The Liar Syndrome.

Albert A. Johnstone

ABSTRACT.

This paper examines the various Liar paradoxes and their near kin, Grelling's paradox and Gödel's Incompleteness Theorem. All are found to spring from circular definition – whether of statements, predicates, or sentences – a manoeuvre that generates the fatal disorders of the Liar syndrome: semantic vacuity, semantic incoherence, and predicative catalepsy. Afflicted statements, such as the Liar statement, fail to be genuine statements, and hence say nothing – a point that invalidates the arguments on which the various paradoxes rest. Formal systems are found to require disambiguators to distinguish the pseudo-statements from their genuine doubles. Gödel's Theorem is shown to be fallacious, and measures are proposed to correct the conceptual mistakes on which it is based.

I. Rationale.

'The Liar' is the name often given to a family of paradoxes generated by Liar statements, that is, statements attributing to themselves a negative evaluation of their own truth status.[1] A paradigm instance of a paradox of the sort is the one generated by the statement, this very statement is false. Regarding the latter it may be argued that while it must be either true or false, it can be neither. If it is true, then since it says it is false, it must be false. Alternatively, if it is false, then since it says something true, it must be true. Apparently, each of the two options available entails a contradiction.

Traditionally the Liar has been linked to the concept of truth.[2] Bertrand Russell drew the conclusion from the Liar that there are stratified orders of truth, a view echoed by Alfred Tarski with his claim that a consistent language requires

[1] Their reputed progenitor, the dialectical dialogue, *The Liar*, attributed to the Megarian philosopher, Eubulides, allegedly focused on a paradox in question form: Is a man who says, "I am lying," saying something that is true or something that is false? See Diogenes Laertius, II, 108. For the English translation see Diogenes Laertius, *Lives of Eminent Philosophers*, translated by R. D. Hicks (Cambridge, England: William Heinemann, 1970), Vol. I, p. 237.

[2] Eubulides reputedly used the Liar to attack the correspondence theory of truth newly formulated by Plato and Aristotle.

the services of a separate metalanguage for truth evaluations of its statements. Subsequent discussion has tended to seek a remedy from paradox in postulated levels of truth, or in context-dependent truth predicates.[3]

As will be shown in what follows, the Liar is but one of the more arresting and troublesome results of implicit definitional circularity, an error at the basic level of sense-making. A Liar statement is self-referential, with the consequence that it designates itself through an expression defined in terms of the statement itself. Consequently, it suffers from what may be termed 'the Liar Syndrome', a constellation of three semantic disorders, each of which alone suffices to invalidate a claim to genuine statementhood. The Liar Syndrome and its attendant semantic nonsense may be engendered by circular definition not only of designating expressions as in the case of self-referential statements, but also of predicates, in particular when precipitous substitutions yield predications having circularly defined criteria. As a result, the syndrome is to be found in the statement generating the Grelling paradox, and, in a less obvious way in the Gödel sentence essential to Gödel's thesis of the incompleteness of formalized arithmetic. Substantiation of these various claims is the goal of this article.

II. Self-Referential Statements.

A good place to begin is with self-reference. A self-referential statement is one that says something about itself, a particular statement. Sometimes the expression 'self-referential' is used to characterize a statement that asserts something about the sentence it is using. Properly speaking, a statement of the latter sort is not self-referential at all since it says something about a particular

[3] See Alfred North Whitehead and Bertrand Russell, *Principia Mathematica*, Second Edition (London: Cambridge University Press, 1960), pp. 60-64. See also Bertrand Russell, *An Enquiry into Meaning and Truth* (London: George Allen and Unwin, 1940), p. 62. Tarski claimed that everyday language is inconsistent, and that in any language with the normal laws of logic, if the concept of truth is universal, then antinomies make the language inconsistent. See, Alfred Tarski, 'The Concept of Truth in Formalized Languages', in *Logic, Semantics, Metamathematics* translated by J. H. Woodger (Oxford: The Clarendon Press, 1956), pp. 153-4, 158, 164-5. Tyler Burge has proposed treating semantic predicates as indexical rather than arrayed in a hierarchy of languages each with its own concept of truth. See his 'Semantical Paradox', in *Recent Essays on Truth and the Liar Paradox*, ed. Robert L. Martin (Oxford: The Clarendon Press, 1984), p. 100. Charles Parsons has endorsed Burge's approach, while claiming that the price of concept-consistence in language is inconsistency. See, 'The Liar Paradox', in his *Mathematics in Philosophy* (Ithica, N.Y.: Cornell University Press, 1983), pp. 264-7. More recently still in his extensive survey of the issue, Keith Simmons has argued that there is no universal truth predicate. See his *Universality and the Liar* (Cambridge, England: Cambridge University Press, 1993), p. 181.

string of words, and not about itself, the statement. To avoid confusion in what follows such statements will always be characterized as sententially self-referential.

