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

SELF-REFERENCE IN ARITHMETIC II

VOLKER HALBACH
Oxford University
and
ALBERT VISSER
Utrecht University

Abstract. In this sequel to *Self-reference in arithmetic I* we continue our discussion of the question: What does it mean for a sentence of arithmetic to ascribe to itself a property? We investigate 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 expressing formulae are obtained. In this second part we look at some further examples. In particular, we study sentences apparently expressing their Rosser-provability, their own Σ^0_n-truth or their own Π^0_n-truth. Finally we offer an assessment of the results of both papers.

This is the second part of a two-part paper. In the first part *Self-reference in arithmetic I* we asked what it means for a sentence of arithmetic to ascribe a property to itself. In this first part we then focused on the property of provability in a given artihmetical system. In this sequel we look at further properties such as Rosser-provability, Σ_n-truth and Π_n-truth.

To facilitate referencing of previous definitions and results, sections are numbered consecutively through both parts. When we refer to definitions or results in the first part, we only provide the section number without indicating that they are found in the first part. This second part presupposes acquaintance with the first part.

§7. Further examples for the intensionality of self-reference. Löb's theorem has attention drawn away from the problem of the 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 provably equivalent.

Of course, Löb's theorem, that is, Theorem 5.13, which eliminates all intensionality from self-reference, applies only to certain predicates, most notably, to the canonical provability predicate. For other formulae – whether they are supposed to express provability or another property – we cannot expect something analogous. As we have seen, noncanonical provability predicates like $\mathrm{Bew}_\Pi(x)$ are susceptible to intensionality phenomena and their fixed points can vary in their properties, even if they satisfy 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. We look first at Rosser provability and then at partial truth predicates.

Received Dec. 23, 2013

© Association for Symbolic Logic, 2014
doi:10.1017/S175502031400029X

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

7.1. On Rosser provability. The Rosser provability predicate is defined as follows.

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

Here $B(y, x)$ strongly represents[1] the relation that y is a proof of x, and $\dot{\neg}$ represents the 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 fixed points of $\mathrm{Bew}^R(\dot{\neg} x)$ are (variants of) Σ_1-Rosser sentences. Henkin's problem for Rosser provability becomes:

> *What property does the sentence that says about itself that it is Rosser provable have? Is it provable, refutable or independent?*

As in the case of the Henkin problem for canonical provability, the singular *the sentence* is misleading: Even granted that the notion of self-reference is unclear, there will not be just one such sentence. It's also far from clear that the answer to the question is the same for all such sentences. At any rate, the different fixed points of the Rosser provability predicate have different properties. It is known for a long time that the Rosser provability predicate – unlike the canonical provability predicate – has nonequivalent fixed points. So in answering this question about the Henkin problem for Rosser provability, one will have to focus on self-referential fixed points and make essential use of their self-referentiality in determining whether they are provable, refutable or independent.

OBSERVATION 7.1. *Let Σ be consistent and let it contain* Basic *plus the axioms stating that $<$ is a linear ordering. All Σ-provable as well as all Σ-refutable formulae are fixed points of the Rosser provability predicate.*

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 (†) $\forall z < \overline{n} \neg B(z, \dot{\neg}\ulcorner\psi\urcorner)$ holds 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, $\Sigma \vdash \neg B(\overline{k}, \ulcorner\psi\urcorner)$, as all these sentences are Δ_1.

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

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

Since $<$ is linear, we may conclude that $p \leq \overline{n}$. But then, by the axioms of Robinson arithmetic, we find $\bigvee_{k \leq n} p = \overline{k}$. Let $\overline{k} = p$. 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, for example, 0=0 as well as 0≠0 are fixed points of the Rosser provability predicate.

[1] See Section 2.2 for a definition of *strong representation*.

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

QUESTION 7.2. *What is the status of the fixed point of* $\mathrm{Bew}^R(x)$ *that is obtained by the Gödel construction? Moreover, does* $\mathrm{Bew}^R(x)$ *have fixed points that are independent?*

Kurahashi (2014) considered Rosser provability predicates constructed from provability predicates that satisfy the Löb derivability and some additional conditions. He showed that some of them have no independent fixed points and that some of them have also independent ones. This implies that also the fixed points obtained by the canonical Gödel construction from these Rosser provability predicates cannot be independent and thus we have a partial answer to the last part of Question 7.2.

Rosser sentences themselves are subject to various intensionality phenomena. In particular, Rosser sentences can be construed as Σ_1- or Π_1-sentences:

REMARK 7.3. *Fixed points of* $\neg\mathrm{Bew}^R(x)$ *and* $\mathrm{Bew}^R(\dot{\neg}\, x)$ *are respectively the* Π_1*-Rosser sentences and the* Σ_1*-Rosser sentences.*

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 nonequivalent fixed points connected to alternative provability predicates, the reader is referred to (Visser 1989) and (Shavrukov 1994).

