THE REVIEW OF SYMBOLIC LOGIC
Volume 7, Number 4, December 2014

SELF-REFERENCE IN ARITHMETIC I

VOLKER HALBACH
Oxford University
and

ALBERT VISSER
Utrecht University

> The net result is to substitute articulate hesitation for inarticulate certainty. Whether this result has any value is a question which I shall not consider.
>
> Russell (1940, p. 11)

Abstract. A Gödel sentence is often described as a sentence saying about itself that it is not provable, and a Henkin sentence as a sentence stating its own provability. We discuss what it could mean for a sentence of arithmetic to ascribe to itself a property such as provability or unprovability. The starting point will be the answer Kreisel gave to Henkin's problem. We describe how the properties of the supposedly self-referential sentences depend on the chosen coding, the formulae expressing the properties and the way a fixed points for the formulae are obtained. This paper is the first of two papers. In the present paper we focus on provability. In part II, we will consider other properties like Rosser provability and partial truth predicates.

§1. Introduction. 'We thus have a sentence before us that states its own unprovability.' This is how Gödel describes the kind of sentence that has come to bear his name.[1] Ever since, sentences constructed by Gödel's method have been described in lectures and logic textbooks as saying about themselves that they are not provable or as ascribing to themselves the property of being unprovable.

The only 'self–reference–like' feature of the Gödel sentence γ that is used in the usual modern proofs of Gödel's first incompleteness theorem is the derivability of the equivalence $\gamma \leftrightarrow \neg\mathrm{Bew}(\ulcorner\gamma\urcorner)$; in other words, the only feature needed is the fact that γ is a (provable) fixed point of $\neg\mathrm{Bew}(x)$.[2] But the fact that a sentence is a fixed point of a certain formula expressing a certain property does by no means guarantee that the sentence ascribes that property to itself, as we shall argue in what follows; and whether γ is also

Received: Dec. 23, 2013.

[1] The German original reads: 'Wir haben also einen Satz vor uns, der seine eigene Unbeweisbarkeit behauptet.' Of course Gödel (1931, p. 175) refers to provability in a specific system.

[2] In what follows, we will often talk about *fixed points* when we mean fixed points that can be shown to be fixed points in the relevant theory, that is, *provable* fixed points. As a referee correctly pointed out, Gödel did not prove the equivalence $\gamma \leftrightarrow \neg\mathrm{Bew}(\ulcorner\gamma\urcorner)$ in his original paper.

doi:10.1017/S1755020314000288

self-referential, or 'states its own unprovability' in any sense whatsoever is not relevant for proof of the first incompleteness theorem.[3]

Löb's theorem and related results are more intensional than Gödel's first incompleteness theorem or Tarski's theorem on the undefinability of truth, in the sense that the former are more sensitive to the choice of the formula expressing provability. The proof of Löb's theorem also relies only on the existence of a fixed point of a certain formula. Whether this fixed point also states something about itself is not relevant for the proof. In *this* respect at least, Löb's theorem, the second incompleteness theorem and so on are still *extensional*. Below we give examples of results which even fail to be extensional in this respect. They are intensional with respect to how fixed points are obtained. Since we take into account more possible sources of intensionality, in particular, the way how fixed points are constructed, our notion of extensionality differs from others found in the literature, for instance, Feferman's (1960).[4]

Generally, mathematical logicians have come to focus on questions and results that do not rely on the notion of self-reference; and thus a deeper analysis of the somewhat elusive notion of self-reference is not required. Instead they have become interested in sentences that can be proved to be fixed points of certain formulae. The development in metamathematics has led away from self-reference as a fundamental concept, because '[t]he notion of a sentence's expressing something about itself has not proven fruitful.'[5]

However, the notion of self-reference played a crucial in the questions that drive the development of metamathematics. Certain questions cannot been asked in a sensible way without using the notion of self-reference. For instance, it is natural to ask whether the sentences that state their own provability in a given system like Peano arithmetic are provable or not in that system. This question, usually known as Henkin's Problem, cannot be reformulated extensionally as the question about the status of fixed points of the canonical provability predicate, because not all fixed points of the standard provability state their own provability. The sentence $0=0$, for instance, is a trivial example of such a fixed point. The observation that the provable sentence $0=0$ is a fixed point of the provability predicate clearly does not settle the original problem whether sentences stating their own provability are provable or not. Therefore, the notion of *stating one's own provability* in the original question cannot be eliminated by the notion of being a fixed point. What one wants to know is whether the *self-referential* fixed points are provable or not. Questions of this kind led to the proof of results like Löb's theorem and other results that are no longer intensional with respect to how fixed points are constructed.

As we shall show, there are many further questions that essentially rely on the notion of self-reference. For instance, given a Π_1-predicate $\mathrm{Tr}_{\Pi_1}(x)$ expressing truth for Π_1-sentences, one can construct a sentence stating its own Π_1-truth. As any Π_1-sentence is a fixed point of $\mathrm{Tr}_{\Pi_1}(x)$, we see that fixed points can be provable, refutable and independent. The interesting question whether Π_1-sentences saying about themselves that they are Π_1-truth are provable, refutable or independent, and, in the latter case, whether they are true or not.

[3] We use the label *self-referential* in a very loose way. See the Note on Terminology on p. 673 below.

[4] See p. 680 for further remarks on Feferman's treatment of intensionality.

[5] This quote is taken from (Smoryński 1991, p.122). He gives an illuminating account of the historical development, in which his claim is substantiated.

In philosophical discussions, the notion of self-reference in metamathematics still assumes a prominent role. Here self-reference in formal languages is a topic in its own right. Some authors, among them Heck (2007) and Milne (2007), focus on self-reference in metamathematics and ask, for instance, whether a given mathematical sentence really states its own unprovability (or some other property). Even more prominently, the notion of self-reference is used in the analysis of the paradoxes, and it is often hoped that the metamathematical notion of self-reference in an arithmetized setting sheds some light also on self-reference in informal discourse. As our discussion will show, the question of whether it does is a delicate matter. A salient example is the discussion about the question whether Yablo's (1993) Paradox in its original informal version is self-referential. Authors apply metamathematical tools to answer the question whether the Yablo sentences are self-referential or not. But even for metamathematical sentences we lack a full analysis of self-reference; the use of some fixed-point property in itself isn't a sufficient criterion for self-reference. So we think that a better understanding of self-reference in metamathematics would facilitate this discussion.

> *A note on terminology*. In this paper we try to analyze the informal metamathematical predicate *ascribes property P to itself* as applied to sentences of arithmetic. If P is provability, for instance, we take this phrase to be equivalent to *states its own provability*, *says of itself that it is provable*, *predicates provability of itself* and so on. If it is clear which property P is meant, we also say that the sentence is *self-referential* without explicitly specifying the property. Moreover, the term *self-reference* has been used in different ways and may be misleading. Presumably *self-predication* cannot so easily be understood in a deviant way and therefore would be preferable. However, since the term *self-reference* has become common parlance, it will be used here and our observations in this paper are explicitly intended as a contribution towards the discussion about self-reference understood in the sense of self-predication.

§2. Intensionality. To begin with, we will look at the method that is usually thought to yield self-referential sentences of arithmetic. To obtain an arithmetical sentence that, according to common parlance, ascribes a certain property such as provability or unprovability to itself, one proceeds in three stages: First, the expressions of the language are coded in the numbers; second, a formula expressing the property is determined; finally, a self-referential sentence is constructed from this formula.

At each of the three stages, choices have to be made. They impinge on the properties of the sentences that supposedly ascribe some property to themselves. Often a result is loosely stated and the reader is left with the task of filling in the details of the construction of the self-referential sentence. If different choices actually do make a difference and the coding, the representing formula or the construction for self-reference have to be chosen carefully, we call a result intensional. Thus, corresponding to the three stages there are at least three sources of intensionality.

These three sources of intensionality are not independent of each other, and a choice made at an earlier stage will have effects on the availability of choices at later stages. Examples will be presented below.

Of course, the three sources of intensionality themselves depend on further parameters, in particular, on the language and the formal system. We are interested only in theories that are 'sufficiently strong'. To explicate the notion of sufficient strength, we introduce a theory that we call Basic. The language of Basic is the language of arithmetic extended with function symbols for all primitive recursive functions. The theory R, introduced by Tarski, Mostowski & Robinson (1953), contains the recursive axioms for addition and multiplication only in their numeralwise versions. The theory Basic is then R extended with all true identities of the form $t = \overline{n}$, where t is a closed term and $\overline{n}$ the numeral of n. These additional identities do not add power since they can be proved in a definitional extension of R, and Basic doesn't 'know' anything about the behaviour of primitive recursive functions outside the standard domain. In what follows we will focus on theories Σ that are sufficiently strong in the sense that they extend Basic and are formulated in the same language as Basic.