Self-referential statements of the simpler sort (to which subsequent discussion will be confined) employ two specific types of grammatical expression: a semantic predicate (i.e., a predicate used to attribute a property such as truth, falsity, provability, meaning, possibility, necessity, probability),[4] and a statement-designator (i.e., a referring expression such as 'that statement', or 'this' or 'what you just said' or 'his contention', or 'her statement to the press'). Interestingly enough, in a notational system devoid of statement-designators (of which imagistic thinking is one example), although semantic predication is not thereby precluded, self-reference is impossible, as is also the Liar. This fact suggests that the latter may be simply an aberrant artifact of the notational device of statement-designators.

To clarify matters, let us examine a quite benign instance of self-reference, the one found in a statement such as the following that predicates truth of itself: "this very statement [the one I am now making] is true". In this statement, the designator, 'this very statement', purportedly refers to a statement. It is not unreasonable to ask what the statement is to which the designator refers. The obvious answer, and the only cogent one available, is that the expression, 'this very statement', refers to the statement, this very statement is true. While accurate enough, the answer is quite unilluminating in that it repeats the very expression about which the question is raised. As an explanation of what is meant, it is circular since what is to be explained is included as part of the proposed explanation.

From a slightly different perspective, the question as to what statement the designator designates may be considered a request for a definition of the designator, 'this very statement'. The answer states what is meant by the designator, and in the process indicates the sentence for which the designator goes proxy. The proffered elucidation is not quite a definition in the usual sense since the two, designator and sentence, differ grammatically, and so are not intersubstitutable in all contexts. Nevertheless, it is a definition of some sort since the two are intersubstitutable (roughly) in contexts in which the sentence functions as a noun, which is the case in particular in the case being considered. We might represent the definition as follows, using, to avoid confusion, a special definition sign '$=_{ds}$' to be read as 'is by designation', and

[4] Some confusion may arise because of the practice in formal systems, contrary to English usage, of attributing truth or falsity to sentences rather than to statements. Readers are advised that this article is written in English.

letting the name 'p' replace the designator, 'this very statement'.

(1) $$p =_{ds} p \text{ is true}$$

Now, clearly such a definition is circular in that the *definiendum* appears in the *definiens*. The expression to be defined is defined in terms of itself, and thus left incompletely defined. Clearly too, repeated attempts to complete the definition by substitution of relevant definitions for undefined terms in the *definiens* will not be successful. The designated statement contains an unbridgeable semantic gap that precludes its being a completely meaningful whole. It suffers, one might say, from chronic *semantic vacuity*. As a result, it cannot be cogently evaluated for truth or falsity. Indeed, it is quite unclear where to begin or how to proceed with any such evaluation. Since the statement cannot be judged to be true or to be false, on the classical view of a proposition or statement as anything that may be true or false, it fails to qualify as a genuine statement.[5] It nevertheless gives the impression of being a statement, and for this reason might be termed 'a pseudo-statement'.

A second radical disorder surfaces when it is asked what the statement says. The obvious answer, and the only one available, is that the statement says that the statement itself is true. While the answer might superficially sound cogent enough, it is in fact semantically incoherent. According to the answer, what is said by the statement is identical with what is said by the statement that the statement is true. The statement is alleged to be identical in meaning with something said about the statement, or, otherwise stated, the two intersubstitutable expressions flanking the equal sign in (1) are alleged to be identical in meaning. Such a claim is absurd – at least on an everyday sense of 'meaning' as distinct from that of some cherished philosophical account. An affirmation cannot be identical in meaning with an attribution of truth to that very affirmation. While the two are equivalent in truth value in most cases, they differ semantically: one is about the other, and predicates something of the other.[6] Thus, the answer to the question posed postulates a semantically absurd situation, one in which two expressions differing in meaning are nevertheless ruled to say exactly the same thing. A self-referential attribution of truth is *semantically incoherent.*

[5] Russell of course proposed a similar ban with his 'vicious circle principle' that stated that "no totality can contain members defined in terms of itself." Bertrand Russell, 'Mathematical Logic as based on the Theory of Types', in *Logic and Knowledge* (London: George Allen & Unwin, 1956), p. 75; See also *Principia Mathematica*, p.37.

[6] Proponents of a redundancy account of truth who find the argument dubious, are requested to substitute some other semantic predicate.

Yet a third disorder afflicts a self-attribution of truth. For the predication of an attribute to make sense, it must be possible in principle also to predicate meaningfully the negation of the attribute. In the present instance, since truth is predicated, it should also be possible, in principle at least, to attribute falsity or untruth (even if falsely). However, the very possibility of an attribution of falsity is excluded. Suppose that in the statement, this statement is true, 'true' was replaced by 'false', yielding the statement, this statement is false. As a result of the substitution, the designator, 'this statement', could no longer refer to its intended referent, this statement is true; it could only refer to the different statement, this statement is false. Thus, it is not possible in principle for the predication of truth to be replaced by one of falsity with regard to the statement in question. Such a situation violates one of the essential conditions for meaningful predication. The result is predicative failure and the creation of a pseudo-statement afflicted with what might be termed 'predicative catalepsy'.