7.2. 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 nonequivalent fixed points. A formula $\varphi(x)$ for which all sentences are fixed points would be a total truth predicate, but, by Tarski's theorem on the undefinability of truth, such a predicate cannot exist in a consistent system. However, there are partial truth predicates for sentences with limited quantifier complexity. More formally, for each n, there is a truth predicate for all Σ_n-sentences, and similarly for Π_n-sentences.

To dot our i's and to cross our t's, we need to pay attention to a detail concerning the definition of Σ_n and Π_n. In the most narrow sense, Σ_1-formulae, for instance, consist in a block of unbounded existential quantifiers – or even just one existential quantifier – followed by a Δ_0-formula. We could liberalize this definition in various ways, allowing, say, closure under conjunction and disjunction, or allowing negations of Π_1-formulae to count as Σ_1-formulae and vice versa. We could even allow closure under bounded quantification; this last move, however, 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 function in the intended way. 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 & Pudlák (1993) or Kaye (1991). There is some extra detail in developing these partial truth predicates, since we have function symbols for all primitive recursive functions in our language. However, as 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 satisfying Kreisel's Condition, that is, they cannot even weakly represent the sets of all Σ_n- and all Π_n-truths, respectively. So the question arises in which sense they are proper true truth-predicates.

Usually logicians resort to the *meaning postulate* approach, as mentioned in Section 2.2, and 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

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

all Σ_n-sentences φ. Π_n-truth predicates are defined analogously.[2] 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 actually 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 a sentence that says of itself that it is Σ_n-true is provable, refutable, independent and, in the latter case, whether it's true or false. 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 as well as 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 similar to Henkin sentences, but their properties are more affected by intensionality phenomena than those of Henkin sentences. The fixed points of provability predicates just satisfying Löb's derivability conditions as meaning postulates are by Löb's theorem all provably equivalent. Hence all fixed points of the canonical provability predicate are equivalent. In constrast, the fixed points of the canonical partial truth predicates $\mathrm{Tr}_{\Sigma_n}(x)$ and $\mathrm{Tr}_{\Pi_n}(x)$ are not. We will show that the status of truth-teller sentences is also highly sensitive to the first source of intensionality, that is, to the choice of 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.

7.3. 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 $\mathrm{Bew}_{\mathrm{I}\Sigma_1}(x)$ and $\mathrm{Tr}_{\Sigma_1}(x)$ yields a PA-provable and a PA-refutable sentence, respectively.

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

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

THEOREM 7.4. $\mathsf{PA} \vdash \mathrm{Bew}_{\mathrm{I}\Sigma_1}(\ulcorner\sigma\urcorner) \leftrightarrow \sigma$ *for all* Σ_1*-formulae* σ.

Proof. The left-to-right direction $\mathsf{PA} \vdash \mathrm{Bew}_{\mathrm{I}\Sigma_1}(\ulcorner\sigma\urcorner) \rightarrow \sigma$, that is, local reflection, is well-known and can be obtained by formalising the cut-elimination theorem for $\mathrm{I}\Sigma_1$ in PA and proving reflection in the usual way, outlined, for instance, in (Kreisel & 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 $\mathrm{Bew}_{\mathrm{I}\Sigma_1}(x)$ is a truth predicate for the set of Σ_1-sentences in PA but not in $\mathrm{I}\Sigma_1$.

[2] 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.

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

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

THEOREM 7.5. *Let σ be a Σ_1-sentence. Then the following are equivalent:*

(i) *σ is true.*
(ii) $\mathrm{I}\Sigma_1 \vdash \sigma$
(iii) $\mathsf{PA} \vdash \sigma$
(iv) $\mathrm{I}\Sigma_1 \vdash \mathrm{Bew}_{\mathrm{I}\Sigma_1}(\ulcorner\sigma\urcorner) \leftrightarrow \sigma$

The easy proof uses Löb's theorem. The fixed point $d(\mathrm{Bew}_{\mathrm{I}\Sigma_1}(x))$ obtained from the predicate $\mathrm{Bew}_{\mathrm{I}\Sigma_1}(x)$ by the canonical diagonalization procedure is Σ_1; thus it is of the appropriate complexity and can be called a *truth teller*. Since the fact that canonical diagonalization works can be verified in $\mathrm{I}\Sigma_1$, we find, by the above theorem, that $\mathsf{PA} \vdash d(\mathrm{Bew}_{\mathrm{I}\Sigma_1}(x))$. 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.[3] We assume we use a *monotone Gödel coding*. By this we mean a coding where the code of a sequence is always greater than the code of any member of the sequence, the code of a numeral of a number is always greater than that number, and so on. Thus, for instance, the code of a formula is greater than the code of all terms contained.

