Journal of Philosophical Logic (2023) 52:1005–1052
https://doi.org/10.1007/s10992-022-09696-y

Varieties of Self-Reference in Metamathematics

Balthasar Grabmayr[1] · **Volker Halbach**[2] · **Lingyuan Ye**[3]

Received: 9 November 2021 / Accepted: 20 December 2022 /Published online: 14 March 2023

Abstract
This paper investigates the conditions under which diagonal sentences can be taken to constitute paradigmatic cases of self-reference. We put forward well-motivated constraints on the diagonal operator and the coding apparatus which separate paradigmatic self-referential sentences, for instance obtained via Gödel's diagonalization method, from *accidental* diagonal sentences. In particular, we show that these constraints successfully exclude refutable Henkin sentences, as constructed by Kreisel.

Keywords Self-reference · Arithmetic · Intensionality · Diagonalization · Well-foundedness · Uniformity · Henkin sentence

1 Self-Reference and Diagonalization

1.1 Diagonal Sentences

It is common to talk about *the* Gödel sentence, *the* Henkin sentence, *the* liar sentence, and other allegedly self-referential sentences. We take these labels as abbreviations

We thank Albert Visser and two referees for valuable comments and suggestions. Volker Halbach gratefully acknowledges the support of his work on this paper by the Leverhulme Trust. Balthasar Grabmayr thanks the Azrieli Foundation for their generous support. The order of the authors is alphabetical. The authors contributed equally to this work.

✉ Volker Halbach
volker.halbach@philosophy.ox.ac.uk

Balthasar Grabmayr
balthasar.grabmayr@gmail.com

Lingyuan Ye
ye.lingyuan.ac@gmail.com

[1] Department of Philosophy, University of Haifa, Haifa, Israel

[2] University of Oxford and New College, Oxford, UK

[3] ILLC, University of Amsterdam, Amsterdam, Netherlands

of definite descriptions of sentences: By "*the* Gödel sentence" we mean the sentence stating its own unprovability, by "*the* Henkin sentence" the sentence asserting its own provability, by "*the* liar sentence" the sentence stating its own falsity, and so on.

In each case, it is obvious that the definite article is incorrect. In the case of "the Gödel sentence", it is obvious that there are many different ways of defining an unprovability predicate $\neg\mathsf{Bew}(x)$ from which the Gödel sentence is constructed. The formula $\mathsf{Bew}(x)$ will depend on the particular theory and its language, the chosen Gödel coding, and then on how the provability predicate is defined relative to that coding. The problems of choosing a reasonable coding relative to the theory under consideration and then defining a suitable provability predicate relative to the coding and theory are well-known, thoroughly studied, and increasingly better understood.

When logicians talk about *the* Gödel sentence, they often assume that a language and a sound theory (or family of theories) have been fixed, and a "natural" provability predicate has been been chosen. They often assume that this ensures that $\mathsf{Bew}(x)$ satisfies the Löb derivability conditions. The usual justification for speaking about *the* Gödel sentence is then that all diagonal or fixed-point sentences of $\neg\mathsf{Bew}(x)$ are provably equivalent. That is, if Σ is some suitable system of arithmetic, $\Sigma \vdash \varphi_1 \leftrightarrow \neg\mathsf{Bew}(\ulcorner\varphi_1\urcorner)$ and $\Sigma \vdash \varphi_2 \leftrightarrow \neg\mathsf{Bew}(\ulcorner\varphi_2\urcorner)$ imply $\Sigma \vdash \varphi_1 \leftrightarrow \varphi_2$. Moreover, all these sentences are provably equivalent to the consistency statement for Σ. Provable equivalence, however, is a very coarse grained kind of equivalence, too coarse grained for justifying the definite article in "*the* Gödel sentence".