Thus, a self-referential attribution of truth fails on three counts to achieve genuine statementhood. The failure is clearly in no way dependent on the fact that the specific semantic predicate of truth is involved. It follows instead from the more general fact that the statement is decreed to be semantically identical (with due allowance for grammatical differences) with a statement attributing some semantic predicate to that very statement. A circular definition of the sort may be represented as follows, where 'p' represents some statement designator, and 'Ψ' any semantic predicate:

(2) $$p =_{ds} \Psi p$$

The inevitable result of such an operation is threefold semantic nonsense: chronic semantic vacuity, an absurd semantic identity, and a cataleptic semantic predication. The nonsense generated may reasonably be expected at some point to disrupt peaceful communication.

III. Paradox.

The Liar paradox in its simpler forms arises when a statement self-referentially says of itself that it is false. Since the statement is an instance of a self-referential statement defined as in (2), it should fail to qualify as a genuine statement. To make the point, it suffices to ask, reasonably enough, what statement is meant by the statement-designator through which the Liar statement designates itself. The only cogent answer to the question may be fairly represented as a definition as follows, with 'p' functioning as the statement-designator:

(3) $$p =_{ds} p \text{ is false}$$

A circular definition of the sort clearly generates the three above-mentioned semantic disorders: it leaves the Liar statement semantically vacuous in that the 'p' being defined appears in the *definiens*; it postulates an absurd semantic identity, the identity of what the statement says with what a denial of the statement says; it makes the statement predicate of falsity cataleptic in that it excludes the conceivability of alternative predicates. The second disorder takes the particularly virulent and attention-attracting form of an inconsistency that generates a paradox.

Inconsistency-producing definitions are not, of course, the exclusive preserve of statements. Conceivably, nouns could be analogously defined, and with analogous intolerable results.[7] For instance, the proper name 'Gorg' might be defined as the name for any individual that is not Gorg. It might then be reasoned that if any individual is Gorg, then it is not Gorg, and if it is not Gorg, it is Gorg; since any individual must be either Gorg or not Gorg, it has to be both Gorg and not Gorg. A Parmenidean or Kantian metaphysical manoeuvre might then generate either a case for the nonexistence of all individuals in the phenomenal world, or an exhaustive proof of Antirealism. Arguments of the sort are rarely heard, even from philosophers. Self-contradictory noun-definitions are rigorously banned from serious discourse, and indeed rightly so. Parity of reasoning should find self-contradictory definitions of statement-designators to be equally illicit, and their products to merit ostracization from the community of discourse.

Since self-referential statements such as the Liar statement are afflicted with the Liar Syndrome, they do not qualify as genuine statements. Recognition of this fact may reasonably be expected to resolve the Liar paradoxes. However, a word of caution is in order since the import of the fact is all too easily overlooked.

For instance, it is insufficient simply to conclude that since the Liar statement is a pseudo-statement, it is neither true nor false. While this conclusion entails the failure of Bivalence, and thus invalidates the argument in the classical version of the Liar paradox, nevertheless, as Robert Martin has pointed out, the simple Liar argument need not depend on the assumption of Bivalence.[8]

[7] As Robert Koons rightly points out, a paradox need not depend on self-reference. See Robert C. Koons, *Paradoxes of belief and strategic rationality* (Cambridge: Cambridge University Press, 1992), pp. 14 *et seq.*

[8] The case may be argued as follows: p, as defined in (3), must be either false or not false; if it is false, then since it says it is false, what it says is true, and hence it must be true, and consequently, not false; if it is not false, then what it says is false, and consequently, it must be false. Thus the failure of Bivalence does not resolve all paradoxes. See Robert Martin, 'Introduction', in *Recent Essays on Truth and the Liar Paradox*, ed. Robert R. Martin (Oxford: Clarendon Press, 1984), pp. 2-3.

Furthermore, a contradiction may be generated by defining 'p' as follows to designate the statement that p is neither true nor false:

(4) $p =_{ds}$ p is neither true nor false

It may be argued that since self-referential statements are neither true nor false, and since p is self-referential, p is neither true nor false; yet, since it says of itself that it is neither true nor false, what it says is true, and so p must be true, contradicting the claim that it is neither true nor false.

Likewise, it is insufficient simply to point out that the Liar statement is not meaningful since it may be argued that if self-referential statements are meaningless, a statement that predicates meaninglessness of itself says something true, and hence cannot be meaningless. In addition, as has often been pointed out in the literature, from the fact that a meaningless statement is untrue further paradoxes termed "Strengthened Liar Paradoxes"[9] are easily generated from the following definition and a line of reasoning that echoes that of the Simple Liar Paradox:

(5) $p =_{ds}$ p is untrue

To avoid all such complications, the conclusion it is important to draw from the above findings is that self-referential statements do not state anything. The conclusion is obvious enough. If a statement is neither true nor false, it is not a statement, and not being a statement, it cannot state anything. This conclusion is crucial to countering the above arguments. In all these arguments, appeal is invariably made at some point to what the statement allegedly says. For instance, in the classical form of the Liar, it is argued that if the statement is false, then it says something true, and hence must be true. Similar manoeuvres occur in the Strengthened Liar as derived from (5), in the contradiction argument derived from (4), and in Martin's revised Liar argument that makes no appeal to Bivalence. All these arguments assume that a statementally self-referential statement may be deemed to say something, and indeed, found to say something that is true. The assumption is, of course, absurd. The statements are all semantically defective, suffering from three fatal disorders each of which precludes their being genuine statements. As a result, what they say is something nonsensical – something semantically vacuous, semantically incoherent, and predicatively cataleptic. The statements can at best create the illusion of being genuine statements, at best *seem* to say something meaningful and hence something that is true. All the above-mentioned arguments are unsound since

[9] Bas van Frassen introduces the term in 'Presupposition, Implication, and Self-Reference'. *Journal of Philosophy* 65 (1968): p. 147.

they rely on the false assumption that a self-referential statement says something, without which assumption the arguments collapse. Their collapse takes with them the apparent contradictions and paradoxes to which they lead.

