in $\mathscr{O}$. Nor, it is perhaps worth noting, did we assume that all T-sentences for $\mathscr{O}$ were truths of the meta-language $\mathscr{M}$, since, similarly, we did not assume that they could be expressed in $\mathscr{M}$: The object-language $\mathscr{O}$ could be arbitrarily weaker or stronger in expressive power than the meta-language $\mathscr{M}$, and the proof would still go through.

to that Gödel numbering. More generally, then, we might say: Any argument exploiting self-reference that can be formalized in PA^+ using the strong form of the diagonal lemma can be formalized in PA_S using the non-standard Gödel numbering I am about to describe, for the strong form of the diagonal lemma holds for it, too.[28]

Fix some typical Gödel numbering of the formulae of the language of PA_S: For definiteness, let it be the Gödel numbering given in **?**, ch 15. This Gödel numbering is a function $b(A)$ from expressions to natural numbers. Of course, b is a computable function whose inverse is also computable. Moreover, there is a computable enumeration ϕ of the formulae of the language of PA_S that contain just the variable x free,[29] whose inverse is also computable. We can, in fact, take this enumeration to be determined by b, for we can effectively order the formulae by their Gödel numbers and then take $\chi_n(x)$ to be the n^{th} formulae in this ordering that contains just x free. The inverse is also clearly computable: Given a formula $A(x)$ containing just x free, one need only determine its Gödel number and then determine how many other formulae containing just x free precede it.

We now define a new Gödel numbering as follows. If, for some i, A is the formula $\chi_i(\mathbf{2i}+\mathbf{1})$—that is, if A is the result of substituting the numeral for $2i+1$ for x in the i^{th} formula in the enumeration mentioned above—let its new Gödel number be $2i+1$. If A is not $\chi_i(\mathbf{2i}+\mathbf{1})$ for any i, let its new Gödel number be twice its old one. That is:

$$g(A) = \begin{cases} 2i+1 & \text{, if } A \text{ is } \chi_i(\mathbf{2i}+\mathbf{1}) \\ 2 \times b(A) & \text{, otherwise} \end{cases}$$

It should be clear that g is, indeed, a Gödel numbering: It is a one-one function from expressions to natural numbers, and, in virtue of how it was defined, it is a computable function whose inverse is also computable. Moreover, the strong form of the diagonal lemma is provable with respect to this Gödel numbering: For every formula $A(x)$ with just x free, there is quite definitely a term t such that PA_S proves: $t = \ulcorner A(t) \urcorner^g$, where $\ulcorner A(t) \urcorner^g$ abbreviates the numeral that denotes $g(A(t))$. For $A(x)$ is $\chi_i(x)$ for some i, so we may simply take t to be $\mathbf{2i}+\mathbf{1}$. Since the Gödel number of $A(\mathbf{2i}+\mathbf{1})$ is $2i+1$, what is being claimed is simply that PA_S proves: $\mathbf{2i}+\mathbf{1} = \mathbf{2i}+\mathbf{1}$, which it most certainly does.[30]

[28]The basic idea here is mentioned in passing by **?**, p. 80. I discovered it for myself when trying to reconstruct a line of thought mentioned by **?**, p. 693, fn 6. I have since learned that Albert Visser develops Kripke's suggestion in more detail in as yet unpublished work (**?**).

[29]As usual, the proof does not actually depend upon the assumption that $A(x)$ contains only x free. So we can apply this reasoning to formulae of the form $A(x,\mathbf{y})$ to show that, for each such formula, there is a term $t_{A(x,\mathbf{y})}$ such that $\mathrm{PA}_S \vdash t_{A(x,\mathbf{y})} = \ulcorner A(t_{A(x,\mathbf{y})},\mathbf{y}) \urcorner$, whence of course we have that $\mathrm{PA}_S \vdash A(t_{A(x,\mathbf{y})},\mathbf{y}) \equiv A(\ulcorner A(t_{A(x,\mathbf{y})},\mathbf{y}) \urcorner,\mathbf{y})$.

