THE BULLETIN OF SYMBOLIC LOGIC
Volume 9, Number 1, March 2003

GÖDEL, TARSKI, CHURCH, AND THE LIAR

GYÖRGY SERÉNY

The fact that Gödel's famous incompleteness theorem and the archetype of all logical paradoxes, that of the Liar, are related closely is, of course, not only well known, but is a part of the common knowledge of the community of logicians. Indeed, almost every more or less formal treatment of the theorem makes a reference to this connection. Gödel himself remarked in the paper announcing his celebrated result (cf. [7]):

> The analogy between this result and Richard's antinomy leaps to the eye; there is also a close relationship with the 'liar' antinomy, since ... we are ... confronted with a proposition which asserts its own unprovability.

In the light of the fact that the existence of this connection is commonplace it is all the more surprising that very little can be learnt about its exact nature except perhaps that it is some kind of similarity or analogy. There is, however, a lot more to it than that. Indeed, as we shall try to show below, the general ideas underlying the three central theorems concerning internal limitations of formal deductive systems can be taken as different ways to resolve the Liar paradox. More precisely, it will turn out that an abstract formal variant of the Liar paradox, which can almost straightforwardly inferred from its original ordinary language version, is a possible common generalization of (both the syntactic and semantic versions of) Gödel's incompleteness theorem, the theorem of Tarski on the undefinability of truth, and that of Church concerning the undecidability of provability. We feel that, in addition to its outcome, the process of deriving abstract forms of such important mathematical results from a common language expression is in itself interesting and enlightening. On the other hand, our investigations will shed some light on the limits one cannot get beyond solely on the basis of the Liar paradox.

To obtain the formal resolution of the Liar paradox that can be considered as the common generalization of the theorems concerned, we shall reformulate it in a step–by–step manner in four main stages. First we shall seek an ordinary language equivalent of the paradox in a form that shows clearly its logical structure, and then we shall directly translate the expression we have

Received September 25, 2001; revised September 9, 2002.

obtained into a formal language. The detailed analysis of the formal version of the paradox will lead to its natural generalization in the third stage. By applying this result to the logical systems used to describe mathematical theories, we shall make explicit the connection between the age old paradox and its modern reincarnations. The examination of possible ways to generalize our results will finally reveal the limits of our project set by the Liar paradox. Let us start with the origins.

§1. The paradoxes.

1.1. The Liar paradox. As is well known, the Liar paradox consists in the fact that the truth of the sentence 'This sentence is false.' (the *Liar* from now on) implies its falsity and *vice versa*. One of the innumerable attempts to get rid of the paradox is the claim according to which it only pretends to be a declarative sentence of the subject–predicative form since it has not really got a subject. Certainly, in order for a noun phrase to play the role of the subject of a meaningful sentence, it has to be a complete linguistic entity. In the case of the Liar, however, the noun phrase under consideration, the sentence itself, is just *in statu nascendi* at the point where the need for the subject appears. Thus, according to this criticism, without a proper subject, the sentence is not paradoxical but simply meaningless; it cannot express a statement. Using Ryle's words (see [15]):

> If unpacked our pretended assertion would run 'The current statement {namely that the current statement [namely that the current statement (namely that the current statement ... '. The brackets are never closed; no verb is ever reached; no statement of which we can even ask whether it is true or false is ever adduced.

The significance of this analysis from our point of view is obvious. It demonstrates that, due to the central role the meaning of the demonstrative adjective 'this' plays in it, the paradox, in the above form, is not an ideal object for a straightforward formalization. Indeed, in a formal system, 'the function and the meaning of an expression should depend exclusively on its form' (Tarski [21]). Since our aim is to find a natural language equivalent of the Liar which can directly be formalized, we should make explicit the self-reference the paradox is based on.[1] The help comes from the analysis of another famous paradox, which is just built on the linguistic phenomenon of being about itself.

1.2. Quine's paradox. The paradox providing notions in terms of which we can formulate a more formal version of the Liar was devised by Grelling (cf. [9]). An adjective is called *autological* if the property denoted by the

[1] Let us stress that we do *not* examine the question to what extent this manoeuvre may contribute to the solution of any proper philosophical problem related to the Liar paradox. The only thing that we claim is that this step takes us closer to the possible *formalization* of the paradox.

adjective holds for the adjective itself, otherwise it is called *heterological*. For example, 'short', 'English', 'polysyllabic' are autological, while 'long', 'German', 'monosyllabic' are heterological. Well then, what about the adjective 'heterological' itself? Clearly, it is true of itself just in case it is false of itself. This fact can be utilized to formulate a sentence with the same paradoxical nature as the Liar:

(1) 'heterological' is heterological.

This paradoxical sentence is immune to the kind of criticism that the Liar was open to since it does have a proper subject. Of course, it relies heavily on the meaning of the technical term in it, so, in order to obtain a sentence that stands on its own feet, the term 'heterological' must be eliminated. Since, by definition, 'heterological' means 'yields a falsehood when applied to itself', we have

'yields a falsehood when applied to itself'
yields a falsehood when applied to itself.

On the other hand, this sentence suggests that—at least in this particular case—

(2) 'applied to itself' is equivalent to 'appended to its own quotation',

which leads to the masterpiece Grelling-type version of the Liar due to Quine ([14], p. 7):

'yields a falsehood when appended to its own quotation'
yields a falsehood when appended to its own quotation.

Having given due credit to the ingenuity of *Quine's paradox*,[2] we should notice that the fact it conveys the intended meaning is based on a contingent characteristic of the English language. Indeed, since the usual word–order in declarative sentences is 'subject + verb + . . . ' in English, if there is no indication to do otherwise, a noun–phrase at the beginning of a sentence, however complicated it may be, is automatically taken as the subject of the sentence by a speaker of English. Thus in English, the application of an adjective (or adjectival phrase) to an object can be a simple attachment of the adjective (or adjectival phrase) to the name of the object in some cases. Since Quine's paradox exploits this specific feature of the English language,

[2]Some minor doubts, however, could be raised about the paradoxical nature of this sentence. Perhaps someone can insist that this sentence must, in fact, be false. He or she may argue as follows. Unlike the Liar, which asserts only its own falsehood *and nothing else*, Quine's paradox can be interpreted as a conjunction asserting both its own falsehood *and* that it is the subject of the sentence (that is, the given string of words: 'yields a falsehood when appended to its own quotation') that yields its falsehood. Since this conjunction cannot be true, it is necessarily false, so its first conjunct is true. Consequently, the second one must be false.

it cannot exhibit its logical structure, which is what we actually need. To attain our goal, therefore, we ought to look for an abstract 'language–free' procedure to execute the application of an adjective to an object.

1.3. Findlay's paradox. Nearly every language as well as the majority of the usual logical systems have the common formal tool of *substitution* for handling the application of an adjective (or adjectival phrase) to an object. Indeed, let us associate with any adjective (or adjectival phrase) a sentence asserting that the object referred to by its only indefinite constituent, denoted usually by a single letter and called (logical) variable, possesses the property expressed by the adjective(al phrase).[3] Given such a sentence, we can execute the application of the adjective (or adjectival phrase) to an object by substituting the name of the object for the variable in the sentence. The substitution is a formal, entirely language–independent operation in the sense that the proposition expressed by its result does not depend on the specific structure of the sentence to which it is applied.[4] So we can describe self-reference in language–independent abstract terms by changing the transformation in (2) to substitution in a suitable way (and using the fact that putting a linguistic object between quotation marks is only a particular way to give a name to it[5]):

(3) 'applied to itself' is equivalent to
'its name is substituted for the variable in it'.

Accepting this interpretation, however, has its consequences. It implies directly that a linguistic entity, in order for it to be meaningfully applied to itself, has to be a sentence having a single variable in it. Therefore, we should modify our original notion of heterologicality accordingly.[6] So we define *a sentence with a single variable* (not an adjective or adjectival phrase) to be heterological if its application to itself yields a false sentence, that is, using (3), we shall say that:

> a sentence with a single variable is *heterological* if the new sentence obtained by substituting the name of the sentence for the variable in it is false.

Now we should find the sentence corresponding to (1), which expresses the result of applying this new notion to itself. Since the sentence with a single

[3]The fact that we also use the word 'sentence' to denote linguistic entities that contain variables cannot cause any confusion. If we would like to stress that the sentence considered contains a variable, then we shall refer to this fact explicitly.