It should be noted that the above resolution of the paradoxes cannot be rightly accused of appealing to semantic concepts that themselves generate new paradoxes of the sort Simmons terms "the Revenge Liar."[10] It would be a misunderstanding to claim that paradoxes may be generated by a statement that denies its own meaningfulness, or, as Gupta suggests, that denies of itself that it is 'stably true'[11], or again, as Simmons suggests, that denies of itself a universal truth-predicate.[12] Any such statement that attempted to deny its own truth would be statementally self-referential. It would not say anything, and so could not generate paradoxes by saying something true.[13]

IV. Ambiguity.

A feature of self-referential statements that seriously hinders their being recognized as pseudo-statements is that the sentences used to make such statements may also be used to make different statements, genuine statements that evaluate the self-referential statements. Consider, for instance, the sentence, 'this statement is neither true nor false'. It may be used to make a self-referential statement about itself (hence one that is a pseudo-statement, and as such, neither true nor false), or it may be used to make a non-self-referential statement about another statement, the self-referential one (and hence be a genuine statement that is in fact true). Sentences of the sort lead what may be termed 'a double life',[14] given that they may make either of two quite different statements.[15] In view of a widespread reluctance to acknowledge the ambiguity, it is perhaps wise to devote a moment to the point.

[10] Keith Simmons, *Universality and the Liar* (Cambridge: Cambridge University Press, 1993), p. 7.

[11] Anil Gupta, 'Truth and Paradox', in Robert L. Martin, ed. *Recent Essays on Truth and the Liar Paradox* (Oxford: Oxford University Press, 1984) pp. 175-235, p. 233.

[12] Simmons, *Universality and the Liar*, pp. 100, 159, 181.

[13] It is only fair to note that many authors have disapproved of a blanket condemnation of statemental self-reference. They have done so on the grounds that many sentences involving self-reference are not problematic. Space precludes discussion of the matter here, but it should be noted that even if such a claim could be substantiated, it would be irrelevant to present concerns: it could not convert the self-referential attributions considered above into genuine statements.

[14] The term is borrowed from Albert A. Johnstone, 'Self-reference, the double life and Gödel'. *Logique et Analyse*, 93, March 1981, pp. 35-47.

[15] An Austinian account of such ambiguity is made the focus of investigation in Jon Barwise and John Etchemendy, *The Liar: An Essay on Truth and Circularity* (New York: Oxford University Press, 1987).

There are two persuasive indications that the one sentence may be used to make two distinct statements, a self-referential statement and a non-self-referential one. The first is that the statement made may have different truth-values. When the sentence is used to make a statement saying something about itself, the statement made is neither true nor false; when the sentence is used to make a statement evaluating a statement other than itself (which statement is self-referential), the statement made is true or false as the case may be. The fact that one and the same statement cannot have simultaneously two different truth values should suffice to establish that despite their being expressed by the one sentence, the two are indeed two distinct statements.

A second indication that the two statements are distinct statements may be seen from the fact that they differ semantically. While it may seem plausible to claim that the intended referent of the statement-designator is in each case the same, closer scrutiny finds otherwise. In the case of the genuine statement, the intended referent of the statement-designator is the self-referential statement, and that referent does not include the predicate that is then predicated of the statement. In the case of the self-referential statement, the intended referent of the statement-designator is the statement itself that the designator is helping to make, and thus the referent includes in principle both the referent of the statement-designator and the semantic predicate being predicated of that referent. The referent differs structurally in the two cases. An adequate semantic analysis should distinguish between the two – a fact that implies that the two are distinct statements.

For philosophers hardy enough to risk contamination from introspective activities, the difference in intended reference becomes obvious with an attempt to think meaningfully each of the two statements. Self-reference involves an exercise akin to tail-chasing, an attempt to refer to the object of one's own act of referring coupled to the as yet unexpressed evaluation of that object. No referential gymnastics of the sort are present when one meaningfully expresses the genuine statement.

