SELF-REFERENCE IN ARITHMETIC*

Volker Halbach and Albert Visser

draft
15th December 2013

> 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

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 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 point for the formula is obtained. Some further examples of self-referential sentences are considered such as sentences that 'say of themselves' that they are Σ^0_n-true (or Π^0_n-true) and their formal properties are investigated.

*Volker Halbach's work was supported by the Arts & Humanities Research Council AH/H039791/1. He thanks Christopher von Bülow, Cezary Cieśliński, Kentaro Fujimoto, Graham Leigh, Dan Isaacson, Arthur Merin and Lavinia Picollo for valuable comments and suggestions.

1. Introduction

'We thus have a sentence before us that states its own unprovability.' This is how Gödel describes the sentence 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 not provable.

The only 'self-reference like' feature of the Gödel sentence γ that is used in the proof of the 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 fixed point, modulo provable equivalence, of $\neg\mathrm{Bew}(x)$.[2] But 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. But whether γ is also self-referential or 'states its own unprovability' in whatever sense is not relevant for Gödel's proof.[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. But also the proof of Löb's theorem only relies on the existence of some fixed points of a certain formulae. 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 even fail to be extensional in this respect. Since we take into accunt more sources for intensionality, our notion of extensionality siffers from others found in the literature, for instance, Feferman's (1960).[4]

[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 he relevant theory.

[3] We use the label *self-referential* in a very loose way. When we call a sentence *self-referential* we mean that it ascribes to itself the property under consideration. Which property is meant should be clear from the context. This is inaccurate, because a sentence may well ascribe to itself another property but not the one under consideration and be self-referential with respect ot this other property. But it would be very cumbersome to avoid *self-referential* and talk about *ascribes to itself the property expressed the formula so-and-so*.

[4] See p. 13 for further remarks on Feferman's treatement of intensionsality.

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 sences that can be proved to be fixed points of certain formulae. The development in metamathematics has lead 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]

At least for *certain* questions in metamathematics, the notion of self-reference cannot be easily avoided. For instance, it is natural to ask whether the sentence that states its own provability is provable or not. This question becomes pointless if reformulated extensionally as a question about fixed points of the (standard) provability predicate, modulo provable equivalence, because $0 = 0$, for instance, is a trivial example of such a fixed point. So the question under consideration is intensional in the sense that it requires the notion of *stating its own provability*. So even though self-reference may not be a central notion of modern metamathematics, its study has led to results – most remarkably Löb's theorem – that are in themselves no longer dependent on the notion of self-reference.

In philosophical discussions, the notion of self-reference in metamathematics 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 sheds some light also on the problems of reference and intensionality in informal discourse. As our discussion will show, whether it does, is a delicate matter. A salient example where the possibility of such an application is urgent is the discussion about Yablo's (1993) paradox where metamathematical tools are used to answer the question whether the Yablo sentences are self-referential or not. But even for metamathematical sentences we lack a

[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.

full analysis of sef-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 the 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 specifying explicitly the property. This is somewhat sloppy, because the sentence may not ascribe the property P to itself but rather some other property or properties. Moreover, the notion of 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. So any result about such a sentence is relative to or intensional with respect to these choices. 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 a later stage. Examples will be presented below.

Of course, the three source of intensionality depend themselves 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 will call Basic. The language of Basic is the language of arithmetic extended with function symbols for all primitive recursive functions. The Tarski–Mostowski–Robinson theory R, introduced by Tarski et al. (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: Whether an arithmetical formula expresses a property of syntactical objects depends on the chosen coding 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, a 'reasonable' Gödel coding is constructed in Appendix A such that for each formula $\varphi(v)$ there is a unique number m such that $\varphi(\overline{m})$ has code m and, consequently, $\varphi(\overline{m})$ is at least a fixed point of $\varphi(v)$ and ascribes to itself the property expressed by $\varphi(v)$ according to the standards of certain authors. 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 be bypassed, as diagonalization is built into the coding schema for all formulae. Whether this coding leads to truly self-referential statements is another delicate matter.

For further parts of the paper, other assumptions on the coding are required. For instance, in our analysis of arithmetical truth tellers constructed with partial truth predicates, the provability or refutability of these sentences will be shown under certain assumptions on the coding scheme.

In the context of weak theories certain coding schemes are more appropriate than others. There Gödel's or Kleene's method of coding are not sufficiently effective. In such cases we assume that a suitable coding scheme 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 scheme. We follow the usual practice and assume some 'standard' coding scheme without setting out the details, unless we explicitly introduce further assumptions.

2.2. Second Source of Intensionality: Expressing a property.

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).

Here we do not attempt to give a full answer, but we will state some assumptions that give some indication of the shape a possible answer could take. First, we assume that a 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 logically equivalent, may express the same property. Thirdly, even provably equivalent formulae may fail to express the same property, where provability may be understood as provability in some designated theory.

Different kinds of criteria have been used to argue that a certain 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 property.[6]

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 with associated numerals $\overline{n}$:

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

In metamathematics Kreisel's Condition became the formal notion notion of weak representability:[7] A formula $\varphi(x)$ is said to *weakly represent* a set S of numbers if and only if the following equivalence holds:

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

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

It is obvious that Kreisel's Condition can hardly yield an adequate explication of expressing a syntactic property, because it is purely extensional: Properties expressed by provably equivalent formulae would always be identical, if this condition were taken as a full explication. Even if the formulae are not provably equivalent, they can still express the same property, according to Kreisel's Condition, as is witnessed by a canonical provability predicate and its associated Rosser–provability predicate. This is in conflict with our intensional stance on syntactic properties.

The applicability of Kreisel's Condition is also very limited. For notions such as Π_1-truth there is no representing formula even though most logicians believe that there is an arithmetical formula expressing the notion of Π_1-truth.

[6] 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.

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

But not only would Π_1-truth be not expressible; a canonical Π_1-truth predicate that provably only applies to Π_1-sentences would express not Π_1-truth; it would rather express the property of being a Σ-provable 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 expressing a property relative to a theory. If this theory is unsound, then the criterion may yield unwanted consequences. For instance, relative to the theory $\mathrm{PA} + \neg\,\mathrm{Con}_{\mathrm{PA}}$, which is consistent by Gödel's Second Incompleteness theorem, the canonical provability predicate doesn't express provability in PA.

At any rate, Kreisel's Condition is neither sufficient nor necessary as a criterion of expressing a syntactic property. As alternative or additional criterion, which we call the *meaning postulates* or *conditions*, 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 or partial truth, what is often meant by saying 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 (and that perhaps also the compositional axioms are provably satisfied for these sentences).

So the general *meaning postulates* or *conditions* criterion would read as follows: A formula of arithmetic expresses a property of syntactic objects if and only if the formula satisfies certain principles or axioms associated with the property, relative to a coding scheme.

In the case of provability, Löb's derivability conditions have been used as meaning postulates, although in themselves they will hardly suffice because they are satisfied by the formula $x = x$ and, even if they are 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 **?**.

Moreover, it is far from being clear how the meaning postulates become associated with the property. In fact, in many cases the postulates were only dis-

covered 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 a formula expressing provability; the development of the derivability conditions started in Gödel's paper, was continued by Hilbert and Bernays (1939) and reached their completion in the work of Löb (1955).

Like Kreisel's Condition, the meaning postulate criterion is relative to a theory. However, both 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 we mentioned above; and vice versa in any consistent theory with enough coding we have a predicate that expresses provability according to the Kreisel conditions and 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. Syntactic properties are usually given by an informal metamathematical description of that property. Often logicians expect a formula to express the syntactic property if the formula structurally resembles the metamathematical description of the property. It's difficult to explain what such a structural resemblance could consist in.

The resemblance criterion is highly sensitive to the chosen coding scheme. The similarity between the arithmetical formula and the description of the syntactic property could be seen as a comparison between the definition of a set of syntactic objects and of a set of numbers. Both definitions will involve not just the syntactic objects and numbers in the respective sets, but also some further intermediary objects and numbers, that need to correspond to each other. This correspondence is obviously sensitive to the coding as well.

One way to sidestep the sensitivity to coding 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.

In practice we seem to recognize the relevant resemblances in many specific cases, but we lack a general account of resemblance that is required. Usually

the resemblance criterion seems to be applied in claims to the effect that a certain formula expresses a property *in a natural way*. In what follows, *canonical* provability predicates, for instance, will be assumed to resemble in their 'salient' features the definition of the informal provability predicate in metamathematics. In what follows all three criteria mentioned will be applied and 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 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 $\varphi(x)$. To this end, usually Gödel's diagonal construction or a variant thereof is employed.

In this paper we assume that there are *some* paradigmatic cases of self-reference, usually established via Gödel's diagonal method using. But even for the canonical diagonalization method it is not accepted without exceptions that it establishes self-reference in the intended way. For instance, Heck (2007) and Milne (2007) have raised some sceptical worries. We share some of the doubts and do not assume that all so-called Gödel sentences found in textbooks 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 be thought to be self-referential nevertheless. In Appendix A a coding is sketched in which no third stage is needed, because diagonalization is already built into the coding. In his initial answer to Henkin's problem, Kreisel presented a sentence that may be thought to be self-referential but that hasn't been obtained from the formula $\varphi(x)$ expressing provability in the usual way.

If a sentence γ says about itself that it has property P and P is expressed by the formula $\varphi(x)$, then 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). 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, we call it extensional (with respect to the third source of intensionality); if it does depend on the choice of the fixed point, it is intensional.[8]

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

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 according to the standard model is not elementarily definable; and the set of provable fixed points is only recursively enumerable but not decidable.

OBSERVATION 1. Let a formula $\varphi(x)$ be given. Then 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 truth cannot be expressed by a formula of Σ according to the observation.

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

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

OBSERVATION 2. Assume Σ extends Basic, as defined on p. 5. The set of all provable fixed points of any given formula $\varphi(x)$, that is, the set of all sentences ψ with $\Sigma \vdash \varphi(\ulcorner\psi\urcorner) \leftrightarrow \psi$ is not recursive.

The proof is reminiscent of Curry's paradox and McGee's (1992) trick, gives us 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 gives applied to a formula ψ 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. 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 all *provable* fixed points of a given formula are all provably equivalent in a certain theory. But it's obvious that for other formulae such as the partial truth predicate for Σ_1-sentences or the Rosser provability predicate this is not the case.[9]

Hence any sentence that says about itself that it possesses the property expressed by $\varphi(x)$ will be a fixed point of that formula; but in order 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 now.

This completes the description of the three stages in which, according to the standard method, a sentence is obtained that ascribes a property to itself. None of the three stages yields a unique output: There are many different Gödel codings; given a coding, every property that can be expressed at all by some formula

[9] See pages 29 and 32 below.

can be expressed by many different formulae; and, given a coding and a property expressed by a formula, different fixed can be constructed such that their fixed-point property is provable in the theory is question.

Different choices of the coding, the formula and the method of obtaining presumed self-reference can each yield sentences with different properties. Therefore the three stages very roughly correspond to 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 also self-reference has intensional aspects. 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 on 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 where reached before the problem was solved by Löb. So we will begin with the situation in the early 1950s.

If the Gödel sentence is the sentence that states its own unprovability, it is natural to consider the sentence that states its own *provability* and to investigate whether it is 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 a fixed-point 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 fixed point of $\mathrm{Bew}(x)$, if $\mathrm{Bew}(x)$ weakly represents provability, because both 1=1 and thus $\mathrm{Bew}(\ulcorner 1=1 \urcorner)$ will be provable. But 1 = 1 doesn't 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 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?