[4]There are also other possibilities with the same advantages, cf. e.g., [18], but neither of them seems to be as natural and widely used as substitution.

[5]This is the usual practice, which we also tacitly assumed already.

[6]By the way, without this modification, Grelling's paradox, which is based on the paradoxical nature of an *adjective*, cannot be expected to provide a faithful reformulation of the Liar paradox, which is a statement about a *sentence*.

variable associated with the adjective 'heterological' is the sentence 'x is heterological', by (3), the sentence corresponding to the self–application of this adjective is:

(4) 'x is heterological' is heterological.

By our new definition of heterologicality above, 'x is heterological' takes the following form: 'the new sentence obtained by substituting the name of the sentence x for the variable in it is false', which, in turn, by (4), provides *Findlay's paradox*:[7]

> the new sentence obtained by substituting the name of the sentence 'the new sentence obtained by substituting the name of the sentence x for the variable in it is false' for the variable in it is false.

It can directly be checked that this sentence indeed says of itself that it is false (and says nothing else) since it is built up in such a way that if we perform the substitution described in it, then we obtain the sentence itself, which is stated to be false. Moreover, it is a perfectly well–formed sentence. Finally, its paradoxical character does not depend on any contingent linguistic fact, it is entirely language–independent so it can immediately be translated to almost any language, be it natural or formal. Thus, at last, having an impeccable common language version of our original paradox, we can turn to embedding it to an abstract formal logico–linguistic system in order to examine the possible ways to resolve it.

§2. The formal resolution of the paradoxes.

2.1. The abstract system of Smullyan. Findlay's paradox implicitly contains the minimal requirements an abstract system within which the paradox can be formulated should meet. Clearly, it has to contain the formal versions of elements of natural languages used for setting down the paradox. These are the formal versions of any string of letters (they will be called expressions), those of the (meaningful) sentences that can be true or false, and the sentences having single variables (in formal terms the elements of these two sets together are called formulas). Further, such a system should have a set of objects that will play the role of the names of the linguistic objects, and two mappings for representing the procedure of naming and that of the substitution. Finally, the notion of truth can be represented formally by an arbitrary set of formal sentences. The elements of this set will be considered to be true.

[7] Findlay used sentences of the same structure to examine informally Gödel's incompleteness theorem (cf. [6]). Quine also formulated a sentence of this kind (cf. [13], p. 307). Nevertheless, both authors used the notion of non-theoremhood instead of falsehood since, doing just the opposite of what we do, they looked for the common language formulation of the formal Gödel sentence. It is not clear, however, how one can interpret the sentences they obtained in the absence of any natural language counterpart of the purely technical notion of non-theoremhood.

Let us set the notation. ‘ iff’ stands for ‘if and only if’. For any sets X, Y, and $Z \subseteq X$, and for any function $f : X \longrightarrow Y$, we set $X \sim Y = \{x \in X : x \notin Y\}$, $X \times Y = \{(x, y) \colon x \in X, y \in Y\}$, $f^*Z = \{y \in Y \colon f(x) = y$ for some $x \in Z\}$. If f is one–one, the inverse of f will be denoted by f^{-1}. Now we are ready to give the definition, which is a slightly modified version of the notion appearing in [19] (p. 5).

Definition. $\mathcal{S} = \langle E, S, F, N, g, s\rangle$ is an *abstract formal system* if

(i) $\emptyset \neq S \subseteq F \subseteq E$ and N is an arbitrary set such that $F \cap N = \emptyset$. E is the set of *expressions*, while F, S, and N are called the sets of *formulas*, *(formal) sentences*, and *(formal) names* respectively. The elements of $F \sim S$ are the *proper formulas*. We set $\sim H = F \sim H$ for any $H \subseteq F$ and $\sim X = N \sim X$ for any $X \subseteq N$.

(ii) g is a one–one mapping from F onto N. g is the *naming function of* $\mathcal{S}$. For any subset of F, we shall denote the image of this subset under g by the boldface version of the letter denoting the subset concerned, that is, e.g., if $H \subseteq F$, then g^*H will be denoted by $\boldsymbol{H}$. Later we shall need the simple fact that $g^*(\sim H) = \sim g^*H = \sim\boldsymbol{H}$ for any $H \subseteq F$.

(iii) s is a mapping from $F \times N$ into S. s is the *substitution* in $\mathcal{S}$. We shall denote $s(\varphi, n)$ by $\varphi[n]$ for any $\varphi \in F$, $n \in N$.

Now we try to formulate the formal version of Findlay’s paradox in $\mathcal{S}$. Let $T \subseteq S$ be arbitrary. We shall consider T to be the set of true sentences of $\mathcal{S}$ and, by the same token, the sentences outside T to be the false ones. Further, in the case of any common language sentence s having a single variable, we shall abbreviate the result of substituting a linguistic phrase q for the variable in s by $s(q)$. The name of any common language expression e will be denoted by $\ulcorner e \urcorner$ and we shall abbreviate the phrase ‘the new sentence obtained by substituting the name of the sentence x for the variable in it is false’ by p. Using these notations and denoting Findlay’s paradox by f, we finally have

$$f = p(\ulcorner p \urcorner). \tag{5}$$

Now, the formal name of the formula x is $g(x)$ and the result of the (formal) substitution of this formal name into the formula x is $x[g(x)]$. Therefore, $x[g(x)]$ is the formal version of the phrase

(6) ‘the new sentence obtained by substituting the name of the sentence x for the variable in it’.

Consequently, using the terminology provided by $\mathcal{S}$, p corresponds to the following sentence, which will be denoted by $\overline{p}$:

$$x[g(x)] \notin T.$$

It should be noticed that the phrase (6) has a variable in it so it does not denote a unique sentence; the object it refers to may change with the name of the sentence that is substituted for this variable. This property of indefiniteness is, of course, inherited by its formal version $x[g(x)]$ as well as by the phrase $\overline{p}$, in which this formal version is contained, so we obtain a sentence without a variable, which can be true or false, only after substituting the (informal) name of a formula $\varphi \in F$ for the variable in $\overline{p}$.

Now, of course, $\overline{p}$ is generally not an expression *belonging to* $\mathcal{S}$, but one *about* $\mathcal{S}$! It is still a natural language reformulation of p in terms of the abstract formal system $\mathcal{S}$. Consequently, not belonging to the set of expressions of $\mathcal{S}$, it does not possess a formal name (it is not in the domain of the naming function g), so there is no formal version of $\ulcorner\overline{p}\urcorner$. It follows, then, that we cannot continue the process of formalization, which, by (5), would be the substitution of this formal name for the variable in $\overline{p}$ in order to obtain the formal version of f. Consequently, we get to a crucial point in the reconstruction of the paradox within our formal framework.

2.2. The Formal Liar. The way out of our difficulties lies in the possibility of looking at the situation from another perspective. The fact that prevents us from continuing our procedure can, from the opposite point of view, be taken as a condition under which this procedure can really be accomplished.[8] Indeed, though $\overline{p}$ does not belong to $\mathcal{S}$, there might exist elements of $\mathcal{S}$ that can, in some way, represent it, that is, play the role of $\overline{p}$ *within* $\mathcal{S}$! Of course, these elements, in some sense, should 'express the same state of affairs' as $\overline{p}$. Now let us recall that $\overline{p}$ is a sentence with a variable, so it cannot be true or false. We obtain a sentence that can have a truth value only after substituting the (informal) name of a $\varphi \in F$ for the variable in $\overline{p}$. Those *formal* expressions that have the corresponding property are just the proper formulas of $\mathcal{S}$. So we are seeking a formula, say π, 'to the same effect' as $\overline{p}$, which obviously means that π and $\overline{p}$ have to be true at the same time, in other words, they have to be true exactly for the same formulas, that is, by the definition of $\overline{p}$,

$$\text{for any } \varphi \in F, \quad \pi[g(\varphi)] \in T \text{ iff } \varphi[g(\varphi)] \notin T,$$

or equivalently,

$$\text{for any } n \in N, \quad \pi[n] \in T \text{ iff } g^{-1}(n)[n] \notin T. \tag{7}$$

Our arguments above made it clear that the existence of such a formula characterizes an abstract formal system in a fundamental way guaranteeing that a given object (the set of formulas satisfying $\overline{p}$, i.e., making $\overline{p}$ true if their names substituted for the variable in it) can be discussed within the system. Thus the expansion of this property to any set of objects is related to the expressive power (or, as it is often called, the strength) of a formal system.