For speakers of everyday English, the ambiguity causes little confusion. Which of the two statements is meant is easily indicated by various intonational and contextual cues, by the addition of qualifying expressions, or by replacement with alternative unambiguous formulations. In a formal system powerful enough to contain self-referential attributions, analogous measures of some sort must be taken if sentences are to reflect essential statemental structure, and if the confusions attendant upon the ambiguity are to be avoided. The system will require some notational device, a *disambiguator*, to indicate which of the two statements is meant. In cases of self-reference, a left square

bracket might be introduced to indicate the intended scope of the statement designator (e.g., [Tp, for the self-referential reading of the sentence, 'this statement is true'). The square bracket would in fact indicate that the statement to its right was a pseudo-statement (or that the sentence to the right was faulty). In cases of non-self-referential evaluation of a self-referential statement, the usual notation might be used (e.g., Tp, for the non-self-referential reading of the sentence, 'This statement is true'). A disambiguator of the above sort would function like the cross on the door of a plague-stricken house. It would extend the existing ban on circular definitions of predicates and named items to encompass circular definitions of statements. Since the stigmatized statements would be neither true nor false, the formal system in which it functioned would also have to be three-valued.[16]

V. Criterially Circular Predications.

As might be expected, the Liar Syndrome can in principle be generated not only when a noun term is defined in terms of itself, but also when a predicate is given a circular definition. A simple-minded instance of the latter might be a definition that stipulates the color-predicate 'blorange' to mean 'not blorange', thus engendering all three disorders of the Liar Syndrome as well as the paradox that all colored objects have two contradictory properties. The syndrome may also be generated through less flagrantly illicit manoeuvres. A particular individual may be mistakenly assumed to be a candidate for membership in a particular class, whereas in fact for that individual the necessary and sufficient conditions for possession of the defining property of the class are vacuously defined in terms of possession of the defining property.

One well-known example of the latter sort is to be found in the Grelling paradox. To launch the paradox, a predicate-word 'heterological' is coined to characterize words that do not apply to themselves (words such as 'French', which is not a French word, or 'long', which is not a long word). The question is then raised as to whether the word 'heterological' applies to itself. It is argued that 'heterological' must either be heterological or not. Yet, if it is heterological, then by that very fact it applies to itself, from which it follows, by the definition of the term, that it does not apply to itself, which is to say that it is not heterological. On the other hand, if it is not heterological, then it does not apply to itself, from which it follows, by definition, that it is heterological.

The key to the dissolution of the paradox is the fact that it is improper to use the word 'heterological' with regard to itself. For words such as 'thoughtful'

[16] Perhaps a variation on the weak three-valued system in S. C. Kleene, *Introduction to Metamathematics* (New York: Van Nostrand, 1950), p. 334.

or 'short' or 'red', there are independent criteria for determining whether they apply to themselves, but in the case of 'heterological', the criterion for the word applying to itself is that the word doesn't apply to itself. Such a criterion is circular in that the necessary and sufficient conditions for its applying are defined in terms of its applying. A definition of what it means for 'heterological' to be heterological is chronically vacuous, semantically absurd, and predicatively cataleptic. Given the absence of any criteria to settle the matter of truth, both the claim that 'heterological' is heterological and its denial are undecidable. As such, they are neither true nor false, hence not genuine statements.

A more complex example of circular predication, one that is useful in discussing Gödel's Theorem, may be generated by varying the Grelling scenario slightly. Let us suppose that numbers are assigned in some orderly way to the various classes of numbers (even numbers, prime numbers, and so on). Some of these class-numbers will qualify for membership in the class they name, others will not. For instance, if the number 27 should happen to be the class-number associated with the class of numbers that are divisible by 3, then 27 is a member of the class it numbers in virtue of its having the property necessary for membership, that of being divisible by 3. Various classes of class-numbers may also be generated: the class of class-numbers that qualify as members of the class they number, the class of class-numbers that do not so qualify, the class of class-numbers of which it is provable that they qualify, and so on. Each of these classes of class-numbers may also be assigned a number as its class-number. Interestingly enough, the question of whether the class-number of the class of class-numbers itself belongs to the class it numbers will in each case have no answer – and that because, as in the case of the Grelling Paradox, the criterion of membership turns out to be circular. Let us see how.

Consider the case of the class N, the class of all those class-numbers for which it is not provable that they possess the defining characteristic of the class they number. In general, for any class of numbers (such as the even numbers) it is quite possible to determine whether it is not provable that its class-number possesses the defining characteristic of the class it numbers (that of being even) – and hence possible to determine whether its class-number is in the class N. Letting 'E' and 'e' represent the class and its number, '~P' a predicate that attributes unprovability, and '$\equiv_{df}$' mean 'is equivalent by definition to', the membership conditions may be represented as follows:

(6) $$Ne \equiv_{df} \sim P(Ee)$$

Now, suppose the class N is assigned the number n. The question arises as to

whether n qualifies for membership in N. In contrast to the above case, the conditions the class-number n must satisfy to be a member of the class it numbers, N, are circular, a point that readily emerges from the following representation of them:

(7) $\quad Nn \equiv_{df} \sim P(Nn)$

To be a member of the class of class-numbers, N, the class-number n must be not provably a member of the class it numbers, which in its case is N. The class-number has the required characteristic of the class of class-numbers if and only if it is unprovable that it has the required characteristic. Such a criterion is clearly unworkable, and makes the question of membership undecidable. As a result, the statement of membership is afflicted with all three disorders of the Liar Syndrome, and is not a genuine statement.