2.1. First source of intensionality: coding. The coding is the bridge between properties of numbers and properties of syntactic objects such as formulae and terms. The choice of coding is primary in the sense that the satisfaction of the other two tasks depends on it. It depends on the depends on the chosen coding whether an arithmetical formula expresses a property of certain syntactical objects, and thus also whether a formula ascribes a syntactical property to itself.

As an extreme example of the effect of coding on the last stage, the construction of a self-referential sentence from a given formula, in the appendix to the second part *Self-reference in arithmetic II* of this paper a 'reasonable' Gödel coding is constructed such that for each formula $\varphi(v)$ there is a unique number m such that $\varphi(\overline{m})$ has code m; whence $\varphi(\overline{m})$ is at least a fixed point of $\varphi(v)$, and presumably even ascribes to itself the property expressed by $\varphi(v)$.[6] So by choosing the coding in a clever way, the entire last stage, that is, the construction of a self-referential statement from φ, can completely bypassed, as diagonalization is built into the coding schema for all formulae. Whether this kind of coding leads to truly self-referential statements is another delicate matter.

For some results in the second part of this paper, other assumptions about the coding are required. Arithmetical truth-tellers, that is, sentences stating their own truth via partial truth predicates, provide an example. As we shall show, these sentences are provable under certain assumption about the coding schema, and they can be refutable under others.

Whether a coding schema is appropriate, depends also on the theory. In the context of weak theories certain standard and straightforward coding schemata are less appropriate than other more sophisticated schemata. In the case of weak theories, Gödel's or Kleene's methods of coding, for instance, are not sufficiently effective. In cases of weak theories we assume that a suitable coding schema is employed.

However, for most of the time we will work in sufficiently strong theories and most parts of the paper are fairly stable with respect to the chosen coding schema. We follow the usual practice and assume some 'standard' coding schema without setting out the details, unless we explicitly introduce further assumptions.

2.2. Second source of intensionality: expression of properties. What does it mean for a formula of arithmetic to express a certain syntactical property? This question is notoriously difficult to answer and has been investigated by many logicians, among them Feferman (1960), Auerbach (1985) and Franks (2009).

[6] In the appendix to the second part of this paper the numeral $\overline{m}$ of m is defined as the so-called *efficient* numeral of m not as the usual numeral. But this is only a technically convenient convention. In the main body of the papers the numerals are the usual ones.

Here we do not attempt to give a full answer, but we will state *some* assumptions that give an indication of the shape an answer could take. First, we assume that every formula of arithmetic with one free variable *does* express an arithmetical property, which in turn may relate to a property of sentences or other syntactic objects. Secondly, different formulae, even when they are not provably equivalent in some designated theory, may express the same property. Thirdly, even provably equivalent formulae may fail to express the same property.

Different kinds of criteria have been used to argue that an arithmetical formula expresses a specific syntactic property. The purely extensional notion of representation, introduced by Kreisel (1953), has occasionally been used as a first approximation towards an answer to the question of what it means for a formula to express a certain property.[7]

KREISEL'S CONDITION. *A formula $\varphi(x)$ is said to express a property P (in a system Σ) if and only if the following condition is met for all numbers n:*

$$\Sigma \vdash \varphi(\overline{n}) \textit{ iff } n \textit{ has property } P.$$

As usual, $\overline{n}$ is the numeral associated with the number n.

In metamathematics Kreisel's Condition became the formal notion of weak representability:[8] A formula $\varphi(x)$ is said to *weakly represent* a set S of numbers if and only if the following equivalence holds:

$$\Sigma \vdash \varphi(\overline{n}) \text{ iff } n \in S.^{9}$$

For the property of being a Σ-provable sentence such a formula $\varphi(x)$ exists, as long as Σ is consistent, recursively enumerable, and sufficiently strong – even if the theory Σ is disturbingly unsound, as is shown in (Visser 2014).

If one adopts an intensional stance on syntactic properties – as we do –, Kreisel's Condition can hardly yield an adequate explication of the notion of expressing a syntactic property. Kreisel's Condition cannot distinguish between distinct properties that are weakly represented by the same formula. For instance, by the standards of Kreisel's Condition, the same formulae express canonical provability and Rosser provability over a sound theory.

The applicability of Kreisel's Condition is also very limited. For notions such as Π_1-truth there is no weakly representing formula, even though most logicians believe that there is an arithmetical formula expressing Π_1-truth. But not only would Π_1-truth be not expressible, if Kreisel's Condition were a full adequate analysis of the expression of a property; a canonical Π_1-truth predicate that provably only applies to Π_1-sentences would not express Π_1-truth; it would rather express the property of being a Σ-provable

[7] Kreisel (1953) didn't state his condition for arbitrary properties, but rather only for provability:

> A formula $\mathfrak{P}(a)$ is said to express provability in Σ if it satisfies the following condition: for numerals a, $\mathfrak{P}(a)$ can be proved in Σ if and only if the formula with number a can be proved in Σ.

[8] Feferman (1960) introduced and used the term 'to numerate' for 'to weakly represent'.

[9] In contrast, a formula $\varphi(x)$ *strongly represents* a set S if and only if $\varphi(x)$ weakly represents S and, in addition, the equivalence

$$\Sigma \vdash \neg\varphi(\overline{n}) \text{ iff } n \notin S$$

obtains.

Π_1-sentence, because, for any Π_1-sentence φ, $\mathrm{Tr}_{\Pi_1}(\ulcorner\varphi\urcorner)$ is Σ-provable if and only if φ is Σ-provable, as $\mathrm{Tr}_{\Pi_1}(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$ holds for all Π_1-sentences. Hence, for such more complicated properties like Π_1-truth other criteria for expressing a property have to be used instead of Kreisel's Condition.

Kreisel's Condition makes the notion of a formula's expressing a property relative to a theory. If the theory is unsound, then the criterion may yield unwanted consequences. For instance, relative to the theory $\mathsf{PA} + \neg\mathrm{Con}_{\mathsf{PA}}$, which is consistent by Gödel's Second Incompleteness theorem, the canonical provability predicate does not express provability in PA.

At any rate, Kreisel's Condition is neither sufficient nor necessary as a criterion of expressing a syntactic property. As an alternative or additional criterion, which we call the *conditions* or *meaning postulates* criterion, we could say that formula expresses a certain syntactic property if, verifiably within the theory, the formula satisfies certain conditions or meaning postulates.

In the case of partial truth, what is often meant by saying that Π_1-truth is expressible or definable in an arithmetical system Σ is the observation that there is a formula $\varphi(x)$ such that the equivalences $\varphi(\ulcorner\psi\urcorner) \leftrightarrow \psi$ are provable in Σ for all Π_1-sentences ψ.[10] These equivalences are then the meaning postulates that have to be provably satisfied.

So, the *meaning postulates* or *conditions* criterion would take the following form: A formula of arithmetic expresses a property P of syntactic objects if and only if the formula satisfies certain principles or axioms associated with P, relative to a coding schema.

In the case of provability, Löb's derivability conditions have been used as meaning postulates for the property of provability, although in themselves they will hardly suffice, because they are satisfied by the formula $x = x$ and, even if combined with Kreisel's Condition, they still admit formulae as presumed provability predicates that can hardly be accepted as genuine provability predicates, as is shown in (Visser 2014).

Moreover, it is far from clear how the meaning postulates can be associated with the property. In fact, in many cases the postulates were only discovered once a formula already taken to express a certain property had been analyzed in detail. In the case of Löb's conditions, Gödel first defined a formula that was generally thought of as expressing provability; the development of the derivability conditions started in Gödel's paper, was continued by Hilbert and Bernays (1939) and reached its completion in the work of Löb (1955).

Like Kreisel's Condition, the meaning postulates criterion is relative to a theory. However, the two criteria are not always compatible: In certain unsound theories the canonical provability predicate for PA expresses provability on the meaning-postulate account, but not according to Kreisel's Condition, as remarked above; and vice versa, in any consistent theory with enough coding we have a predicate that expresses provability according to Kreisel's Condition, but fails to do so according to the Löb conditions viewed as meaning postulates.