The canonical truth predicate $\mathrm{Tr}_{\Sigma_1}(x)$ is of the form $\exists y\, \vartheta(y, x)$ with a formula $\vartheta(y, x)$ containing only bounded quantifiers.[4] In a nutshell, $\exists y\, \vartheta(y, x)$ says that there is a sequence y of triples each consisting of a formula, a finite variable assignment and a truth value; moreover, the sequence of triples follows the usual Tarskian rules. Suppose x is of the form $\ulcorner\exists\urcorner * v * z$, where '$*$' is our arithmetization of concatenation. In this case the penultimate element of the sequence y will be a triple $(z, s, \overline{1})$, 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 7.6. *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 $\mathsf{PA} \vdash \forall y\, (\vartheta(y, \ulcorner\exists v\, \sigma(v)\urcorner) \rightarrow \exists v < y\, \sigma(v))$ holds.*

If we keep everything standard, then the Σ_1-truth teller becomes refutable in PA. For the proof we use the assumption above, the monotonicity of the coding, and the fact that the fixed-point sentence satisfies the Kreisel–Henkin Criterion for self-reference; the latter holds because the Σ_1-truth teller is obtained by a diagonal operator with the Kreisel–Henkin property in the sense of Definition 5.4. Of course Gödel's canonical diagonal operator has this property.

[3] Vann McGee and Albert Visser have independently communicated this observation to me. V.H.

[4] 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.

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

THEOREM 7.7. *Suppose we employ a standard, monotone Gödel coding. If d is a diagonal operator with the Kreisel–Henkin property,* $\mathsf{PA} \vdash \neg d(\mathrm{Tr}_{\Sigma_1}(x))$ *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 $t = \ulcorner \exists y\, \vartheta(y, t) \urcorner$, Assumption 7.6 above combined with (a), gives us $\exists z < y_0\, \vartheta(z, t)$. But this contradicts (b). Hence our supposition that $\exists y\, \vartheta(y, t)$ must fail. □

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

Theorem 7.7 can be generalized to Σ_n with $n > 1$.[5] 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 k-tuples typically having witnesses as a component; y may range 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 have for Σ_{n+1} sentences $\exists v\, \sigma(v)$ the following property analogous to Assumption 7.6:

ASSUMPTION 7.8. $\mathsf{PA} \vdash \forall y\, (\vartheta(y, \ulcorner \exists v\, \sigma(v) \urcorner) \rightarrow \exists v < y\, \sigma(v))$ *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 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 analogous assumptions again, we can generalize the above result, using the same proof idea:

THEOREM 7.9. *Suppose we employ a standard, monotone Gödel coding. If d is a diagonal operator with the Kreisel–Henkin property,* $\mathsf{PA} \vdash \neg d(\mathrm{Tr}_{\Sigma_{n+1}}(x))$ *obtains.*

Before constructing a counterexample to Assumption 7.8, 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 then deleting all 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 $\sim \mathrm{Tr}_{\Sigma_n}(\dot{\sim} x)$.[6] If Tr_{Σ_n} is in prenex normal 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. Under Assumption 7.8 the Π_n-truth tellers are provable. The proof is a variation of the proofs of Theorems 7.7 and 7.9.

THEOREM 7.10. *Suppose we employ a standard, monotone Gödel coding. If d is a diagonal operator with the Kreisel–Henkin property and* Tr_{Π_n} *is defined as described above,* $\mathsf{PA} \vdash d(\mathrm{Tr}_{\Pi_n}(x))$ *obtains for all $n > 0$.*

---

[5] We thank Graham Leigh for pointing out to us that Theorem 7.7 generalizes to higher n for many reasonable truth predicates.

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

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

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 infer $\neg \sim \mathrm{Tr}_{\Sigma_n}(\dot{\sim} t)$ 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 $\vartheta(y_0, \dot{\sim} \ulcorner \mathrm{Tr}_{\Pi_n}(t) \urcorner)$, which is $\vartheta(y_0, \dot{\sim} \ulcorner \sim \mathrm{Tr}_{\Sigma_n}(\dot{\sim} t) \urcorner)$, that is, $\vartheta(y_0, \dot{\sim} \ulcorner \sim \exists y\, \vartheta(y, \dot{\sim} t) \urcorner)$. Using Assumption 7.8 we conclude $\dot{\exists} v < y_0\, \vartheta(v, \dot{\sim} t)$. This contradicts the supposition that y_0 is minimal. □

We have claimed that for $n > 1$ there are reasonable Σ_n-truth predicates for which Assumption 7.8 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:

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

Therefore, a witness for a Σ_{n+1}-sentence is also a witness for its Σ_{n+1}-truth and vice versa. So, clearly Assumption 7.8 is violated because under this assumption the smallest witness for a proper Σ_{n+1}-sentence is always smaller than a 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 normal 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 normal form. So we can simply concentrate on sentences in prenex normal 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 normal form, further tricks would have to be applied. As 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

$$\exists y\, \forall v\, {<}\, x\, \forall a\, {<}\, x\, \big(x = \dot{\exists} v a \rightarrow \mathrm{Tr}_{\Pi_n}(a(\dot{y}/v))\big). \tag{1}$$

Here $x = \dot{\exists} v\, a$ expresses that x is a sentence and that x is 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 (1) itself cannot serve as a 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.[7] For the universal quantifiers this 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 we have, for $n > 1$, constructed Σ_n- and, if we use the tricks from above again, also Π_n-truth predicates that do not conform to Assumption 7.8. Defining truth predicates along the lines of (1) 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 7.9 and 7.10 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 tellers remain refutable and the Π_n-truth tellers provable. It seems that another proof idea would be

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

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

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 noncanonical diagonal operators and deviant partial truth predicates, Henkin's trick in the proof of Kreisel's Observation can be applied again to produce another example of the intensionality of self-reference. Henkin's trick then yields a Σ_n-truth predicate with a provable and a refutable truth teller.