It would obviously be fallacious to argue that since the statement on the right-hand side of (7) says of itself that it is not provable, it says something true. While it is certainly true that the statement is not provable given that it is neither true nor false, the alleged claim of unprovability made by the statement itself is an integral part of the pseudo-statement. Such a statement says nothing and consequently cannot say that it is itself unprovable. A cogent assessment of the statement's unprovability can only be made by a further statement, one that has not been ruled by definition to be equivalent to the statement of which it states the unprovability. In point of fact, the sentence appearing on the right-hand side of (7) is ambiguous. In the context of (7) it makes a pseudo-statement; outside that context the sentence may be correctly used to make a true statement about the statement on the left-hand side of (7).

A similar situation holds when instead of the semantic predicate of unprovability, any semantic predicate appears. A criterially circular predication of the sort may be represented as follows, where 'Nn' represents some predication, and 'Ψ' any semantic predicate:

(8) $\quad Nn \equiv_{df} \Psi(Nn)$

Such a situation might well have been expected in view of the striking structural resemblance between (8) and (2). Whereas in (2) a noun term for a statement is given a circular definition, in (8) the predicate criteria are defined circularly. The following instance of (2) corresponds to (7), an instance of (8):

(9) $\quad p =_{ds} \sim Pp$

VI. Gödel and Sentential Self-Reference.

Kurt Gödel's well-known theorem, widely termed 'Gödel's Theorem', demonstrates that any formal system of classical two-valued logic augmented with the axioms of arithmetic and a portion of its own metalanguage will contain sentences that are undecidable in the system-sentences for which neither they nor their negations are provable within the system.[17] The metalinguistic evaluations are made possible through a provability predicate defined syntactically as membership in the set of sentences that are immediate consequences of the axiom-sentences. Since the provability predicate applies to sentences rather than statements, to avoid confusion it is better termed 'a derivability predicate'. The undecidable sentence figuring in the theorem, the Gödel sentence, says that a particular sentence, itself, is not derivable. Thus, the undecidable sentence responsible for the incompleteness apparently states something true, its own underivability.[18]

The paradoxical line of reasoning central to the theorem also strongly resembles the one found in the Liar. It runs roughly as follows: if the sentence were derivable, it would have to be true, hence say something true, and hence, as it says, not be derivable – which contradicts the assumption of its derivability; if the negation of the sentence were derivable, since the sentence states its underivability, it would have to be not underivable, hence derivable – with the result that both the sentence and its negation would be derivable. As with the Liar, each of two possible alternatives generates a contradiction, although in the present case the consequence is not paradox but incompleteness.

The Gödel sentence figuring in Gödel's proof[19] states that a particular number satisfies a particular one-place propositional function that defines a class of numbers. In Gödel's formal system a number is assigned as a name to each class of numbers according to its rank in an ordering of the various classes of numbers. Roughly characterized, the Gödel sentence states that a particular class-number (the class-number of the class of class-numbers for which the sentences stating that they possess the defining characteristics of the classes they number are not derivable) has the defining characteristic of the class it numbers (that of the non-derivability of the sentence stating its possession of the defining characteristic of the class it numbers).

Clearly, since the Gödel sentence, on its intended interpretation, states the

[17] See Kurt Gödel, 'On formally undecidable propositions of *Principia mathematica* and related systems I', in Jean van Heijenoort, ed., *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931* (Cambridge, Mass.: Harvard University Press, 1967), pp. 596-616.
[18] Gödel, 'Formally Undecidable Propositions', pp. 599, 615 note 67.
[19] In Gödel's notation, '17 Gen r'.

underivability of a certain string of symbols, rather than the unprovability of what is said, it is not vulnerable to the reasoning presented earlier against statemental self-reference. Sentential self-reference is widely and plausibly esteemed to be a harmless operation. In this spirit, Saul Kripke has contended that by interpreting elementary syntax in number theory, "Gödel put the issue of the legitimacy of self-referential sentences beyond doubt; he showed that they are as incontestably legitimate as arithmetic itself."[20] Kripke is obviously right when 'a legitimate sentence' is taken to mean a formula of the formal system that is a well-formed formula according to the formation rules of the system. However, the important issue is whether such sentences are legitimate in the sense that they make good sense on their intended interpretation, rather than express dubious statements that inadequate formation rules have failed to exclude. Gödel makes no attempt to show that the interpreted Gödel sentence makes sense (nor does Kripke); he seems simply to assume that it makes sense given that it is well-formed according to the rules of the system. The assumption hardly commands automatic endorsement, since, as we saw earlier with statemental self-reference and criterially circular predication, sentences considered to be well-formed may in fact express nonsense. The issuing of a certificate of legitimacy should be contingent upon the results of closer scrutiny of the meaning of the Gödel sentence.

To clarify matters, let 'E' and 'e' represent some normal class of numbers (such as the class of even numbers) and its class-number, and let '~D' represent an underivability predicate. Let 'N' and 'n' represent respectively the class and class-number of all classes such that the sentence stating that the number has the defining property for membership in the class it numbers, is not derivable. The necessary and sufficient conditions for each of the two class-numbers, e and n, to be members of the class of class-numbers, N, may then be stated respectively as follows:

(10) $\quad Ne \equiv_{df} \sim D(\text{'Ee'})$

(11) $\quad Nn \equiv_{df} \sim D(\text{'Nn'})$

The statement of membership conditions in (10) is clearly not circular. The same is not obviously the case for the statement of membership conditions in (11). Indeed, on further inspection, the alleged legitimacy of the Gödel sentence, the left-hand side of (11), becomes quite suspect.