[30]Moreover, the proof can be adapted to deal with multiple formulae simultaneously. We begin by (primitive recursively) enumerating the finite sets of formulae. Suppose the first set contains $A(x,y)$ and $B(x,y)$. Then we will want to assign the Gödel numbers 1 and 3 to $A(\mathbf{1},\mathbf{3})$ and $B(\mathbf{1},\mathbf{3})$, respectively, and so forth, using the odd numbers in this process and leaving the even numbers for expressions not otherwise assigned a Gödel number. It is rather harder to

It should be clear that all syntactic notions that are numeralwise representable using the old Gödel numbering b are also numeralwise representable using the new one g. In fact, there is a simple relationship between formulae that represent syntactic notions with respect to b and ones that represent those same syntactic notions with respect to g: For example, a number n is a new Gödel number of a sentence if, and only if, the number that is the *old* Gödel number of the expression[31] whose *new* Gödel number is n is itself the *old* Gödel number of a sentence.

More formally, consider the function $b \circ g^{-1}$: Its value for a number x is the old Gödel number of the formula (if any) whose new Gödel number is x. This function is intuitively computable. By Church's thesis, then, it is recursive, so it is representable in PA_S by some formula $\mathrm{bog}(x,y)$. And if $\mathrm{synt}(x)$ represents some syntactic notion with respect to b, then $\exists y(\mathrm{bog}(x,y) \wedge \mathrm{synt}(y))$ represents it with respect to g. For example, suppose $\mathrm{sent}(x)$ represents "x is a sentence" with respect to b: That is, PA_S proves: $\mathrm{sent}(\mathbf{n})$, if $n = b(A)$ for some sentence A, and refutes it otherwise. Then $\exists y(\mathrm{bog}(x,y) \wedge \mathrm{sent}(y))$ represents "x is a sentence" with respect to g: That is, PA_S proves: $\exists y(\mathrm{bog}(\mathbf{n},y) \wedge \mathrm{sent}(y))$, if $n = g(A)$ for some sentence A, and refutes it otherwise. For since $\mathrm{bog}(x,y)$ represents $b \circ g^{-1}$ and, obviously, $b(A) = b \circ g^{-1}(n)$, PA_S proves: $\forall x[\mathrm{bog}(\mathbf{n},x) \equiv x = \mathbf{b}(\mathbf{A})]$. If A is a sentence then PA_S proves: $\mathrm{sent}(\mathbf{b}(\mathbf{A}))$, and so proves: $\mathrm{sent}(\mathbf{b}(\mathbf{A})) \wedge \mathrm{bog}(\mathbf{n},\mathbf{b}(\mathbf{A}))$, and so proves: $\exists y(\mathrm{bog}(\mathbf{n},y) \wedge \mathrm{sent}(y))$. If A is not a sentence PA_S proves: $\neg\mathrm{sent}(\mathbf{b}(\mathbf{A}))$, and so proves: $\forall x(\mathrm{bog}(\mathbf{n},x) \rightarrow \neg\mathrm{sent}(x))$, that is, $\neg\exists x(\mathrm{bog}(\mathbf{n},x) \wedge \mathrm{sent}(x))$. Syntactic facts that were formalized and proven using the old Gödel numbering can therefore easily be reformalized and reproven using the new one.

It's theft rather than honest toil, to be sure, but it works.

3.4. **A Comparison.** Harvey Friedman and Michael Sheard (**?**) provide a complete catalog of the consistent and inconsistent subsets of a large collection of intuitive principles about truth.[32] They assume as a background theory PA^+, rather than PA_S, and they assume three further principles specific to the truth-predicate, which are the universal closures of:[33]

---

write down a formula describing the new Gödel numbering in this case, but it should be clear on reflection that the method described will work.

[31] If any, of course, but let us not worry about what to do with this case.

[32] See also **?**, which continues the theme.

[33] The point of assuming these principles is that they guarntee that Base_T regards PA^+ as true: If A is a theorem of PA^+, then $\mathrm{Base}_T \vdash T(\ulcorner A \urcorner)$.

That might suggest that the basic truth-principles are strong indeed, and, in the form in which Friedman and Sheard state them, they are. Here's an illustration of that strength. Reason in Base_T. Suppose A is a theorem of PA^+. Then there are axioms $A_1,\ldots,A_n$ of PA^+ such that $A_1 \wedge \cdots \wedge A_n \rightarrow A$ is valid. By (A11), $T(\ulcorner A_1 \wedge \cdots \wedge A_n \rightarrow A \urcorner)$, and so by (A10): $T(\ulcorner A_1 \wedge \cdots \wedge A_n \urcorner) \rightarrow T(\ulcorner A \urcorner)$. So Base_T proves that, if every *conjunction of* axioms of PA^+ is true, then every theorem of PA^+ is true.