We think that Henkin's formulation is an attempt to ask whether a formula stating its own provability is provable or not. Henkin avoided a direct appeal to self-reference and did not ask whether a sentence stating its own provability is provable or not; but his formulation of the question still leaves some space for interpretation. Presumably, Henkin had Gödel's construction in mind 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 worthwhile. It sheds light on the question whether 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 to understand 'express' in the sense of Kreisel's Condition in the sense of 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.[10] Then the following hold:

a) There is a formula $\mathrm{Bew}_{\mathrm{I}}(x)$ and a term t_1 such that the following three conditions are satisfied:
   (i) $\mathrm{Bew}_{\mathrm{I}}$ weakly represents provability in Σ.
   (ii) $\Sigma \vdash t_1 = \ulcorner \mathrm{Bew}_{\mathrm{I}}(t_1) \urcorner$
   (iii) $\Sigma \vdash \mathrm{Bew}_{\mathrm{I}}(t_1)$

b) Similarly, there is a provability predicate $\mathrm{Bew}_{\mathrm{II}}(x)$ and a term t_2 such that
   (i) $\mathrm{Bew}_{\mathrm{II}}$ weakly represents provability in Σ.
   (ii) $\Sigma \vdash t_2 = \ulcorner \mathrm{Bew}_{\mathrm{II}}(t_2) \urcorner$
   (iii) $\Sigma \vdash \neg \mathrm{Bew}_{\mathrm{II}}(t_2)$

The examples employed by Kreisel in the proof are of some interest. In particular, the example for $\mathrm{Bew}_{\mathrm{I}}(t_1)$ foreshadows Kreisel's (1974) proof of Löb's theorem, as was pointed out by Smoryński (1991). Henkin suggested simpler examples that 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 the second 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.

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

This is a direct consequence of the Friedman–Goldfarb–Harrington Theorem.[11] Using the canonical diagonal construction (or any other method), 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:

$$\Sigma \vdash \varphi \text{ iff } \Sigma \vdash \mathrm{Bew}_{\mathrm{II}}(\ulcorner \varphi \urcorner) \tag{3}$$

If φ is different from $t_2 \neq t_2 \wedge \mathrm{Bew}(t_2)$ 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. (See Löb (1955) or, e.g. Boolos (1993).)

As Löb's theorem wasn't known, Henkin and Kreisel had to use a different construction.[12] 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

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

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

[12] Note also that the Kreisel–Henkin construction works in some very weak cases where it is not clear that we have Löb's theorem.

In the next section we will investigate whether Henkin's question and Kreisel's examples can be seen as was has come to be called a Henkin sentence, that is, a sentence that states its own provability.

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 answered Henkin's question, which is not explicitly formulated as a question about a sentencestating its own provability.

We think that, if $\mathrm{Bew_I}(x)$ expresses provability, then $\mathrm{Bew_I}(t_1)$ 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 $\mathrm{Bew_{II}}(t_2)$ as well. 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 $\mathrm{Bew_I}(a)$ nor $\mathrm{Bew_{II}}(a)$] expresses the propositional function *a is provable*.' Thus Henkin's dismissal of Kreisel's answer is based on his rejection of what we have called Kreisel's Condition for provability; Henkin doesn't believe that any formula weakly representing provability expresses provability.

Henkin's rejection of Kreisel's and his own examples of contrived provability predicates is well motivated and the focus on canonical provability predicates leads on to Löb's theorem. This part of the story is well known.

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 $\mathrm{Bew_I}(x)$ and $\mathrm{Bew_{II}}(x)$ are assumed to express provability, $\mathrm{Bew_I}(t_1)$ and $\mathrm{Bew_{II}}(t_2)$ are Henkin sentences, that is, sentences stating there own provability.

It is not obvious that $\mathrm{Bew_I}(t_1)$ and $\mathrm{Bew_{II}}(t_2)$ say of themselves that they are provable, even if Kreisel Condition is accepted. Kreisel had not only used non-canonical provability predicates; he had also employed a non-canonical way of

obtaining the fixed points of his provability predicates.[13] In particular, he had *not* applied Gödel's construction to his provability predicates to obtain their fixed points. If he had applied the usual Gödel 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 3 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 are obtained. Both presumably tried to avoid murky formulations such as 'states its own provability' on the one hand; on the other hand, they didn't intend to distract from the intuitive appeal of the question by referring to a fixed point that is obtained by the method used by Gödel, because Gödel's construction is a trick after all, a means to an end, and the method gains it's intuitive appeal by the comparison with self-referential constructions in natural language.

Whatever the motives were, Henkin and Kreisel merely required that the fixed point is of the form $\varphi(t)$ where $\varphi(x)$ is a formula expressing provability and $t = \ulcorner \varphi(t) \urcorner$ is provable.[14]

We are interested in the question whether Henkin's way of stating the question, if it is read in Kreisel's way, is really a question about a sentence stating its own provability. If the question is answered to 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:

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

[14] 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 already be found in Henkin's question in a nutshell. 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 and '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, the criterion can merely function as a *sufficient* criterion.

If a sentence is of the form $\varphi(t)$ and $t = \ulcorner\varphi(t)\urcorner$ obtains, then the sentence says of itself that it has property P. Therefore, if t is obtained in the usual way by the Gödel construction, then $\varphi(t)$ says of itself that it has property P.

The phrase 't that has (the code of) $\varphi(t)$ as its value' doesn't stipulate whether $t = \ulcorner\varphi(t)\urcorner$ must be provable in Σ or merely only 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 any 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 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 a true sentence $\psi(s)$ and then the value of s may be $\varphi(t)$ again. In at least some cases of this kind, one want to say that $\varphi(t)$ *indirectly* ascribes to itself the property expressed by $\psi(x)$. We do not want to go further into the intricacies of indirect self-reference and explicitly state that here in this paper self-reference is always understood *direct* self-reference. Similar remarks apply to related notions.

In semantic approaches to the semantic paradoxes, authors often don't specify a deductive system and work with languages containing special constants and with a designated model. 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 as on the construction just outlined.[15] 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 Appendix A, for every $\varphi(x)$ there is an n such that $\varphi(\overline{n})$ has n as its code. 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 S . . . S0, terms of the form $1 + \ldots + 1$ 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.

[15] 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 most textbooks and Gödel himself prove the first incompleteness theorem with out such a term in the language.

OBSERVATION 3. 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.[16] 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.[17]

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 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 can be chosen by Gödel's 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.[18] ⊣

The formula $\mathrm{Bew}_{\mathrm{II}}(x)$ expresses the 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 expressing 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 that concerns the expression of properties but also to the third source that concerns self-reference, at least if Kreisel's Condition and the Kreisel–Henkin Criterion are adopted.

Generally, the Kreisel–Henkin Criterion for self-reference can't narrow down the set of all self-referential sentences such that questions about sentences as-

[16] 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.

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

[18] For Löb's theorem, see Löb (1955) or, e.g. Boolos (1993).

cribing to themselves a property via a fixed coding and a fixed formula expressing 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.

We don't take a definite stance whether the Kreisel–Henkin Criterion is adequate. Many authors seem at least to sympathize with a criterion similar to it.[19]

Examples such as the sentences $\mathrm{Bew_I}(t_1)$ and $\mathrm{Bew_{II}}(t_2)$, 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. $\mathrm{Bew_I}(t_1)$, 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, $\mathrm{Bew_I}(t_1)$ says of itself that it has both properties.

As in the case of the second source of intentionality, 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 non-canonical method for obtaining a fixed-point of his provability predicates was indispensable.

[19] 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. However, a general criterion is hardly ever explicitly discussed. However, a number of authors, e.g. (Priest, 1997, p. 236), seem to rely implicitly on criteria akin to the Kreisel–Henkin Criterion. (Milne, 2007, p. 210) is more cautious.

5. *Refutable and independent Henkin sentences obtained by canonical diagonalization*

If we are interested in Henkin sentences that do not only satisfy the Kreisel–Henkin Criterion for self-reference but that are – unlike $\mathrm{Bew}_{\mathrm{II}}(t_2)$ – obtained from applying the usual Gödel construction to a chosen provability predicate, then Kreisel's Observation doesn't provide an answer. Kreisel's construction, however, can be finessed to produce such a sentence. The provability predicates from which this sentences is obtained by the canonical diagonal construction that expresses again provability in Σ by Kreisel's Condition.

THEOREM 4. 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. 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 6. 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[20], a formula with the same variables but not x free that satisfies the following condition:

$$\Sigma \vdash d(\varphi(x)) \leftrightarrow \varphi(\ulcorner d(\varphi(x)) \urcorner) \tag{4}$$

DEFINITION 7. 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.

[20] We allow that x occurs zero times.

It is easy to see that the Kreisel–Henkin property implies satisfaction of Equation 4. 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 some primitive recursive diagonal operator d; the canonical one will suffice. The diagonal operator d is represented in Σ by $\dot{d}$, so for any formula ψ we have:

$$\Sigma \vdash \dot{d}(\ulcorner\psi\urcorner) = \ulcorner d(\psi)\urcorner. \tag{5}$$

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 sentences $\varphi^\gamma(x)$ satisfying the following condition:

$$\Sigma \vdash \varphi^\gamma(x) \leftrightarrow \big(x \neq \dot{d}(\ulcorner\varphi^\gamma(x)\urcorner) \wedge \varphi(x)\big) \vee \big(x = \dot{d}(\ulcorner\varphi^\gamma(x)\urcorner) \wedge \gamma\big) \tag{6}$$

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, that is, weak representability, both $\varphi^\gamma(x)$ and $\varphi(x)$ express provability, if $\varphi(x)$ is a provability predicate. In the general case we only have this for all cases except for the crucial one:

LEMMA 8. $\Sigma \vdash x \neq \dot{d}(\ulcorner\varphi^\gamma(x)\urcorner) \to (\varphi^\gamma(x) \leftrightarrow \varphi(x))$

For the crucial case we can prove the following claim:

LEMMA 9. $\Sigma \vdash d(\varphi^\gamma(x)) \leftrightarrow \gamma$.

Proof. We have:

$$\begin{aligned}
\Sigma \vdash d(\varphi^\gamma(x)) &\leftrightarrow \varphi^\gamma(\ulcorner d(\varphi^\gamma(x))\urcorner) && \text{diagonal property (4)}\\
&\leftrightarrow (\ulcorner d(\varphi^\gamma(x))\urcorner \neq \dot{d}(\ulcorner\varphi^\gamma(x)\urcorner) \wedge \varphi(\ulcorner d(\varphi^\gamma(x))\urcorner)) \vee\\
&\quad (\ulcorner d(\varphi^\gamma(x))\urcorner = \dot{d}(\ulcorner\varphi^\gamma(x)\urcorner) \wedge \gamma) && \text{def. of } \varphi^\gamma(x)\\
&\leftrightarrow \gamma
\end{aligned}$$

⊣

As in the proof of Kreisel's Observation on p. 15, fix some formula $\mathrm{Bew}(x)$ weakly representing Σ-provability in Σ. We verify that the predicate $\mathrm{Bew}^\gamma(x)$ satisfies Kreisel's Condition.

LEMMA 10. $\mathrm{Bew}^{\gamma}(x)$ expresses Σ-provability by Kreisel's Condition, that is, it weakly represents provability in Σ.

Proof. 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 8; if ψ is $d(\mathrm{Bew}^{\gamma}(x))$ the claim follows from the fixed point property of $d(\mathrm{Bew}^{\gamma}(x))$, that is, (4) in Definition 6. ⊣

Note that Lemma 10 really says that the Kreisel property is preserved by the $(\cdot)^{\gamma}$ construction. We summarize our insights in a theorem.

THEOREM 11. Suppose Σ is an arithmetical theory that contains Basic. Let d be a diagonal operator and let γ be a sentence of the language. Then there is a predicate Bew^{γ} which satisfies the Kreisel property for Σ such that

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

Theorem 4 is then proved by choosing the canonical diagonal operator as d, $\mathrm{Bew}(x)$ as $\varphi(x)$ and γ as $0=1$. Theorem 5 is proved by choosing the canonical diagonal operator as d, $\mathrm{Bew}(x)$ as $\varphi(x)$ and γ as as an independent sentence. This concludes the proofs of Theorems 4 and 5.

We can use the ideas above to strengthen Observation 3. 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 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, whenever x occurs in φ, the Gödel number of $d_n(\varphi)$ is strictly larger than n.

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

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

$$\Sigma \vdash \dot{d}(\overline{n}, \ulcorner\psi\urcorner) = \ulcorner d_n(\psi)\urcorner. \tag{7}$$

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 sentences $\varphi^\gamma(x)$ satisfying the following condition:

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

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 12. $\Sigma \vdash \forall y < \overline{n}\ \overline{n} \neq \dot{d}(y, \ulcorner\varphi^\gamma(x)\urcorner) \to (\varphi^\gamma(\overline{n}) \leftrightarrow \varphi(\overline{n}))$

LEMMA 13. $\Sigma \vdash d_n(\varphi^\gamma(x)) \leftrightarrow \gamma(\overline{n})$.