For instance, it might be found tempting to argue as follows in favor of the

[20] Saul Kripke, 'Outline of a Theory of Truth'. *The Journal of Philosophy* 72 (1975): pp. 690-716, p. 692.

claim that Nn, the left-hand side of (11), should make perfectly good sense. What it states is equivalent to what is stated by the right-hand side, the underivability of a particular string of symbols, 'Nn'. Since a string of symbols is either derivable from the axiom-strings or not, a statement asserting it is not derivable must be meaningful, and hence be a genuine statement. Given the equivalence of the right-hand and left-hand statements, the Gödel sentence must also express a genuine statement. However, such a line of reasoning begs the point at issue. The question is whether the sentence 'Nn' makes sense. If it does not, then the left-hand statement of (11) does not, and so neither does the statement equivalent to it, the right-hand side of (11). The latter must then be a pseudo-statement, one that appears to assert the underivability of a particular string of symbols, but one that in fact cannot assert anything. Thus, in assuming that the right-hand side of (11) asserts something, the argument presupposes what it purports to establish.

For the same reason, it would be fallacious to claim (as Gödel does) that the Gödel sentence states something true, its own underivability, and then to argue that since it states something true, the left-hand statement of the equivalence must also be true, and hence a genuine statement. If the sentence makes a pseudo-statement, it states nothing, and so cannot state anything true. Such an argument simply assumes (as Gödel does) that the sentence makes a genuine statement, and so fails to show that it does.

In point of fact, there are two excellent reasons for thinking the sentence cannot make a genuine statement. First, the predication on the left-hand side of (11) is meaningful only if the statement on the right-hand side is meaningful. The latter is meaningful only if the string of symbols 'Nn' is a string of symbols that expresses a meaningful statement. If the string 'Nn' expressed nonsense, then since it is also the Gödel sentence, the latter would not make a meaningful statement. Thus, the meaningfulness of the predication, Nn, is conditional upon the meaningfulness of the statement expressed by 'Nn', which is to say, itself. As a result, the predication is criterially circular. The situation echoes that of the Grelling paradox: the attribution of a particular predicate to a particular individual fails to make sense. In the case of the Gödel sentence the circularity is less apparent because the relevant statement is defined in terms of its sentence rather than in terms of itself. However, the shift from statement to sentence fails to avoid circularity since the question still arises as to whether the particular string of symbols is legitimate in the sense of expressing a genuine statement.

The second reason for thinking the sentence illegitimate is no less decisive. If the formalization of arithmetic-plus-metalanguage is to be considered a faithful rendition of arithmetic-plus-metalanguage in English, its translation

back into English must make good sense. The exception could only be a situation where the formal system employs some peculiar idiom in order to correct an incoherent English one. Such appears not to be the case. It is true that in English one speaks of the provability of statements rather than of the derivability of sentences, but one manages to do so without collapsing into incoherence. Talk of sentences being true, or false, or derivable, has its source in what is convenient for logicians, and not in the incoherence of some English idiom. In these circumstances, the only cogent translation of the Gödel sentence back into English is a statement asserting its own unprovability, as in (7). Such a statement is a pseudo-statement afflicted with the Liar Syndrome, the negative effects of which are neutralizable in English with appropriate precautions.

Thus, the Gödel sentence is properly judged to be illegitimate. It makes a pseudo-statement, and consequently should never have been admitted into a formal system that is two-valued, and hence unequipped to accommodate such sentences. Moreover, since a pseudo-statement says nothing, the argument in Gödel's Incompleteness Theorem fails, appealing as it does at two crucial points to what the statement says.

The theorem cannot be rescued by an appeal to the services of the simplified version of Gödel sentence suggested by Kripke, a sententially self-referential sentence constructed through the use of proper names for sentences.[21] The definition of such a sentence may be represented as follows, with 'n' representing a sentence name:

(12) $\qquad n =_{ds} \sim D'n'$

Clearly, the statement expressed by 'n' has nothing to do either with arithmetic or with the metalanguage of arithmetic, so its presence in a system of formalized arithmetic is quite unwarranted. In addition, a definition as in (12) succumbs to charges analogous to those directed above against (11). First of all, 'n' is a meaningful name of a sentence in a two-valued system only if the right-hand side of (12) is a sentence that expresses a meaningful statement, and the latter is the case only if the 'n' on the right-hand side is the name of a sentence that expresses a meaningful statement. Thus, the meaningfulness of the name 'n' has been made to depend in circular fashion upon the name 'n' being meaningful. The situation is not unlike that of declaring the word 'Gerg' to be a name for the word 'Gerg', whereas prior to a definition it is a mere string of letters, and not a word. Likewise, in (12) 'n' may name a name only if 'n' is already a name and hence designates something.

