SELF-REFERENCE, THE DOUBLE LIFE AND GÖDEL

Albert A. JOHNSTONE

It is widely held that semantically self-referential statements are adequately legitimized both by custom and syntax, and that, although on occasion they may emit an aura of oddity, they remain steadfastly and tediously innocuous. These misapperceptions are reinforced by an appreciation of the obvious advantage in structural simplicity accruing to formal systems which assimilate semantic to syntactic self-reference. Among their less felicitous and less noticed consequences should be counted both the *ad hoc* tenor of solutions proposed to the Liar Paradoxes, and the inexhaustible succession of obstacles encountered in the construction of formal systems incorporating their own metalanguages. The present paper attempts to establish this claim. More specifically, it argues (I) that any semantically self-referential statement is undecidable and not a genuine statement, (II) that any such statement has a double identical with it in sentential form, and hence is accompanied by paradox, and (III) that the various truth paradoxes may all be resolved in a trivalent formal system following a procedure equally applicable to Gödel's undecidable sentence.

I

Semantic self-reference is attained in English through two syntactic devices of the language, the noun-clause construction and the statement-designator. The first of these is any construction composed of a statement-predicate together with a sentence appropriately transformed into a noun-clause for the occasion. Instances include constructions used to attribute truth or falsity, to make logical and epistemological evaluations of necessity, of likelihood, of possibility or of demonstrability, and more complex constructions involving verbs of assertion, belief, hope, assumption, suspicion, recollection, or some other verb of attitude. The second device, the statement-designator, is used to refer to a statement or to a putative state of affairs, and goes proxy for a noun-clause in the noun-clause construction.

Instances include the demonstrative pronoun, «that,» as found in concise locutions such as «That's a lie,» «I know that,» and a wealth of more cumbersome referring expressions of the form, «the statement which...,» «what he said,» and so on. Self-reference arises when the designator is used to refer to the very statement which the combination of statement-predicate and – designator is being used to make. The wide choice of statement-predicates and – designators makes self-reference available for almost any taste.

Interest has tended to gravitate almost exclusively towards self-referential statements of one particular type: those giving rise to paradox. This circumstance has had an adverse effect on the proper understanding not only of self-reference but of the paradoxes themselves. The point is perhaps best made by considering a simple version of the paradox termed that of the Simple Liar.

The paradox arises with a statement of the form, q is false, where «q» is some manner of statement-designator used to refer to the statement itself. The paradoxical feature of such a statement is that it can neither be said to be true nor be said to be false, for in either case a contradiction results: if true, then the statement must be false, and yet if false, then since it states a truth, it must be true. The dilemma as to whether the statement is true or false is generally resolved by ruling that it is neither. This is done on the grounds that the other possible rulings lead directly to a contradiction. The ruling has, however, a distinctly *ad hoc* flavor to it, that of a measure dictated by the purely practical consideration of excluding contradiction. This impression is heightened when paradoxical statements are compared with self-referential statements of the type, q is true. The two types are very similar in structure, use and content. Yet the latter are generally considered to be genuine statements, whereas the former are, by the above ruling, declared to be neither true nor false, and *ipso facto* ruled not to be authentic statements at all. The impression produced by the disparity in status is that the measure taken is one of bald expediency, imposed to be sure by the gravity of the situation, but for this very reason an intrusion of practical politics into what might be expected to be an eminently reasonable pursuit.

Closer attention to the general phenomenon of semantic self-reference brings a quite different viewpoint to bear on the matter. Admittedly, a self-referential statement of the form, q is true, looks benign enough on a cursory glance. It may quite consistently be declared either to be true or to be false, for no paradox results in either case. In addition, as Kripke maintains, it is arbitrary to assign truth rather than falsity to the statement, or falsity rather than truth (1). Either alternative may be chosen with as much or as little reason, so that the choice between them is left to the dictates of personal whim. Yet this seemingly felicitous situation is deceptive, for surely there is something defective in a statement which may with as much reason be declared to be false as to be true. What can such a statement possibly state if both truth and falsity may indifferently be predicated of it? It must be that the purported statement states nothing; where truth and falsity change nothing, nothing is being enunciated, not even a tautology. It follows, of course, that the only appropriate ruling to be adopted with respect to such utterances is that they are neither true nor false.