OBSERVATION 7.11. *Assume again that a standard, monotone Gödel coding is used. For each n there is a Σ_n-truth predicate $\sigma_n(x)$, sentences τ_1 and τ_2 such that both sentences τ_1 and τ_2 ascribe to themselves the property expressed by $\sigma_n(x)$ according to the Kreisel–Henkin Criterion and τ_1 is provable while τ_2 is refutable.*

Proof. Let $\mathrm{Tr}_{\Sigma_n}(x)$ be some Σ_n-truth predicate satisfying Assumption 7.8, 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.

For d the the canonical diagonal operator, one can show that $d(x = t \vee \mathrm{Tr}_{\Sigma_n}(x))'$ is refutable by Theorem 7.9. □

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

7.5. Nonstandard truth predicates constructed in a different way. After having shown that by merely varying the method for obtaining a truth teller one can obtain provable as well as a refutable truth tellers, we are going to show in this section that by merely varying the truth predicate but adhering to the canonical diagonal operator d, one can also obtain both provable and refutable truth tellers. This is achieved by applying 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 5.5 and 5.6:

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

Our formula φ^γ has to satisfy (4) in Section 5, 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 or Π_n), respectively, then so is φ^γ. Fortunately this is easily arranged. We can rewrite the formula

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

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

in the prescribed strict Σ_n-(Π_n-)form, obtaining a formula $\eta(x, y)$ and then apply the canonical diagonal construction with respect to 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 $\beta := d(\mathrm{Tr}^{\gamma}_{\Sigma_n}(x))$. 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, the situation is analogous to that of provability: In order to show that a truth teller can be provable or refutable, we don't need to invoke a deviant diagonal operator *and* a noncanonical formula expressing a property simultaneously. The refutable truth teller is obtained by keeping the diagonal operator and the partial truth predicate canonical; and we only need a deviant partial truth predicate in order to show that a Σ_n-truth teller can also be provable. A noncanonical diagonal operator is dispensable.

§8. Uniform diagonal operators. We think that Gödel's diagonalization method and certain variants of it produce paradigmatically self-referential sentences, *if* there is any self-reference in metamathematics at all. In this speculative section we consider a condition that is satisfied by these paradigmatically self-referential sentences, but not by some other fixed points that satisfy the Kreisel–Henkin Criterion.

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 7.7, 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, of 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 satisfying the Kreisel–Henkin Criterion will include and apply to the truly self-referential fixed points.

There remain also doubts about whether *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 $\mathrm{Bew}_{\Pi}(t_2)$ from

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

Kreisel's Observation is refutable, while applying the canonical diagonal operator to $\mathrm{Bew}_{\Pi}(x)$ yields a provable fixed point, as has been shown in Observation 4.1. The fixed point $\mathrm{Bew}_{\Pi}(t_2)$ is not obtained from $\mathrm{Bew}_{\Pi}(x)$ by a slight variant of the canonical diagonalization method; rather $\mathrm{Bew}_{\Pi}(x)$ it is constructed in a such a way that t_2 'happens' to be a fixed point satisfying the Kreisel–Henkin Criterion. The sentence $\mathrm{Bew}_{\Pi}(t_2)$ surely refers to (the code of) a sentence that happens to be $\mathrm{Bew}_{\Pi}(t_2)$. But we are not sure whether this implies that $\mathrm{Bew}_{\Pi}(t_2)$ states of itself in the strictest sense that it has the property expressed by the formula $\mathrm{Bew}_{\Pi}(x)$. One could also distinguish between two kinds of self-reference: $\mathrm{Bew}_{\Pi}(t_2)$ 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.

In natural language there are also different forms of self-reference and some seem to be stronger or more paradigmatic forms of self-reference than others. The 'strongest' form of self-reference is presumably obtained via first-person personal pronouns. A person, who has just knocked over the glass, may say 'I am wounded' or the 'The person who has knocked over the glass is wounded'. The first sentence expresses a *de se*-belief; it is clearly self-referential. The second sentence is also self-referential. By uttering the sentence 'The person who has knocked over the glass is wounded', he makes again a claim about himself; but he may not even be aware that he is making a claim about himself. Moreover this second sentence is only accidentally about the speaker. Self-reference via the first-order pronoun 'I' seems in some sense stronger and more paradigmatic than self-reference via a definite description.[8]

Of course, pronouns are not available in the language of arithmetic; and self-reference cannot be in the same way contingent as in natural language. But still a sentence obtained by applying a uniform method, such as the canonical diagonal method, seems to exhibit a less accidental and stronger form of self-reference than the sentence $\mathrm{Bew}_{\Pi}(t_2)$. The canonical fixed point $d(\mathrm{Bew}_{\Pi}(x))$ seems to come closer to expressing a *de se*-claim than $\mathrm{Bew}_{\Pi}(t_2)$, in as far as this term is appropriate in our context.