Still further criteria may not fail so obviously, but are vague or unclear. This is the case with what we would like to call the *resemblance* criterion. A syntactic property is usually given by an informal metamathematical description of that property. Often, logicians expect a formula to express a syntactic property if the formula structurally resembles the

[10] There are further candidates for meaning postulates for truth. In particular, one could also demand that a truth predicate has to satisfy the *compositional* axioms for truth, that is, axioms stating that truth commutes with connectives and quantifiers on the relevant class of sentences.

metamathematical description of the property. However, it's difficult to explain what such a structural resemblance could consist in.

An informal metamathematical description of provability in a specific system Σ, for instance, will invoke other properties, like the property of being a formula of the formal arithmetical language. The metamathematical definition of arithmetical formulae involves a definition of a string, sequence or tree of symbols of the arithmetical language.

'High level' properties such as provability can structurally resemble an arithmetical formula in a recognizable way. But the definitions of these higher level properties depend on the definitions of 'lower level' properties such as being a string of symbols or, even lower, of being a code of a certain symbol.

In the case of lower level properties there is much less of a recognizable resemblance between the metamathetical definition and a corresponding arithmetical formula. At the very basic level, there is no resemblance but only stipulation: The property of being the negation symbol will hardly bear any resemblance with the corresponding arithmetical formula, which will just say that x is the numerical code of the negation symbol $\neg$. Which number codes $\neg$ is a matter of stipulation. Therefore, an arithmetical formula resembles a description of a syntactic property only relative to a coding schema. But also at the next higher level, it's hard to see how the definitions of sequences of symbols can resemble arithmetical formulae. There are various ways to code sequences of numbers and again the choice of the method seems to be to a great extent a matter of stipulation. To add to the complication, it is not even clear whether formulae are best understood as sequences of symbols, or parsing trees or still something else. Only once all these choices have been made, it makes sense to invoke the resemblance criterion.

One way to sidestep the sensitivity to all these choices and stipulations would be to employ a syntax theory that directly describes syntactic objects. However, we wonder whether such an approach would not also necessitate arbitrary choices in the construction of the formulae expressing syntactic properties such as the property of being a formula or provable in a fixed deductive system. Within the theory of syntax we would still have to decide whether formulae are understood as sequences, trees or something else. To apply the resemblance criterion one needs to assume that these decisions and choices have been made and fixed in some way.

But even after fixing these choices, the resemblance of the required kind remains an elusive. In practice we seem to recognize the relevant resemblances of an arithmetical formula with the informal metamathematical description of a property in many specific cases; but we lack a general account of resemblance. Usually the resemblance criterion seems to be applied in claims to the effect that a certain formula expresses a property *in a natural way*.

Despite of all the defects and problems of the resemblance criterion, it seems almost indispensable for some purposes. When one tries to explain what a 'canonical' or 'natural' provability predicate is (given a coding schema for symbols and their sequences), it seems hard to avoid an appeal to 'resemblance'. In what follows, *canonical* provability predicates will be assumed to resemble in their 'salient' features the definition of the informal provability predicate in metamathematics.

In what follows all three types of criteria mentioned will be applied and further discussed.

2.3. Third source of intensionality: self-reference. As laid out above, the construction of a sentence that ascribes to itself a certain property P usually proceeds in three stages: In the first step, a coding of the syntactic objects is fixed. Then a formula $\varphi(x)$

is picked that expresses the property P relative to the chosen coding. In the third and final stage, a sentence is constructed that, in common parlance, ascribes to itself the property P via the formula $\varphi(x)$. Usually Gödel's diagonal construction or a variant thereof is employed to this end.

In this paper we assume that there are paradigmatic cases of self-reference, usually established via Gödel's diagonal method. But even for the canonical diagonalization method it is not universally accepted that it establishes self-reference in the intended way. For instance, Heck (2007) and Milne (2007) have raised some worries. We share some of their doubts and do not assume that all so-called Gödel sentences found in textbooks really say of themselves that they are not provable. We shall also look at sentences that have not been obtained via Gödel's classical construction, but may neverless be thought to be self-referential. For example, in the appendix to the second part of this paper, a coding is sketched in which no third stage is needed, because diagonalization is already built into the coding. Kreisel, in his initial answer to Henkin's question, presented a sentence that may be thought to be self-referential but that hasn't been obtained in the usual way from a formula $\varphi(x)$ expressing provability.

Assume that a coding has been fixed and the property P is expressed by the formula $\varphi(x)$. Then, if a sentence γ says about itself that it has property P via $\varphi(x)$, γ must be a fixed point of $\varphi(x)$, that is γ must be equivalent to $\varphi(\ulcorner\gamma\urcorner)$ (in a sense to be specified). If γ were not equivalent to $\varphi(\ulcorner\gamma\urcorner)$, γ would not ascribe the property expressed by $\varphi(x)$ to γ, that is, to itself. In other words, the fixed-point property is a necessary condition for self-reference.

If a result under consideration doesn't depend on the choice of the fixed point, that is, it remains true whichever fixed point of $\varphi(x)$ we use, we call the result extensional (with respect to the third source of intensionality); if it does depend on the choice of the fixed point, it is intensional.[11]

Now one may wonder to what extent the fixed-point requirement narrows the choice of the set of sentences that may be said to assign to themselves the property P. That is, whether for a given formula expressing P there are many fixed points.

It's not hard to see that, for any given formula $\varphi(x)$, there are always many fixed points to choose from. The set of formulae that are fixed points of a given formula according to the standard model is not elementarily definable; and the set of provable fixed points is only recursively enumerable, but not decidable. We prove these claims in the following two observations.

OBSERVATION 2.1. *For any formula $\varphi(x)$, there is no formula $\chi(x)$ such that for all formulae ψ the following is true in the standard model:*

$$\chi(\ulcorner\psi\urcorner) \leftrightarrow \big(\varphi(\ulcorner\psi\urcorner) \leftrightarrow \psi\big) \tag{1}$$

The proof is a generalization of Tarski's theorem on the undefinability of truth: Truth is a predicate with a very simple set of fixed points, because *all* sentences are fixed points. Hence, according to the observation, no formula of Σ can express truth.

[11] Our notion of extensionality should be distinguished from other closely related notions of extensionality, such as uniqueness of fixed points, and from the following definition of extensionality: If $\gamma \leftrightarrow \gamma'$ is provable, then also $\varphi(\ulcorner\gamma\urcorner) \leftrightarrow \varphi(\ulcorner\gamma'\urcorner)$ is provable.

Proof. Assume there is such a formula $\chi(x)$. Then, by propositional logic, (1) would imply

$$\big(\chi(\ulcorner\psi\urcorner) \leftrightarrow \varphi(\ulcorner\psi\urcorner)\big) \leftrightarrow \psi,$$

and therefore $\chi(x) \leftrightarrow \varphi(x)$ would be a truth predicate, whose existence contradicts Tarski's theorem on the undefinability of truth. □

So, for any formula $\varphi(x)$, the set of its fixed points is not arithmetically definable. Next we show that, for a sufficiently strong system Σ, the set of sentences such that $\Sigma \vdash \varphi(\ulcorner\zeta\urcorner) \leftrightarrow \zeta$ cannot be recursive.

OBSERVATION 2.2. *Assume Σ extends* Basic*, as defined on p. 674. For any given formula $\varphi(x)$, the set of all provable fixed points, that is, the set of all sentences ψ with $\Sigma \vdash \varphi(\ulcorner\psi\urcorner) \leftrightarrow \psi$, is not recursive.*[12]

The proof is reminiscent of Curry's paradox and McGee's (1992) trick. The following proof yields the somewhat stronger conclusion that the set of provable fixed points of φ is complete recursively enumerable.

Proof. There is a primitive recursive function that, applied to a formula ψ, gives a formula γ_ψ with the following property:

$$\Sigma \vdash \gamma_\psi \leftrightarrow (\varphi(\ulcorner\gamma_\psi\urcorner) \leftrightarrow \psi)$$

By propositional logic, this implies the following claim:

$$\Sigma \vdash \psi \leftrightarrow (\varphi(\ulcorner\gamma_\psi\urcorner) \leftrightarrow \gamma_\psi)$$

Thus a sentence ψ is provable iff γ_ψ is a provable fixed point of $\varphi(x)$. Hence, if the set of provable fixed points of $\varphi(x)$ were decidable, the set of Σ-provable sentences would be decidable. □

Of course, this leaves open the possibility that the provable fixed points of a given formula are all provably equivalent in a certain theory. This is the case for provability predicates satisfying the Löb derivability condition. For other formulae such as Rosser provability predicates or partial truth predicates like the truth predicate for Σ_1-sentences, this is not the case.[13]

Any sentence that says about itself that it possesses the property expressed by $\varphi(x)$ must be a fixed point of that formula; but in order for a sentence to be truly self-referential further conditions will have to be met. Such a condition was implicitly used by Henkin and Kreisel in an exchange we are going to describe in the next section.