Analogous conclusions might be drawn regarding other types of semantically self-referential statements – q is possible, or b believes q – although in the case of the latter considerable discussion would be required. Any decision to declare such statements true rather than false, or false rather than true, turns out on examination to be equally arbitrary. In this respect, the statements resemble vacuous and nonsensical statements of the kinds made with sentences such as, «The present King of France is bald,» or «T'was brillig.» Such utterances are undecidable, and hence most aptly characterized as neither true nor false. Analogous too are self-predications of falsity which behave in similar fashion in regard to truth conditions; that is, they too have as little reason to be declared true as to be declared false. In this perspective, of course, a ruling to the effect that they are neither ceases to be *ad hoc*.

The undecidability of semantically self-referential statements is a result, and hence a symptom of a fundamental semantic deficiency, that of chronic vacuity. It is quite impossible to give a satisfying account of the meaning of such statements. In everyday practice when a statement contains a statement-designator, a request for clarification can be and generally is satisfied by replacing the designator with the

(1) Saul KRIPKE, «Outline of a Theory of Truth,» *The Journal of philosophy*, 72 (1975), 709.

noun-clause for which it goes proxy and then repeating the statement. For example, any substitution into a self-referential statement of the form, George believes q, yields the equally obscure utterance, George believes that George believes q. More generally, where «q» has been stipulated to designate Fq, no finite number of substitutions, however great, can suffice to define «q»: the n^{th} substitution will yield a sequence of symbols consisting of $n + 1$ «F»'s and one «q» still in need of definition. Elucidation is unattainable. To be sure, an illusion of meaningfulness is created by the presence of the meaningful statement-predicate, F. However, if the practice of defining «q» as «Fq» produced meaningful statements, it would be reasonable to expect analogous results in the limiting case, that where the sentential sign was declared to denote the statement it is used to make. In such a case «Q» would be defined to denote Q. Here the illusion of meaningfulness disappears in a paroxysm of self-reference. English syntax, wedded as it is to the noun-clause construction, precludes the attainment of this higher form of vacuity.

II

The decision to declare Liar statements neither true nor false gives rise directly to a further paradox, termed by van Fraassen the Strengthened Liar Paradox (2). Here it suffices to have the designator, «q», refer to a statement of the form, q is untrue, to make possible an argument analogous to that of the Simple Liar: if q is true, then since it says it is untrue, it must be untrue; if q is untrue, then since it says it is untrue, it states a truth and hence must be true. As Skyrms puts the matter, the embarrassing fact about the Strengthened Liar statement is that the decision to declare it untrue (a point on which everyone agrees), leads directly to a contradiction (3).

An interesting point about the Strengthened Liar argument as stated above is that it may easily be adapted to apply to any self-referential evaluations with the exception of those using egocentric expressions.

(2) Bas C. VAN FRAASSEN, «Presupposition, Implication and Self-Reference,» *The Journal of Philosophy*, 65 (1968), 147.

(3) Brian SKYRMS, «Notes on Quantification and Self-Reference,» *The Paradox of the Liar*, ed. Robert L. Martin (New Haven: Yale University Press, 1970), p. 68.

A semantically self-referential evaluation is of the form, Fq. Now, on our above ruling Fq is neither true nor false. Fq however attributes some property to q. Where the attribution is such that it may either accord or conflict with the ruling that Fq is neither true nor false, then a contradiction arises. If the attribution is in agreement, then Fq states a truth, and hence must be true. If the attribution is in disagreement, then Fq states a falsehood, and hence must be false. In both cases it is either true or false, and thus contravenes the ruling that it is neither.