Proof. We reason in Σ as follows:

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

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

LEMMA 14. $\mathrm{Bew}^\gamma(x)$ expresses provability in Σ according to Kreisel's Condition; that is, the following holds for every ψ:

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

Proof. If ψ is different from $d(n, \mathrm{Bew}^\gamma(x))$, for all n, the claim follows from Lemma 12; 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 result in a theorem.

THEOREM 15. Suppose Σ extends Basic. Let $(d_n)_{n\in\omega}$ be a primitive recursive, semi-injective, expansive sequence of diagonal operators. Let $\gamma(x)$ be a formula of the language with only x free. Then there is a predicate Bew^γ that has the Kreisel property such that, for each n, $\Sigma \vdash d_n(\mathrm{Bew}^\gamma(x)) \leftrightarrow \gamma(\overline{n})$.

In Observation 3 it was shown that by applying to different diagonal operators with the Kreisel–Henkin property to a certain provability predicate, a provable and a refutable Henkin sentence can be obtained. The theorem above will give us infinitely many diagonal operators that yield infinitely many nonequivalent Henkin sentences, if Kreisel's Condition and the Kreisel–Henkin Criterion are accepted.

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

THEOREM 16. Assume that Σ is is sufficiently strong and $\mathrm{Bew}(x)$ is the canonical provability predicate.[21] Then the following obtains for all sentences φ:

$$\Sigma \vdash \mathrm{Bew}(\ulcorner\varphi\urcorner) \to \varphi \text{ iff } \Sigma \vdash \varphi$$

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 equivalent.

Of course, it is not so easy to say in general 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, because Löb's theorem will strike irrespective of

[21] In our context a theory is sufficiently strong if it extends Basic and an appropriate variant of S^1_2.

whether a fixed point of the provability predicate says about itself that it is provable or not. Hence a criterion for self-reference or 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. *Further examples for the intensionality of self-reference*

Löb's theorem has drawn away attention from the problem of intensionality of self-reference: a criterion such as the Kreisel–Henkin Criterion for self-reference isn't required to answer Henkin's question for the canonical provability predicate because – under reasonable assumptions on the theory – all fixed points of the standard provability predicate are provable and thus equivalent.

Löb's theorem, which eliminates all intensionality from self-reference, is specific to the canonical provability predicate. For other formulae we cannot expect something analogous. As we have seen, non-canonical provability predicates like $\mathrm{Bew}_{\mathrm{II}}(x)$ are susceptible to intensionality phenomena and fixed points can vary in their properties, even if they satisfies the Kreisel–Henkin Criterion for self-reference. We shall now look at further formulae that lend themselves to questions similar to Henkin's question about provability. First we look at Rosser provability and then at partial truth predicates.

6.1. *On Rosser provability*

The Rosser provability predicate is defined as follows.

$$\mathrm{Bew}^{R}(x) \coloneqq \exists y \left(B(y, x) \wedge \forall z < y \neg B(z, \dot{\neg} x) \right)$$

Here $B(y, x)$ strongly represents the relation that y is a proof of x and $\dot{\neg}$ represents that function that gives, when applied to a sentence, its negation. For the sake of definiteness, let's say that we employ the canonical representations. Gödel fixed points of $\neg\, \mathrm{Bew}^{R}(x)$ are Π_1 Rosser sentences. Jeroslow (1973) fixed points of $\mathrm{Bew}^{R}(\dot{\neg} x)$ are (variants of) Σ_1-Rosser sentences. Henkin's problem for Rosser provability becomes:

Is the sentence that says about itself that it is Rosser provable, refutable, or independent?

In this case, one will have to exploit the self-referentiality of the sentence. Unlike in the case of standard provability, the fixed point property does not suffice to show that the Henkin sentence is provable or refutable or independent. This follows from the long-known observation that the Rosser provability predicate has non-equivalent fixed points.

OBSERVATION 17. Let Σ be consistent and let it contain Basic plus the axioms stating that $<$ is a linear ordering. The set of fixed points of the Rosser provability predicate for Σ contains all Σ-provable and Σ-refutable formulae.

Proof. Assume there is a Σ-proof n of ψ. Then, $\Sigma \vdash B(\overline{n}, \ulcorner \psi \urcorner)$ holds. Since Σ is assumed to be consistent, there is, *a fortiori*, no proof smaller than n that is a proof of $\neg\psi$, so (†) $\forall z < \overline{n}\, \neg B(z, \dot{\neg} \ulcorner \psi \urcorner)$ is true as well. The sentence (†) is Δ_1 and thus provable. So, we get:

$$\Sigma \vdash B(\overline{n}, \ulcorner \psi \urcorner) \wedge \forall z < \overline{n}\, \neg B(z, \dot{\neg} \ulcorner \psi \urcorner)$$

By existential weakening, $\Sigma \vdash \mathrm{Bew}^R(\ulcorner \psi \urcorner)$ follows. So, $\Sigma \vdash \psi \wedge \mathrm{Bew}^R(\ulcorner \psi \urcorner)$, and thus, as desired, $\Sigma \vdash \psi \leftrightarrow \mathrm{Bew}^R(\ulcorner \psi \urcorner)$.

Now assume $\Sigma \vdash \neg\psi$ and let n be a proof of $\neg\psi$. We may conclude that $\Sigma \vdash B(\overline{n}, \dot{\neg} \ulcorner \psi \urcorner)$. Since Σ is consistent, we have, for all k, $\Sigma \vdash \neg B(\overline{k}, \ulcorner \psi \urcorner)$.

We reason in Σ. Suppose $\mathrm{Bew}^R(\ulcorner \psi \urcorner)$. Let p witness $\mathrm{Bew}^R(\ulcorner \psi \urcorner)$, so we have

$$(\ddagger)\quad B(p, \ulcorner \psi \urcorner) \wedge \forall z < p\, \neg B(z, \dot{\neg} \ulcorner \psi \urcorner).$$

Since, $<$ is linear, we may conclude that $p \leq \overline{n}$. But then, by the R-axioms, we find $\bigvee_{k \leq n} p = \overline{k}$. Let $p = \overline{k}$. Then, by (‡), $B(\overline{k}, \ulcorner \psi \urcorner)$. Quod non, since we have $\neg B(\overline{k}, \ulcorner \psi \urcorner)$. So, we may conclude $\neg \mathrm{Bew}^R(\ulcorner \psi \urcorner)$.

We return to the real world. We have found that $\Sigma \vdash \neg\psi \wedge \neg \mathrm{Bew}^R(\ulcorner \psi \urcorner)$. Hence, as desired, $\Sigma \vdash \psi \leftrightarrow \mathrm{Bew}^R(\ulcorner \psi \urcorner)$. ⊣

So $0 = 0$ as well as $0 \neq 0$ are fixed points.

QUESTION 18. What is the status of the fixed point of $\mathrm{Bew}^R(x)$ that is obtained by the Gödel construction? Does $\mathrm{Bew}^R(x)$ have fixed points that are independent?

Fixed points of $\neg\,\mathrm{Bew}^R(x)$ and $\mathrm{Bew}^R_{\dot\neg\, x}$ are respectively the Π_1 Rosser sentence and the Σ_1 Rosser sentence (in the last case modulo a small detail). For the treatment of the (non)uniqueness of the Rosser sentences, see the classical paper Guaspari and Solovay (1979) (and von Bülow 2008) and, for a different approach, Voorbraak (1989).

For more examples of non-equivalent fixed points connected to alternative provability predicates, the reader is referred to Visser (1989) and Shavrukov (1994).

6.2. Partial truth predicates

Sentences stating of themselves that they are provable are Henkin sentences; sentences stating of themselves that they are true are truth tellers. Any formula deserving the label of a truth predicate will have many non-equivalent fixed points. A formula $\varphi(x)$ for which all sentences are fixed points would be a total truth predicate, but such a predicate cannot exist in a consistent system by Tarski's theorem on the undefinability of truth. However, there are partial truth predicates for classes of sentences with limited quantifier complexity.

To dot our i's and to cross our t's, we need to pay attention a detail concerning the definition of Σ_n and Π_n. In the most narrow one, e.g. Σ_n-formulae first have a block of unbounded existential quantifiers – or even just one existential quantifiers — and then a Δ_0-formula. We could liberalize this definition in various ways allowing, say, closure under conjunction and disjunction and allowing negation of Π_1 formulae to be Σ_1-formulae and vice versa. We could even allow closure under bounded quantification; however note that this last move has costs: we need collection principles to prove the equivalence to definitions of the more narrow kind and to make truth predicates for formulae satisfying the liberal definition work. For our present treatment, we work with the narrow definition.

We consider partial truth predicates $\mathrm{Tr}_{\Sigma_n}(x)$ and $\mathrm{Tr}_{\Pi_n}(x)$ for Σ_n- and Π_n-sentences defined in a similar way as the truth predicates in the textbooks by

Hájek and Pudlák (1993) or Kaye (1991). There is some extra detail in developing the partial truth predicate since we have function symbols for all primitive recursive functions in our language. However since we work in PA these details are easy. We could for example first eliminate the primitive recursive terms from the given sentence and then apply the truth predicate for formulae only involving successor, plus and times. The truth predicates $\mathrm{Tr}_{\Sigma_n}(x)$ for $n > 1$ and $\mathrm{Tr}_{\Pi_n}(x)$ for $n \geq 1$ are no longer predicates with the Kreisel property, that is, they cannot weakly represent the set of all Σ_n- and Π_n-truths. So the question arises in which sense they are true truth-predicates.

Usually logicians resort to the *meaning postulate* approach, as mentioned on p. 8, and often call a formula σ_n a Σ_n-truth predicate if and only if $\mathsf{PA} \vdash \sigma_n(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$ holds for all Σ_n-sentences φ. Π_n-truth predicates are defined analogously.[22] We will follow this convention, but by calling a formula a Σ_n-truth predicate we don't intend to commit ourselves to the claim that it expresses the property of Σ_n-truth.

For the moment being, however, assume that $\sigma_n(x)$ is a Σ_n-truth predicate that expresses truth for Σ_n-sentences. Moreover we assume that $\sigma_n(x)$ is itself Σ_n. Then one can ask whether the sentence that says of itself that it is Σ_n-true is provable, refutable or independent. We call such a sentence a Σ_n-truth teller.

Since $\sigma_n(x)$ is a Σ_n-truth predicate, every Σ_n-sentence is a fixed point of $\sigma_n(x)$. So clearly there are provable and refutable fixed points and most fixed points don't say of themselves that they are Σ_n-true. However, if the canonical diagonal operator d is applied to $\sigma_n(x)$ we obtain a Σ_n-sentence that is a Σ_n-truth teller. Of course applying the canonical diagonal operator to $\neg\sigma_n(x)$ doesn't yield a liar sentence, because $d(\sigma_n(x)$ is Π_n but not Σ_n. Similar remarks apply to Π_n-truth tellers.

Truth tellers are of course similar to Henkin sentences, but their status is subject more to intensionality phenomena than Henkin sentences. The fixed points of canonical provability predicates or just those satisfying Löb's derivability conditions as meaning postulates are all equivalent by Löb's theorem, while the fixed points of the canonical partial truth predicates $\mathrm{Tr}_{\Sigma_n}(x)$ and $\mathrm{Tr}_{\Pi_n}(x)$ are

[22] In addition to the Tarski equivalences the partial truth and satisfaction predicates can be required to satisfy the compositional axioms for truth for all sentences of the relevant class of sentences.

not. We will show that the status of truth teller sentences is also more sensitive to the coding schema.

Before we turn to the canonical partial truth predicates and truth tellers obtained by the canonical diagonal operators, we look at the special case of Σ_1-truth, because in this case we still have formulae that express Σ_1-truth by the Kreisel Condition, that is, formulae weakly representing Σ_1-truth.

6.3. On Σ_1-truth

For Σ_1-sentences *truth* and *provability in sufficiently strong arithmetical systems* are coextensive properties. But we will show that applying the canonical diagonal operator to $\mathrm{Bew}_{\mathsf{I}\Sigma_1}(x)$ and $\mathrm{Tr}_{\Sigma_1}(x)$ yields a PA-provable and a PA-refutable sentence.