[8] Tarski examined first systematically the possibility of eliminating the paradox along these lines (cf. [20]).

DEFINITION. Let $\mathcal{S}$ be an abstract formal system, and let $T \subseteq S$, $X \subseteq N$ be arbitrary. We say that X is *T-representable (in $\mathcal{S}$)* if there is a $\varphi \in F$ such that for every $n \in N$,

$$\varphi[n] \in T \text{ iff } n \in X.$$

The formula φ is said to *T-represent X (in $\mathcal{S}$)*.

Now if π T-represents the set $\{n \in N : g^{-1}(n)[n] \notin T\}$, that is, condition (7) holds, then we can consider π as a representative of $\overline{p}$ within $\mathcal{S}$. But $\overline{p}$ is just the reformulation of p in terms of $\mathcal{S}$, that is, π can be considered to be the formal version of p. This means that we can continue the process of formalization to obtain the formal version of $f = p(\ulcorner p \urcorner)$. Indeed, in this case, the formal object corresponding to $\ulcorner p \urcorner$ is $g(\pi)$, so the one corresponding to f is $\pi[g(\pi)]$. This is the formal sentence of $\mathcal{S}$ corresponding to Findlay's paradox, which in turn is nothing other than a reformulation of the Liar. Consequently, we have obtained the formal version of the Liar. With the notations above,

the *Formal Liar* is the formal sentence $\pi[g(\pi)]$.

This is the formal sentence we sought. Like its informal counterpart, it turns out to be true iff it is false witnessing the fact that the formula π which it is built on cannot exist. The statement to this effect is the resolution of our paradox in a formal setting. The proof is also the formalization of the informal argument.

PROPOSITION (Liar Theorem). *Let $\mathcal{S}$ be an abstract formal system and $T \subseteq S$. The set $\{n \in N : g^{-1}(n)[n] \notin T\}$ is not T-representable.*

PROOF. Suppose that, on the contrary, there is a $\pi \in F$ such that π T-represents the set $\{n \in N : g^{-1}(n)[n] \notin T\}$, that is, for any $n \in N$, $\pi[n] \in T$ iff $g^{-1}(n)[n] \notin T$. Let $\lambda = \pi[g(\pi)]$ and $m = g(\pi)$. Then $\lambda = \pi[g(\pi)] \in T$ iff $\pi[m] \in T$ iff $g^{-1}(m)[m] \notin T$ iff $g^{-1}(g(\pi))[g(\pi)] \notin T$ iff $\lambda = \pi[g(\pi)] \notin T$, which is a contradiction.

So we have the formal version of the Liar at last. To be entirely sincere, the paradox in this form is not too exciting since it is not easy to interpret in informal terms, or, simply, to understand what it is all about. In order to grasp its essence, we should make it more transparent by giving a name to the fundamental notion lying at the heart of the whole matter.

§3. The Generalized Liar Theorem. The Liar theorem as a formalized version of the Liar paradox opens up the possibility of analyzing precisely the notions the original paradox is built on. Before turning to this analysis, however, we should make a short digression since there is an important point to stress here. In the course of formulating the definition of representability, we have taken for granted that the answer to the question whether a given statement is satisfactory, agreeable, acceptable etc. or not depends on its *truth*. The truth of a statement, on the other hand, means some kind of proper correspondence or adequacy between the state of affairs and the statement concerned.[9] So far so good. But let us stop here for a moment. Truth is, by no means, the only concept connecting facts to statements in some way or other, and so it is not the only basis for formulating the definition of representability. In fact, there could exist and do exist other 'measures of adequacy' than truth that lend themselves to formalization more or less readily. It is enough to think of the notions of probability or confirmability (to different extents). These are, clearly, 'weaker' notions than truth. In the other direction, we can find the 'stronger' ones the most obvious choice of which is, of course, that of provability. To be more precise, it would be better to speak about provability in plural since one can introduce, in a very natural way, a couple of different notions of provability widening or narrowing the circle of stipulations as to which types of logical derivations would be accepted as a proof. We shall actually touch upon one of these possible modifications of the generally accepted notion of provability below.

Anyway, it is natural to generalize our treatment to include any one of these possible 'measures of adequacy' and not to consider our definition above to be the only possible definition of representability, but, rather, only one of them. Accordingly, different sets of formal sentences will be considered to stand for sets of informal sentences satisfying different possible adequacy conditions and, until further notice, no special meaning will be associated with them. As a reminder of this fact, we repeat our representability definition using a letter in it without any connotation.

Definition. Let $\mathcal{S}$ be an abstract formal system, and let $A \subseteq S$, $X \subseteq N$ be arbitrary. We say that X is *A-representable* (*in* $\mathcal{S}$) if there is a $\varphi \in F$ such that for every $n \in N$,

$$\varphi[n] \in A \text{ iff } n \in X.$$

We say that φ *A-represents* X (*in* $\mathcal{S}$) or X is *A-represented by* φ (*in* $\mathcal{S}$).

Now we can turn back to the Liar theorem. It is, no doubt, the expression $g^{-1}(n)[n]$ which prevents us from interpreting it informally in an easily comprehensible way. Everything would seem much simpler if we were able to get rid of it. Certainly, we cannot avoid handling it in some way or other. The reason is that $g^{-1}(n)[n]$ is just the formal version of the phrase 'the new sentence obtained by substituting n into the sentence named by it', that is, by (3), the formalization of the expression 'the new sentence obtained by applying the sentence named by n to itself'. In other words, $g^{-1}(n)[n]$ is nothing else than the formalization of *self-reference* and, clearly, due to its central role in the paradox, it will appear in any formulation of the Liar

[9] The result of this informal argument is independent of the particular theory of truth one happens to prefer.

blurring the resulting picture unless we restrict ourselves to formal systems that can circumvent it. We shall do just that. In other words, we shall only examine those formal systems that contain the formulas needed to express self-reference within the system.[10]

DEFINITION. Let $\mathcal{S}$ be an abstract formal system and $A \subseteq S$. $\mathcal{S}$ is *self-referential with respect to A* if for any $X \subseteq N$,

$$\{n \in N : g(g^{-1}(n)[n]) \in X\} \text{ is } A\text{-representable}$$
$$\text{whenever } X \text{ is } A\text{-representable.}$$

Self-referentiality corresponds to the condition formulated in the *fixed–point* (or *diagonal*) lemma of mathematical logic. It plays a central role in internal limitations of formal systems. On the one hand, as we shall see, it will occur as a condition in every statement that we shall formulate concerning these limitations. On the other hand, most of the usual formal systems of arithmetic are actually self-referential.

With the notion of self-referentiality at our disposal, we get to the last step of our formalization process. Recognizing that the set T plays a double role in the Liar theorem, it is very natural to modify the wording and proof of this statement in the only obvious way to obtain a theorem and its proof about the relation between *two* sets of formal sentences in place of a less general proposition being about only a single one. This modification of the Liar theorem, together with the application of self-referentiality, yields the central result of our investigations. Indeed, the statement below, which is the generalized formal version of the Liar paradox, will turn out to be the main link between the original paradox and the theorems on the limitations of formal reasoning.

THEOREM 1 (Generalized Liar Theorem). *Let $\mathcal{S}$ be an abstract formal system, $A \subseteq F$, $B \subseteq S$. Let $\mathcal{S}$ be self-referential with respect to B and suppose that $\sim\!\boldsymbol{A}$ is B-representable. Then there is a $\lambda \in S$ such that*

$$\lambda \in B \text{ iff } \lambda \notin A,$$

which, in turn, implies that

$$S \cap A \neq B.$$

PROOF. Since $\sim\!\boldsymbol{A}$ is B-representable and $\mathcal{S}$ is self-referential with respect to B, $\{n \in N : g(g^{-1}(n)[n]) \notin \boldsymbol{A}\}$ is also B-representable, that is, there is a $\pi \in F$ such that for any $n \in N$, $\pi[n] \in B$ iff $g(g^{-1}(n)[n]) \notin \boldsymbol{A}$. Let $m = g(\pi)$ and $\lambda = \pi[g(\pi)]$. Then, on the one hand, $\lambda \in S$. On the other hand, $\lambda = \pi[g(\pi)] \in B$ iff $\pi[m] \in B$ iff $g(g^{-1}(m)[m]) \notin \boldsymbol{A}$ iff $g^{-1}(m)[m] \notin A$

[10]The most obvious examples of formal systems with this property are those induced by natural languages, which are rich enough to contain the expression corresponding to the general self-reference itself (cf. (6)).

iff $g^{-1}(g(\pi))[g(\pi)] \notin A$ iff $\lambda = \pi[g(\pi)] \notin A$. That is, we have a $\lambda \in S$, such that $\lambda \in B$ iff $\lambda \notin A$, which was to be proved.