[21] Kripke, 'Truth', p. 693.

Secondly, a formal system that is a formalization of the arithmetic and metalanguage contained in a natural language should in principle be translatable back into that language if it is to be considered a proper formalization of what it purports to formalize. Since the only cogent translation back into English of the concept of derivability is that of provability, the interpreted Gödel sentence becomes a nonsensical self-evaluation of unprovability as in (9) above.

Thus, the shift from statemental self-reference to sentential self-reference is, from the point of view of present concerns, of less than dubious utility. Statements that are self-referential and predicates that are criterially circular in the sentential mode may be represented as follows, where the predicate 'Φ' represents any sentential semantic predicate:

(13) $\quad p =_{ds} \Phi'p'$

(14) $\quad Nn \equiv_{df} \Phi'Nn'$

(13) is, as it were, the sentential rendition of (2), while (14) is that of (8). When transformed into their sentential correlates, the pseudo-statements that instantiate (2) and (8) become sentential evaluations that instantiate (13) and (14). Certainly, in discussing formal systems it may be useful to speak of sentences rather than of the statements they make, but otherwise the transformation yields no significant gain. If syntax faithfully reflects semantics, as it should, the formation rules of the system must screen for definitions and instantiations that generate sentences expressing statements afflicted with the Liar Syndrome. Contradiction is the price of failure to do so.

Any system that contains both semantic predicates of some sort (of truth, provability, possibility, necessity) and names or designators of statements, sentences, or classes, must, if it is to avoid unnecessary problems, screen for failures of instantiation and substitution *salva significatio*. It must be suitably equipped either with formation rules that eliminate any resulting nonsensical and irrelevant statements, or with a notation that prevents confusion of the pseudo-statements with the genuine statements that evaluate them. The system that figures in Gödel's Theorem fails to do any of this.

VII. Implications.

The puzzles attendant upon self-reference have over the years generated a wide variety of extravagant claims. Although in view of the above findings the error of these claims is obvious enough, a brief spelling out of the obvious is perhaps not amiss.

The widespread tenet that a formal language cannot contain its own

metalanguage without generating paradoxes is quite overstated. It is true only of certain formal languages, those lacking the machinery necessary either to eliminate certain pseudo-statements or to accommodate them in a three-valued system equipped with disambiguators. The Liar provides no grounds to speak, as has Hilary Putnam, of "giving up the idea that we have a single unitary notion of truth applicable to any language whatsoever...,"[22] and hence of giving up any notion of a God's Eye View of the world, and embracing a general Antirealist or non-Objectivist account of human knowledge. Indeed, it would be astounding to find such claims warranted. English has been serving as its own metalanguage for an impressive length of time without requiring the services of hermetic levels of truth, and without collapsing into incoherence.

Gödel's Theorem is often understood to show that any system of formalized arithmetic must be incomplete. In addition, it is not infrequently touted to have other far-reaching implications. John Stewart, for one, has argued that Gödel's Theorem undermines an Objectivist epistemology and supports transduction, the view that subject and object exist only in their relationship to each other.[23] Michael Dummett deems the theorem to show "that no formal system can ever succeed in embodying all the principles of proof that we should intuitively accept"[24]. Likewise, Roger Penrose takes it to show that in mathematical thinking "the role of consciousness is non-algorithmic," and that "human understanding and insight cannot be reduced to any set of computational rules."[25]

As concluded above, Gödel's Theorem is made possible by a failure to either exclude or accommodate sentences that express pseudo-statements on their intended interpretation. Such a situation provides no obvious support for the claim that mathematics has no firm foundation, and hence none for Antifoundationalism or for Antirealism. Nor does it reveal some deep feature of mathematical thinking, a feature that eludes capture in a formal system.

[22] Hilary Putnam, 'Realism with a Human Face', in *Realism with a Human Face*, ed. James Conant (Cambridge, Mass.: Harvard University Press, 1990), p. 17.

[23] John Stewart, 'The Biology of Cognition', presented at the international symposium titled *On the Origin of Cognition*, San Sebastien, Spain, December, 1996.

[24] Michael Dummett, 'The Philosophical Significance of Gödel's Theorem', in *Truth and Other Enigmas* (Cambridge, Mass.: Harvard University Press, 1978), p. 200.

[25] Roger Penrose, *The Emperor's New Mind* (New York: Oxford University Press, 1989), pp. 110, 416; *Shadows of the Mind* (New York: Oxford University Press, 1994), p. 64. Hofstadter echoes these themes with his claim that the theorem shows that "the full power of human mathematical reasoning eludes capture in the cage of rigor." See Douglas R. Hofstadter, *Metamagical Themas: Questing for the Essence of Mind and Pattern* (New York: Basic Books, 1985), p. 8.

Such a feature may well exist, but evidence for it must be sought elsewhere. Finally. it cannot reasonably be claimed to reveal some remarkable capacity of the human mind: self-reference. The latter simply generates nonsense. A capacity to lapse into nonsense, however proficiently exercised, is hardly a very awe-inspiring human trait.

Departement of Philosophy
University of Oregon
bertjohnst@yahoo.com