Finding a diagonal sentence of $\neg\mathsf{Bew}(x)$ is not trivial, and this might give the impression that these diagonal sentences are all mere trivial variations of each other. However, this is not the case. Whether a sentence is a diagonal sentence of $\neg\mathsf{Bew}(x)$ is not decidable. In fact, the set of diagonal sentences of any formula is undecidable [10, observation 2.2]. Of course, the provable equivalence of all diagonal sentences of $\neg\mathsf{Bew}(x)$ means that we do not have to distinguish between them as long as we are only interested in their provability, and their properties analyzable in standard (propositional) provability logic. However, if this is taken as justification for talking about *the* Gödel, one could also talk about *the* theorem of Peano arithmetic, as there is only one such theorem up to provable equivalence.

In the case of "*the* Henkin sentence", the problems are more blatant: By Löb's theorem all diagonal sentences of the provability predicate are provable and thus provably equivalent with each other and *the* theorem of Peano arithmetic. Clearly, they are not just trivial variations of each other, and the definite description "*the* Henkin sentence" cannot apply to them.

If formulas other than provability predicates satisfying the Löb conditions are considered, the definite article is even less plausible, because diagonal sentences of a given formula may behave in very different ways and, in particular, fail to be provably equivalent. For instance, we can look at *the* truth teller sentence constructed from the Σ_n-truth predicate $\mathrm{Tr}_{\Sigma_n}(x)$ for some $n \geq 1$. The formula $\mathrm{Tr}_{\Sigma_n}(x)$ itself is a Σ_n formula. Applying some canonical diagonalization procedure to $\mathrm{Tr}_{\Sigma_n}(x)$ thus yields a Σ_n-sentence (see [10] for a more detailed discussion). The point of a truth predicate for a class C of sentences is that *all* sentences in C are fixed-point sentences. $0 = 0$ and $0 = 1$ will be diagonal sentences of any partial truth predicate $\mathrm{Tr}_{\Sigma_n}(x)$. Thus, if we are interested in the question whether the Σ_n-truth teller sentence is provable,

refutable, or independent for a given n, we have to ask about very specific diagonal sentences – in contrast to the case of Gödel and Henkin sentences with a provability predicate satisfying the Löb conditions.

1.2 Self-Reference

In the case of Σ_n-truth we cannot dodge the question of what *the* truth teller sentence is by proving a general theorem about all diagonal sentences of the Σ_n-truth predicate, because all Σ_n-sentences are fixed-points, and some will be provable, while others are not. The same applies to provability for which the Löb derivability conditions may fail. These provability predicates need not be highly contrived: We can consider cut-free provability or Rosser provability. We may consider further formulas expressing other properties, possibly formulated in proper extensions of the language of arithmetic with a primitive truth or necessity predicate, for instance.

We do not expect that we can narrow down the class of diagonal sentence until only one sentence is left that deserves to be called "*the* sentence asserting its own P", where P is the property expressed by the formula. However, we may still hope to be able to narrow down the class so that all remaining sentences behave in the same way. We may even hope to narrow down the class for some formulas to a point where all remaining sentences are so similar (given certain restrictive choices) that we are warranted to use the singular and talk about *the* sentence asserting its own P.

Perhaps we will never be able to arrive at a suitable class of sentences, but rather realize that the status of the sentences depends on accidental choices in the coding or the definition of the formula in some haphazard way. In other cases we may arrive at a fairly stable result, although we may not be as lucky as in the case of canonical provability with the Löb conditions where *all* diagonal sentences behave in the same way; but we may be able to establish a result that applies to a sufficiently interesting class of diagonal sentences.

In the case of the formula $\exists y\, \mathsf{SS0} \times y = x$ expressing that x is even, for instance, there is little hope to obtain a stable result about sentences stating their even parity. Even the status of diagonal sentences obtained in some canonical way will depend on the coding and the method of diagonalization in a haphazard way. "Metatheoretic" properties such as provability and truth will generally yield more stable results. By metatheoretic properties we mean here properties expressible in some non-arithmetized metatheory such as the theories described in [9]; but we do not attempt here to make this distinction sharp.

Continuing to ask about self-referential sentences, even if the equivalence of all fixed-points fails, can lead to important insights. Ironically, the most striking example is the discussion that leads up to the discovery of Löb's theorem. Kreisel [16] replied to Henkin's [13] question: "[...] the answer to Henkin's question depends on which formula is used to 'express' the notion of *provability in* Σ". Kreisel regretted that he did not keep asking. Löb, in contrast, continued asking and proved his celebrated theorem [18] by imposing further restrictions on the formula expressing provability. Hence, by specifying such restrictions on how a property may be expressed and how self-reference is obtained, one can achieve definitely noteworthy results.