Of course, in such an abstract wording, this theorem does not seem to tell much about the relevant formal systems of mathematics. Before applying it to the logical notions that are the usual objects of metamathematical investigations, however, we would like to illustrate the way it can be used by applying it to the notion of provability *in less than a given number of steps* not usually examined despite its very interesting intuitive meaning.[11]

Mimicking the informal systems of mathematics, we can define the set of provable sentences in an abstract formal system $\mathcal{D}$ along the following lines. After selecting a set of sentences that are called axioms and a set of rules for inferring an expression from a finite set of other ones, a *proof* is defined as a finite sequence of expressions every term of which is an axiom or can be inferred from the set of terms preceding the given one. A sentence is defined to be *provable*, in other words, to be a *theorem*, if it is the last term of some proof. Now, for any positive integer n, let us define the set P_n as that of the sentences provable in less than n steps, that is, appearing as the last term of a proof consisting of less than n expressions. Further, let P be the set of all provable sentences of $\mathcal{D}$. Clearly, $P_n \subseteq P$ for any n. Thus if we interpret A and B in the Generalized Liar theorem as P_n and P respectively, then we have

PROPOSITION. *Let $\mathcal{D}$ be self-referential with respect to P. For any positive integer n, if $\sim\boldsymbol{P}_n$ is P-representable, then*

$$P \sim P_n \neq \emptyset.$$

The proof of the fact that in the most important cases of the systems of formalized arithmetic the conditions of the proposition actually hold for any positive integer is essentially simple. Yet, requiring a lot of formal calculations it does not fit into our non-technical exposition.[12] Nevertheless,

[11] There are, however, important results related to this notion. Perhaps the most interesting ones are those that provide partial confirmation of Kreisel's conjecture because they shed light on the nature of arithmetical incompleteness. The conjecture asserts that, in Peano arithmetic, if all the numerical instances $\varphi(n)$ of a formula $\varphi(x)$ are provable and there is a uniform bound to the lengths of their shortest proofs, then the universal sentence $(\forall x)\varphi(x)$ is also provable. Parikh, Baaz, and Pudlák have proved Kreisel's conjecture for different versions of arithmetic (cf. [12] and [1]).

[12] Self-referentiality can directly be checked. As far as P-representability of $\sim\boldsymbol{P}_n$ is concerned, in the systems we are interested in, for every n, P_n can be described as a set of the last terms of *finitely many* proof *schemas*. In fact, the axioms are given as a finite set of formula schemas and the result of the application of any of the finitely many inference rules to a finite set of formula schemas is again a formula schema. To obtain the formula P-representing the complement of a set consisting of names of formulas belonging to a finite set of formula schemas is, in turn, only a matter of a straightforward construction. (A less direct way to show P-representability of $\sim\boldsymbol{P}_n$ is to refer to Church's thesis on the equvalence of all formalizations of the informal notion of decidability. Indeed, since P_n is a finite set of formula schemas, the problem whether or not the name of a given formula belongs to $\boldsymbol{P}_n$ is decidable.

the conclusion of the proposition is worth spelling out in a little more detail since its informal interpretation reveals an interesting feature of the system considered. It states that for any positive integer n, however large it is, there are theorems of the system that can only be proved in more than n steps, that is, there are arbitrary long proofs establishing new theorems. This, on the one hand, is a kind of practical incompleteness since there are theorems of the system whose demonstration needs an astronomical period of time, and consequently, although they are theoretically provable, will never be proved for banal practical reasons: mankind simply has not and will never have enough time to prove them. From another point of view, this very property corresponds formally to the 'unboundedness' of the formal system considered, that is, to the fact that there are new theorems being arbitrarily 'far'[13] from the axioms, thus asserts formally its inexhaustibility, complexity, or richness.

Let us turn now to those special abstract formal systems that constitute the conceptual framework for investigating the theoretical aspects of mathematical activity.

§4. Limitations of logical systems. In order to apply the Generalized Liar theorem to the formal systems of mathematics, we should formulate the minimum requirements any ordinary logic has to comply with. Obviously, such a system ought to have two distinguished sets of sentences that can be interpreted as those of provable and true sentences respectively and the system has to be able to express the basic logical operation of negation. Further, we shall need some notions to describe the basic properties of such systems.

DEFINITION. Let $\mathcal{S}$ be an arbitrary abstract formal system.

(i) Let $P, T \subseteq S$ be arbitrary. We say that $\mathcal{L} = (\mathcal{S}, P, T, ')$ is a *logical system* if $'$ is a mapping from F into F such that
 (a1) $\varphi \in S$ iff $\varphi' \in S$
 (a2) $\varphi \in T$ iff $\varphi' \notin T$ whenever $\varphi \in S$, and, for any $n \in N$,
 (b1) $\varphi[n] \in T$ iff $\varphi'[n] \notin T$
 (b2) $\varphi[n] \in P$ iff $\varphi''[n] \in P$
 (b3) $\varphi'[n] \in P$ iff $(\varphi[n])' \in P$.

A possible formalization of this fact, on the other hand, contains P-representability of $\sim\boldsymbol{P}_n$ as a necessary condition (cf. our remarks on page 20).)

[13]As a matter of fact, to say that two sentences are *far* from each other is more than just a suggestive metaphor. Actually, the function assigning to any pair of sentences the length of a shortest proof deriving the second sentence from the first one is a possible generalization of the notion of distance, which can be extended to sets of sentences in an obvious way. In fact, it corresponds to a notion in functional analysis (called *quasi–pseudometric*) satisfying, among others, the triangle inequality.

P and T are called the sets of *provable* and *true* sentences of $\mathcal{S}$ respectively and φ' is the *negation of* φ. We shall use the following notation for any $H \subseteq F$: $H' = \{\varphi \in F : \varphi' \in H\}$.

(ii) Let $\mathcal{L} = (\mathcal{S}, P, T, ')$ be a logical system.
   (a) $\mathcal{L}$ is *consistent* if $P \cap P' = \emptyset$, otherwise it is *inconsistent*.
   (b) $\mathcal{L}$ is *complete* if $P \cup P' = S$, otherwise it is *incomplete*.
   (c) $\mathcal{L}$ is *sound* if $P \subseteq T$.

Obviously, consistency is one of those conditions that any meaningful logical system should satisfy since this means that it is free of contradiction: for any sentence and its negation, *at most* one of them belongs to the set of provable sentences. Moreover, clearly, the set of provable sentences has to be chosen in such a way that the system remain sound; only the true sentences are permitted to be provable. On the other hand, our main concern is to examine the completeness of formal systems, the property which is, in some sense, a maximal requirement, a dual of the consistency. In complete systems, for an arbitrary sentence and its negation, *at least* one of them does belong to the set of provable sentences. The importance of this notion partly stems from the fact that, since all the interesting formal systems are sound, any one of them is complete iff $P = T$,[14] that is, *all* true sentences are indeed provable.

Now it is immediate from the definitions of consistency and completeness that a logical system $\mathcal{S}$ is complete *and* consistent iff $S \sim P = P'$.[15] Clearly, the relation between this equation and the conclusion of the Generalized Liar theorem needs no comment; they together yield an abstract version of Gödel's incompleteness theorem to the effect that, under suitable conditions concerning self-referentiality and representability, *consistency and completeness exclude each other.* What is more, restricting ourselves to logical systems, the Generalized Liar theorem takes a form that yields the abstract versions of three basic theorems concerning the limitations of formal methods. This fact demonstrates that these theorems are *all* different manifestations of the *same* logical principle:[16]

THEOREM 2. *Let $\mathcal{L} = (\mathcal{S}, P, T, ')$ be an arbitrary logical system.*

(i) (a) *Suppose that $\mathcal{L}$ is self-referential with respect to P and $\boldsymbol{P}$ is P-representable.*

If $\mathcal{L}$ is consistent, then $\mathcal{L}$ is incomplete.