To illustrate matters, let us take the self-evaluation, q is true. Since q is self-referential, it is neither true nor false; in particular it is not true. Consequently any statement to the effect that q is true is mistaken, and hence false. Yet q itself is precisely such a statement. Hence q must be false, which contradicts the claim that it is neither true nor false. Analogous arguments might be constructed for self-evaluations of the form q is necessary, q is provable, q is possible, and so on. (The extension of the procedure to all semantically self-referential statements is obvious, but raises issues unhelpful in the present context.)

It is perhaps tempting to construe these various arguments as *reductio ad absurdum* of the conclusion that self-referential statements are neither true nor false. However to do so would require showing that the considerations leading to that conclusion are in error. The fallacy resides rather in the above argument form. Indeed the latter houses a rather curious phenomenon; to wit, the assigning of each of two distinct truth evaluations to the statement, Fq. On the one hand Fq is said to be neither true nor false on the ground that it is self-referential. On the other, it is said to be true (or to be false) on the ground that what it states agrees (or disagrees) with the ruling that Fq, which is to say q, is neither true nor false. The second evaluation is clearly unwarranted. Fq is ruled to be neither true nor false on the ground that it fails to make a genuine statement. It cannot then subsequently be accorded a second truth evaluation on the basis of what it purportedly states. Since it is not a genuine statement, if fails to state anything.

This observation, however, leaves a central point unexplained: how is it possible in the argument to assign Fq a second evaluation in so convincing a manner? The answer would seem to be that there is a second statement, Fq, a double of the first, and it is this second statement which has the second evaluation. Consider the following: the self-referential statement, q, is given a truth status (neither true nor false); now a statement which predicates some truth status of q (or entails one) must itself be either true or false, true if it rightly states q's truth status, false if it does not; one such statement will be a statement of the form, Fq, a non-self-referential statement predicating F of q. Thus there are in fact two distinct statements, Fq, the first predicating F of itself, the second predicating F of the first. The proof that they are two distinct statements is the fact that they have distinct truth statuses: the first is neither true nor false, the second is one of true and false.

The notion of the sentential ambiguity of Liar statements has received a less than enthusiastic reception since it was first put forth by Eric Toms([4]) some three decades ago. This is all the more surprising in that the phenomenon of a double life is clearly to be found at work in alternative versions of the Strengthened Liar argument. The key manoeuvre in these arguments is a play on the ambiguity of «Fq,» a treating of the self-referential statement as if it were the non-self-referential one, or *vice versa*. Consider, for instance, a second version of the argument: since q is untrue, the statement that q is untrue is true; but since the statement that q is untrue is the statement q, q is true, which contradicts the assumption that it is untrue. The non-self-referential statement, q is untrue, which acts as premiss of the argument, is subsequently taken to be the self-referential statement, q is untrue. In an adequate notation, one where some notational distinction marked the difference between the two statements, such a move would be flagrantly invalid.

A third version of the argument merits particular attention. Here it is argued that the statement, q is untrue, is untrue, and that consequently, by the Double Negation Equivalence, q is true. The invalid manœuvre is the substitution: here a truth predicate, used non-self-referentially, is substituted for two untruth predicates, the second of which is used self-referentially. The self-referential predication is thus being treated as if it were a non-self-referential one. Such a procedure is incorrect given the difference in truth status of the two predications. It is made to appear correct by the ambiguous notation.

([4]) Eric Toms, «The Liar Paradox,» *Philosophical Review* 65 (1956), 546.

III

It would seem then that insofar as the natural language, English, is concerned, the various Liar paradoxes are susceptible of solution. It suffices to note the vacuity of semantically self-referential statements, and the double life led by sentences used to make self-evaluations. The question now to be considered briefly is whether these conclusions have any application to formal systems. Let us turn first to the task of constructing a formal system devoid of paradoxes but containing both Liar statements and statements giving truth estimates of such statements. A possible system may be briefly outlined as follows.