In the present paper we do not delve into the intricacies of what it means for a formula of arithmetic to express some property P and ask on which diagonal sentences we should focus, once a formula expressing P is given. These fixed-point sentences should be self-referential.

Defining what it means for a sentence to be self-referential is notoriously difficult. Self-reference may be thought to be reducible to aboutness by the following definition: A sentence is self-referential if, and only if it is about itself. But then presumably the sentence $\forall x\, x = x$ is self referential because it states the self-identity of everything, including the sentence itself. Halbach and Leigh [9] and Picollo [20] provide more comments on self-reference via quantification.[1] Following [10] and [12], we consider only self-reference via a closed term. That is, to be self-referential, more precisely to ascribe a property to itself, that sentence must contain a closed term that refers to that sentence.[2]

Before we provide a precise definition of self-reference via terms, some technical preliminaries are in order. Let $\mathcal{L}_0$ contain the logical symbols $=, \neg, \wedge, \forall$ together with the constant symbol $\mathsf{0}$, the unary function symbol S and the binary function symbols $+$ and $\times$. Let $\mathcal{L}$ be an effective extension of $\mathcal{L}_0$ that contains a function symbol for each p.r. function, which may also contain further constants, functions symbols or predicates which we do not specify explicitly. Let Σ be a consistent recursive enumerable $\mathcal{L}$-theory which contains R together with all true identities of closed $\mathcal{L}$-terms.

The name of a string in $\mathcal{L}$ is given by a numbering and a numeral function. We call an injective and effective function which maps $\mathcal{L}$-expressions to numbers a *numbering*. We write # for standard numberings. We call an injective function $\nu : \omega \to \mathrm{ClTerm}$ that maps each number to a closed term of $\mathcal{L}$ which has the same value a *numeral function*. A numbering α and a numeral function ν induce a *naming function* $\ulcorner - \urcorner$, which is the composition $\nu \circ \alpha$. In order to make α and ν explicit, we also sometimes write $\ulcorner \varphi \urcorner^{\alpha,\nu}$ for $\ulcorner \varphi \urcorner$. If ν is the standard numeral function, i.e., $\nu(n) = \overline{n}$ for all $n \in \omega$, we sometimes also write $\ulcorner \varphi \urcorner^{\alpha}$. Finally, $\equiv$ denotes syntactic identity between expressions.

Definition 1.1 (Kreisel–Henkin Criterion for Self-Reference) Under a specific coding, a sentence says of itself that it has the property expressed by the formula $\varphi(x)$, if it is of the form $\varphi(t)$ where t is a closed term that satisfies the following condition,

$$\Sigma \vdash t = \ulcorner \varphi(t) \urcorner$$

[1] Perhaps different forms of self-reference via quantification are somehow reducible to self-reference via a closed term. In standard first-order logic, quantifiers range over the entire domain, and restrictions are expressed in Frege's way with connectives. Frege's insight that the binary quantifiers of syllogistic logic are expressible with a unary quantifier makes the notion of aboutness difficult to capture. Here we remain agnostic about self-reference via quantification and concentrate on the Kreisel–Henkin criterion below.

[2] For discussions and applications of the criterion see [9–11].

Therefore, if t refers to $\varphi(t)$, that is, if it has the code of $\varphi(t)$ as its value, Σ will prove $t = \ulcorner\varphi(t)\urcorner$. Of course, $\Sigma \vdash t = \ulcorner\varphi(t)\urcorner$ implies $\Sigma \vdash \varphi(t) \leftrightarrow \varphi(\ulcorner\varphi(t)\urcorner)$ and $\varphi(t)$ is thus also a diagonal sentence of $\varphi(x)$.

We say that a sentence has the Kreisel–Henkin property if, and only if it is of the form $\varphi(t)$ and ascribes to itself the property expressed by $\varphi(x)$ in the sense of the Kreisel–Henkin criterion. We also call such term t a fixed-point term of $\varphi(x)$.