[14] Assuming soundness, $P \neq T$ implies that there is a $\sigma \notin P$ such that $\sigma \in T$ (iff $\sigma' \notin T$), that is, $\sigma' \notin P$. On the other hand, by definition, $T \cup T' = S$. Therefore, if $P = T$, then $P \cup P' = S$.

[15] i.e., $\{P, P'\}$ is a *partition* of S.

[16] Smullyan, in [19], investigated a wide variety of abstract theorems concerning the limitations of formal systems. It is interesting, however, that he has apparently not recognized their common logical structure or, at least, has not found it to be worth examining.

(b) *Suppose that* $\mathcal{L}$ *is self-referential with respect to* T *and* $\boldsymbol{P}$ *is* T*-representable.*

$$\text{If } \mathcal{L} \text{ is sound, then } T \sim P \neq \emptyset.$$

(ii) *Suppose that* $\mathcal{L}$ *is self-referential with respect to* T*. Then*

$$\boldsymbol{T} \text{ is not } T\text{-representable.}$$

(iii) *Suppose that* $\mathcal{L}$ *is self-referential with respect to* P*. Then*

$$\sim\!\boldsymbol{P} \text{ is not } P\text{-representable.}$$

Proof. We list the substitutions needed to obtain the different items from the Generalized Liar theorem:

(i) (a) $A = \sim P$, $B = P'$
(i) (b) $A = P$, $B = T$
(ii) $A = T$, $B = T$
(iii) $A = P$, $B = P$.

Going into a little more detail, it directly follows from the definitions that, generally, for any set $C \subseteq S$, if $\boldsymbol{C}$ is T-represented by $\varphi \in F$, then $\sim\!\boldsymbol{C}$ is T-represented by $\varphi' \in F$, and if $\boldsymbol{C}$ is P-represented by $\varphi \in F$, then $\boldsymbol{C}$ is P'-represented by $\varphi' \in F$. Further, self-referentiality of $\mathcal{S}$ with respect to P implies its self-referentiality with respect to P'. Using these facts where needed, we have:

(i) (a) $\sim g^*(\sim P) = \sim\sim\boldsymbol{P} = \boldsymbol{P}$, so, by the Generalized Liar, $S \sim P \neq P'$, which is what we need since $S \sim P \supseteq P'$ iff $\mathcal{S}$ is consistent and $S \sim P \subseteq P'$ iff $\mathcal{S}$ is complete.
(i) (b) $\mathcal{S}$ is sound, thus $P \subseteq T$ and, by the Generalized Liar, $P \neq T$.
(ii) By the Generalized Liar, T-representability of $\sim\!\boldsymbol{T}$ implies $T \neq T$.
(iii) By the Generalized Liar, P-representability of $\sim\!\boldsymbol{P}$ implies $P \neq P$.

The statements of the theorem above can be considered to be the abstract versions of theorems of Gödel, Tarski, and Church concerning a single logical system. Indeed, in addition to its first main item consisting of the abstract syntactic and semantic versions of Gödel's incompleteness theorem describing the relation between *provability and refutability*[17] and that between *provability and truth*, respectively, it contains, as its second and third main items, a generalization of Tarski's theorem on the *undefinability of truth*,[18] and a statement that, as we shall see below, can be interpreted as an abstract variant of Church's theorem on the *undecidability of provability* restricted to a single logical system. Moreover, covering every sensible cast of roles

[17]A sentence is refutable if its negation is provable. In fact, this proposition is the abstract version of an improvement of Gödel's original theorem due to Rosser (cf. e.g., [8] pp. 64–5 or [10] pp. 143 and 145). It is worth noting that this is a purely syntactic statement having essentially nothing to do with the notion of truth.

[18]Informally this result can be interpreted as stating that the notion of being true in $\mathcal{S}$ cannot be defined within $\mathcal{S}$. As to the original theorem, see e.g., [4] p. 176 or [10] p. 151.

for the Generalized Liar theorem, its proof indicates that the theorems of Gödel, Tarski, and Church are just the only possible relevant results given in terms of truth and provability alone that can be considered as some direct reformulations of the Liar paradox within the conceptual framework of modern mathematical logic. Let us underline that this theorem is, in a sense, exhaustive: it contains some general versions of all the known main theorems on truth and provability (concerning the limitations of deductive methods) that can be formulated at the level of abstraction implicitly defined by the formalization of the ordinary language Liar.[19] It describes the general *conceptual structure* of these results making explicit the different roles played, so to speak, by some *general* laws of logic (represented by the Liar) on the one hand, and, on the other, by the special features of the particular system concerned. From this general point of view, the essence of Gödel's original result on the incompleteness of formal arithmetic consists in demonstrating that the conditions of the theorem concerning representability and self-referentiality really hold in that special case, so that the theorem can indeed be applied to the logical systems of arithmetic.[20]

Let us now apply our theorem above in a nonstandard way in the opposite direction. Presburger has shown that additive number theory (a weakened version of the standard Peano arithmetic obtained by the omission of multiplication) is complete (cf. e.g., [5], p. 43). In this system the set of all expressions can be enumerated, and the question whether a given term of the resulting sequence is a provable sentence or not can be decided.[21] So the naming function g can be defined by recursion in the following way. For any expression e, we set $g(e)$, according to whether e is a provable sentence or not, to be the smallest even or odd natural number, respectively, that has not already been chosen. Now $P = T$, since the system considered is sound and complete. On the other hand, both the set of all even natural numbers

[19]The only missing important theorem on internal limitations of formal systems, perhaps the most important one, Gödel's theorem on the unprovability of consistency (according to which if arithmetic is consistent, then 'it cannot prove its own consistency'), cannot be taken as a different theorem at the level of our formalization since it is, essentially, the formal version of the incompleteness theorem itself given at a still deeper level, *within* formal arithmetic.

[20]One of Gödel's ingenious discoveries, the so called Gödel numbering, is the method of associating numbers with expressions in such a way that the resulting naming function has the most important property of the quotation as the naming of ordinary languages, viz. that the name of an expression contains every important information of the named object (clearly, the quoted version of a common language expression is ideal in this respect). The logical system endowed with such a naming function inherits the expressive power of natural languages, among others, their self-referentiality and the fact that $\boldsymbol{P}$ is representable. Loosely speaking, the expressive power of arithmetic can be compared to that of the natural languages since it can also be forced to talk about itself.

[21]This simply follows from the fact that the set of all proofs can be enumerated and, by completeness, for *every* sentence, either the sentence itself or its negation must occur as a last term of a proof.

and that of all odd ones *are* representable[22]. Consequently, in this system $\boldsymbol{T}$, $\boldsymbol{P}$, and $\sim\!\boldsymbol{P}$ are both P- and T-representable (so, among others, truth *is* definable[23]) and, therefore, by the theorem, this system is self-referential with respect to neither T nor P.

This example, together with the facts that, on the one hand, Gödel's incompleteness result provides examples of self-referential logical systems in which $\boldsymbol{P}$ is representable (thus the systems are incomplete) and, on the other hand, our previous theorem shows that $\boldsymbol{P}$ is not representable in those self-referential logical systems in which $P = T$ (since these systems are complete, sound, and consistent), yields a complete description of the mutual correspondence between the general properties of self-referentiality, representability and completeness.

At the same time, the theorem above is only about a single logical system and it does not tell anything about the relevant properties of possible extensions, that is, about the hereditary character of incompleteness, undefinability, and undecidability. More precisely, since the set of true sentences is—by definition—maximal, this question only arises in connection with the statements concerning the set of provable sentences, that is, those about incompleteness and undecidability (of provability).

Let us first examine the question of undecidability since there is a yet unexploited apparent possibility to modify the Generalized Liar theorem just in this direction. For there are two ways to apply the Generalized Liar theorem to logical systems and, by its direct application, we have only used one of these two possibilities. Actually, we can choose another approach making use of the notion of negation to define a symmetric version of the concept of representability. With this new property at our disposal, before applying the Generalized Liar theorem to logical systems, we can first modify the claim and proof of the theorem itself (again in the only obvious way simply supplementing the original proof by its counterpart in which the (generalized) Formal Liar sentence is replaced by its negation). The symmetric version of the theorem obtained in this way yields a faithful abstract version of Church's theorem.

Definition. Let $\mathcal{L} = (\mathcal{S}, P, T, ')$ be a logical system, and let $A \subseteq S$, $X \subseteq N$ be arbitrary.