But we can also prove in Base_T that every conjunction of axioms of PA^+ is true: For any x and y , $x \dot{\rightarrow} (y \dot{\rightarrow} x \dot{\wedge} y)$ is valid and so, by (A11), true; but then logic and (A10) yield: $T(x) \wedge T(y) \rightarrow T(x \dot{\wedge} y)$, and induction and (A12) yield our conclusion. So we can prove in Base_T the *generalization* that every theorem of PA^+ is true—not just, of every theorem, that it is true. It follows that Base_T proves that, if there is some sentence in the language of PA^+that is not

(A10) $T(\ulcorner x \to y \urcorner) \wedge T(x) \to T(y)$

(A11) $\text{valid}(x) \to T(\text{ucl}(x))$

(A12) $\text{axiom}(x) \to T(x)$

where $\text{valid}(x)$ formalizes: x is the Gödel number of a valid first-order formula; $\text{ucl}(x)$: the universal closure of the formula with Gödel number x; and $\text{axiom}(x)$: x is the Gödel number of an axiom of PA^+. The result of adding these principles to PA^+ is the theory Friedman and Sheard call Base_T.

Our (Not_F) and (Disq_F) comprise one of the axiom sets Friedman and Sheard show to be inconsistent: These are equivalent to their $\{T\text{-Cons}, T\text{-Comp}, T\text{-Del}, T\text{-Rep}\}$.[34] As we have seen, however, these principles cannot be proved inconsistent in PA_S unless we select a very particular Gödel numbering with which to work. So it is reasonable to ask two questions about Friedman and Sheard's treatment: To what extent do their results depend upon the choice of PA^+ as opposed to PA_S? To what extent do their results require the use of (A10), (A11), and (A12)—which, together, we may call the 'basic truth-principles'?

The answer is that, *in the presence of the basic truth-principles*, their results do not depend upon the choice of PA^+ as opposed to PA_S. Indeed, Friedman and Sheard do not actually make use of the strong form of the diagonal lemma in their arguments. Rather, their proofs of inconsistency assume only the existence of the sort of sentence that would be delivered by the weaker form of the diagonal lemma, namely, a sentence Λ such that PA^+ proves:

(19) $\Lambda \equiv \neg T(\ulcorner \Lambda \urcorner)$

It then follows from the basic truth principles that Base_T proves:

(34) $T(\ulcorner \Lambda \urcorner) \equiv T(\ulcorner \neg T(\ulcorner \Lambda \urcorner) \urcorner)$,

---

true, then PA^+ is consistent. So any set of truth-theoretic principles strong enough to imply the antecedent of this claim will, if added to Base_T, prove $\text{Con}(\text{PA}^+)$, and there are ostensibly very weak such sets of principles.

The contrast with the usual situation regarding Tarskian truth-theories is striking: We do *not* need induction for formulae containing T for this argument. In effect, the basic truth-principles are doing the work induction on such formulae normally does in 'trivial' consistency proofs based on Tarskian theories of truth.

But the above argument would collapse if the basic truth-principles were weakened so that they took a schematic form rather than a quantified form. So (A11), for example, would become: $\text{valid}(t) \to T(\text{ucl}(t))$, for each term t. Even in schematic form, the basic truth-principles imply all instances of $T(\ulcorner A \wedge B \urcorner) \equiv T(\ulcorner A \urcorner) \wedge T(\ulcorner B \urcorner)$ and therefore imply all instances of our (Bicon) from section 3.2: $T(\ulcorner A \equiv B \urcorner) \equiv [T(\ulcorner A \urcorner) \equiv T(\ulcorner B \urcorner)]$. (The basic truth-principles do *not* imply all instances of the corresponding principle for disjunction.) So: Since PA^+ proves (19), Base_T proves $T(\ulcorner \Lambda \equiv \neg T(\ulcorner \Lambda \urcorner) \urcorner)$, whence (34) follows from the relevant instance of (Bicon). The schematic forms will thus be enough for most, and possibly all, of Friedman and Sheard's results, though the proof of the inconsistency of T-Rep, T-Del, T-Intro, and T-Elim would need to be checked carefully.

[34] Our (Not_F) is equivalent to the conjunction of T-Cons, which is $\neg[T(t) \wedge T(\dot{\neg}t)]$, and T-Comp, which is $T(t) \vee T(\dot{\neg}t)$; our (Disq_F) is equivalent to the conjunction of T-Del, which is $T(\ulcorner T(t) \urcorner) \to T(t)$, and T-Rep, which is $T(t) \to T(\ulcorner T(t) \urcorner)$.