2 Accidental Self-Reference

In the presence of suitable function symbols, Gödel's diagonalization method enables us to find a suitable closed term t_φ for each $\varphi(x)$ such that $\Sigma \vdash t_\varphi = \ulcorner\varphi(t_\varphi)\urcorner$. "Suitability" needs to be understood relative to the coding scheme used. However, instead of using a systematic method for arriving at diagonal sentences that satisfy the Kreisel–Henkin criterion, for any given $\varphi(x)$, we could also try a brute force method by enumerating all closed terms of the language and browse through them until we have found the first term t_φ with $\Sigma \vdash t_\varphi = \ulcorner\varphi(t_\varphi)\urcorner$.

We may strike it lucky and the first t_φ could be one that would have been generated by a systematic method; but we could also stumble upon some diagonal sentence with the Kreisel–Henkin property "accidentally". We could rig the game and set up the coding in such a way that there is an easy to find t_φ. We can even use numerals $\mathsf{S}\cdots\mathsf{S0}$ as the only closed terms and use the coding schema from [26]. The question is whether all these diagonal sentences with the Kreisel–Henkin property are all also obtainable with the usual Gödel diagonal or some similar systematic method and, if there are other such diagonal sentences whether they differ in their properties from the diagonal sentences obtained by some reasonable systematic method.

Let us call a fixed-point sentence not obtained by systematic method an *accidental* diagonal sentence. Of course, this is not (yet) a precise definition, and we have given no evidence that there are indeed any accidental diagonal sentences.[3] Before trying to make the distinction between accidental and non-accidental diagonal sentences precise, we provide examples of clearly accidental fixed-point sentences that behave in ways that are very different from those of generated by the usual systematic methods.

If we ask about *the* Σ_n-truth teller or *the* Henkin sentence, we will select a sentence for the diagonal sentences of a given predicate that ascribe to themselves the relevant property by the Kreisel–Henkin criterion. But we may then still be left with accidental and non-accidental diagonal sentences with rather different properties. In this case we would select those obtained by some systematic method and not the accidental diagonal sentences. If there is *the* Σ_n-truth teller or *the* Henkin sentence it will have been arrived at by a systematic method not by some quirk in the coding or some clever trick that works not generally, but only for the predicate in question.

Accidental diagonal sentences may satisfy the Kreisel–Henkin criterion because of some very specific feature of the formula that is being diagonalized. An accidental

[3] Carnap [6] distinguished between accidental and functional self-reference. We are using "accidental" in a different, but related sense.

diagonal sentence may be chosen in a very *ad hoc* way to obtain a specific result. However, if we ask about self-referential sentences such as truth teller sentences, we are more interested in knowing the properties of diagonal sentences that have been constructed in a straightforward way and not in those obtained by some trickery.

In this paper we do not attempt to provide a thorough defence for preferring non-accidental fixed-point sentences. Before one can enter this discussion, we need to show that it is possible to come up with a precise distinction; we also need to provide examples of accidental diagonal sentences with the Kreisel–Henkin property that behave differently from all non-accidental ones.

First we provide some examples of accidental diagonal sentences taken from the literature. The first example is Kreisel's [16] refutable Henkin sentence.[4] Of course, Kreisel had to employ a deviant provability predicate. He claimed that whether the Henkin sentence is provable or not depends on the way provability is expressed. However, it also depends on how the formula expressing provability is diagonalized. Only accidental diagonal sentences of Kreisel's provability predicate are refutable; those obtained in a systematic and uniform way are provable, as we are going to show.