To this end we consider the pair of theories $\mathsf{I}\Sigma_1$ and PA. We stipulate that in our versions of $\mathsf{I}\Sigma_1$ and PA we have the recursion equations for all primitive recursive functions.

Let $\mathrm{Bew}_{\mathsf{I}\Sigma_1}(x)$ be a predicate naturally representing provability in $\mathsf{I}\Sigma_1$. Then $\mathrm{Bew}_{\mathsf{I}\Sigma_1}(x)$ is a truth predicate in PA for Σ_1-sentences in the following sense:

THEOREM 19. $\mathsf{PA} \vdash \mathrm{Bew}_{\mathsf{I}\Sigma_1}(\ulcorner\sigma\urcorner) \leftrightarrow \sigma$ for all Σ_1-formulae σ.

Proof. The left–to–right direction $\mathsf{PA} \vdash \mathrm{Bew}_{\mathsf{I}\Sigma_1}(\ulcorner\sigma\urcorner) \rightarrow \sigma$, that is, local reflection, is well-known and can be obtained by formalising the cut elimination theorem for $\mathsf{I}\Sigma_1$ in PA and proving reflection in the usual way outlined in e.g. Kreisel and Lévy (1968). See also Ono (1987). The right–to–left direction is formalised Σ_1-completeness. ⊣

For the left–to–right direction it is essential to work in a system that exceeds the strength of the system encoded in the provability predicate. This is not needed for the converse direction. Hence $\mathrm{Bew}_{\mathsf{I}\Sigma_1}(x)$ is a truth predicate for the set of Σ_1-sentences in PA but not in $\mathsf{I}\Sigma_1$.

The sentences $0 = 0$ and $0 = 1$ are fixed points of $\mathrm{Bew}_{\mathsf{I}\Sigma_1}(x)$ in PA. And thus fixed points of $\mathrm{Bew}_{\mathsf{I}\Sigma_1}(x)$ can be refutable or provable in PA. If we consider the predicate $\mathrm{Bew}_{\mathsf{I}\Sigma_1}(x)$ over $\mathsf{I}\Sigma_1$ the situation is dramatically different. We have:

THEOREM 20. Let σ be a Σ_1-sentence. Then the following are equivalent:

(i) σ is true.

(ii) $\mathsf{I}\Sigma_1 \vdash \sigma$.

(iii) $\mathsf{PA} \vdash \sigma$.

(iv) $\mathsf{I}\Sigma_1 \vdash \mathrm{Bew}_{\mathsf{I}\Sigma_1}(\ulcorner\sigma\urcorner) \leftrightarrow \sigma$

The easy proof uses Löb's theorem. The fixed point $d(\mathrm{Bew}_{\mathsf{I}\Sigma_1}(x))$ obtained from the predicate $\mathrm{Bew}_{\mathsf{I}\Sigma_1}(x)$ by the canonical diagonalization procedure is Σ_1; thus it is of the appropriate complexity and can be called a *truth teller sentence*. Since the fact that canonical diagonalization works can be verified in $\mathsf{I}\Sigma_1$, we find, by the above theorem, that $\mathsf{PA} \vdash d(\mathrm{Bew}_{\mathsf{I}\Sigma_1}(x))$. Hence there is a truth predicate for the set of Σ_1-sentences with a provable fixed point obtained by standard diagonalization and the existence of a provable Σ_1-truth teller is established.

In contrast, the fixed point obtained by canonical diagonalization of the usual Σ_1-truth predicate is refutable as we will show next.[23] We assume we use *a monotone Gödel coding*. By this we mean a coding where the code of a formula is greater than the code of all terms contained in it and where the code of a sequence is greater than the code of any member of the sequence and so on.

The truth predicate $\mathrm{Tr}_{\Sigma_1}(x)$ is of the form $\exists y\, \vartheta(y, x)$ for a formula $\vartheta(y, x)$ not containing any unbounded quantifier.[24] In a nutshell, $\exists y\, \vartheta(y, x)$ says that there is a sequence of triples of formulae, finite variable assignments, and truth values with certain properties. Suppose x is of the form $\ulcorner\exists\urcorner * v * z$, where '$*$' is our arithmetization of concatenation. In this case the prefinal element of the sequence y will be a triple $(s, z, \overline{1})$, where s codes an assignment that assigns a witness w of x to the variable v. An assignment is coded either as a finite set of pairs or as a sequence. In all cases we get: $w < s < y$. Thus the following assumption seems natural:

ASSUMPTION 21. If $\exists v\, \sigma(v)$ is a Σ_1-sentence, that is, if the formula $\sigma(v)$ contains no unbounded quantifier and only the variable v is free in $\sigma(v)$, then

[23] Added by Volker Halbach: Vann McGee and Albert Visser have independently communicated this observation to me.

[24] Our truth predicate could be a truth predicate for Σ_1-sentences narrowly defined, that is, of the form $\exists \vec{x}\, A$, where A is Δ_0, or of a more liberal kind, where these formulae are e.g. closed under conjunction and disjunction, etc.

$\mathsf{PA} \vdash \forall y\,(\vartheta(y, \ulcorner \exists v\, \sigma(v) \urcorner) \to \exists v < y\, \sigma(v))$ holds.

If we keep everything standard, then the Σ_1-truth teller becomes refutable in PA. For the proof we use assumption above, the monotonicity of the coding and that the fixed-point sentence satisfies the Kreisel–Henkin Criterion for self-reference, that is, it is obtained by a diagonal operator with the Kreisel–Henkin property in the sense of Definition 7. Of course the Gödel's canonical diagonal operator has this property.

THEOREM 22. Suppose we employ a standard, monotone Gödel coding. If d is a diagonal operator satisfying the Kreisel–Henkin Criterion, $\mathsf{PA} \vdash \neg\, d(\mathrm{Tr}_{\Sigma_1}(x))$ obtains.

Proof. The truth teller $d(\mathrm{Tr}_{\Sigma_1}(x))$ is of the form $\exists y\, \vartheta(y, t)$ where t is a term denoting this very sentence and $t = \ulcorner \exists y\, \vartheta(y, t) \urcorner$ is true and, hence, PA-provable.

We reason in PA. Suppose $\exists y\, \vartheta(y, t)$. Let y_0 be the smallest witness of $\exists y\, \vartheta(y, t)$. So (a) $\vartheta(y_0, t)$ and (b) $\forall z < y_0 \neg\, \vartheta(z, t)$. Since $t = \ulcorner \exists y\, \vartheta(y, t) \urcorner$, our assumption above combined with (a) gives us $\exists z < y_0\, \vartheta(z, t)$. But this contradicts (b). Hence our assumption that $\exists y\, \vartheta(y, t)$ must fail. ⊣

6.4. More truth tellers

In this section we look at partial truth predicates for Σ_n-sentences with $n > 1$ and Π_n-sentences with $n \geq 1$ and truth-teller sentences constructed from these truth predicates. To work comfortably with the wider class of truth predicates we employ PA with the recursion equations for for all primitive recursive functions here.

Theorem 22 can be generalized to Σ_n with $n > 1$.[25] If the truth predicate for Σ_{n+1} is constructed in a fairly straightforward way, it will be of the form $\exists y \vartheta(y, x)$ where y ranges over a k-tuple typically having a witness as a component; y may range of over triples having a variable assignment, a formula and a truth value as components. As in the case of Σ_1-truth, ϑ will then we will have for Σ_{n+1} sentences $\exists v\, \sigma(v)$ the following property analogous to Assumption 21:

[25] We thank Graham Leigh for pointing out to us that Theorem 22 generalizes to higher n for many reasonable truth predicates.

ASSUMPTION 23. $\mathsf{PA} \vdash \forall y\,(\vartheta(y, \ulcorner \exists v\, \sigma(v) \urcorner) \to \exists v < y\, \sigma(v))$ obtains for all Σ_n-formulae $\sigma(v)$ with at most v free.

This assumption is more problematic for Σ_n with $n > 1$ than that for Σ_1. There are reasonable Σ_{n+1}-truth predicates that do not satisfy this condition as we shall show in a moment. However, if we make the same assumptions again, we can generalize the above result, using the same proof idea:

THEOREM 24. Suppose we employ a standard monotone Gödel coding. If d is a diagonal operator satisfying the Kreisel–Henkin Criterion, $\mathsf{PA} \vdash \neg\, d(\mathrm{Tr}_{\Sigma_{n+1}}(x))$ obtains.

Before constructing a counterexample to Assumption 23, we look at the behaviour of the canonical Π_n-truth tellers.

Let $\sim \varphi$ be result of 'pushing in' the negation symbol in $\neg\varphi$ as far as possible and possibly deleting double occurrences of $\neg$, so that the negation symbol is only in front of atomic formulae. If φ is in prenex normal form, then $\sim \varphi$ is in prenex normal form and logically equivalent to $\neg\varphi$. We write $\dot{\sim}$ for the function symbol naturally corresponding to $\sim$.

Given a truth predicate $\mathrm{Tr}_{\Sigma_n}(x)$ for Σ_n-sentences, we define a corresponding Π_n-truth predicate $\mathrm{Tr}_{\Pi_n}(x)$ as $\sim \mathrm{Tr}_{\Sigma_n}(\dot{\sim} x)$.[26] If Tr_{Σ_n} is in prenex form, then Tr_{Π_n} is also in prenex normal form. If Tr_{Σ_n} is a Σ_n-truth predicate in the sense that $\mathsf{PA} \vdash \mathrm{Tr}_{\Sigma_n}(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$ holds for all Σ_n-sentences, then Tr_{Π_n} is a Π_n-truth predicate as well. Under the Assumption 23 the Π_n-truth tellers are provable. The proof is a variation of the proof of Theorems 22 and 24.

THEOREM 25. Suppose we employ a standard, monotone Gödel coding. If d is a diagonal operator satisfying the Kreisel–Henkin Criterion and Tr_{Π_n} is defined as described above, $\mathsf{PA} \vdash d(\mathrm{Tr}_{\Pi_n}(x))$ obtains for all $n > 0$.

Proof. We reason in PA. Assume $\neg d(\mathrm{Tr}_{\Pi_n}(x))$, that is, $\neg \mathrm{Tr}_{\Pi_n}(t)$ for some term t with $t = \ulcorner \mathrm{Tr}_{\Pi_n}(t) \urcorner$. Using the definition of Tr_{Π_n} we conclude $\neg \sim \mathrm{Tr}_{\Sigma_n}(\dot{\sim} t)$ and, by logic, $\mathrm{Tr}_{\Sigma_n}(\dot{\sim} t)$. Assume $\mathrm{Tr}_{\Sigma_n}(x)$ is of the form $\exists y \vartheta(y, x)$, then we have $\exists y\, \vartheta(y, \dot{\sim} t)$. Therefore there is a minimal y_0 such that $\vartheta(y_0, \dot{\sim} t)$ and thus

[26] One may want to modify $\mathrm{Tr}_{\Pi_n}(x)$ so that it is provable false of all sentences not in Π_n. This doesn't affect the argument below.

$\vartheta(y_0, \dot{\sim}\ulcorner \mathrm{Tr}_{\Pi_n}(t)\urcorner)$, which is $\vartheta(y_0, \dot{\sim}\ulcorner \sim \mathrm{Tr}_{\Sigma_n}(\dot{\sim} t)\urcorner)$ and $\vartheta(y_0, \dot{\sim}\ulcorner \sim \exists y\, \vartheta(y, \dot{\sim} t)\urcorner)$. Using Assumption 23 we conclude $\exists v < y_0\, \vartheta(v, \dot{\sim} t)$. This contradicts the assumption that y_0 is minimal. ⊣

We claimed that for $n > 1$ there are reasonable Σ_n-truth predicates such that Assumption 23 is not satisfied. We show how to define a Σ_{n+1}-truth predicate $\exists y \vartheta(y, x)$ from Tr_{Σ_n} such that the following holds for all Σ_{n+1}-sentences:

$$\mathsf{PA} \vdash \forall y\, (\vartheta(y, \ulcorner \exists v\, \varphi\urcorner) \leftrightarrow \varphi(y)) \tag{8}$$

So a witness for a Σ_{n+1}-sentence is also a witness for its Σ_{n+1}-true and vice versa. So clearly Assumption 23 is violated because under this assumption the smallest witness for a proper Σ_{n+1}-sentence always has to be smaller than any witness for its truth. We will require $\exists y \vartheta(y, x)$ to be a Σ_{n+1}-formula (and thus to be in prenex form).