This completes the description of the three stages in which a sentence that ascribes a property to itself is obtained, if the usual textbook procedures are followed. None of the three stages yields a unique output: There are many different Gödel codings; given one such coding, every property that can be expressed at all by some formula can be expressed by many different formulae; and given a coding and a property expressed by a formula, different fixed points can be constructed such that their fixed-point property is provable in the theory in question.

[12] See Blanck (2011) for a strengthening of this result and a detailed investigation of sets of provable fixed points.

[13] Rosser provability predicates or partial truth predicates will be treated in Section 6 of the second part of this paper. See also Solovay's (1985) analysis of Hofstadter's explicit Henkin sentences.

Different choices of the coding, the expressing formula and the method of obtaining presumed self-reference can each yield sentences with different metalogical properties. Therefore, the three stages correspond very roughly with three different dimensions of intensionality.

The problems of intensionality in the first two dimensions are fairly well studied. For example, Feferman (1960, p. 35) classified the applications of the method of arithmetization 'as being *extensional* if essentially only numerically correct definitions are involved, or *intensional* if the definitions must more fully *express* the notions involved [...]'. This corresponds to our first two stages.

The main point of our paper is to show that self-reference, too, has intensional aspects.[14] This means that at least for certain natural questions, it does not only matter whether a sentence is a fixed point of a formula expressing the property in question, but also whether the sentence says about itself (in a sense to be discussed) that it has the property expressed by the formula.

§3. Henkin's problem and Kreisel's answer. At least at one point in the development of metamathematics, a question essentially involving the notion of self-reference initiated a development that led to a fundamentally new and important result, namely Löb's theorem. Ironically, the solution of the problem, which was found only after some detours, implied that the notion of self-reference is actually irrelevant to the problem. Nevertheless something is to be learnt from following the dead ends that were reached before the problem was solved by Löb. So we will begin with the situation in the early 1950s.

If a Gödel sentence is a sentence that states its own unprovability, it is natural to consider also sentences that state their own *provability*, and to investigate whether they are independent like the Gödel sentence or provable or refutable.

The problem whether a sentence stating its own provability is provable or not is intensional with respect to all three sources of intensionality, in particular, also with respect to the third: To ask the question whether a sentence stating its own provability is provable or not, it does not suffice to ask about the status of the fixed points of a formula expressing provability, that is, of a formula τ such that $\tau \leftrightarrow \mathrm{Bew}(\ulcorner\tau\urcorner)$ is provable. Clearly, $1=1$ is a provable fixed point of $\mathrm{Bew}(x)$, if $\mathrm{Bew}(x)$ weakly represents provability, because both $1{=}1$ and thus $\mathrm{Bew}(\ulcorner 1=1\urcorner)$, and thus also the corresponding biconditional, will be provable. But $1{=}1$ does not say of itself that it's provable, unless a very peculiar coding is used. The question is about a sentence that *says of itself that it's provable*, not just about arbitrary fixed points of the provability predicate. Thus the notion of self-reference is required to state the problem and cannot be substituted with a question about the provability or refutability of arbitrary fixed points of the provability predicate.

Sentences stating their own provability are nowadays known as *Henkin sentences*. Henkin (1952) himself did not pose his question directly in terms of self-reference and used a different formulation for ruling out 'accidental' fixed points such as $1=1$:

> If Σ is any standard formal system adequate for recursive number theory, a formula (having a certain integer q as its Gödel number) can be constructed which expresses the proposition that the formula with Gödel number q is provable in Σ. Is this formula provable or independent in Σ?

[14] Cf. Skyrms (1984).

We think that Henkin's formulation is an attempt to ask whether a formula stating its own provability is provable or not. Henkin did not ask whether a sentence stating its own provability is provable or not. He avoided a direct appeal to self-reference; but his formulation of the question still leaves some space for interpretation. Presumably, Henkin had in mind Gödel's construction for obtaining the said sentence, but he didn't explicitly appeal to it.

Nowadays Löb's theorem is seen by most logicians – including Kreisel – as the only pertinent answer to Henkin's question. However, we think a second look at Kreisel's first attempt (1953) to answer Henkin's question is worth while. It lets us see more clearly if Henkin's formulation is an adequate rendering of the question whether a sentence stating its own provability is provable or independent. In his paper Kreisel summarized his reply to Henkin in the following way:

> We shall show below that the answer to Henkin's question depends on which formula is used to 'express' the notion of *provability in* Σ.

Kreisel proposed understanding 'to express' in the sense of Kreisel's Condition, that is, as weak representability. Kreisel constructed two sentences that are both supposed to satisfy Henkin's condition; one of them is provable, the other refutable:

KREISEL'S OBSERVATION. *Let Σ be a consistent theory that extends* Basic.[15] *Then the following holds:*

a) *There is a formula* $\mathrm{Bew_I}(x)$ *and a term* t_1 *such that the following three conditions are satisfied:*

   (i) $\mathrm{Bew_I}$ *weakly represents provability in* Σ.
   (ii) $\Sigma \vdash t_1 = \ulcorner \mathrm{Bew_I}(t_1) \urcorner$
   (iii) $\Sigma \vdash \mathrm{Bew_I}(t_1)$

b) *Similarly, there is a formula* $\mathrm{Bew_{II}}(x)$ *and a term* t_2 *such that*

   (i) $\mathrm{Bew_{II}}$ *weakly represents provability in* Σ.
   (ii) $\Sigma \vdash t_2 = \ulcorner \mathrm{Bew_{II}}(t_2) \urcorner$
   (iii) $\Sigma \vdash \neg \mathrm{Bew_{II}}(t_2)$

The examples employed by Kreisel in the proof are of some interest. In particular, the example for $\mathrm{Bew_I}(t_1)$ foreshadows Kreisel's proof of Löb's theorem in (Kreisel & Takeuti 1974), as was pointed out by Smoryński (1991). Henkin suggested simpler examples which are mentioned by Kreisel (1953) in footnotes. We will use Henkin's examples and refer the reader to Smoryński's paper for an exposition of Kreisel's original examples.

Proof. We start with a proof for part (b). Fix some predicate $\mathrm{Bew}(x)$ that weakly represents Σ-provability in Σ. In case Σ is Σ_1-sound, a standard arithmetization of provability will do. In the unsound case, one uses the theorem that any recursively enumerable set is weakly representable in a consistent recursively enumerable extension of the Tarski–Mostowski–Robinson theory R. This is a direct consequence of the Friedman–Goldfarb–Harrington Theorem.[16] Using the canonical diagonal construction (or any other method),

[15] Kreisel stipulated that the theory be Σ_1-sound, but that demand is superfluous.

[16] See, for instance, (Visser 2005) for a discussion.

one obtains a term t_2 satisfying the following condition

$$\Sigma \vdash t_2 = \ulcorner t_2 \neq t_2 \wedge \mathrm{Bew}(t_2)\urcorner \tag{2}$$

and defines $\mathrm{Bew}_{\mathrm{II}}(x)$ as

$$x \neq t_2 \wedge \mathrm{Bew}(x).$$

Condition b(ii), that is, $\Sigma \vdash t_2 = \ulcorner \mathrm{Bew}_{\mathrm{II}}(t_2)\urcorner$ is then obviously satisfied by the choice (2) of t_2. Since Σ refutes $t_2 \neq t_2 \wedge \mathrm{Bew}(t_2)$, item b(iii) is satisfied as well.

It remains to verify b(i), which is the claim that $\mathrm{Bew}_{\mathrm{II}}(x)$ weakly represents Σ-provability. In other words we must establish the following equivalence for all formulae φ:

$$\Sigma \vdash \varphi \ \text{ iff } \ \Sigma \vdash \mathrm{Bew}_{\mathrm{II}}(\ulcorner\varphi\urcorner).$$

If φ is different from $t_2 \neq t_2 \wedge \mathrm{Bew}(t_2)$ and whence $\Sigma \vdash \ulcorner\varphi\urcorner \neq t_2$, this is obvious from the definition of $\mathrm{Bew}_{\mathrm{II}}(x)$, using the fact that Bew weakly represents provability in Σ. In the other case the left-hand side of the equivalence is refutable, and so is the right-hand side by (2). This concludes the proof of part (b) of Kreisel's Observation.