Let the language, L3, be a propositional calculus having the following peculiarities:

(a) L3 contains four primitive operators, two unary operators, «~» and «T» (the latter to be interpreted as a truth predicate), and two binary operators, one for conjunction and one for disjunction.

(b) L3 contains a finite number of proxy formulas, «Q_1», «Q_2», ..., «Q_n», each of which is a wff, and each of which is defined in terms of some well-formed formula or wff for which it goes proxy. These stipulations allow the formation of wffs which are self-referential in the interpreted system. (It would suffice, for instance, to define «Q» as the wff, «Q».

(c) L3 contains a device termed a self-reference filter, symbolized as «S» with an appropriate subscript. The defintion of any proxy sentence, «Q», takes the following form:

(1) $Q =_{df} S_Q P$

where «$S_Q P$» is to be read as «P as stated by Q.» The definition of «Q» resulting from the successive replacement of any proxy sentences in «P» (except for the proxy sentence, «Q») by their definitions will be termed the ultimate definition of «Q».

The following valuation rules map the non-atomic wffs of L3 onto the elements of the set {t, f, u} (to be interpreted respectively in terms of truth, falsity and undecidability).

(i) ~ P is assigned t iff P is assigned f, f iff P is assigned t, u iff P is assigned u.

(ii) TP is assigned t iff P is assigned t, f iff P is assigned f or u.

(iii) S_QP is assigned u iff the expression giving the ultimate definition of «S_QP» contains «Q»; it is assigned the value assigned P in the remaining cases.
(iv) P & R is assigned t iff both P and R are assigned t, u iff at least one of P and R is assigned u, f in the remaining cases.
(v) P v R is assigned t iff at least one of P and R is assigned t, u iff both P and R are assigned u, f in the remaining cases.

By the above valuation rules the statement, Q, or S_QP, is equivalent to P if and only if the ultimate definition of «Q» does not contain the expression «Q». Thus if Q is self-referential, «Q» and «P» are not interchangeable, and the self-reference filter is not removable. It might further be noted that the two distinct statements which an English Liar sentence may be used to make will be rendered in this notation by distinct sentences. For instance a statement, Q, which says of itself that it is untrue is symbolized as follows:

(2) $S_Q{\sim}TQ$.

A statement predicating untruth of the self-referential Q is symbolized indifferently as,

(3) $\sim TQ$

or again as,

(4) $\sim TS_Q{\sim}TQ$

The presence of the self-reference filter in (2) blocks all three versions of the Strengthened Liar argument mentioned earlier.

Let us now see whether the method used in L3 for undecidable statements is applicable in a system containing Gödel's undecidable sentence. Let us take the system with which Gödel works, the quantificational calculus of *Principia Mathematica* enlarged with a successor function, a constant to be interpreted as the number zero, and an appropriate set of arithmetic axioms. Let us suppose that by an assignment of prime numbers to the elementary symbols of the system, any formula may be assigned a unique Gödel number, and consequently that the syntax of the system is expressible in the system itself(5). In particular, let the system contain a provability predicate (for short «Bew x») defined in terms of the existence of a sequence of wffs each satisfying certain conditions, and a one-place predicate formulated metalinguistically as «Sub (F_n x, n)» to be interpreted as «the sentence resulting from the substitution of the numeral for the number, n, into the one-place predicate with Gödel number n.» Let the latter predicate have q as its Gödel number. The language then contains a sentence written metalinguistically as «~ Bew Sub (F_q x, q).» Since this sentence is the sentence obtained by substituting the numeral for q into the predicate with Gödel number q, it is itself the sentence, «Sub (F_q x, q),» which it states to be unprovable. It is such a sentence that Gödel shows to be undecidable by showing that both the assumption that it is provable and the assumption that its denial is provable lead to contradiction.