(a) Any $X \subseteq N$ is said to be *strongly A-representable* (*in* $\mathcal{L}$) if X and $\sim\!X$ are A-represented by φ and φ' respectively for some $\varphi \in F$. φ is said to *strongly A-represent X* (*in* $\mathcal{L}$).

(b) $\mathcal{L}$ is *strongly self-referential with respect to A* if for any $X \subseteq N$, $\{n \in N : g(g^{-1}(n)[n]) \in X\}$ is strongly A-representable whenever X is strongly A-representable.

[22] The system contains the formulas $(\exists y)(y + y = x)$ and $(\forall y)(y + y \neq x)$.

[23] That is, we have a very natural example of a system which 'can define its own truth'; actually, this is a more transparent and self-explanatory example than Myhill's one, cf. [11].

THEOREM 3 (Symmetric Generalized Liar Theorem). *Let* $\mathcal{L} = (\mathcal{S}, \boldsymbol{B}, T, ')$ *be a logical system and* $A \subseteq F$. *Let* $\mathcal{L}$ *be strongly self-referential with respect to* B *and suppose that* $\boldsymbol{A}$ *is strongly* B*-representable. Then there is a* $\lambda \in S$ *such that*

(i) $\lambda \in B$ iff $\lambda \notin A$,
(ii) $\lambda' \in B$ iff $\lambda \in A$.

PROOF. Since $\boldsymbol{A}$ is strongly B-representable and $\mathcal{L}$ is strongly self-referential with respect to B, there is a $\pi \in F$ such that for any $n \in N$, $\pi[n] \in B$ iff $g(g^{-1}(n)[n]) \in \boldsymbol{A}$ and $\pi'[n] \in B$ iff $g(g^{-1}(n)[n]) \notin \boldsymbol{A}$. Let $m = g(\pi')$ and $\lambda = \pi'[g(\pi')]$. Then $\lambda \in S$ and $\lambda = \pi'[g(\pi')] \in B$ iff $\pi'[m] \in B$ iff $g(g^{-1}(m)[m]) \notin \boldsymbol{A}$ iff $g^{-1}(m)[m] \notin A$ iff $g^{-1}(g(\pi'))[g(\pi')] \notin A$ iff $\lambda = \pi'[g(\pi')] \notin A$. On the other hand, $\lambda' = (\pi'[g(\pi')])' \in B$ iff $\pi''[g(\pi')] \in B$ iff $\pi[g(\pi')] \in B$ iff $\pi[m] \in B$ iff $g(g^{-1}(m)[m]) \in \boldsymbol{A}$ iff $g^{-1}(m)[m] \in A$ iff $g^{-1}(g(\pi'))[g(\pi')] \in A$ iff $\lambda = \pi'[g(\pi')] \in A$, which concludes the proof.

Now, both the notion of P-representability and that of its strong version provide genuine characterizations of logical systems from a 'practical' point of view. For in each standard logical system describing the arithmetic of natural numbers, P-representability of the set of all names of sentences provable in another logical system (satisfying some natural conditions) is a formal counterpart of the informal property that, in this other system, provability can be defined through a decidable set of axioms.[24] Hence, in

[24]Decidability of the set of axioms means that there exists a purely mechanical rule to decide the question whether a sentence is an axiom or not. Now the formal property we are talking about is indeed a possible choice as a formal version of this informal notion in the sense that, generally, P-representability of the set of all names of provable sentences does imply the existence of a decidable set of axioms. First, in all logical systems formalizing real mathematical practice (in which the notion of proof is defined the way we described briefly on page 13), the set of all provable sentences is enumerable just in case there exists a decidable set of axioms (cf. [16] p.138). (In one direction, this connection is trivial. Indeed, let us suppose the existence of such a set of axioms. Now, in the systems concerned, the set of all finite sequences of expressions—which themselves are finite sequences of symbols chosen from a finite vocabulary—can be enumerated applying a purely mechanical method. Furthermore, the rules of inference—just as in the world of informal mathematics—are defined in such a way that the fact whether or not a given sequence of expressions is a proof from a decidable set of axioms can be decided in a finite number of steps. Therefore an enumeration of the set of all finite sequences of expressions gives rise to an enumeration of the set of proofs and, by the same token, that of the set of all provable sentences.) Secondly, the axioms of all standard arithmetical systems *are*, indeed, given in a decidable way. Consequently, in a standard arithmetical system $\mathcal{A} = (\mathcal{S}, P, T, ')$ (in which the set N of names is a subset of natural numbers), the set of provable sentences can be enumerated, which, in turn, provides an enumeration of any P-representable set of natural numbers (since the question of whether or not a sentence is the result of the substitution of a given number in a given formula is also decidable). Now let $\mathcal{M}$ be an arbitrary logical system formalizing a segment of real mathematics such that the set of names in $\mathcal{M}$ is a subset of N. Then the naming function in $\mathcal{M}$ can be defined in such a way that each value of its inverse can be computed in a finite number of steps (cf. footnote 20). Thus, if Q is the set of provable sentences of $\mathcal{M}$ and $\boldsymbol{Q}$ is

the presence of a logical system $\mathcal{L} = (\mathcal{S}, P, T, ')$ that has already been fixed in some way or other, it is natural to choose P-representability of the set of all names of sentences provable in a logical system $\mathcal{M} = (\mathcal{S}, Q, T, ')$ as the abstract notion characterizing provability in $\mathcal{M}$ from the point of view of its definability via a decidable set of axioms. On the other hand, in the usual logical systems of arithmetic (in which the set of names is a subset of natural numbers), strong P-representability of an arbitrary set of natural numbers is one of the several equivalent properties devised to formalize the informal notion of decidability of the question whether or not a number belongs to the set considered.[25] This fact suggests an obvious way to define the abstract version of decidability (of provability) concerning the logical systems we are interested in.

Definition. Let $\mathcal{L} = (\mathcal{S}, P, T, ')$ be a logical system. For any $Q \subseteq S$, we shall denote by $\mathcal{L}_Q$ the quadruple $(\mathcal{S}, Q, T, ')$ whenever it is a logical system. In particular, $\mathcal{L}_P = \mathcal{L}$. If $Q \supseteq P$, then we say that $\mathcal{L}_Q$ is an *extension* of $\mathcal{L}$. Further

(i) $\mathcal{L}_Q$ is $\mathcal{L}$-*axiomatizable* if $\boldsymbol{Q}$ is P-representable.
(ii) $\mathcal{L}_Q$ is $\mathcal{L}$-*decidable* if $\boldsymbol{Q}$ is strongly P-representable, otherwise it is $\mathcal{L}$-*undecidable*.

Proposition. *Let $\mathcal{L} = (\mathcal{S}, P, T, ')$ be a logical system which is strongly self-referential with respect to P. Then any consistent extension of $\mathcal{L}$ is $\mathcal{L}$-undecidable.*

Proof. Let $\mathcal{L}_Q$ be a consistent extension of $\mathcal{L}$ and suppose that $\mathcal{L}_Q$ is $\mathcal{L}$-decidable. Then we can apply the Symmetric Generalized Liar with $A = Q$ and $B = P$, that is, there is a $\lambda \in S$ such that

$$\text{(i)}\ \lambda \in P \text{ iff } \lambda \notin Q \qquad\qquad \text{(ii)}\ \lambda' \in P \text{ iff } \lambda \in Q$$

Now, $\lambda \in P$ leads to a contradiction since $\lambda \in P \subseteq Q$ implies $\lambda \in Q$ and, at the same time, by (i), $\lambda \notin Q$. Consequently, $\lambda \notin P$, which in turn, again by (i), implies $\lambda \in Q$. Then, by (ii), $\lambda' \in P \subseteq Q$, that is, $\lambda' \in Q$. Therefore $\lambda \in Q'$. But this means that $\lambda \in Q \cap Q'$ contradicting the consistency of $\mathcal{L}_Q$, which concludes the proof.

P-representable, then $\boldsymbol{Q}$ and hence Q is enumerable, thus Q is defined by a decidable set of axioms.