We turn to case (a). If we assume that our theory is Σ_1-sound and sufficiently strong (e.g., if it extends the arithmetical version of Buss' theory S^1_2), then the canonical provability predicate can be used as $\mathrm{Bew}_{\mathrm{I}}(x)$ and t_1 can be obtained in any way, including the usual Gödel diagonal construction. Claim a(iii) follows then by Löb's theorem.[17]

As Löb's theorem was then not known, Henkin and Kreisel had to use a different construction.[18] Henkin suggested the following construction. He picked a term t_1 such that

$$\Sigma \vdash t_1 = \ulcorner t_1 = t_1 \vee \mathrm{Bew}(t_1)\urcorner$$

and defines $\mathrm{Bew}_{\mathrm{I}}(x)$ as

$$x = t_1 \vee \mathrm{Bew}(x). \qquad \square$$

§4. The Kreisel–Henkin Criterion for self-reference. Before considering the question whether Kreisel's Observation has any bearing on the question whether a sentence stating its own provability is provable or not, we investigate whether Kreisel really answered Henkin's question, which is not explicitly formulated as a question about sentences stating their own provability.

We think that, if $\mathrm{Bew}_{\mathrm{I}}(x)$ does expresses provability, then $\mathrm{Bew}_{\mathrm{I}}(t_1)$ is a formula (with Gödel number q) 'which expresses the proposition that the formula with Gödel number q is provable in Σ'. An analogous remark applies to $\mathrm{Bew}_{\mathrm{II}}(t_2)$. Consequently Kreisel would have answered Henkin's question.

Henkin himself, however, didn't accept Kreisel's answer. In his review (1954) of Kreisel's (1953) answer, he rejected Kreisel's assumption that a formula satisfying Kreisel's Condition, that is, a formula weakly representing provability, always expresses provability, claiming that 'it seems fair to say that in one sense, at least, neither formula [that is, neither $\mathrm{Bew}_{\mathrm{I}}(a)$ nor $\mathrm{Bew}_{\mathrm{II}}(a)$] expresses the propositional function *a is provable*.' Thus Henkin's dismissal of Kreisel's answer is based on a rejection of Kreisel's Condition applied to

[17] See (Löb 1955) and, for a modern account, (Boolos 1993).

[18] Note that the Kreisel–Henkin construction works for some very weak theories, too, where it is not clear that we have Löb's theorem.

provability; Henkin doesn't believe that every formula weakly representing provability already expresses provability.

Kreisel (1953) used contrived formulae to express provability; in his reply Henkin suggested simpler, but still contrived formulae that can be used to prove Kreisel's Observation. In the end Henkin (1954) rejected all these formulae as provability predicates. This rejection is well motivated. The focus on more canonical provability predicates led on to Löb's theorem. This part of the story is well known.[19]

But we would like to ask whether Kreisel's Observation can shed any light on the possible properties of Henkin sentences, if Kreisel's Condition is accepted. That is, we wonder whether, if $\mathrm{Bew}_{\mathrm{I}}(x)$ and $\mathrm{Bew}_{\mathrm{II}}(x)$ are, intuition notwithstanding, assumed to express provability, $\mathrm{Bew}_{\mathrm{I}}(t_1)$ and $\mathrm{Bew}_{\mathrm{II}}(t_2)$ are Henkin sentences, that is, sentences stating their own provability.

It is not obvious that $\mathrm{Bew}_{\mathrm{I}}(t_1)$ and $\mathrm{Bew}_{\mathrm{II}}(t_2)$ say of themselves that they are provable, even if Kreisel's Condition is accepted. Kreisel had not only used noncanonical provability predicates; he had also employed a noncanonical way of obtaining the fixed points of his provability predicates.[20] In particular, he had *not* applied Gödel's construction to his provability predicates to obtain their fixed points. If he had applied the usual diagonal construction to $\mathrm{Bew}_{\mathrm{II}}(x)$, he would have obtained a provable sentence rather than the desired refutable sentence $\mathrm{Bew}_{\mathrm{II}}(t_2)$. This is the content of our Observation 4.1 below.

When Henkin posed his question, he presumably had the canonical provability predicate and the standard Gödel fixed-point construction in mind. The evidence for this conjecture is that he used the singular 'this formula' when he asked whether the formula is provable; he didn't ask whether *any* formula satisfying the description is provable. That Kreisel hadn't used the canonical provability predicate was sufficient for rejecting his answer as besides the point. But it is surprising that neither Henkin nor Kreisel really remarked on the way the fixed points of the provability predicates in Kreisel's Observation are obtained. Both presumably tried, on the one hand, to avoid murky formulations such as 'states its own provability'; on the other hand, they didn't intend to distract from the intuitive appeal of the question by referring specifically to a fixed point that is obtained by the Gödel method, because Gödel's construction is a trick after all, a means to an end.

Whatever the motives were, Henkin and Kreisel merely required that the fixed point be of the form $\varphi(t)$, where $\varphi(x)$ is a formula expressing provability and $t = \ulcorner\varphi(t)\urcorner$ is provable.[21]

We are interested in the question whether Henkin's way of stating the question, if read in Kreisel's way, is really a question about a sentence stating its own provability. If the question is answered in the affirmative, then Henkin and Kreisel just used a mathematically precise rendering of self-reference. This way of turning the notion of self-reference into a mathematically precise notion can then be captured in the following criterion for self-reference:

[19] Of course, it turned out that only some global properties of canonical predicates not shared by Kreisel's and Henkin's examples are relevant. That does not contradict the fact that these global properties were motivated by canonical examples of provability predicates. The canonical predicates were the ladder that could be thrown away.

[20] Kreisel seems to hint at this feature of his construction at the end of his paper.

[21] Contra Smoryński (1991, p. 114) we don't think it was Kreisel who 'relaxed the stricture that φ be *constructed* to express its own provability', as Smoryński puts it, but that this relaxation can be traced back to Henkin's question. It is doubtful, however, whether Henkin intended this relaxation, as we remarked above.

KREISEL–HENKIN CRITERION FOR SELF-REFERENCE. *Let a formula $\varphi(x)$ expressing a certain property P in Σ and a closed term t be given. Then the formula $\varphi(t)$* says of itself that it has property *P iff t has (the code of) $\varphi(t)$ as its value.*

We would like to use the names of Kreisel and Henkin for this criterion, even though neither Henkin nor Kreisel explicitly put forward such a criterion for self-reference, self-predication or 'saying about itself' in exactly this way.

As we have formulated it, the criterion applies only to formulae of the form $\varphi(t)$, where $\varphi(x)$ expresses the property in question. Since the criterion doesn't state anything about sentences of a different form, for instance, quantified sentences, the criterion can merely function as a *sufficient* criterion for self-reference: A sentence says of itself that it has property P, if it is of the form $\varphi(t)$ and $t = \ulcorner\varphi(t)\urcorner$. Therefore, if t is obtained from some formula $\varphi(x)$ in the usual way by the Gödel construction, then $\varphi(t)$ says of itself that it has property P.

The condition 't has (the code of) $\varphi(t)$ as its value' leaves open whether $t = \ulcorner\varphi(t)\urcorner$ must be provable in Σ or merely true (in the standard model). But since equations of this kind are decidable in the theories under consideration, we don't have to commit ourselves to a particular stance on this. In other languages, one will have to make a decision.

The observation that for any formula $\varphi(x)$ there is a term t such that $t = \ulcorner\varphi(t)\urcorner$ is known as the *Strong Diagonal Lemma*. We don't know who formulated it first. Heck (2007) surmises that it made its first appearance in (Jeroslow 1973), but it is sufficiently clear from Kreisel's (1953) answer to Henkin's problem that Kreisel was fully aware of it, as he uses it in his construction.

It would have been more precise to call the Kreisel–Henkin Criterion a criterion for *direct* self-reference. If $\varphi(x)$ is a (partial) truth predicate, a sentence $\varphi(t)$ says that the value of the term t is (the code of) a true sentence $\psi(s)$, and then the value of s may be (a code of) $\varphi(t)$ again. In at least some cases of this kind, one would want to say that $\varphi(t)$ *indirectly* ascribes to itself the property expressed by $\psi(x)$. We do not want to delve deeper into the intricacies of indirect self-reference and therefore explicitly state that here in this paper self-reference is always understood as *direct* self-reference. Similar remarks apply to related notions.

Occasionally in semantic analyses of the paradoxes, terms designating sentences in which they occur are obtained in a trivial way: To mimic the effect of the diagonal lemma, for each formula $\varphi(x)$ a special constant c is added and interpreted in such a way that it has $\varphi(c)$ (or its code) as its value in the designated model. This semantic approach also yields self-reference in the sense of the Kreisel–Henkin Criterion if the term *value* in the criterion is understood in a semantic way as the value of c in the model.