For similar reasons, Base_T proves:[35]

(35) $\quad T(\ulcorner\neg\Lambda\urcorner) \equiv T(\ulcorner T(\ulcorner\Lambda\urcorner)\urcorner)$.

The only non-logical facts Friedman and Sheard use in the derivations of the main inconsistencies are (19), (34), and (35) (**?**, pp. 14–15). But these will all be provable in PA_S plus the basic-truth principles, too, since PA_S obviously proves (19). So, in particular, (Not_F) and (Disq_F) are inconsistent in PA_S plus the basic truth-principles, no matter what Gödel numbering we employ.[36]

The basic truth-principles, however, are not dispensible. Working in PA^+, we can find a term λ such that PA^+ proves: $\lambda = \ulcorner\neg T(\lambda)\urcorner$. If we now take Λ to be the sentence $\neg T(\lambda)$, then we will be able to prove (19) and (34). So most of Friedman and Sheard's proofs of inconsistencies go through in PA^+. But one of Friedman and Sheard's proofs assumes that there is a Λ for which just (35) can be proven, and two of them assume that there is a single sentence Λ for which both (19) and (35) can be proven.[37] I do not myself see how to construct either sort of sentence in PA^+. More significantly, Friedman and Sheard's proof that our (Disq_F)—which is their T-Rep and T-Del conjoined—is inconsistent with the two rules

(T-Intro) $\quad A \vdash T(\ulcorner A\urcorner)$
(T-Elim) $\quad T(\ulcorner A\urcorner) \vdash A$

makes such heavy use of the basic truth-principles—in particular, of (A10)—that it is difficult to see how the proof could be constructed without them. The moral, then, is that the basic truth-principles are highly non-trivial assumptions. Perhaps it is not so surprising that the difference in which we have been interested vanishes, at least to some extent, in their presence.[38]

But however that may be, it is clear that, if we were to work in PA_S using a standard, regular Gödel numbering, and if we did *not* assume the basic truth-principles, our catalog of consistent and inconsistent theories would look somewhat different from what we would get if we worked in PA^+ and made use of the strong diagonal lemma: At least one collection of truth-principles would be differently classified, namely: $\{T\text{-Cons}, T\text{-Comp}, T\text{-Del}, T\text{-Rep}\}$.

4. Philosophical Reflections

The technical situation is thus reasonably clear: There are Gödel numberings with respect to which the relevant formalizations of (Not_F) and (Disq_F) are inconsistent with PA_S, and there are others

[35] Negate both sides of (19) and then reason as in the previous note.

[36] The model given above to demonstrate the consistency of (Not_F) and (Disq_F) with PA_S verifies (A10)—if a conditional is even and so is its antecedent, then so is its consequent—but does not verify (A11). Whether it verifies (A12) will depend upon the exact formulation of PA^+, in particular, upon the formulation of induction.

[37] The theories in question are $\{T\text{-Comp}, T\text{-Rep}, \neg T\text{-Intro}\}$—which is the one for which we need just (35)—$\{T\text{-Cons}, \neg T\text{-Elim}, T\text{-Del}\}$ and $\{T\text{-Comp}, T\text{-Rep}, T\text{-Elim}\}$. I do not know whether these are consistent with PA^+.

[38]

with respect to which the relevant formalizations are consistent. It would probably be best if I simply made that point and were satisfied. But I am a philosopher, and so I am by nature driven to ask another question.

As we saw above, the following two principles

(Not) The negation of a sentence A is true iff A itself is not true.

(Disq) A sentence of the form $\ulcorner t \text{ is true} \urcorner$ is true iff t denotes a sentence that is itself true.

are intuitively inconsistent: That is to say, there is a compelling informal proof of their inconsistency. For all that it is informal, however, the proof is a perfectly legitimate proof, and we may wish to ask about it, as we ask about other informal proofs, in what sorts of systems it can be formalized. Now, we have certainly seen that, in some sense, the informal proof of a contradiction from (Not) and (Disq) can be formalized in PA_S: There are non-standard Gödel numberings one can use to formalize that proof. But I find myself wanting to say that *that* proof is a cheat and that the informal proof cannot really be formalized in PA_S.