The Σ_{n+1}-truth predicates we define sensibly apply only to sentences in prenex normal form. The Σ_{n+1}-truth predicate will be in prenex form. So we can simply concentrate on sentences in prenex form and we will still be able to formulate truth teller sentences. To obtain more general truth predicates that apply to other sentences not in prenex form further tricks would have to be applied. Like above, the coding schema is assumed to be monotone.

Assume we are given a Π_n-truth predicate. This can be $\sim \mathrm{Tr}_{\Sigma_n}(\dot{\sim} x)$ as above. The idea is to define Σ_{n+1}-truth of a Σ_{n+1}-sentence $\exists v \psi(v)$ as the claim that there is a Π_n-true instance of $\psi(n)$. So $\mathrm{Tr}_{\Sigma_{n+1}}(x)$ will be defined as a formula equivalent to

$$\exists y\, \forall v < x\, \forall a < x \left(x = \dot{\exists} v a \rightarrow \mathrm{Tr}_{\Pi_n}(a(\dot{y}/v))\right). \tag{9}$$

Here $x = \dot{\exists} v a$ expresses that x is a sentence and the existential quantification of the formula a with respect to the variable v; $a(\dot{y}/v)$ stands for the result of formally substituting the variable v with the numeral of y.

The formula (9) itself cannot serve as Σ_{n+1}-truth predicate because it is not yet in prenex form. The unrestricted quantifiers in $\mathrm{Tr}_{\Pi_n}(a(\dot{y}/v))$ need to be moved in front of the bounded quantifiers.[27] For the universal quantifiers this

[27] Alternatively we can let the bounded quantifiers be 'eaten' by the outer universal quantifier

is straightforward. For existential quantifiers (in the case $n > 2$) the collection principle can be employed. For this we need the appropriate instances of the induction schema, which are all available in PA.

Thus for $n > 1$ we have constructed Σ_n- and, if we use the tricks from above again, also Π_n-truth predicates that do not conform to Assumption 23. Defining truth predicates along the lines of (9) also doesn't appear to be too artificial. Moreover, that the witness of an existential formula and the witness for the claim that it is true are the same may be seen as a desirable feature of a truth predicate. Consequently one may conjecture that Theorems 24 and 25 depend on somewhat arbitrary features of the partial truth predicates. If the partial truth predicates are defined in the way just outlined and these features are removed, we don't know whether the corresponding Σ_n-truth teller remain refutable and the Π_n-truth tellers provable. It seems that another proof idea would be required and thus the problem of arithmetical truth tellers leaves some open questions, even for quite natural partial truth predicates and canonical diagonalization.

We conclude this section with an application of the Kreisel–Henkin trick from Kreisel's Observation to partial truth instead of provability. If we consider diagonal sentences that are not obtained by the standard diagonal operator and deviant partial truth predicates, Henkin's trick from in our proof of Kreisel's Observation can be applied again to produce another example of the intensionality of self-reference.

Henkin's trick the yields a Σ_n-truth predicate with a provable and a refutable truth teller.

OBSERVATION 26. Assume again that a standard, monotone Gödel coding is used. For each n there is a Σ_n-truth predicate $\sigma_n(x)$, a sentence τ_1 and a sentence τ_2 such that both sentences τ_1 and τ_2 ascribe to themselves the property expressed by $\sigma_n(x)$ by the Kreisel–Henkin Criterion and τ_1 is provable while τ is refutable.

Proof. Let $\mathrm{Tr}_{\Sigma_n}(x)$ be some Σ_n-truth predicate satisfying Assumption 23 and

of Tr_{Π_n} using the fact that we have a pairing function. Yet alternatively we can replace $a(\dot{y}/v)$ by a function that delivers the appropriate substitution instance of a if a is of the right form and $0 = 1$ otherwise and drop the bounded quantifiers entirely.

write the formula $x = x \vee \mathrm{Tr}_{\Sigma_n}(x)$ in strict Σ_1-form and call the resulting formula $(x = x \vee \mathrm{Tr}_{\Sigma_n}(x))'$. By Gödel's diagonal lemma there is a term t with the following property:

$$\mathsf{PA} \vdash t = \ulcorner (t = t \vee \mathrm{Tr}_{\Sigma_n}(t))' \urcorner$$

Now the predicate $\sigma_n(x)$ is defined as $(x = t \vee \mathrm{Tr}_{\Sigma_n}(x)))'$. It's easy to verify that $\sigma_n(x)$ is a Σ_n-truth predicate, that is, $\mathsf{PA} \vdash \sigma_n(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$ for all Σ_n-sentences φ. In particular, if φ is $t = t \vee \mathrm{Tr}_{\Sigma_n}(t)$, then both sides of the equivalence are obviously provable and therefore the equivalence is provable.

As the sentence $(t = t \vee \mathrm{Tr}_{\Sigma_n}(t))'$ is provable and satisfies the Kreisel–Henkin Criterion, it can serve as τ_1.

If the canonical diagonal operator d is applied to the formula $(x = t \vee \mathrm{Tr}_{\Sigma_n}(x))'$ one can show that $d((x = t \vee \mathrm{Tr}_{\Sigma_n}(x))'$ is refutable by Theorem 24. ⊣

The second part of the proof of Kreisel's Observation can be used to produce analogous examples for Π_n-truth predicates.

6.5. *Nonstandard truth predicates done in a different way*

After having shown that by only varying the method of obtaining a truth teller one can obtain provable and refutable truth tellers, we are going to show in this next section that by merely varying the truth predicate but adhering to the canonical diagonal operator d one can obtain provable and refutable truth tellers. This is achieved by applying the the results of Section 5 to truth predicates.

We remind the reader of a result of Section 5. Suppose Σ extends Basic. Let γ be a sentence. For any formula φ, we constructed a formula φ^γ with the following properties, that is, Lemmata 8 and 9:

1. $\Sigma \vdash x \neq d(\ulcorner \varphi^\gamma(x) \urcorner) \rightarrow (\varphi^\gamma(x) \leftrightarrow \varphi(x))$.
2. $\Sigma \vdash d(\varphi^\gamma(x)) \leftrightarrow \gamma$.

Our formula φ^γ has to to satisfy Equation 6, to wit:

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

Our main desideratum is that if φ and γ are Σ_n (Π_n), then so is φ^γ. Fortunately this is easily arranged. We can rewrite the formula

$$\big(x \neq \dot{d}(y) \wedge \varphi(x)\big) \vee \big(x = \dot{d}(y) \wedge \gamma\big)$$

in the prescribed strict Σ_n- (Π_n-) form, say obtaining $\eta(x, y)$ and then apply the canonical diagonal construction w.r.t. y to $\eta(x, y)$, The resulting formula will still have the strict Σ_n– (Π_n-) form.

We consider $\mathrm{Tr}^\gamma_{\Sigma_n}$. We assume that Tr_{Σ_n} is in the strict Σ_n form and that γ is Σ_n. Consider α in Σ_n. Let $\beta \coloneqq d(\mathrm{Tr}^\gamma_{\Sigma_n}(x))$. If $\alpha \neq \beta$, we find:

$$\Sigma \vdash \mathrm{Tr}^\gamma_{\Sigma_n}(\ulcorner \alpha \urcorner) \leftrightarrow \mathrm{Tr}_{\Sigma_n}(\ulcorner \alpha \urcorner)$$

and hence $\Sigma \vdash \mathrm{Tr}^\gamma_{\Sigma_n}(\ulcorner \alpha \urcorner) \leftrightarrow \alpha$. If $\alpha = \beta$, we find by the fixed point property: $\Sigma \vdash \mathrm{Tr}^\gamma_{\Sigma_n}(\ulcorner \alpha \urcorner) \leftrightarrow \alpha$. So $\mathrm{Tr}^\gamma_{\Sigma_n}$ is a truth predicate for Σ_n. Moreover we have: $\Sigma \vdash \mathrm{Tr}^\gamma_{\Sigma_n}(\ulcorner \beta \urcorner) \leftrightarrow \gamma$, and hence $\Sigma \vdash \beta \leftrightarrow \gamma$. Thus, for a given diagonal operator d, we can find a Σ_n-truth predicate σ_n such that $d(\sigma_n(x))$ is provable, refutable or undecidable via appropriate choices for γ. Similarly for the Π_n-case.

So as in the case of provability, we have been able to eliminate the simultaneous use of a deviant diagonal operator *and* a non-canonical formula expressing a property in order to show that a truth teller can be provable or refutable. We only need a deviant partial truth predicate in order to show that a Σ_n-truth teller can also be provable.

7. *Uniform diagonal operators*

We think that Gödel's diagonalization method and certain variants of it produce paradigmatic self-referential sentences, *if* there is any self-reference in metamathematics at all. The philosophical challenge is to explain why these fixed point yield self-reference, while certain other fixed points do not.

The Kreisel–Henkin Criterion provides at least a partial explanation in terms of reference. If a sentence $\varphi(t)$ satisfies the Kreisel–Henkin Criterion then t refers to (the code of) the sentence $\varphi(t)$ and the theory can prove this in the sense that $\Sigma \vdash t = \ulcorner \varphi(t) \urcorner$. The criterion also allows us to generalize certain results. To show the refutability of the Σ_1-truth teller in Theorem 22, we don't

have to retreat to the canonical diagonal operator, but can prove a claim about all fixed points satisfying the Kreisel–Henkin Criterion. We think that this kind of generalization should increase the significance of the results, because it shows that the result does not depend on petty details of the diagonalization method. In this sense it also allows one to extensionalize results to a certain degree.

However all this does not imply that the Kreisel–Henkin Criterion is also an adequate analysis of self-reference or, more precisely, self-attribution of properties. In this respect the Kreisel–Henkin Criterion may be similar to Kreisel's Condition: Kreisel's Condition, that is, weak representability is hardly a satisfactory analysis of what it means for a formula to express provability. It is at best a necessary condition for the expression of provability in sound systems. For many purposes we just need to assume that a formula satisfies Kreisel's Condition and need not care whether the formula 'really' expresses provability. In the same way, all sentences of the form $\varphi(t)$ truly saying about themselves that they have the property expressed by $\varphi(x)$ will satisfy the Kreisel–Henkin Criterion and therefore all results on fixed-points with the Kreisel–Henkin Criterion will include and apply to the truly self-referential fixed points.

But there remain also doubts that *all* fixed points $\varphi(t)$ satisfying the Kreisel–Henkin Criterion really say of themselves that they have the property expressed by $\varphi(x)$. The criterion doesn't rule out certain 'deviant' fixed points. The sentence $\mathrm{Bew}_{\mathrm{II}}(t_2)$ from Kreisel's Observation is refutable, while applying the canonical diagonal operator to $\mathrm{Bew}_{\mathrm{II}}(x)$ yields a provable fixed point, as has been shown in Observation 3. The fixed point $\mathrm{Bew}_{\mathrm{II}}(t_2)$ is not arrived at from the predicate $\mathrm{Bew}_{\mathrm{II}}(x)$ by a slight variant of the canonical diagonalization method, but rather $\mathrm{Bew}_{\mathrm{II}}(x)$ has been been constructed in a such a way that t_2 'happens' to be a fixed point satisfying the Kreisel–Henkin Criterion. The sentence $\mathrm{Bew}_{\mathrm{II}}(t_2)$ surely refers to (the code of) a sentence that happens to be $\mathrm{Bew}_{\mathrm{II}}(t_2)$. But we are not sure whether that implies that $\mathrm{Bew}_{\mathrm{II}}(t_2)$ states of itself that it has the property expressed by the formula $\mathrm{Bew}_{\mathrm{II}}(x)$.[28]

Thus among the fixed points satisfying the Kreisel–Henkin Criterion one can

[28] One could also distinguish between two kinds of self-reference: $\mathrm{Bew}_{\mathrm{II}}(t_2)$ would be *de iure* self-referential while the application of the canonical diagonal operator to a formula gives what one could call *de facto* self-reference. The distinction is reminiscent of van Fraassen's (1970) betweeen *accidental* and *functional* self-reference in natural language.

still distinguish between more or less contrived. What seems to be deviant about $\mathrm{Bew}_{\mathrm{II}}(t_2)$ is, very loosely speaking, that it hasn't been arrived at by some general method that yields applied to a formula a fixed point. In order to obtain more robust results, we may try to impose more conditions on the fixed points beyond the Kreisel–Henkin Criterion that rule out sentences such as $\mathrm{Bew}_{\mathrm{II}}(t_2)$.