If a formula $\varphi(t)$ satisfying the Kreisel–Henkin Criterion is obtained via the canonical diagonal lemma, the term t will be complex, in contrast to the constants in the construction just outlined.[22] Under most textbook coding schemata the function symbols for zero, successor, addition, and multiplication will not suffice to construct the term t. However, the term t can be a mere numeral if the coding is chosen in an appropriate way. Under the coding that is constructed in the appendix to the second part of this paper, there is for every

[22] By the *canonical diagonal lemma* we mean the straightforward construction of such a term t, based on Gödel's idea. We do not want to imply that the construction of a Gödel sentence proceeds in this way in most textbooks. In fact, we surmise that authors in most textbooks, as indeed Gödel himself did, prove the first incompleteness theorem without such a term in the language.

formula $\varphi(x)$ an n such that $\varphi(\overline{n})$ has n as its code.[23] Moreover, the elementary properties of the coding can be verified in a sufficiently strong theory Σ (which is not possible on the semantic approach involving the constants c).

For any given property expressed by a formula $\varphi(x)$ there are infinitely many sentences saying about themselves that they have the property, at least according to the Kreisel–Henkin Criterion. For instance, there are trivial and not very exciting variations on Gödel's diagonal construction. Rather than formally substituting numerals (containing the symbol for 0 and successor symbols) terms of the form $1 + \cdots + 1$ can be substituted.

More interestingly, there can be sentences $\varphi(t_1)$ and $\varphi(t_2)$ with highly different properties that both ascribe to themselves the property expressed by $\varphi(x)$ according to the Kreisel–Henkin Criterion. In fact, the formula $\mathrm{Bew}_{\mathrm{II}}(x)$ from Kreisel's Observation can be used as an example:

OBSERVATION 4.1. *Suppose* Σ *is a consistent theory of arithmetic that extends both* Basic *and* S^1_2 *– in other words,* Σ *should be strong enough to verify Löb's theorem.*[24] *There is a formula* $\mathrm{Bew}_{\mathrm{II}}(x)$ *weakly representing provability in* Σ *and terms* t_2 *and* t *such that both sentences* $\mathrm{Bew}_{\mathrm{II}}(t_2)$ *and* $\mathrm{Bew}_{\mathrm{II}}(t)$ *satisfy the Kreisel–Henkin Criterion and* $\mathrm{Bew}_{\mathrm{II}}(t)$ *is provable while* $\mathrm{Bew}_{\mathrm{II}}(t_2)$ *is refutable.*[25]

Proof. The formula $\mathrm{Bew}_{\mathrm{II}}(x)$ is the formula from Kreisel's Observation, built from the canonical provability predicate $\mathrm{Bew}(x)$ in case Σ is Σ_1-sound or from a suitable other predicate satisfying the Kreisel Condition, that is, weak representability, otherwise. The refutability of $\mathrm{Bew}_{\mathrm{II}}(t_2)$ is established in the proof of Kreisel's Observation.

The term t is the one obtained by the usual diagonal construction. With this choice of t, $t \neq t_2$ is provable under reasonable assumptions about the coding. Moreover, if $t = \ulcorner\varphi\urcorner$, we have $\Sigma \vdash \varphi \leftrightarrow \mathrm{Bew}(\ulcorner\varphi\urcorner)$. So, $\Sigma \vdash \varphi$ follows by Löb's theorem. □

The formula $\mathrm{Bew}_{\mathrm{II}}(x)$ expresses provability according to Kreisel's Condition; and both sentences $\mathrm{Bew}_{\mathrm{I}}(t_1)$ and $\mathrm{Bew}_{\mathrm{II}}(t_2)$ say of themselves that they are provable. Therefore, Kreisel's assessment that 'the answer to Henkin's question depends on which formula is used to 'express' the notion of *provability in* Σ' is at least misleading. By the standards Kreisel employed back then, it also depends on how the sentence asserting its own provability is constructed from a given formula expressing provability. Therefore, the answer to Henkin's question is subject to intensionality phenomena not only with respect to the second source of intensionality, which concerns the expression of properties, but also to the third source, which concerns self-reference, at least if Kreisel's Condition and the Kreisel–Henkin Criterion are adopted.

Generally, the Kreisel–Henkin Criterion for self-reference cannot narrow down the set of all self-referential sentences such that questions about sentences ascribing to themselves a property via a fixed coding and a fixed formula for the property yield a unique answer. There are many different ways to obtain self-referential statements in the sense of the Kreisel–Henkin Criterion. We shall discuss more examples of intensionality arising from the third source soon.

[23] As mentioned above, in the mentioned appendix $\overline{n}$ is (re-)defined as the efficient numeral of n.

[24] The most elegant way to formulate such theories is to demand that we add the recursion equations for the p-time computable functions to Basic, using the function symbols that are already present.

[25] We will strengthen this result in the next section.

We don't take a definite stance on whether the Kreisel–Henkin Criterion is adequate. Many authors seem at least to sympathize with a criterion similar to it.[26]

Examples such as the sentences $\mathrm{Bew}_{\mathrm{I}}(t_1)$ and $\mathrm{Bew}_{\mathrm{II}}(t_2)$, however, cast some doubt on the adequacy of the Kreisel–Henkin Criterion. Both sentences are self-referential in the weak sense that they ascribe certain properties to themselves. $\mathrm{Bew}_{\mathrm{I}}(t_1)$, for instance, ascribes to itself the property expressed by the formula $x = x \vee \mathrm{Bew}(x)$, but, at least to us, it is not so obvious that it also states about itself that it has the property expressed by the predicate $x = t_1 \vee \mathrm{Bew}(x)$. However, according to the Kreisel–Henkin Criterion, $\mathrm{Bew}_{\mathrm{I}}(t_1)$ says of itself that it has both properties.

As in the case of the second source of intensionality, that is, the expression of a property by a formula, it is hard to specify general criteria, but one can retreat to a default position by invoking a 'canonical' construction. In the case of self-reference, the canonical construction is Gödel's diagonal method and, at least for the purposes of this paper, we assume that the canonical method yields a paradigmatic case of self-reference. Before dipping into a more general discussion of the Kreisel–Henkin Criterion and possible improvements of it, we ask whether Kreisel's use of a noncanonical method for obtaining a fixed point of his provability predicates was indispensable.

§5. Refutable and independent Henkin sentences obtained by canonical diagonalization. If we are interested in Henkin sentences which not only satisfy the Kreisel–Henkin Criterion for self-reference but which are – unlike $\mathrm{Bew}_{\mathrm{II}}(t_2)$ – obtained by applying the usual Gödel construction to a chosen provability predicate, then Kreisel's Observation doesn't provide an answer. Kreisel's construction can, however, be finessed to produce such a sentence.

THEOREM 5.1. *There is a provability predicate* $\mathrm{Bew}_2(x)$ *weakly representing provability in* Σ *such that its fixed point obtained by the usual diagonal construction is refutable.*

Lavinia Picollo used a variation of the construction to show that there are also independent Henkin sentences of this kind and suggested to us the following observation:

THEOREM 5.2. *There is a provability predicate* $\mathrm{Bew}_3(x)$ *weakly representing provability in* Σ *such that its fixed point obtained by the usual diagonal construction is neither provable nor refutable.*

We give a unified treatment of both theorems. We assume that Σ is a recursively enumerable theory extending Basic.

DEFINITION 5.3. *A* diagonal operator d *(for* Σ*) is a primitive recursive function that returns, when applied to a formula of the language of* Σ *with a designated variable x free,*[27] *a formula with the same variables but not x free that satisfies the following condition:*

$$\Sigma \vdash d(\varphi(x)) \leftrightarrow \varphi(\ulcorner d(\varphi(x)) \urcorner) \tag{3}$$

[26] For instance, Heck (2007, p. 19) writes: 'So suppose that $\mathcal{O}$ [an interpreted language with the truth predicate] contains a truly self-referential liar sentence, that is, that $\mathcal{O}$ contains a term λ that denotes the sentence $\ulcorner \neg T\lambda \urcorner$.' In the discussion on whether Yablo's paradox is self-referential, the availability of criteria for self-reference is crucial. A general criterion is hardly ever explicitly discussed. However, a number of authors, for instance, Priest (1997, p. 236), seem to rely implicitly on criteria akin to the Kreisel–Henkin Criterion. Milne (2007, p. 210) is more cautious.