Observation 2.1 *Let* $\mathsf{Bew}(x)$ *be a provability predicate that weakly represents provable sentences, viz. for any sentence* ψ, $\Sigma \vdash \psi$ *iff* $\Sigma \vdash \mathsf{Bew}(\ulcorner \psi \urcorner)$. *Let* t *be a term satisfying the Kreisel–Henkin criterion with respect to the formula* $x \neq x \wedge \mathsf{Bew}(x)$, *i.e.,*

$$\Sigma \vdash t = \ulcorner t \neq t \wedge \mathsf{Bew}(t) \urcorner.$$

Define another predicate $\mathsf{Bew}_K(x)$ *to be the following one*

$$\mathsf{Bew}_K(x) :\equiv x \neq t \wedge \mathsf{Bew}(x).$$

Then $\mathsf{Bew}_K(x)$ *also weakly represents provability, and* $\mathsf{Bew}_K(t)$ *is a refutable sentence stating its own provability with respect to the Kreisel–Henkin criterion.*

Proof We show that $\Sigma \vdash \varphi$ iff $\Sigma \vdash \mathsf{Bew}_K(\ulcorner \varphi \urcorner)$ by distinguishing two cases. If $\varphi \not\equiv t \neq t \wedge \mathsf{Bew}(t)$, then $\Sigma \vdash \ulcorner \varphi \urcorner \neq t$. Hence,

$$\Sigma \vdash \mathsf{Bew}_K(\ulcorner \varphi \urcorner) \text{ iff } \Sigma \vdash \mathsf{Bew}(\ulcorner \varphi \urcorner).$$

The claim then follows from the fact that $\mathsf{Bew}(x)$ weakly represents provable sentences. If $\varphi \equiv t \neq t \wedge \mathsf{Bew}(t)$, we have $\Sigma \vdash \neg\varphi$ and $\Sigma \vdash \neg\mathsf{Bew}_K(\ulcorner \varphi \urcorner)$. Thus, $\mathsf{Bew}_K(x)$ also weakly represents provability.

According to the definition, $\mathsf{Bew}_K(t) \equiv t \neq t \wedge \mathsf{Bew}(t)$, hence by the assumption

$$\Sigma \vdash t = \ulcorner \mathsf{Bew}_K(t) \urcorner.$$

This shows that t is also a fixed-point term with respect to $\mathsf{Bew}_K(x)$, and obviously $\mathsf{Bew}_K(t)$ is refutable. □

[4] Kreisel's provability predicate was different, and the version here is due to Henkin, who was the referee for Kreisel's paper.

Intuitively, the deviant Henkin sentence $\mathsf{Bew}_K(t)$ is not the result of some systematic fixed-point construction. Rather, t is already contained in $\mathsf{Bew}(x)$ and "happens" to be its fixed-point term [11, p. 701].

We can start from a provability predicate $\mathsf{Bew}(x)$ satisfying the Löb derivability conditions. It is not hard to see that applying the usual canonical diagonalization method yields a term s distinct from t and that the resulting Henkin sentence $\mathsf{Bew}_K(s)$ is provable [10, observation 4.1].

We can generalize the above method of obtaining fixed-point sentences:

Observation 2.2 *Let $\varphi(x)$ be a formula with one designated free variable x. Suppose there are n (marked) free occurrences of x in φ ($n \geq 1$). Given any fixed-point term t of φ, i.e.*

$$\Sigma \vdash t = \ulcorner \varphi(t) \urcorner,$$

and given any proper subset $S \subset \{1, 2, \ldots, n\}$, let φ^t_S be the formula obtained by substituting the i-th occurrence of the free variable x in φ by t, for every $i \in S$ ($\varphi^t_\emptyset \equiv \varphi$). Then t is also a fixed-point term of φ^t_S.

Proof Since S is a proper subset, φ^t_S still contains at least one free occurrences of x, hence is still a formula with the only free variable x. Now according to our definition it is easy to see that

$$\varphi(t) \equiv \varphi^t_S(t).$$

This in particular means that the codes of $\varphi^t_S(t)$ and $\varphi(t)$ coincide. Since t satisfies the Kreisel–Henkin criterion with respect to $\varphi(x)$, we also have

$$\Sigma \vdash t = \ulcorner \varphi^t_S(t) \urcorner.$$

Hence, t is also a fixed-point term of $\varphi^t_S(x)$. □

Kreisel's original refutable Henkin sentence in [16] – not Henkin's simplified version $\mathsf{Bew}_K(t)$ above – can be obtained from this observation with $n = 2$.