So, among the fixed points satisfying the Kreisel–Henkin Criterion one can still distinguish between more or less contrived ones. What seems to be deviant about $\mathrm{Bew}_{\Pi}(t_2)$ 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, beyond the Kreisel–Henkin Criterion, more conditions on fixed points, so that sentences such as $\mathrm{Bew}_{\Pi}(t_2)$ are ruled out.

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

DEFINITION 8.1. *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:[9]

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

[8] van Fraassen's (1970) distinction between *accidental* and *functional* self-reference in natural language, which was used by him for an analysis of the paradoxes.

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

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

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

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

OBSERVATION 8.2. *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, $d(\varphi(t, x))$ is the formula $\varphi(t, \dot{d}\ulcorner\varphi(t, x)\urcorner)$. By assumption $d(\varphi(t, x))$ is the formula $\varphi(t, t)$. Since $\varphi(t, \dot{d}\ulcorner\varphi(t, x)\urcorner)$ and $\varphi(t, t)$ are identical expressions by assumption, the term t must be the expression $\dot{d}\ulcorner\varphi(t, x)\urcorner$. Hence the term t would contain a numeral for a formula that in turn contains the term t, contradicting monotonicity of the coding. □

If the diagonal operator yields only a fixed point that satisfies the Kreisel–Henkin Criterion then t and $\dot{d}\ulcorner\varphi(t, x)\urcorner$ 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 at how to narrow down the choice of diagonal operators, if a result cannot be proved for all fixed points satisfying the Kreisel–Henkin Criterion.

§9. 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 with their defining equations as axioms. Self-reference has been discussed in many other settings: The language may lack appropriate function symbols, or the language may contain symbols going beyond those of arithmetic by containing, for instance, 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 only touch upon 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 sufficient conditions for a sentence to ascribe some property to itself. Those conditions 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. However, still 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 symbols for zero and successor, suffice if the coding is carefully chosen. In the appendix we construct a Gödel coding along with a diagonal operator with the Kreisel–Henkin property that relies only on numerals as the terms. However, if a monotone coding is employed and the language doesn't contain the appropriate function symbols – like the the usual language of PA – then

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

there are no formulae that ascribe a property to themselves 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 operator with the Kreisel–Henkin property are absent, as is the case in Peano arithmetic 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. One might surmise that, in a sense, the Kreisel–Henkin Criterion already captures an important aspect of self-reference in arithmetic, and thus one could try to generalize 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 formula $\varphi(t)$ by eliminating the function symbols in $\varphi(t)$ ascribes to itself the same properties as $\varphi(t)$. The elimination can be carried out by applying one of the usual methods of transforming a formula that contains function symbols into a provably equivalent formula where the functions are expressed using quantification and appropriate predicate expressions. However, self-reference is too intensional and therefore may not be preserved under this transformation. The new formula ψ will still refer to $\varphi(t)$ and instead of itself, except in some special fortunate cases. Therefore, one will have to adapt the method of elimination in a more sophisticated way.[10]

However, we don't assume that such an elimination necessarily preserves self-reference, even if carried out properly, nor, 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.[11]

We now turn from languages that are properly contained in that of Basic to languages that properly extend it. Many remarks carry over in a straightforward way. In particular, we think that many of our remarks carry over 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 negated truth predicate. However, for arriving 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). Therefore, at least some of the phenomena we have studied here arise also in these wider contexts.

§10. Summary: Henkin sentences and truth tellers. Before trying to draw 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 $\mathrm{Bew}_{\mathrm{II}}(x)$, $\mathrm{Bew}_2(x)$, $\mathrm{Bew}_3(x)$, and $\mathrm{Bew}^R(x)$ are each defined from some given provability predicate. We assume that the

[10] 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.

[11] Heck (2007) has raised some worries about the possibility of appropriately expressing self-reference in languages lacking 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.'

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

canonical provability predicate $\mathrm{Bew}(x)$ is always 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 all 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
Bew_Π (Kreisel–Henkin)	p	p,r	p,r
Bew_2 (Theorem 5.1)	r	p,r	p,r
Bew_3 (Theorem 5.2)	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 Σ_n- and Π_n-sentences, respectively, with $n \geq 1$; the formulae Tr_{Σ_n} are assumed to satisfy Assumption 7.8, and the formulae Tr_{Π_n} are defined from them in the way indicated in Section 7.4. The formula $\mathrm{Bew}_{\mathrm{I}\Sigma_1}$ is viewed here as a truth predicate for Σ_1-sentences and restricted to such. All fixed points are assumed to be of the complexity concerned, that is, fixed points of Σ_n-truth predicates that are not Σ_n are not taken into account, and similarly for Π_n.

	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 7.11	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	r,i	p,r,i

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

The truth predicates $\mathrm{Tr}^{\gamma_1}_{\Sigma_n}$ and $\mathrm{Tr}^{\gamma_2}_{\Sigma_n}$ have been defined for arbitrary γ in Section 7.5. 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.