[27] We allow that x occurs zero times.

DEFINITION 5.4. *A diagonal operator d has the* Kreisel–Henkin property *if $d(\varphi(x))$ is of the form $\varphi(t)$, where t is a closed term and $t = \ulcorner d(\varphi(x))\urcorner$ is true.*

The Kreisel–Henkin property trivially implies satisfaction of Condition (3). The *canonical* diagonal operator is the function that is obtained 'in the usual way', as found, for instance, in Smoryński's (1985). Clearly, this operator has the Kreisel–Henkin property.

For the following definitions we fix a primitive recursive diagonal operator d; the canonical one will suffice. The diagonal operator d is represented in Σ by $\underset{\cdot}{d}$, so we have for any formula ψ:

$$\Sigma \vdash \underset{\cdot}{d}(\ulcorner\psi\urcorner) = \ulcorner d(\psi)\urcorner$$

Let γ be any sentence of the language. Given a formula $\varphi(x)$ (e.g., the canonical provability predicate) and using some diagonalization function (not necessarily d), we obtain a formula $\varphi^{\gamma}(x)$ satisfying the following condition:

$$\Sigma \vdash \varphi^{\gamma}(x) \leftrightarrow \big(x \neq \underset{\cdot}{d}(\ulcorner\varphi^{\gamma}(x)\urcorner) \wedge \varphi(x)\big) \vee \big(x = \underset{\cdot}{d}(\ulcorner\varphi^{\gamma}(x)\urcorner) \wedge \gamma\big) \tag{4}$$

We note that if γ and $\varphi(x)$ are Σ_1, then so is $\varphi^{\gamma}(x)$, if $\varphi^{\gamma}(x)$ has been obtained by applying the canonical diagonal operator.

As in the case of Henkin's and Kreisel's formulae, it is possible to show that, according to Kreisel's Condition, if $\varphi(x)$ is a provability predicate, both $\varphi^{\gamma}(x)$ and $\varphi(x)$ express provability; that is, they both weakly represent provability. Obviously $\varphi^{\gamma}(x)$ and $\varphi(x)$ agree on all numbers except the code of $d(\varphi^{\gamma}(x))$:

LEMMA 5.5. $\Sigma \vdash x \neq \underset{\cdot}{d}(\ulcorner\varphi^{\gamma}(x)\urcorner) \rightarrow (\varphi^{\gamma}(x) \leftrightarrow \varphi(x))$

For the crucial case we can prove the following claim:

LEMMA 5.6. $\Sigma \vdash d(\varphi^{\gamma}(x)) \leftrightarrow \gamma$

Proof. We have:

$$\begin{aligned}
\Sigma \vdash d(\varphi^{\gamma}(x)) &\leftrightarrow \varphi^{\gamma}(\ulcorner d(\varphi^{\gamma}(x))\urcorner) \quad \text{diagonal property (3)}\\
&\leftrightarrow \big(\ulcorner d(\varphi^{\gamma}(x))\urcorner \neq \underset{\cdot}{d}(\ulcorner\varphi^{\gamma}(x)\urcorner) \wedge \varphi(\ulcorner d(\varphi^{\gamma}(x))\urcorner)\big) \vee\\
&\qquad \big(\ulcorner d(\varphi^{\gamma}(x))\urcorner = \underset{\cdot}{d}(\ulcorner\varphi^{\gamma}(x)\urcorner) \wedge \gamma\big) \quad \text{def. of } \varphi^{\gamma}(x)\\
&\leftrightarrow \gamma
\end{aligned}$$

□

Now, as in the proof of Kreisel's Observation on p. 681, fix some formula $\mathrm{Bew}(x)$ weakly representing Σ-provability in Σ. We verify that $\mathrm{Bew}^{\gamma}(x)$ satisfies Kreisel's condition.

LEMMA 5.7. *The formula* $\mathrm{Bew}^{\gamma}(x)$ *expresses Σ-provability in the sense of Kreisel's Condition, that is, it weakly represents provability in Σ.*

Proof. For arbitrary formulae ψ, we need to show the equivalence $\Sigma \vdash \psi$ iff $\Sigma \vdash \mathrm{Bew}^{\gamma}(\ulcorner\psi\urcorner)$. If ψ is different from $d(\mathrm{Bew}^{\gamma}(x))$, the claim follows from Lemma 5.5; if ψ is $d(\mathrm{Bew}^{\gamma}(x))$, the claim follows from the fixed-point property of $d(\mathrm{Bew}^{\gamma}(x))$, that is, from (3) in Definition 5.3. □

Note that Lemma 5.7 really says that the satisfaction of Kreisel's Condition is preserved by the $(\cdot)^{\gamma}$ construction, that is, if $\varphi(x)$ satisfies Kreisel's Condition, then so will $\varphi^{\gamma}(x)$. We summarize our insights in a theorem:

THEOREM 5.8. *Suppose Σ is an arithmetical theory that contains* Basic. *Let d be a diagonal operator and let γ be a sentence of the language of Σ. Then there is a predicate* Bew^{γ} *satisfying Kreisel's Condition for Σ such that the following holds:*

$$\Sigma \vdash d(\mathrm{Bew}^{\gamma}(x)) \leftrightarrow \gamma$$

Theorem 5.1 is then proved by choosing the canonical diagonal operator as d, $\mathrm{Bew}(x)$ as $\varphi(x)$ and $0=1$ as γ. Theorem 5.2 is proved by again choosing the canonical diagonal operator as d, $\mathrm{Bew}(x)$ as $\varphi(x)$, but an independent sentence as γ. This concludes the proofs of Theorems 5.1 and 5.2.

We can use the ideas above to strengthen Observation 4.1. Let's say that a sequence of diagonal operators $(d_n)_{n\in\omega}$ is *primitive recursive* if there is a binary primitive recursive function d such that $d(n,x) = d_n(x)$, for all n, x in ω. We say that $(d_n)_{n\in\omega}$ is *semi-injective* if, whenever x occurs freely in φ and $n \neq m$, we have $d_n(\varphi(x)) \neq d_m(\varphi(x))$. Finally we say that $(d_n)_{n\in\omega}$ is *expansive* if, for all n, whenever x occurs freely in φ, the Gödel number of $d_n(\varphi(x))$ is strictly larger than n.

It is easy to construct a semi-injective and expansive primitive recursive sequence $(d_n)_{n\in\omega}$ of diagonal operator, where each d_n has the Kreisel–Henkin property. For example, for each n, we can take the 'internal variable' of the usual fixed-point construction as a variable v_n, where the Gödel number of v_n exceeds n. Of course, the variables v_n have to be kept distinct from other variables occurring in φ. Clearly all such details can be taken care of. Alternatively, we can add superfluous material to the formula expressing the substitution function.

Let $(d_n)_{n\in\omega}$ be a semi-injective and expansive primitive recursive sequence of diagonal operators. The sequence is represented in Σ by $\underset{\cdot}{d}$, so for any formula ψ, we have:

$$\Sigma \vdash \underset{\cdot}{d}(\overline{n}, \ulcorner\psi\urcorner) = \ulcorner d_n(\psi)\urcorner$$

Let γ be a formula containing x free. Given a formula $\varphi(x)$ (e.g., the canonical provability predicate) and using some diagonalization function (not necessarily d), we obtain a formula $\varphi^{\gamma}(x)$ satisfying the following condition:

$$\begin{aligned}\Sigma \vdash \varphi^{\gamma}(x) \quad \leftrightarrow \quad & \forall y < x \left(x \neq \underset{\cdot}{d}(y, \ulcorner\varphi^{\gamma}(x)\urcorner) \wedge \varphi(x)\right) \vee \\ & \exists y < x \left(x = \underset{\cdot}{d}(y, \ulcorner\varphi^{\gamma}(x)\urcorner) \wedge \gamma(y)\right)\end{aligned}$$

We note that if γ is Σ_1, then so is $\varphi^{\gamma}(x)$, if $\varphi^{\gamma}(x)$ has been obtained by the canonical diagonal method.