Consider a different question. It is sometimes said that Tarski proved that the naïve theory of truth, given by the scheme

(36) 'S' is true iff S,

is inconsistent with arithmetic. This claim is based upon the somewhat more precise claim that Tarski's Theorem shows that the scheme

(37) $T(\ulcorner A \urcorner) \equiv A$,

which we may regard as a formalization of (36), is classically inconsistent with PA_S (and, in fact, with far weaker theories). I think it is reasonable to credit Tarski with both results. But it is essential to its being reasonable to report Tarski's Theorem in these ways that his proof does not depend upon the Gödel numbering used. Tarski's Theorem would be of far less interest if it had to be stated as: For *some* Gödel numberings, the relevant formalization of (37) is inconsistent, although for some it is not. If that were all Tarski had shown, then I would not wish to report him as having shown that (37) is inconsistent with PA_S nor as having shown, in this way,[39] that (36) is inconsistent with arithmetic. Of course, what Tarski actually showed is that, if you add a one-place predicate T to the standard language of formal arithmetic and take as axioms all instances of a formalization of (37), then, no matter what Gödel numbering is in use, the resulting theory is inconsistent, so long as certain very weak arithmetic assumptions are in place.[40]

And so, similarly, if we ask whether these two principles

[39]The caveat reflects that fact that, if Tarski had proven that (37) was inconsistent with PA^+, then I might still wish to regard him as having shown that (36) is inconsistent *with arithmetic*. The issue here, though, is what one counts as arithmetic, not what one counts as a proof of the inconsistency of some informal claim.

[40]And so long as the logic remains classical. See **?**.

(Not$_F$) $\quad T(\ulcorner\neg A\urcorner) \equiv \neg T(\ulcorner A\urcorner)$

(Disq$_F$) $\quad T(t) \equiv T(\ulcorner T(t)\urcorner)$

are inconsistent with PA$_S$, then, or so it seems to me, we should answer "No". Not only are we unable to show that the relevant formalizations of these principles will be inconsistent no matter what Gödel numbering we use, the *most natural* Gödel numberings are ones with respect to which their formalizations are *not* inconsistent.[41] And for that same reason, I find myself wanting to deny that the informal argument that demonstrates the inconsistency of (Not) and (Disq) can be formalized in PA$_S$. Perhaps I would feel differently if the situation were reversed, that is, if the only Gödel numberings with respect to which (Not$_F$) and (Disq$_F$) were jointly consistent were very unnatural. But that is not the actual situation.

5. Closing

I have tried to answer two questions here. The first is in what kinds of theories truly self-referential reasoning can be formalized. The answer is that it can be formalized in theories in which the strong form of the diagonal lemma can be proven. One such theory is PA$^+$; the strong form of the diagonal lemma can be proven in PA$_S$, too, but only if we use a non-regular Gödel numbering. The second question was whether there is any mathematical significance to the sort of phenomenon illustrated by the examples with which we began. I have argued that there is. There are, for example, collections of intuitively plausible principles concerning truth that are inconsistent with PA$^+$ but consistent with PA$_S$, if we use a standard Gödel numbering. So it is at least of *some* mathematical significance whether we work in a language in which true self-reference is possible. It is not yet clear how significant the phenomenon is, but it certainly seems worth asking whether there are similar examples in other areas of logic in which self-reference is important.[42]

References

Current address: Department of Philosophy, Box 1918, Brown University, Providence RI 02912 USA

[41] Whether it is possible to give appropriate mathematical content to the presently vague term "most natural" is a nice question. Of course, we could just require that the Gödel numbering be regular, but what would be our motivation? The most complete study I know of this kind of question is in **?**. I have not yet absorbed Feferman's ideas well enough to be sure how they bear upon our investigation here, but my impression is that they would support the thought that any 'natural' Gödel numbering in PA$_S$ would have to be regular.

[42] Thanks to Albert Visser for extensive, and extremely helpful, comments on an earlier version of this paper. Thanks to John Burgess, Michael Glanzberg, and Vann McGee for discussions of these matters during the long time it took me to sort them out, and to Peter Koellner for helping me refine these ideas during two long sessions in front of his blackboard. Comments by anonymous referees helped clarify the paper at important points.

I presented some of this material as part of a series of seminars I gave in St Andrews in February 2004. Thanks to everyone there, espcially Agustín Rayo and Stewart Shapiro, for their comments. That visit was arranged by Crispin Wright and supported by the British Academy and Arché, the AHRC Research Centre for the Philosophy of Logic, Language, Mathematics and Mind. Thanks to both for their support, which is much appreciated.