§11. 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 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 aren'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, beyond that of self-reference 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 have become more suspicious of self-reference, and some have dismissed questions concerning 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, in mathematics and the sciences. Therefore, they shouldn't 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 which methods count as canonical. 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 on 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 7.9 on the provability of Σ_n-truth tellers we assume that the fixed points of the truth predicates satisfy the Kreisel–Henkin Criterion. 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 of intensionality, that is, the way of expressing a property, proved fruitful. Thus, we believe that we should strive for an analysis of what it means for a formula to attribute a property of itself.

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

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 wherever we know that all sentences ascribing 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 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 and the problem of expressing a property in the formal language. Once deviant provability predicates are excluded, the third source of intensionality, which concerns the way self-reference is obtained, is also blocked.

The case of provability, however, is very special. In most other cases we cannot as easily escape the intricacies of the intensionality of self-reference: Fixed points of a formula will usually behave in 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 7.1. In order to obtain results about sentences ascribing such properties to themselves 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 via 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 condition, too. However, we still have serious doubts whether it even provides a sufficient condition for self-reference. There is surely something awkward about fixed points satisfying the Kreisel–Henkin Criterion 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 7.9 on the provability of Σ_n-truth tellers for $n \geq 1$ we assume that the fixed points of the partial truth predicates satisfy the Kreisel–Henkin Criterion. If all Σ_n-sentences that say about themselves that they are Σ_n-true do so in virtue of the Kreisel–Henkin Criterion, and if Assumption 7.8 and the monotonicity of coding are accepted, then Theorem 7.9 answers the question about whether Σ_n-truth tellers are provable.

Theorems 7.9 and 7.10 demonstrate that questions about partial-truth tellers are the sensitive to all three sources of intensionality. In the two theorems we make assumptions concerning the coding, concerning specific properties of the formulae expressing Σ_n- or Π_n-truth and, of course, concerning 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 7.9 and 7.10 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 4.1 it

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

was noted that the formula $\mathrm{Bew}_{\mathrm{II}}$ does have diagonal sentences $\mathrm{Bew}_{\mathrm{II}}(t_2)$ and $\mathrm{Bew}_{\mathrm{II}}(t)$ satisfying the Kreisel–Henkin Criterion, of which the first is provable and the second refutable. Observation 7.11 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 satisfying the Kreisel–Henkin Criterion 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 a formula that we cannot accept as an example that would support an affirmative answer to our question. Consider a formula $\zeta(x)$ that does not contain an occurrence of the successor symbol and that expresses the property that x doesn't contain an occurrence of the successor symbol. 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 occurrences 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 questions in the opposite direction. Assume that the provability or refutability of supposedly self-referential statements of a certain kind in a formal language or of just one such sentence has been established. Then one can try to generalize this result and ask question of the following kind: Can one obtain a sentence with deviating properties by using other coding schemata, other formulae for expressing the property or a different method for obtaining self-reference? Does it suffice to use only a different formula expressing the property concerned or is also another fixed-point operator required. Can one show that the result can be generalized? For instance, can one prove that all sentences satisfying the Kreisel–Henkin Criterion behave in the same way, if certain conditions are imposed on the coding and the formula that expressing the property?

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 7.9 and 7.10 on partial truth tellers rely on the use of a monotone coding schema. It may well be possible to obtain provable Σ_n-truth tellers by using a suitable nonmonotone coding. This would provide us with another example of intensionality arising from the first source. Of course Löb's theorem is a result that establishes the robustness of Löb's answer to Henkin's problem: Once a provability predicate satisfying the Löb derivability conditions is fixed, any fixed point of the provability predicate is provable, irrespective of the coding or of how the fixed point has been obtained.

Here in this paper we have somewhat suppressed intensionality due to coding and focused on intensionality arising from the second and third source. In particular, we have established that in some cases intensionality due to the second source will suffice to change results. Kreisel (1953) obtained provable and refutable Henkin sentences by exploiting simultaneously the second *and* third source of intensionality. Observation 7.11 shows that at least with a deviant provability predicate, we can obtain provable and refutable Henkin

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

sentences by changing the method of diagonalization (without violating the Kreisel–Henkin Criterion). Finally, in Section 5 it was established that by changing the formula expressing provability but using only the canonical diagonal operator, one can obtain both provable and refutable Henkin sentences. Section 7.5. 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 which show that the analysis of self-reference in arithmetic is not as straightforward as it may appear. There are still some mathematical and philosophical questions concerning formulae that are described as making statements about themselves; the answers may be both interesting and fruitful, just as Löb's answer to Henkin's question was.

§12. Appendix: Gödel numbering with built-in diagonalization. The present treatment 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. In this appendix $\overline{n}$ stands for the *efficient* numeral of n, which will be defined below, and not the usual numeral.

We specify a coding schema gn_1 with built-in diagonalization and two variants. For any given formula $\varphi(x)$, there will be a number n such that $\varphi(\overline{n})$ has n as its code, that is, $\mathrm{gn}_1(\varphi(\overline{n})) = n$. It follows that the associated diagonal operator $d_1 : \varphi(x) \mapsto \varphi(\overline{n})$, where n is the code of $\varphi(\overline{n})$, has the Kreisel–Henkin property in the sense of Definition 5.4. The number n can be effectively calculated from the formula $\varphi(x)$, as is required for a diagonal operator.