[25] In a little more detail, the informal property to be formalized is as follows: for the given set of natural numbers, there exists an entirely mechanical method that, for any natural number n, can be used to decide the question whether or not n belongs to the set considered. Again, it can easily be shown that strong P-representability is a possible choice as the formal version of decidability, that is, strong P-representability does imply decidability. Since P-representability of a set implies its enumerability (see our remarks in the previous footnote), strong P-representability of a set guarantees the enumerability of both the set itself and its complement. In order to decide whether a number belongs to such a set, therefore, we simply enumerate the set and its complement simultaneously until our number appears on one of the resulting lists since it must appear on one of them sooner or later.

Having proved the faithful abstract version of Church's theorem on the *hereditary* undecidability of arithmetic,[26] we can make an attempt to generalize the abstract incompleteness theorems in the same direction. Now, this is the first time in our project that we should face the difficulties caused by the limited expressive power of the (meta)language we have chosen by defining logical systems in a fairly general way. In order to circumvent these difficulties, we are forced to introduce some strengthened versions of axiomatizability using generalized versions of completeness and soundness. We shall also need a strengthened version of self-referentiality.

Definition. Let $\mathcal{L} = (\mathcal{S}, P, T, ')$ be a logical system and let $\Delta \subseteq F$ be closed under substitution.

(a) $\mathcal{L}$ is Δ-*complete* if $\Delta \cap T \subseteq P$.
(b) $\mathcal{L}$ is Δ-*sound* if $\Delta \cap P \subseteq T$.

Before going on, let us simplify a little our notation. Having any one of the standard logical systems of arithmetic[27] in mind, let, from now on, $\mathcal{P} = (\mathcal{S}, P, T, ')$ be an arbitrary but fixed logical system and let $\Sigma \subseteq F$ be an arbitrary but fixed set such that $\mathcal{P}$ is Σ-complete.

Definition.

(i) Let $Q \subseteq S$ be arbitrary.
  (a) $\mathcal{P}_Q$ is Σ-*completely* $\mathcal{P}$-*axiomatizable* if $\boldsymbol{Q}$ is P-represented by some $\varphi \in \Sigma$.
  (b) $\mathcal{P}_Q$ is Σ-*soundly* $\mathcal{P}$-*axiomatizable* if
    (1) $\mathcal{P}_Q$ is Γ-sound for some $\Gamma \subseteq F$
    (2) $\boldsymbol{Q}$ is P-represented by some $\varphi \in \Sigma \cap \Gamma$.
(ii) $\mathcal{P}$ is *hereditarily self-referential* if for any $Q \supseteq P$, $\mathcal{P}_Q$ is self-referential with respect to Q.

Though all of these new notions seem rather *ad hoc* from our abstract point of view, the concept of Δ-completeness is the generalization of a genuine arithmetical notion. As a matter of fact, there is, in formal arithmetic, a set of formulas (usually denoted by Σ_1) which satisfies the definition and with respect to which the axiomatizable and completely axiomatizable extensions coincide (cf. e.g., [3] pp. 25 and 44). As far as the notion of hereditary self-referentiality is concerned, in the logical systems of arithmetic, it is an immediate consequence of self-referentiality. (For that matter, the same is also true of strong self-referentiality.) On the other hand, the only justification for introducing the notion of sound axiomatizability is the fact that the application of this condition seems to be the only way to reconstruct—at least in some weakened form—the extended syntactical version of Gödel's

[26] As far as the original theorem is concerned, see e.g., [16] p. 131.

[27] The paradigmatic case is, of course, Peano arithmetic, but, among others, we may think of Robinson arithmetic (see e.g., [4] p. 158), which is a much weaker system.

theorem in such an abstract setting as our present one. The notion, in a slightly different form, appears in a remark of Smorynski (cf. [17] p. 841).

PROPOSITION. (i) *Suppose that $\mathcal{P}$ is hereditarily self-referential. Then any Σ-soundly $\mathcal{P}$-axiomatizable consistent extension of $\mathcal{P}$ is incomplete.*

(ii) *Suppose that $\mathcal{P}$ is self-referential with respect to T. If $\mathcal{P}$ is sound, then any Σ-completely $\mathcal{P}$-axiomatizable sound $\mathcal{P}_Q$ is incomplete.*

PROOF. (i) Let $\mathcal{P}_Q$ be a Σ-soundly $\mathcal{P}$-axiomatizable consistent extension of $\mathcal{P}$. Then there are $\Gamma \subseteq F$ and $\varphi \in \Sigma \cap \Gamma$ such that $\Gamma \cap Q \subseteq T$ and $\boldsymbol{Q}$ is P-represented by φ. By Theorem 2 (i)(a), it is enough to show that $\boldsymbol{Q}$ is also Q-represented by φ. Now, let $n \in N$ be arbitrary. Then $n \in \boldsymbol{Q}$ implies $\varphi[n] \in P \subseteq Q$, so $\varphi[n] \in Q$. On the other hand, let $\varphi[n] \in Q$. Then $\varphi[n] \in \Gamma \cap Q \subseteq T$. Therefore $\varphi[n] \in \Sigma \cap T$, which in turn, by the Σ-completeness of $\mathcal{P}$, implies $\varphi[n] \in P$, that is, $n \in \boldsymbol{Q}$.
(ii) Let $Q \subseteq S$ be arbitrary and suppose that $\mathcal{P}_Q$ is Σ-completely $\mathcal{P}$-axiomatizable and sound. Then there is a $\varphi \in \Sigma$ such that $\boldsymbol{Q}$ is P-represented by φ. We first show that $\boldsymbol{Q}$ is also T-represented by φ so that we can apply Theorem 2 (i)(b) to $\mathcal{P}_Q$. Let $n \in N$ be arbitrary. Then $n \in \boldsymbol{Q}$ implies $\varphi[n] \in P \subseteq T$ ($\mathcal{P}$ is supposed to be sound), so $\varphi[n] \in T$. On the other hand, it follows from $\varphi[n] \in T$ that $\varphi[n] \in \Sigma \cap T$, which in turn, by the Σ-completeness of $\mathcal{P}$, implies $\varphi[n] \in P$, that is, $n \in \boldsymbol{Q}$. Since self-referentiality of $\mathcal{P}$ with respect to T implies the same for $\mathcal{P}_Q$, Theorem 2 (i)(b) can, indeed, be applied to $\mathcal{P}_Q$. Consequently $T \sim Q \neq \emptyset$. In order to conclude the proof, therefore, it is enough to refer to the fact that a sound system is complete just in case all true sentences are provable (see our remarks on p. 15).

In the arithmetic counterpart of the proposition above, both the condition of sound axiomatizability and that of the complete one can be weakened to axiomatizability.[28] In the case of the complete axiomatizability, this fact simply follows from the equivalence between the axiomatizability and complete axiomatizability within arithmetic, so the second part of the proposition is, in a sense, a faithful abstract version of its arithmetical counterpart. In the other case, however, the arithmetical result is really a strictly stronger one than its abstract variant. Actually, the presence of the condition weakening the first part of the proposition above is an indication of the fact that, by formulating it, we have reached the limits of our project set by the Liar paradox. Indeed, being the only one among our abstract versions of the results on the theoretical limitations of formal systems whose arithmetical counterpart can essentially be strengthened, this statement seems to be the final point one can get to solely on the basis of the Liar paradox.[29]

[28] See e.g., [4] p. 179 and [2] p. 100, respectively.