A few remarks are in order here. Firstly, it would be a mistake to claim that Gödel's definitions make provability a property of sentences rather than of statements. It should be noted, on one hand, that in the initial informal presentation of his argument Gödel speaks indifferently of unprovable propositions and of unprovable sentences. On the other hand, where appropriate to do so, Gödel's notation may easily be given an interpretation in terms of statements rather than sentences. In particular the expression, «Sub (F_b x, a),» might without difficulty be interpreted as referring to the statement made by the sentence resulting from a certain substitution, rather than to the sentence itself. These circumstances strongly suggest that if no notational distinction is made between statements and sentences, it is merely for the sake of the convenience procured by the simpler notation. The absence would be easy to rectify should the need be felt (as it would, for instance, if any restrictive measures were taken concerning semantic but not syntactic self-reference). It follows, moreover, that it is misleading to claim (as Kripke does([6]), for example) that Gödel put the legitimacy of self-referential sentences beyond doubt, as if some significant distinction should be made here between self-referential sentences and self-referential statements. Gödel constructed a consistent system in which the formation rules

([5]) See Kurt GÖDEL, «On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems I,» *Frege and Gödel*, ed. Jean van Heijenoort, (Cambridge: Harvard University Press, 1970), pp. 94 *et. seq.*

([6]) KRIPKE, «Truth,» p. 692.

declare well-formed those sentences which, on the intended interpretation of the system, make statements about their own provability. The consistency of the system obviously in no way shows, nor was it meant to show, that self-attributions of provability are meaningful statements. Such statements are no more meaningful in Gödel's notation than in English. In both cases they are simply ruled to be syntactically correct.

Now the odd feature of the Gödel sentence is that, while a given expression presumably stands for the result of a certain substitution, the expression obtained by actually carrying out the substitution indicated is a sentence which itself contains the original expression. This circumstance is what allows it to be said that the statement states something about itself. Such a statement falls under the criticisms sketched in Section I. In addition, the sentence used to make the statement may be expected to lead a double life. If no measures are taken to distinguish the two statements, a self-referential predication of unprovability, which is neither true nor false, will be indistinguishable from its non-self-referential double, which is a true statement. A confusion of this nature is to be found in each half of Gödel's proof of the undecidability of «$\sim$ Bew Sub (F_q x, q) (7).» It is argued, for instance, that if the sentence were provable, it would be true and hence, by its own saying, unprovable. However this conclusion is merely the reading of a self-referential statement as a non-self-referential one, a transition made legitimate by the notation. As to the second half of the proof, the argument is somewhat involved in its formal presentation, but clearly in its informal presentation the Double Negation Equivalence is implicitly used, and so used as to treat a self-referential denial of provability as if it were non-self-referential (in the manner of the third Strengthened Liar argument above).

Let us then introduce the self-reference filter used earlier in L3. Since self-reference arises through certain substitutions, the filter must appear in the Rule of Substitution itself. The latter may be altered in such a way as to have a substitution of «a» into the one-place semantic predicate, «F_bx», yield «SF_ba», rather than «F_ba», in cases where self-reference results. A first approximation is the following rule:

(7) GÖDEL, «Undecidable Propositions,» pp. 99, 100, 89.

(5) Sub (F_b x, a) yields F_b a if and only if «F_b a» does not contain «Sub (F_b x, a)», and yields SF_b a otherwise.

This rule is insufficient, however, to cover more complex cases of self-reference involving two or more statements, where, for instance, each of two statements gives a truth evaluation of the other. Hence (5) should be expanded as follows:

(6) Sub (F_1 x, a) yields F_1 a if and only if there is no sequence of formulas, «Sub (F_1 x, a),» «Sub (F_2 x, a),» ..., «Sub (F_n x, a)» with $n \geqslant 1$ such that the formula resulting from the substitution of «a» in «F_n x» contains the formula «Sub (F_1 x, a)», and for each j other than n, the formula resulting from the substitution of «a» in «F_j x» contains the formula «Sub (F_{j+1}, a)»; Sub (F_1 x, a) yields SF_1 a otherwise.