Moreover, the commonly used 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.

Consider the language of arithmetic with its finite alphabet $\mathcal{A}$. Suppose $\mathcal{A}$ consists of the letters $\mathtt{a}_0, \ldots, \mathtt{a}_{s-1}$ (in some fixed ordering). We extend this language with a fresh constant c. The extended alphabet is called $\mathcal{A}_c$. 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 analogously for $\mathcal{A}_c^\star$.

We enumerate $\mathcal{A}_c^\star$ using the shortlex or radix ordering. This means that after the empty sequence we first enumerate the sequences of length 1 lexicographically, then the sequences of length 2, etc. This is the ordering used in crossword dictionaries. Let α_n be the n-th string in this enumeration. So, α_0 is the empty sequence.

The number of symbols in $\mathcal{A}_c$ is $s + 1$. For each $\mathtt{a}_i$ $(0 \leq i \leq s)$ in $\mathcal{A}_c$ we define an expression $\mathrm{S}_{\mathtt{a}_i}(x)$ with the fixed variable x:

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

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

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

The efficient numeral of the empty sequence α_0 is the constant 0 for zero and thus $\overline{0}$ is just the constant 0. Similarly, the value of the efficient numeral $\overline{n}$ is n for each n. Efficient

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

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 $\beta_n := \mathrm{e}(n) := \alpha_n[c := \overline{n}]$. 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.

OBSERVATION 12.1. *The enumeration* e *has repetitions.*

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

THEOREM 12.2. *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 each of α_m and α_n, and $m < n$ and $\alpha_m[c := \overline{m}] = \alpha_n[c := \overline{n}]$. 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 $\alpha_m[c := \overline{m}]$, 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

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

We can now obtain 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 $\alpha_m[c := \overline{m}]$. But then $\overline{m}$ is initial in $\overline{n}$, which is impossible. For the second way, we note that, since $\overline{m}$ is a subterm of $\overline{n}$, we have $|\overline{m}| < |\overline{n}|$. Ergo, we obtain the contradiction $\left|\alpha_m[c := \overline{n}]\vartheta[c := \overline{n}]\right| > \left|\alpha_m[c := \overline{m}]\right|$. □

We look into three Gödel numberings based on the ideas introduced above.

For a string ϑ in the alphabet $\mathcal{A}$ of the language of arithmetic, we define $\mathrm{gn}_0(\vartheta) := \{m \mid \vartheta = \beta_m\}$. By Observation 12.1, gn_0 is a many-valued Gödel numbering for $\mathcal{A}^\star$. Many-valued Gödel numberings can naturally be combined with many-valued syntactical operations. For instance, we may define the metatheoretic operation that yields, applied to two Gödel codes m and n, the codes of the conjunction of β_m and β_n:

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

The next coding schema gn_1 is single-valued with built-in diagonalization. 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 $\mathrm{gn}_0(\vartheta)$.

The syntactical operations can be defined in the obvious way; for instance, we can define conjunction in the following way:

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

The coding gn_1 is effective, that is, one can effectively determine $\mathrm{gn}_1(\vartheta)$ from ϑ using the following method: A given ϑ must occur as an α_k in our enumeration of strings in $\mathcal{A}_c^\star$. The number k is bounded by $1 + (s+1)^\ell$, where ℓ is the number of symbols in ϑ. Then we

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

search through all α_j with $i \leq k$ and check whether ϑ is β_n, that is, $\alpha_n[c := \overline{n}]$. If we find such an n, we have $\mathrm{gn}_1(\vartheta) = n$; otherwise we reach k and have $\mathrm{gn}_1(\vartheta) = k$.

In the next lemma we show that diagonalization is built into the coding schema gn_1.

LEMMA 12.3. *Let $\varphi(x)$ be a formula in the language of arithmetic. Then we can effectively find the unique m such that $\mathrm{gn}_1(\varphi(\overline{m})) = m$. Thus, the function $d_1 : \varphi \mapsto \varphi[x := \overline{m}]$ is a diagonal operator with the Kreisel–Henkin property in the sense of Definitions 5.4.*

Proof. Let a formula $\varphi(x)$ with at least one free occurrence of the designated variable x be given. Let us say that the formula $\varphi[x := c]$ occurs as α_m in the enumeration of the elements of $\mathcal{A}_c^\star$. Using the definition of gn_1 we conclude $\mathrm{gn}_1(\alpha_m[x := \overline{m}]) = m$.

The function sending each φ to the corresponding self-referential code m is primitive recursive and even elementary recursive. Thus, the diagonal operator $d_1 : \varphi \mapsto \varphi[x := \overline{m}]$ is primitive recursive. □

With the coding gn_1 the diagonal lemma becomes trivial. The identity $\overline{m} = \overline{\varphi(\overline{m})}$ is trivially provable, because $\overline{m}$ and $\overline{\varphi(\overline{m})}$ are actually the same term; and consequently the *same* formula occurs on both sides of the equivalence $\varphi(\overline{m}) \leftrightarrow \varphi(\overline{\varphi\overline{m}})$.