LEMMA 5.9. $\Sigma \vdash \forall y < \overline{n} \;\left(\overline{n} \neq \underset{\cdot}{d}(y, \ulcorner\varphi^{\gamma}(x)\urcorner)\right) \to \left(\varphi^{\gamma}(\overline{n}) \leftrightarrow \varphi(\overline{n})\right)$

LEMMA 5.10. $\Sigma \vdash d_n(\varphi^{\gamma}(x)) \leftrightarrow \gamma(\overline{n})$

Proof. We reason in Σ as follows:

$$\begin{aligned} d_n(\varphi^{\gamma}(x)) &\leftrightarrow \varphi^{\gamma}(\ulcorner d_n(\varphi^{\gamma}(x))\urcorner) \\ &\leftrightarrow \forall y < \ulcorner d_n(\varphi^{\gamma}(x))\urcorner (\ulcorner d_n(\varphi^{\gamma}(x))\urcorner \neq \underset{\cdot}{d}(y, \ulcorner\varphi^{\gamma}(x)\urcorner) \wedge \varphi(\ulcorner d_n(\varphi^{\gamma}(x))\urcorner)) \vee \\ &\qquad \exists y < \ulcorner d_n(\varphi^{\gamma}(x))\urcorner (\ulcorner d_n(\varphi^{\gamma}(x))\urcorner = \underset{\cdot}{d}(y, \ulcorner\varphi^{\gamma}(x)\urcorner) \wedge \gamma(y)) \\ &\leftrightarrow \gamma(\overline{n}) \end{aligned}$$

□

The next lemma shows that $\mathrm{Bew}^{\gamma}(x)$ is a provability predicate by Kreisel's standards.

LEMMA 5.11. $\mathrm{Bew}^{\gamma}(x)$ *expresses provability in* Σ *according to Kreisel's Condition; that is, the following equivalence holds for every* ψ*:*

$$\Sigma \vdash \psi \quad \textit{iff} \quad \Sigma \vdash \mathrm{Bew}^{\gamma}(\ulcorner\psi\urcorner)$$

Proof. If ψ is different from $d(n, \mathrm{Bew}^{\gamma}(x))$, for all n, the claim follows from Lemma 5.9; if ψ is $d(n, \mathrm{Bew}^{\gamma}(x))$, for some n, the claim follows from the fixed-point property of $d(n, \mathrm{Bew}^{\gamma}(x))$. □

We summarize our results in a theorem:

THEOREM 5.12. *Suppose* Σ *extends* Basic. *Let* $(d_n)_{n\in\omega}$ *be a semi-injective and expansive primitive recursive sequence of diagonal operators. Let* $\gamma(x)$ *be a formula of the language with only x free. Then there is a predicate* Bew^{γ} *satisfying Kreisel's Condition such that* $\Sigma \vdash d_n(\mathrm{Bew}^{\gamma}(x)) \leftrightarrow \gamma(\overline{n})$ *for each n.*

In Observation 4.1 it was shown that by applying different diagonal operators with the Kreisel–Henkin property to a given provability predicate, both a provable and a refutable Henkin sentence can be obtained. The construction just outlined will give us infinitely many diagonal operators that yield infinitely many pairwise nonequivalent Henkin sentences, if Kreisel's Condition and the Kreisel–Henkin Criterion are accepted.

It is well known how the story continued after the exchange between Henkin and Kreisel: Löb (1955) proved his celebrated theorem, which we state here somewhat loosely in the following form:

THEOREM 5.13. *Assume that* Σ *is sufficiently strong and* $\mathrm{Bew}(x)$ *is the canonical provability predicate.*[28] *Then the following obtains for all sentences* φ*:*

$$\Sigma \vdash \mathrm{Bew}(\ulcorner\varphi\urcorner) \rightarrow \varphi \quad \textit{iff} \quad \Sigma \vdash \varphi$$

Note that this theorem is independent of soundness conditions. Our theory Σ is even allowed to be inconsistent.

It follows that *any* fixed point of the canonical provability predicate is provable, whether it is self-referential or not. Hence all these fixed points are provably equivalent.

Of course, it is not so easy to say for arbitrary formal systems what their 'natural' provability predicates are; but once the criterion for the expression of provability is strengthened so that any formula meeting the strengthened criterion will satisfy the Löb derivability conditions, then a criterion for self-reference is no longer needed. This is the case because Löb's theorem will hold irrespective of whether a fixed point of the provability predicate says about itself that it is provable. Hence a criterion for self-reference, or a method for sieving out the interesting fixed points of the provability predicate, is not needed for solving Henkin's problem, as long as the provability predicate satisfies the Löb derivability conditions, which only deviant provability predicates such as $\mathrm{Bew}^{\gamma}(x)$ will fail.

§6. Acknowledgments.

Volker Halbach's work was supported by the Arts & Humanities Research Council AH/H039791/1. We thank Rasmus Blanck, Cezary Cieśliński, Kentaro Fujimoto, Graham Leigh, Hannes Leitgeb, Dan Isaacson, Arthur Merin, Lavinia

[28] For the purposes here a theory is said to be sufficiently strong if it extends Basic and an appropriate variant of S^1_2.

Picollo and two referees for valuable comments, and suggestions. We are especially indebted to Christopher von Bülow, who provided numerous corrections.

BIBLIOGRAPHY

Auerbach, D. D. (1985). Intensionality and the Gödel theorems. *Philosophical Studies*, **48**, 337–351.

Blanck, R. (2011). *Metamathematical fixed points*. Philosophical Communications Red series 41, Gothenburg: University of Gothenburg.

Boolos, G. (1993). *The Logic of Provability*. Cambridge: Cambridge University Press.

Feferman, S. (1960). Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, **49**, 35–91.

Franks, C. (2009). *The Autonomy of Mathematical Knowledge: Hilbert's Program Revisited*. Cambridge: Cambridge University Press.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik*, **38**, 173–198.

Heck, R. (2007). Self-reference and the languages of arithmetic. *Philosophia Mathematica*, **15**, 1–29.

Henkin, L. (1952). A problem concerning provability. *Journal of Symbolic Logic*, **17**, 160.

Henkin, L. (1954). Review of G. Kreisel: On a problem of Henkin's. *Journal of Symbolic Logic* ,**19**, 219–220.

Hilbert, D., & Bernays, P. (1939). *Grundlagen der Mathematik II* (second edition: 1970). Berlin: Springer .

Jeroslow, R. (1973). Redundancies in the Hilbert-Bernays derivability conditions for Gödel's second incompleteness theorem. *Journal of Symbolic Logic*, **38**, 359–367.

Kreisel, G. (1953). On a problem of Henkin's. *Indagationes Mathematicae*, **15**, 405–406.

Kreisel, G., & Takeuti, G. (1974). *Formally self-referential propositions for cut free classical analysis and related systems*. Technical report, Warsaw: Polska Akademia Nauk.

Löb, M. H. (1955). Solution of a problem of Leon Henkin. *Journal of Symbolic Logic*, **20**, 115–118.

McGee, V. (1992). Maximal consistent sets of instances of Tarski's schema (T). *Journal of Philosophical Logic*, **21**, 235–241.

Milne, P. (2007). On Gödel sentences and what they say. *Philosophia Mathematica*, **15**, 193–226.

Priest, G. (1997). Yablo's paradox. *Analysis*, **57**, 236–242.

Russell, B. (1940). *An Enquiry into Meaning and Truth*. London: George Allen and Unwin.

Skyrms, B. (1984). Intensional aspects of semantical self-reference. In Martin, R. L., editor. *Recent Essays on Truth and the Liar Paradox*, Oxford: Oxford University Press, 119–131.

Smoryński, C. (1985). *Self-Reference and Modal Logic*. Universitext. New York, Berlin, Heidelberg, and Tokyo: Springer.

Smoryński, C. (1991). The development of self-reference: Löb's theorem. In Drucker, T., editor. *Perspectives on the History of Mathematical Logic*, Boston: Birkhäuser, pp. 110–133.

Solovay, R. (1985). Explicit Henkin sentences. *Journal of Symbolic Logic*, **50**, 91–93.

Tarski, A., Mostowski, A., & Robinson, R. M. (1953). *Undecidable Theories*. Amsterdam: North Holland.

Visser, A. (2005). Faith & Falsity: A study of faithful interpretations and false Σ^0_1-sentences. *Annals of Pure and Applied Logic*, **131**(1–3), 103–131.

Visser, A. (2014). *Jumping in arithmetic*. Logic Group Preprint Series, Vol. 319, Utrecht University, http://www.phil.uu.nl/preprints/lgps/.

Yablo, S. (1993). Paradox without self-reference. *Analysis*, **53**, 251–252.

NEW COLLEGE
OXFORD, OX1 3BN, ENGLAND
E-mail: volker.halbach@new.ox.ac.uk

PHILOSOPHY, FACULTY OF HUMANITIES
UTRECHT UNIVERSITY
JANSKERHOF 13
3512 BL UTRECHT, THE NETHERLANDS
E-mail: albert.visser@phil.uu.nl