Clearly rule (6) is itself statable in the vocabulary of the system. It has the desired effect of separating the self-referential from the non-self-referential versions of the undecidable sentence, for the first is written as «S ~ Bew Sub (F_q x, q)» and the second as «~ Bew-Sub (F_q x, q).» Both prongs of Gödel's proof of undecidability are thereby blocked. On the other hand it is fairly obvious that since S is defined only through its brief appearance in (6), any wff, «SF_b a,» including Gödel's undecidable sentence, will be unprovable, and that moreover, its unprovability will be provable within the system. A formal proof of the point will not be attempted here.

A simpler system might be obtained by a somewhat different approach, that of ruling semantically self-referential statements to be ill-formed. This would involve converting the conditions stated above in (6) into restrictions on the formation of wffs. A simpler system still, of course, is the one used by Gödel. Since it both decrees semantically self-referential statements to be well-formed, and moreover allows the resulting sentences to lead double lives, it must contain sentences which are undecidable within the system itself. To state their undecidability recourse must be had to a separate metalanguage. In this perspective the hierarchy of metalanguages, the self-reference filter, and the restrictions on well-formedness are but three means of achieving the one goal: that of avoiding the confusion of genuine statements with their nonsensical doubles.

Albert A. JOHNSTONE

ADDENDUM TO SELF-REFERENCE, THE DOUBLE LIFE AND GÖDEL

JOHNSTONE

It would be of no avail to attempt to exonerate the Gödel statement from the charge of nonsensicality by contending that in Gödel's system, S, provability is a syntactic rather than a semantic predicate. The proof of this is straightforward enough. For clarity's sake let us coin a predicate and speak of sentences as derivable rather than provable. Clearly, then, to say that a sentence is derivable is analytically equivalent to saying that the statement it makes is provable. Such an equivalence (presupposed throughout Gödel's introductory remarks) may be expressed symbolically as follows:

(7) $Dr\ "P" = Pr\ P$

Now let us assume that the predicate, «Bew x,» ranges over sentences rather than statements. The system, S, then contains a derivability-predicate but no provability-predicate, and consequently is totally inarticulate on questions of provability. Indeed, insofar as the system itself has anything to say on the matter, derivability could be a sentential property analogous to the property of length in having no direct bearing on the property of provability. When the missing predicate is added to the proof-theory of S, then the Gödel statement, Q, which is defined as $\sim$Dr «Q,» is according to (7) analytically equivalent to the statement, $\sim$Pr Q. Yet the equivalence of Q with the denial of the provability of Q is a semantic absurdity. Thus it is irrelevant whether «Bew x» is interpreted as a syntactic or a semantic predicate, for the Gödel statement generates nonsense in either case.

Syntactic correctness does not guarantee meaningfulness. Where he formation rules of a system allow the formulation of nonsense, the only matter to be settled is whether the benefits of quarantining the nonsense are sufficient to warrant the resulting loss in simplicity for the system. The complication introduced by a self-reference filter may be kept to a minimum by having the filter appear exclusively as a prefix to the semantically objectionable statements–like a cross on the door of a pestiferous house. Two modifications of the system then suffice to produce the desired results. Firstly, the rule of substitution is modified to yield prefixed statements in appropriate cases, the exact nature of the modification varying with the modalities of the genesis of self-reference in the system. In a system containing proxy formulas, for instance, it may be ruled that any proxy formula, «Q_i,» which contains «Q_i» in its test-definition, is replaceable not by its definition simpliciter but only by its definition prefixed with a self-reference filter. (The test-definition of «Q_i» is defined to be the expression which results from the successive replacement in the definition of «Q_i» of any proxy sentences with the exceptions of «Q_i» itself and of any «Q_k» appearing in any successive replacement of «Q_k.») Secondly, a single additional axiom-schema is introduced. In a trivalent system in which the semantics for negation are those of L3 above, the schema is the following:

$$\vdash \quad \sim TSP \,\&\, \sim T \sim SP$$

In a bivalent system this axiom reduces to:

$$\vdash \quad \sim SP$$