Before continuing the discussion of gn_1, we define the third Gödel numbering. It is the standard numbering: $\mathrm{gn}_2(\vartheta)$ is the unique m such that $\vartheta = \alpha_m$, that is, $\mathrm{gn}_2(\vartheta) = \max(\mathrm{gn}_0(\vartheta))$. The syntactical operations can be defined in the obvious way, for instance:

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

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.

We can use the codings developed in this appendix, to illustrate an important point. Design choices can be made in part independently of each other. As we will illustrate we can develop a proof predicate in a way that is independent of specific choices concerning the Gödel numbering such that specific choices concerning the Gödel numbering can be plugged in.

Suppose our theory is axiomatized by a schema, say $\mathcal{S}$. We can find schematic formulae $\mathrm{Form}(x)[X, Y, Z \ldots]$ and $\mathrm{Scheme}(x)[X, Y, Z, \ldots]$ and $\mathrm{Bew}(x)[X, Y, Z, \ldots]$ such that, if specific arithmetical formulae At, neg, conj, … are given representing the atomic formulae, the syntactical operation of negation, the syntactical operation of conjunction, …, then $\mathrm{Form}(x)[\mathrm{At}, \mathrm{neg}, \mathrm{conj}, \ldots]$ represents the class of Gödel numbers of formulae for the given Gödel numbering and $\mathrm{Scheme}(x)[\mathrm{At}, \mathrm{neg}, \mathrm{conj}, \ldots]$ represents the class of Gödel numbers which stand for axioms given by the scheme $\mathcal{S}$ and, finally, $\mathrm{Bew}(x)[\mathrm{At}, \mathrm{neg}, \mathrm{conj}, \ldots]$ repesents the class of codes of proofs from $\mathcal{S}$. The definition of e.g. $\mathrm{Bew}(x)[X, Y, Z, \ldots]$ does depend on a number of conventional choices like the choice of the proof system and the choice of a sequence coding, but it does not depend on the choice of the Gödel numbering.[12]

Thus, we can develop in the way described above predicates Bew_i that correspond in a uniform way to the Gödel numberings gn_i.

[12] The ideas of this remark could be made even clearer by viewing the process of arithmetization as a bootstrap executed by developing a sequence of better and better interpretations.

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

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 an 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 have only two options: Either we 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 $G = \neg\mathrm{Bew}_1(\overline{\mathrm{gn}_1(G)})$.

REMARK 12.4. *Clearly, we will have a definable function switch*(x) *such that we have* $\mathrm{switch}(m) = \mathrm{gn}_2(\beta_m)$*, and such that our theory verifies*

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

Note that it does not follow from the assumption that Bew_1 *is intensionally correct with respect to* gn_1*, that also the verifiably extensionally equivalent predicate* $\mathrm{Bew}'_1(x) := \mathrm{Bew}_2(\mathrm{switch}(x))$ *is intensionally correct with respect to* gn_1.

Is conj_2 intensionally correct with respect to 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*:

$$\mathrm{conj}(\sigma, \tau) := \ulcorner(\urcorner * \sigma * \ulcorner\wedge\urcorner * \tau * \ulcorner)\urcorner$$

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 putative necessity of a 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 of 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.

§13. Acknowledgments. Volker Halbach's work was supported by the Arts & Humanities Research Council AH/H039791/1. We thank Rasmus Blanck, Cezary Cieśliński, Kentaro Fujimoto, Graham Leigh, Hannes Leitgeb, Dan Isaacson, Arthur Merin, Lavinia Picollo and two referees for valuable comments and suggestions. We are especially indebted to Christopher von Bülow, who provided numerous corrections.

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press

BIBLIOGRAPHY

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

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

Guaspari, D., & Solovay, R. M. (1979). Rosser sentences. *Annals of Mathematical Logic*, **16**, 81–99.

Hájek, P., & Pudlák, P. (1993). *Metamathematics of First-Order Arithmetic*. Perspectives in Mathematical Logic, Vol. 3. Berlin: Springer.

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

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

Kaye, R. (1991). *Models of Peano Arithmetic*. Oxford Logic Guides. Oxford: Oxford University Press.

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

Kreisel, G., & Lévy, A. (1968). Reflection principles and their use for establishing the complexity of axiomatic systems. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, **14**, 97–142.

Kurahashi, T. (2014). Henkin sentences and local reflection principles for Rosser provability. To appear.

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

Shavrukov, V. Y. (1994). A smart child of Peano's. *Notre Dame Journal of Formal Logic*, **35**, 161–185.

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

Visser, A. (2004). Semantics and the liar paradox. In Gabbay, D., & Guenthner, F., editors. *Handbook of Philosophical Logic* (second ed.), Vol. 11, Heidelberg: Springer, 149–240.

van Fraassen, B. C. (1970). Inference and self-reference. *Synthese*, **21**, 425–438.

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

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

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

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

https://doi.org/10.1017/S175502031400029X Published online by Cambridge University Press