The following condition is supposed to bring out a nice feature of Gödel's canonical fixed points that is lacked by Kreisel's sentence $\mathrm{Bew}_{\mathrm{II}}(t_2)$.

DEFINITION 27. A diagonal operator d is *uniform* iff the following condition is satisfied for each $\varphi(x)$ with a designated variable x free:

$d(\varphi)$ is of the form $\varphi(\dot{d}\ulcorner\varphi\urcorner)$, where $\dot{d}$ represents the function d.

Of course, every uniform diagonal operator d has the Kreisel–Henkin property, that is, the following claim holds:

$$\Sigma \vdash \dot{d}\ulcorner\varphi\urcorner = \ulcorner\varphi(\dot{d}\ulcorner\varphi\urcorner)\urcorner \tag{10}$$

holds.[29] Clearly, the canonical Gödelian diagonal operator is uniform. Kreisel's sentence $\mathrm{Bew}_{\mathrm{II}}(t_2)$, in contrast, cannot be the result of applying a uniform diagonal operator to $\mathrm{Bew}_{\mathrm{II}}(x)$, if the coding is monotone.

Since it doesn't matter for the present purposes that we are dealing with provability predicates, we slightly slightly generalize. The example below can be applied to $\mathrm{Bew}_{\mathrm{II}}(t_2)$ by taking $\varphi(t, x)$ as $x \neq t_2 \wedge \mathrm{Bew}(x)$.

OBSERVATION 28. Let the coding be monotone, t be some term and $\varphi(x, x)$ a formula with two marked (strings of) free occurrences of the variable x. If d is a diagonal operator with $d(\varphi(t, x)) = \varphi(t, t)$, then d is not uniform.

Proof. Assume d is uniform, that is, $d(\varphi(t, v))$ is the formula $\varphi(t, \dot{d}\ulcorner\varphi(t, v)\urcorner)$. By assumption $d(\varphi(t, v)$ is the formula $\varphi(t, t)$. Since then $\varphi(t, \dot{d}\ulcorner\varphi(t, v)\urcorner)$ and $\varphi(t, t)$ are identical expressions, the term t must be the expression $\dot{d}\ulcorner\varphi(t, v)\urcorner$. Hence the term t would contain a numeral for a formula that contains in turn the term t, contradicting monotonicity of coding. ⊣

[29] Cf. also Heck's (2007, p. 9) *Structural Diagonal Lemma*.

If the diagonal operator yields only a fixed point satisfying the Kreisel–Henkin Criterion then t and $\dot{d}^{\ulcorner}\varphi(t,v)^{\urcorner}$ coincide in their values, but they don't have to be the same expression.

We don't expect that, by imposing the uniformity condition, all pathological fixed points can be ruled out. But uniformity may be a first hint to narrow down the choice of diagonal operators, if a result cannot be proved for all fixed points satisfying the Kreisel–Henkin Criterion.

8. Self-reference in other languages

In this paper we have stayed within the realm of arithmetic and, even more specifically, in systems with function symbols for all primitive recursive functions and their defining equations as axioms. Self-reference has been discussed in many other settings: The language may lack appropriate functions symbols or the language may contain symbols going beyond that of arithmetic by containing additional symbols such as a primitive new symbol for truth. Then there are of course theories, set theory being an example, that contain arithmetic only via some interpretation. Obviously questions of the kind we have studied in this paper arise in such settings as well. Here in this section we touch only at some problems and possibilities of generalizing some of our remarks to other settings.

First we look at systems that do not feature function symbols for sufficiently many primitive recursive functions. The Kreisel–Henkin Criterion for self-reference and the improved versions of it in the previous section provide only a sufficient condition for a sentence to ascribe some property to itself. That condition can only be met when suitable closed terms are available. The canonical construction of the strong diagonal lemma with a closed term relies on a function expression for the substitution function. Of course such an expression is not available in the usual language of Peano arithmetic featuring only 0, S, + and $\times$ as function symbols. Even in such languages there can be sentences that are self-referential in virtue of the Kreisel–Henkin Criterion. In fact, even if addition and multiplication are expressed by predicates, numerals alone, that is, the symbol for zero and successor suffice if the coding is carefully chosen. In Appendix A we construct a Gödel coding along with a diagonal operator with the Kreisel–Henkin property that relies only on numerals as the terms. How-

ever, if a monotone coding is employed and the language doesn't contain the appropriate function symbols – like the the usual language of PA –, then there are no formulae that ascribe to themselves any property in virtue of the Kreisel–Henkin Criterion.

Under a monotone coding sentences that are usually thought to be recognizable as Gödel, Henkin and truth-teller sentences can still be constructed, even if the appropriate function symbols and thus a diagonal operators with the Kreisel–Henkin property are absent, as is the case in PA or Zermelo–Fraenkel set theory. In such systems self-reference will be achieved via quantification and the formulae cannot ascribe to themselves any property in virtue of the Kreisel–Henkin Criterion via terms. It may be surmised that the Kreisel–Henkin Criterion already captures an important aspect of self-reference in arithmetic and thus one might try to generalize to the Kreisel–Henkin Criterion to sentences φ that do not contain a term t having φ as its value by the following stipulation: A formula ψ obtained from a formulae $\varphi(t)$ by eliminating the function symbols in $\varphi(t)$ ascribes to itself the same properties as $\varphi(t)$. By eliminating the function symbols we mean one of the usual methods of reformulating a formula that contains function symbols with a provably equivalent formula where the functions are expressed using quantification, other function symbols and appropriate predicate expressions. However, self-reference is too intensional and not preserved under this transformation. The new formula ψ will still refer to $\varphi(t)$ and not to itself, except in some special fortunate cases. Therefore one will have to adapt the method of extension in a more sophisticated way.[30]

However, we don't assume that such an elimination necessarily preserves self-reference, even if carried out properly, and, more generally, that the same self-referential properties are shared by all provably equivalent sentences. At any rate we don't see an obvious way to generalize the Kreisel–Henkin Criterion to sentences without appropriate closed terms.[31]

[30] Of course there are fully relational versions of the Gödel Fixed-Point lemma but there is not clear reason to consider those as self-referential.

[31] Heck (2007) has raised some worries about the possibility of appropriately expressing self-reference in languages lacking the function symbols and concludes on p. 1 that '[t]rue self-reference is possible only if we expand the language to include function-symbols for all primitive recursive functions.'.

We now turn from languages that are properly contained in those of Basic to languages that properly extend its language. Many remarks carry over in a straightforward way. In particular, we think that many of our remarks to extensions of the language with a new primitive unary predicate for truth or necessity. In the discussion of the truth-theoretic paradoxes extensional results often suffice. For instance, the proof of the inconsistency of the full T-schema merely requires a fixed point of the negation of the truth predicate. However, to arrive at certain truth-theoretic paradoxes, it is not sufficient to work with an arbitrary diagonal operator, because in some cases one will require at least a sentence that is self-referential according to the Kreisel–Henkin Criterion. Heck (2007, section 3.2) presents an example. Further examples can be extracted from Burgess (1986) and (Halbach, 1994, p. 313). So at least some of the phenomena we have studied here arise also in these wider contexts.

9. *Summary: Henkin sentences and truth tellers*

Before trying to draw some preliminary conclusions from our observations, we summarize some of our observations on Henkin sentences and truth tellers in two tables.

First we turn to Henkin sentences. The formulae $\mathrm{Bew}_{\mathrm{II}}$, $\mathrm{Bew}_2(x)$, $\mathrm{Bew}_3(x)$ and Bew^R are defined from another provability predicate. We assume that the canonical provability predicate is used for this purpose. In the first column we list various provability predicates. They all express provability in Σ in the sense of Kreisel's Condition, that is, they weakly represent Σ-provability. In the other three columns we describe how different fixed points of these formulae behave. The single letter *p* means that fixed points of the respective kind are provable in Σ; *p,r* means that there are fixed points of this kind that are Σ-provable and others that are Σ-refutable, and so on. The letter *i* stands for *independent* from Σ. The theory Σ is an extension of Basic containing at least S^1_2.

	canonical fixed points	fixed points with the Kreisel–Henkin property	arbitrary fixed points
Bew (canonical provability)	p	p	p
$\mathrm{Bew}_{\mathrm{II}}$ (Kreisel–Henkin)	p	p,r	p,r
Bew_2 (Theorem 4)	r	p,r	p,r
Bew_3 (Theorem 5)	i	p,i	p,i
Bew^R (Rosser provability)	?	?	p,r,?

In the following table we summarize some results on partial truth tellers. All the formulae in the table are partial truth predicates for the class of sentences indicated there in the sense that the T-sentences are provable for all sentences in the given class. A monotone coding schema is assumed. In contrast to the above table the letter *p* now stands for *provable in Peano arithmetic*. The formulae Tr_{Σ_n} and Tr_{Π_n} are the canonical partial truth predicates for $n \geq 1$; the formulae Tr_{Σ_n} are assumed to satisfy the Assumption 23, and the formulae Tr_{Π_n} are defined from them in the way indicated in Section 6.4. The formula $\mathrm{Bew}_{\mathrm{I}\Sigma_1}$ is seen here as a truth predicate for Σ_1-sentences and restricted to such. All fixed points are assumed to be of the relevant complexity.

	canonical fixed points	fixed points with the Kreisel–Henkin property	arbitrary fixed points
$\mathrm{Bew}_{\mathrm{I}\Sigma_1}$	p	p	p,r
Tr_{Σ_n}	r	r	p,r,i
Tr_{Π_n}	p	p	p,r,i
σ_n as in Observation 26	r	p,r	p,r,i
$\mathrm{Tr}^{\gamma_1}_{\Sigma_n}$	p	p,r	p,r,i
$\mathrm{Tr}^{\gamma_2}_{\Sigma_n}$ $(n \geq 2)$	i	i,r	p,r,i

The truth predicates $\mathrm{Tr}^{\gamma_1}_{\Sigma_n}$ and $\mathrm{Tr}^{\gamma_2}_{\Sigma_n}$ have been defined for arbitrary γ in Section 6.5. Any refutable Σ_n-sentence can serve as γ_1; γ_2 is an independent Σ_n-sentence and thus we must assume $n \geq 2$ for this case.

10. Self-reference and intensionality

Although an abundance of claims about sentences ascribing to themselves such properties as truth, falsity or provability can be found in the literature, there is no generally accepted definition of which sentences qualify as self-referential. One possible reaction could be a rejection of the talk about such sentences. However, this would mean that large parts of philosophical logic, philosophy of logic and philosophy of mathematics would have to be rejected. Self-reference and self-predication isn't more elusive than many other notions in the area. If we were to ban intensional notions from these areas, then many notions including that of provability would have to be declared illegitimate as well, because we are not able to define extensionally what it means for a formula to express

provability. The second and third source of intensionality are on a par in this respect.

There are not only philosophical but also mathematical reasons for retaining the notion of self-reference: Questions about self-referential statements have driven progress in logic. They are at the root of Gödel's theorems and Gödel arrived, for all we know, at his proof by thinking about self-reference and self-predication. As mentioned above, logicians became more suspicious of self-reference, and some dismissed questions about sentences stating something about themselves as hopelessly intensional. If Löb, however, had adopted such a sceptical attitude and rejected Henkin's problem as irremediably flawed, he probably would not have proved his theorem. As so often, philosophical notions defying a full formal analysis function as an engine driving progress in logic and, more generally, mathematics and the sciences. Therefore they should not be dismissed, even if they prove somewhat elusive.