[29] Some facts, at least, support this conjecture. There are two standard ways to prove the stronger arithmetic version of this statement. It seems that neither of them can be reconstructed in our abstract framework. The first proof is a constructive one, which, instead of the Gödel sentence (the arithmetical realization of the (generalized) Formal Liar

It is worth noting, on the other hand, that the Symmetric Generalized Liar can also be utilized to reformulate the abstract version of Church's theorem in terms of recursive sets. Any $Z \subseteq N$ is called $\mathcal{P}$-*recursive* if Z is strongly P-representable, and the sets $X, Y \subseteq N$ are called $\mathcal{P}$-*recursively separable* if there is a $\mathcal{P}$-recursive $Z \subseteq N$ such that $X \subseteq Z$ and $Y \cap Z = \emptyset$. Clearly, separability is a symmetric relation. An extension $\mathcal{P}_Q$ of $\mathcal{P}$ is $\mathcal{P}$-*recursively inseparable* if $\boldsymbol{Q}$ and $\boldsymbol{Q}'$ are not $\mathcal{P}$-recursively separable sets. Now, using the Symmetric Generalized Liar, we can essentially repeat the proof of the abstract Church's theorem to obtain:

PROPOSITION. *If $\mathcal{P}$ is strongly self-referential with respect to P, then $\mathcal{P}$ is $\mathcal{P}$-recursively inseparable.*

PROOF. Suppose that, on the contrary, there exists a $\mathcal{P}$-recursive $Z \subseteq N$ that separates $\boldsymbol{P}$ an $\boldsymbol{P}'$. Then Z is strongly P-represented by some $\varphi \in F$. Let $C = \{g^{-1}(n) \in F : \varphi[n] \in P\}$. Then $\boldsymbol{C} = Z$, so we have $\boldsymbol{P} \subseteq \boldsymbol{C}$ and $\boldsymbol{P}' \cap \boldsymbol{C} = \emptyset$ implying $P \subseteq C$ and $P' \cap C = \emptyset$. The application of the Symmetric Generalized Liar to $A = C$ and $B = P$ shows that there is a $\lambda \in S$ such that

$$\text{(i)}\ \lambda \in P \text{ iff } \lambda \notin C \qquad\qquad \text{(ii)}\ \lambda' \in P \text{ iff } \lambda \in C.$$

Now, $\lambda \in P$ leads to a contradiction since $\lambda \in P \subseteq C$ implies $\lambda \in C$ and, at the same time, by (i), $\lambda \notin C$. Therefore, $\lambda \notin P$, which in turn, again by (i), implies $\lambda \in C$. Thus, by (ii), $\lambda' \in P$, that is, $\lambda \in P'$. Hence $\lambda \in P' \cap C = \emptyset$, which, being a contradiction, concludes the proof.

Obviously, $\mathcal{P}$-recursive separability of one of its extensions would imply the same for $\mathcal{P}$ itself, so we have

COROLLARY. *If $\mathcal{P}$ is strongly self-referential with respect to P, then every extension of $\mathcal{P}$ is $\mathcal{P}$-recursively inseparable.*

Let us conclude our analysis by summarizing, in a unified terminology, those propositions that, in an abstract formal framework, can be considered as some more or less direct reformulations of the Liar paradox 'This sentence is false.':

---

sentence), is based on a more complicated formal sentence devised by Rosser. But Rosser's sentence exploits some special properties of the language of arithmetic that do not have abstract counterparts; among others, it is built on the possibility to talk, in this language, about the ordering of natural numbers (cf. e.g., [10] pp. 145–7). The second proof is based on an indirect argument, hence it is non-constructive. Nevertheless, as is generally the case with indirect proofs, this second reasoning is the more elegant one. It exploits the intuitively plausible fact that axiomatizable complete systems of arithmetic are decidable (cf. footnote 21), which in turn, by Church's theorem, immediately yields that any consistent axiomatizable extension of e.g., the Peano arithmetic is incomplete (cf. e.g., [16] p. 132). The formal proof of the decidability of axiomatizable complete systems of arithmetic, however, is based on the internal structure of the formula P-representing provability, so cannot, in principle, be reconstructed in our abstract setting either.

THEOREM 4 (Abstract versions of theorems of Gödel, Tarski, and Church). *Suppose that $\mathcal{P}$ is hereditarily self-referential, self-referential with respect to T, and strongly self-referential with respect to P.*

(i) (a) *Assume that $\mathcal{P}$ is $\mathcal{P}$-axiomatizable. If $\mathcal{P}$ is consistent, then $\mathcal{P}$ is incomplete.*
(b) *Any Σ-soundly $\mathcal{P}$-axiomatizable consistent extension of $\mathcal{P}$ is incomplete.*
(c) *If $\mathcal{P}$ is sound, then any Σ-completely $\mathcal{P}$-axiomatizable sound $\mathcal{P}_Q$ is incomplete.*
In particular, any Σ-completely $\mathcal{P}$-axiomatizable sound extension of $\mathcal{P}$ is incomplete.

(ii) $\boldsymbol{T}$ *is not T-representable.*

(iii) (a) *Any consistent extension of $\mathcal{P}$ is $\mathcal{P}$-undecidable.*
(b) *Any extension of $\mathcal{P}$ is $\mathcal{P}$-recursively inseparable.*

It is worth noting that this theorem also yields the abstract versions of the three basic limitative results on arithmetical truth. Indeed, the abstract version of Tarski's theorem on the 'undefinability of truth' is explicitly contained in the theorem, while the remaining statements, according to which 'arithmetical truth is undecidable and cannot be axiomatized',[30] can be obtained as its immediate consequences. In fact, by definition, $\mathcal{P}_T$ is a sound, complete, and consistent logical system. Therefore, by Theorem 4 (i)(c) and (iii)(a), we have

COROLLARY. *Assume the conditions of the theorem above. If $\mathcal{P}$ is sound, then*

(i) $\mathcal{P}_T$ *is not Σ-completely $\mathcal{P}$-axiomatizable*
(ii) $\mathcal{P}_T$ *is $\mathcal{P}$-undecidable.*

Acknowledgement. I am grateful to the referee for helpful remarks concerning the previous version of the paper and to Ivor Grattan–Guinness for valuable comments on a draft of this paper. The research was supported by Hungarian NSF grant No. T30314 and T035192.

REFERENCES

[1] M. BAAZ and P. PUDLÁK, *Kreisel's conjecture for $L\exists_1$*, ***Arithmetic, proof theory, and computational complexity*** (P. Clote and J. Krajíček, editors), Claredon Press, Oxford, 1993, pp. 30–60.

[2] D. W. BARNES and J. M. MACK, ***An algebraic introduction to mathematical logic***, Springer-Verlag, New York, 1975.

[3] G. BOOLOS, ***The logic of provability***, Cambridge University Press, Cambridge, 1995.

[4] G. BOOLOS and R. C. JEFFREY, ***Computability and logic***, Cambridge University Press, Cambridge, 1992.

[5] C. C. CHANG and H. J. KEISLER, ***Model theory***, North Holland, Amsterdam, 1973.

[30] Cf. e.g., [4] pp. 176 and 179.

[6] J. N. FINDLAY, *Goedelian sentences: A non-numerical approach*, ***Mind***, vol. 51 (1942), pp. 259–265.

[7] K. GÖDEL, ***On formally undecidable propositions of Principia Mathematica and related systems***, Oliver and Boyd, Edinburgh, 1962, translated by B. Meltzer, introduction by R. B. Braithwaite.

[8] R. E. GRANDY, ***Advanced logic for applications***, D. Reidel, Dordrecht, 1979.

[9] K. GRELLING and L. NELSON, *Bemerkungen zu den Paradoxien von Russell und Burali–Forti*, ***Abhandlungen der Fries'schen Schule***, vol. 2 (1907–1908), pp. 300–334.

[10] E. MENDELSON, ***Introduction to mathematical logic***, D. Van Nostrand Company, Inc., Princeton, 1965.

[11] J. MYHILL, *A system which can define its own truth*, ***Fundamenta Mathematicae***, vol. 37 (1950), pp. 190–192.

[12] R. J. PARIKH, *Some results on the length of proofs*, ***Transactions of the American Mathematical Society***, vol. 177 (1973), pp. 29–36.

[13] W. V. QUINE, ***Mathematical logic***, Harvard University Press, Cambridge, 1951.

[14] ———, ***The Ways of Paradox and other essays***, Harvard University Press, Cambridge, 1979.

[15] G. RYLE, *Heterologicality*, ***Analysis***, vol. 11 (1951), no. 3.

[16] J. R. SHOENFIELD, ***Mathematical logic***, Addison–Wesley, Reading, Massachusetts, 1967.

[17] C. SMORYNSKI, *The incompleteness theorems*, ***Handbook of mathematical logic*** (J. Barwise, editor), North–Holland, Amsterdam, 1977, pp. 821–865.

[18] R. M. SMULLYAN, *Languages in which self reference is possible*, ***The Journal of Symbolic Logic***, vol. 22 (1957), no. 1, pp. 55–67.

[19] ———, ***Gödel's incompleteness theorems***, Oxford University Press, New York, 1992.

[20] A. TARSKI, *The semantic conception of truth*, ***Philosophy and Phenomenological Research***, vol. 4 (1944).

[21] ———, *Truth and proof*, ***Scientific American***, vol. 220 (1969), no. 6.

DEPARTMENT OF ALGEBRA
BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS
1111 STOCZEK U. 2. HÉP. 5. EM.
BUDAPEST, HUNGARY
E-mail: sereny@math.bme.hu