It may be hoped that one can escape the problems of intensionality by settling for the 'canonical' methods, that is, a canonical coding, a canonical way of expressing the property under consideration and the canonical way of obtaining a fixed point. But it's far from from clear what the canonical methods are. There are many reasonable codings, different sensible ways to express provability, Σ_n-truth and so on and even on the canonical proof of the diagonal lemma there are variations. It would be odd, however, if the answer to question about the status of self-referential statements depended on the historic development of mathematics and what is seen as the standard proof. It's a challenge to explain what makes the canonical choices most relevant for answering questions about the status of sentences that ascribe certain properties to themselves. The significance of the results that presuppose canonical constructions is limited, if they are just taken as the ones most logicians happen to work with. We need an explanation why these canonical methods yield paradigmatic examples of sentences that ascribe a property to themselves.

Moreover, even partial analyses of self-reference may be useful in generalizing results: In Theorem 24 on the provability of Σ_n-truth tellers we assume that the fixed points of the truth predicates have the Kreisel–Henkin property. Of course we could have proved the result only for the 'canonical' fixed point, but generalizations of this kind seem desirable in the same way generalizations of

the incompleteness theorems in (Feferman, 1960) with respect to the second source intensionality, that is, expressing a property, proved fruitful. Thus we believe that we should strive for an analysis of what it means for a formula to state a property of itself.

Even without a full formal analysis of self-reference in formal systems, many questions about the status of sentences ascribing to themselves a certain property can be answered. This is the case when we know that all sentences that ascribe to themselves a certain property are contained in a certain class of sentences and we can prove a general result about that set. So a sufficient condition for self-reference can enable us to settle a question on self-referential sentences. Canonical provability is a case where a very weak necessary condition will suffice: Once we settle for a provability predicate satisfying the Löb derivability conditions, Löb's theorem applies and all fixed points of the provability predicate are provable. Since all sentences stating their own provability are fixed points, all such sentences are provable.

Therefore when one asks about the status of the sentence that states its own provability (in the fixed system), the intensionality of that question lies solely in the way we express the property of Σ-provability, if we stipulate that provability has to be expressed by a predicate satisfying the Löb derivability conditions. Hence it is not surprising that logicians have focused on the *second* source of intensionality, that is, intensionality arising from the problem of expressing a property in the formal language. Once deviant provability predicates are excluded, the third source of intensionality, that is, intensionality of self-reference, is also blocked.

The case of provability, however, is very special. In most other cases we cannot escape the intricacies of the intensionality of self-reference as easily: Fixed points of a formula will usually behave in different highly disparate ways. Of course, this is most obvious for any formula expressing some truth-like concept, but also for Rosser-provability, as noted in Observation 17. In order to obtain results ascribing properties to themselves via such formulae, a better analysis of self-reference is needed. The Kreisel–Henkin Criterion, as we have formulated it, aims to provide a sufficient condition for self-reference, because it only applies to formulae in which self-reference is attained in virtue of a closed term. But it's status remains controversial. Some, like Heck (2007), are inclined to

think that sentences to which the Kreisel–Henkin Criterion isn't applicable cannot be truly self-referential, so it may well be a necessary *and* sufficient condition. However, we still have serious doubt whether it even provides a sufficient condition for self-reference. There is surely something awkward about fixed points with the Kreisel–Henkin property that have been obtained using Kreisel's (1953) trick. If one shares our scepticism concerning the Kreisel–Henkin Criterion as a necessary condition for self-reference, then uniform diagonal operators may bring us closer to an interesting and formally useful necessary condition for self-reference.

The significance of some of our results depends on whether the Kreisel–Henkin Criterion provides a sufficient condition or even an adequate definition of self-reference. For instance, in Theorem 24 on the provability of Σ_n-truth tellers for $n \geq 1$ we assume that the fixed points of the partial truth predicates have the Kreisel–Henkin property. If all sentences that say about themselves that they are Σ_n-true do so in virtue of the Kreisel–Henkin Criterion, then Theorem 24 answers the question about whether Σ_n-truth tellers are provable, if Assumption 23 and the monotonicity of coding are accepted.

Theorems 24 and 25 on partial-truth tellers demonstrate the sensitivity of questions about self-referential statements to all three sources of intensionality for such formulae. In the two theorems we make assumptions on the coding, specific properties on the formulae expressing Σ_n- or Π_n-truth and, of course, on the fixed points of these formulae. These theorems are in stark contrast to Löb's theorem, which is much more robust and doesn't rely on such specific assumptions.

At least Theorems 24 and 25 are not sensitive to which diagonal operator with the Kreisel–Henkin property is used. There are, however, formulae that are sensitive to exactly which diagonal sentences with the Kreisel–Henkin property are used. In Observation 3 it was noted that the formula $\mathrm{Bew}_{\mathrm{II}}$ does have a diagonal sentences $\mathrm{Bew}_{\mathrm{II}}(t_2)$ and $\mathrm{Bew}_{\mathrm{II}}(t)$ with the Kreisel–Henkin property, of which the first is provable and the second refutable. Observation 26 contains an analogous result for a truth predicate σ_n. However, both formulae $\mathrm{Bew}_{\mathrm{II}}$ and σ_n are contrived and the question arises whether there are natural formulae φ expressing an interesting and relevant property such that φ has two fixed points with the Kreisel–Henkin property of which one is provable and the other refutable.

The problem is highly intensional and hard to make more precise.

We give an example of formula that we could not accept as an example that would support an affirmative answer to our question. Consider a formula $\zeta(x)$ expressing the property that x doesn't contain an occurrence of the successor symbol such that there is no occurrence of the successor symbol in $\zeta(x)$ itself. If we apply the canonical diagonal operator to such a formula, we obtain a formula $\zeta(t)$ where t contains the successor symbol, because there will be occurrence of numerals of Gödel codes in t. Thus $\zeta(t)$ will be refutable. However, we can avoid the use of numerals and the successor symbol and use other closed terms instead; this is possible if we have function symbols for all primitive functions in our language. The resulting fixed point will be provable. However, this formula does not express an interesting and relevant property in the sense of the question. We have to look for less trivial examples. Perhaps there are no such examples, because they are either 'trivial' like $\zeta(x)$ or some trivializing condition is built into them like in $\mathrm{Bew}_{\mathrm{II}}$ and σ_n. Further work is required here.

So far we have looked at the prospects and possibilities of passing from an intensional question about sentences ascribing some property to themselves to formal extensional theorems. There are also question in the opposite direction: Given the provability or refutability of a supposedly self-referential statement, which factors can be tweaked in order to arrive at a different conclusion? By investigating this kind of question, new insights into the robustness of results and the relation between the different sources of intensionality can be gained. For instance, it is known that many intensionality phenomena from the second source can be built into the coding. The proofs of Theorems 24 and 25 on partial truth tellers rely on the use of a monotone coding schema. It may well possible to obtain provable Σ_n-truth tellers by using a suitable non-monotone coding. This would provide us with another example of intensionality arising from the first source.

Here in this paper we have somewhat suppressed intensionality from coding and focused on intensionality from the second and third source. In particular, we have established that in some cases intensionality the second source will suffice to change results. Kreisel (1953) obtained provable and refutable Henkin sen-

tences by exploiting simultaneously the second *and* third source of intensionality. Observation 26 shows that at least with a deviant provability predicate, we can obtain provable and refutable Henkin sentences by changing the method of diagonalization (without losing the Kreisel–Henkin property). Finally, in Section 5 it was established that by changing the formula expressing provability but using only the canonical diagonal operator, one can obtain provable and refutable Henkin sentences. Section 6.5 contains analogous results for partial truth predicates. It would be interesting to see under which circumstances it is possible to shift the effects of intensionality from one source to the other. We know already that there are limits: Provable and refutable Henkin sentences can be obtained by using different provability predicates, but once the canonical provability predicate is fixed, changing the diagonal operator won't affect the provability of the Henkin sentence.

At least we also provided a plethora of entertaining examples that show that the analysis of self-reference in arithmetic is not as straightforward as it may appear. There still are some mathematical and philosophical questions concerning formulae that as described as making statements about themselves; the answers may be both interesting and fruitful, just as Löb's answer to Henkin's question was.

References

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

George Boolos. *The Logic of Provability*. Cambridge University Press, Cambridge, 1993.

John P. Burgess. The truth is never simple. *Journal of Symbolic Logic*, 51:663–81, September 1986.

Solomon Feferman. Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, 49:35–91, 1960.

Solomon Feferman. Intensionality in mathematics. *Journal of Philosophical Logic*, 14:41–55, 1985.

Bas C. Van Fraassen. Inference and self-reference. *Synthese*, 21:425–438, 1970.

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

Kurt Gödel. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik*, 38-38:173–198, 1931.

David Guaspari and Robert M. Solovay. Rosser sentences. *Annals of Mathematical Logic*, 16:81–99, 1979.

Petr Hájek and Pavel Pudlák. *Metamathematics of First-Order Arithmetic*. Perspectives in Mathematical Logic. Springer, Berlin, 1993.

Volker Halbach. A system of complete and consistent truth. *Notre Dame Journal of Formal Logic*, 35:311–327, 1994.

Richard Heck. Self-reference and the languages of arithmetic. *Philosophia Mathematica*, 15:1–29, 2007.

Leon Henkin. A problem concerning provability. *Journal of Symbolic Logic*, 17: 160, 1952.

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

David. Hilbert and Paul Bernays. *Grundlagen der Mathematik II*. Springer, Berlin, 1939. second edition: 1970.

Robert Jeroslow. Redundancies in the hilbert-bernays derivability conditions for Gödel's second incompleteness theorem. *Journal of Symbolic Logic*, 38: 359–367, 1973.

Richard Kaye. *Models of Peano Arithmetic*. Oxford Logic Guides. Oxford University Press, Oxford, 1991.

G. Kreisel and G. Takeuti. Formally self-referential propositions for cut free classical analysis and related systems. Technical report, Warsaw, 1974.

Georg Kreisel. On a problem of Henkin's. *Indagationes Mathematicae*, 15:405–406, 1953.

Georg Kreisel and Azriel Lévy. Reflection principles and their use for establishing the complexity of axiomatic systems. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 14:97–142, 1968.

Martin H. Löb. Solution of a problem of Leon Henkin. *Journal of Symbolic Logic*, 20:115–118, 1955.

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

Peter Milne. On Gödel sentences and what they say. *Philosophia Mathematica*, 15:193–226, 2007.

Hiroakira Ono. Reflection principles in fragments of Peano arithmetic. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 33:317–333, 1987.

Graham Priest. Yablo's paradox. *Analysis*, 57.4:236–242, 1997.

Bertrand Russell. *An Enquiry into Meaning and Truth*. George Allen and Unwin, London, 1940.

Vladimir Yurievich Shavrukov. A smart child of Peano's. *Notre Dame Journal of Formal Logic*, 35:161–185, 1994.

Brian Skyrms. Intensional aspects of semantical self-reference. In Robert L. Martin, editor, *Recent Essays on Truth and the Liar Paradox*, pages 119–131. Oxford University Press, Oxford, 1984.

Craig Smoryński. *Self-Reference and Modal Logic*. Universitext. Springer, New York, Berlin, Heidelberg, and Tokyo, 1985.

Craig Smoryński. The development of self-reference: Löb's theorem. In Thoms Drucker, editor, *Perspectives on the History of Mathematical Logic*. Birkhäuser, Boston, 1991.

Raymond M. Smullyan. Languages in which self reference is possible. *Journal of Symbolic Logic*, 22:55–67, 1957.

Alfred Tarski, Andrzej Mostowski, and Raphael M. Robinson. *Undecidable Theories*. North Holland, Amsterdam, 1953.

Albert Visser. Peano's smart children: A provability logical study of systems with built-in consistency. *Notre Dame Journal of Formal Logic*, 30(2):161–196, 1989.

Albert Visser. Semantics and the liar paradox. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic, second edition*, volume 11, pages 149–240. Springer, Heidelberg, 2004.

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

Christopher von Bülow. A remark on equivalent Rosser sentences. *Annals of Pure and Applied Logic*, 151:62–67, 2008.

Frans Voorbraak. A simplification of the completeness proofs for Guaspari and Solovay's R. *Notre Dame Journal of Formal Logic*, 31(1):44–63, 1989.

Stephen Yablo. Paradox without self-reference. *Analysis*, 53:251–252, 1993.

A. A Gödel numbering with built-in diagonalization

The present construction of a Gödel numbering with built-in self-reference is based on the earlier treatment of the same subject in Visser (2004). The main difference is that the present treatment employs efficient numerals.

As announced above, a coding schema (and two variants) with built-in diagonalization are specified. For any given formula $\varphi(x)$, there will be a number n such that $\varphi(\overline{n})$ has n as its code. The associated diagonal operator yields fixed points that satisfy the Kreisel–Henkin Criterion, because the fixed points are of the form $\varphi(\overline{n})$ and the numeral $\overline{n}$ has n, that is, the code of $\varphi(\overline{n})$ as its value.

Of course the number n can be effectively calculated from $\varphi(x)$. Moreover, the usual syntactic operations can be defined in a straightforward way, so the coding satisfies the usual conditions that a well behaved coding schema is supposed to satisfy. Evidently the coding cannot be monotone.

Consider the language of arithmetic. We suppose its alphabet is $\mathcal{A}$. Suppose $\mathcal{A}_c$ consists of the letters $\mathtt{a}_0, \ldots, \mathtt{a}_{s-1}$ (given in some ordering). We extend this language with a fresh constant c. The extended alphabet is $\mathcal{A}_c$. Say, we treat c as the last of the letters of the extended alphabet. Let $\mathcal{A}^\star$ be the set of strings of letters in $\mathcal{A}$, and, similarly for $\mathcal{A}_c^\star$.

We construct a Gödel numbering as follows. We enumerate $\mathcal{A}_c^\star$ using the shortlex or radix ordering. This means that we first enumerate the sequences of length 0 alphabetically, then the sequences of length 1, etc. This is the ordering used in crossword dictionaries. Let α_n be the n-th string in this enumeration. We take $\mathrm{e}_0(n) \coloneqq \alpha_n$

We will use efficient numerals. We define:

$$\mathrm{S}_{\mathtt{a}_i}(x) \coloneqq \overbrace{\mathrm{S}(\ldots\mathrm{S}(}^{i+1}(x \cdot \overbrace{\mathrm{S}(\ldots\mathrm{S}(}^{s}\,0\,\overbrace{)\ldots)}^{s})\,\overbrace{)\ldots)}^{i+1}.$$

Now consider the number n. Suppose $\alpha_n = a_0 \ldots a_{k-1}$, where the a_j range over $\mathcal{A}_c$. Then we take as efficient numeral $\overline{n}$ for n:

$$\overline{n} \coloneqq \mathrm{S}_{a_{k-1}}(\ldots \mathrm{S}_{a_0}(0)\ldots).$$

Efficient numerals have the convenient property that $\overline{m}$ is a subterm of $\overline{\ell}$ iff α_m is an initial substring of α_ℓ.

We define $\beta_n \coloneqq \mathrm{e}(n) \coloneqq \alpha_n[c \coloneqq \overline{n}]$. This means that $\mathrm{e}(n)$ is the result of substituting $\overline{n}$ for all occurrences of c in α_n. We note that the strings in $\mathrm{e}[\omega]$, the range of e, are strings of letters in $\mathcal{A}$. If ϑ is any string in $\mathcal{A}^\star$, then it is α_m for some m. Clearly $\mathrm{e}(m) = \alpha_m = \vartheta$. So, $\mathcal{A}^\star$ is precisely the range of e.

THEOREM 29. The enumeration e has repetitions.

Proof. Suppose that c occurs in α_m. Clearly for some $n > m$, we have $\alpha_n = \alpha_m[c \coloneqq \overline{m}]$. So, we have $\beta_m = \beta_n = \alpha_n$. ⊣

THEOREM 30. Each string occurs at most twice in the enumeration e.

Let us write $|\vartheta|$ for the number of symbols in ϑ.

Proof. Suppose c occurs at least once in α_m and α_n and $m < n$ and $\alpha_m[c \coloneqq \overline{m}] = \alpha_n[c \coloneqq \overline{n}]$. We note that $\overline{m}$ cannot occur in α_m, since $|\alpha_m| < |\overline{m}|$. Moreover, $\overline{n}$ cannot occur in α_m, since $|\alpha_m| \leq |\alpha_n| < |\overline{n}|$. Since $\overline{n}$ must have at least one occurrence in $\alpha_m[c \coloneqq \overline{m}]$, this occurrence has to overlap with an occurrence of $\overline{m}$. By a unique reading argument it follows that either $\overline{m}$ is a subterm of $\overline{n}$ or vice versa. Since $n > m$, $\overline{m}$ must be a subterm of $\overline{n}$. From this we may conclude that α_m is an initial substring of α_n. Let's say that $\alpha_n = \alpha_m\vartheta$. (Here $\alpha_m\vartheta$ stands for the concatenation of α_m with ϑ, and similarly in what follows). We find that:

$$\alpha_n[c \coloneqq \overline{n}] = \alpha_m[c \coloneqq \overline{n}]\vartheta[c \coloneqq \overline{n}] = \alpha_m[c \coloneqq \overline{m}].$$

We can now get the desired contradiction in two ways. First, suppose α_m starts with ηc, where c does not occur in η. Then both $\eta\overline{n}$ and $\eta\overline{m}$ are initial in $\alpha_m[c \coloneqq \overline{m}]$. But then $\overline{m}$ is initial in $\overline{n}$, quid impossible. For the second way, we note that, since $\overline{m}$ is a subterm of $\overline{n}$, we have $|\overline{m}| < |\overline{n}|$. Ergo: $|\alpha_m[c \coloneqq \overline{n}]\vartheta[c \coloneqq \overline{n}]| > |\alpha_m[c \coloneqq \overline{m}]|$, *quid impossibile*. ⊣

We treat three Gödel numberings based on the ideas introduced above.

For a string ϑ in the alphabet of the language of arithmetic, we define

$$\mathrm{gn}_0(\vartheta) \coloneqq \{m \mid \vartheta = \beta_m\}.$$

This is a many-valued Gödel numbering. We note that many-valued Gödel numberings come naturally with many valued syntactical operations. E.g. we may define:

$$\mathrm{conj}_0(m, n) \coloneqq \mathrm{gn}_0((\beta_m \wedge \beta_n)).$$

For a string ϑ in the alphabet of the language of arithmetic, we define $\mathrm{gn}_1(\vartheta)$ as the smallest m such that $\vartheta = \beta_m$, that is, as the smallest element of $\mathrm{gn}_0(\vartheta)$. Suppose $\vartheta = \alpha_k$. Clearly, the the search for m is bounded by k which is in its turn exponential in the length of ϑ. We note, by the above observations, that, if $\beta = \alpha_m[c \coloneqq \overline{m}]$ and c occurs in α_m, then $\mathrm{gn}_1(\beta) = m$. So, g_n will return the

unique self-referential Gödel number, if there is any. The syntactical operations can be defined in he obvious way. We take e.g.:

$$\mathrm{conj}_1(m, n) \coloneqq \mathrm{gn}_1((\beta_m \wedge \beta_n)).$$

The third Gödel numbering is the standard numbering: $\mathrm{gn}_2(\vartheta) = \mathrm{e}_0^{-1}(\vartheta)$, in other words $\mathrm{gn}_2(\vartheta)$ is the unique m such that $\vartheta = \alpha_m$. Also $\mathrm{gn}_2(\vartheta) = \max(\mathrm{gn}_0(\vartheta))$. The syntactical operations can be defined in he obvious way. We take e.g.:

$$\mathrm{conj}_2(m, n) \coloneqq \mathrm{gn}_2((\beta_m \wedge \beta_n)).$$

Note that we will have, for $i = 1, 2$: if $\mathrm{gn}_i(\vartheta) = m$ and $\mathrm{gn}_i(\eta) = n$, then $\mathrm{gn}_i((\vartheta \wedge \eta)) = \mathrm{conj}_i(m, n)$, as expected of a good functional Gödel numbering.

We can arithmetize the syntactical operations like conj_i defined above in such a way that their elementary properties are verifiable in Elementary Arithmetic.

Suppose our theory is axiomatized by a scheme. As soon as we have appropriate arithmetized syntactical operations like *conjunction*, then we can develop syntax in a uniform way. E.g. we have a schematic formula $\mathrm{Bew}[X, Y, \ldots]$ such that $\mathrm{Bew}[\mathrm{conj}, \ldots]$ will be verifiably a provability predicate provided that the formulae representing the syntactical operations that we plug in for the schematic variables verifiably satisfy some conditions connected with unique reading such as 'a conjunction is never a negation', etcetera. Our schema is described using the given arithmetizations of the syntactical operations. We do not think *intensional correctness* makes much sense for arbitrary computably enumerable collections of axioms.

We submit that the development of syntax using gn_2 is entirely standard. If *any* development produces an intensionally correct representation of provability, then this one does. Since, given that we have the arithmetization of the appropriate syntactical operations, the formalization of provability is uniform, the only point where intensional incorrectness could sneak in, is in the definitions of functions like conj_0 and conj_1. Since the definitions of the conj_i are directly derived from the definition of the gn_i, for $i = 0, 1$, it seems that we just have two options: Either we do accept Bew_i for $i = 0, 1$, as intensionally correct, or we conclude that some Gödel numberings do not support intensionally correct arithmetizations of provability. If we opt for the second, we should try to articulate what it is that precludes intensional correctness …

Let us suppose that we accept, say, Bew_1 as intensionally correct. Consider the formula $\neg\,\mathrm{Bew}_1(c)$. Let this formula be α_g. Then g is the gn_1-Gödel number of $\neg\,\mathrm{Bew}_1(\overline{g})$. So we have an intensionally correct Gödel sentence G where $G = \neg\,\mathrm{Bew}_2(\overline{\mathrm{gn}_1(G)})$.

REMARK 31. Clearly, we will have a definable function switch such that we have $\mathrm{switch}(m) = \mathrm{gn}_2(\beta_m)$, and such that our theory verifies that

$$\forall x \in \mathrm{sent}_1\,(\mathrm{Bew}_1(x) \leftrightarrow \mathrm{Bew}_2(\mathrm{switch}(x))).$$

Note that it does *not* follow from the assumption that Bew_1 is intensionally correct w.r.t. gn_1 that also the verifiably extensionally equal predicate $\mathrm{Bew}_1'(x) \coloneqq \mathrm{Bew}_2(\mathrm{switch}(x))$ is intensionally correct w.r.t. gn_1.

REMARK 32. Is conj_2 intensionally correct w.r.t. the given Gödel numbering? Well, we assume that we have implemented it by first arithmetizing *concatenation*. Our definition of the syntactic operation of conjunction is based on the following definition *in the theory of concatenation*:

$$\mathrm{conj}(\sigma, \tau) \coloneqq \ulcorner(\urcorner * \sigma * \ulcorner\wedge\urcorner * \tau * \ulcorner)\urcorner.$$

Is this definition intensionally correct? People working in the Tarski tradition like John Corcoran and Andrzej Grzegorczyk believe that this definition gives in fact *the essence* of the conjunction operation. But isn't the concatenation format just imposed on us by the necessity of linear representation of syntactic structure? Isn't our basic understanding of the syntax that it is something like a free algebra? For example, do we not see the difference between infix and prefix notation for conjunction as a mere matter of implementation? Note also that we could have implemented the syntax equally well in a theory of finitely branching trees or of finite sets.

The second step is to arithmetize concatenation in the style Smullyan. What this operation is extensionally follows from the chosen Gödel numbering which corresponds to the shortlex ordering. The chosen arithmetical operation is $x \circledast y = x \cdot q^{\ell(y)} + y$, where q is the number of symbols in our alphabet and ℓ is the q-adic length function. Smullyan's clever insight is that one can define $q^{\ell(y)}$ without first defining exponentiation.

Does the question of the intensional correctness of these two steps make sense? Maybe it is simply a matter of stipulation that they are intensionally correct, so that we can judge the other steps to be correct *given* the correctness of the initial steps.

Volker Halbach
University of Oxford
New College
OX1 3BN Oxford
England

Albert Visser
Faculty of Humanities
Utrecht University
Janskerkhof 13
3512 BL Utrecht
The Netherlands