Observation 2.2 has some crucial implications about whether the Kreisel–Henkin criterion alone can be used as a sufficient condition for genuine self-reference. Firstly, if for different subsets S the formula φ^t_S expresses different syntactical properties, then according to the Kreisel–Henkin criterion the formula $\varphi(t)$ self-ascribes several different properties. The number of different proper subsets, or in other words the number of different self-ascribing properties, is equal to $2^n - 1$, which grows exponentially. Secondly, for most proper subsets S, if we apply the usual diagonal construction directly to the formula φ^t_S, we in general obtain a closed term t_S different from t. The two sentences $\varphi^t_S(t_S)$ and $\varphi^t_S(t)$ will not be provably equivalent in general as well. In this sense, $\varphi(t)$ would be an accidental fixed-point for most of these formulas φ^t_S.

We now introduce examples of accidental diagonal sentences which result from contrived codings. In particular, we provide a counterpart of Observation 2.1 on the level of Gödel numberings. That is, by suitably tweaking the coding, we obtain a deviant provability predicate such that the fixed-point property of the resulting Henkin sentence is directly implemented into the coding. As in the case of Kreisel's

provability predicate above, only accidental diagonal sentences thus obtained are refutable, while those constructed in a systematic and uniform manner are provable.

Observation 2.3 *Let # be a standard numbering such that each #-code is positive and even. Let* $\mathsf{Bew}(x)$ *be a provability predicate that weakly represents provability, i.e., for every sentence* ψ, $\Sigma \vdash \psi$ *iff* $\Sigma \vdash \mathsf{Bew}(\ulcorner \psi \urcorner^{\#})$. *We assume that* $\Sigma \vdash \neg\mathsf{Bew}(\overline{n})$, *for each odd n (e.g., this holds for Feferman's* [5] *standard provability predicate). Let* $\widetilde{1} :\equiv \mathsf{S0}$ *and* $\widetilde{m+1} :\equiv (\widetilde{m} + \mathsf{S0})$. *Let m be the smallest odd number such that* $\overline{m}$ *does not occur in* $\mathsf{Bew}(\widetilde{m})$.[5]

We now change our old standard numbering # by defining a new numbering α *as follows*

$$\alpha(\chi) := \begin{cases} m & \text{if } \chi \equiv \mathsf{Bew}(\widetilde{m}); \\ \#\chi & \text{otherwise.} \end{cases}$$

Then $\mathsf{Bew}(x)$ *weakly represents provability relative to* α, *i.e., the set of* α*-codes of* Σ*-theorems. Moreover,* $\mathsf{Bew}(\widetilde{m})$ *is a refutable sentence which states its own provability with respect to the Kreisel–Henkin criterion.*

Intuitively, the refutable Henkin sentence $\mathsf{Bew}(\widetilde{m})$ is accidental. For its fixed-point term $\widetilde{m}$ is not obtained by a systematic method, but rather by a contrived numbering which is specifically tailored for this purpose. Even if $\mathsf{Bew}(x)$ is a canonical provability predicate w.r.t. the numbering #, $\mathsf{Bew}(x)$ does not satisfy Löb's conditions w.r.t. the numbering α.[6] As in the case above, assume that $\mathsf{Bew}(x)$ satisfies Löb's conditions w.r.t. the numbering #. Let s be a fixed-point term of $\mathsf{Bew}(x)$ which is obtained by the usual canonical diagonalization method relative to the coding α. That is, $\Sigma \vdash \mathsf{Bew}(s) \leftrightarrow \mathsf{Bew}(\ulcorner \mathsf{Bew}(s) \urcorner^{\alpha})$. It is then easy to see that s is different to $\widetilde{m}$, and hence that $\mathsf{Bew}(s)$ is a provable Henkin sentence.

Numberings which are designed to immediately provide fixed-points with the Kreisel–Henkin property are sometimes said to have "built-in diagonalization. Paradigmatic examples of such numberings are so-called self-referential numberings" [8, definition 3.3]. It is to be expected that results about axiomatic theories of truth are most stable, because the axioms are formulated relative to a fixed coding, while defined notions such as the usual provability predicate are highly relative to the coding. Heck [12, p. 14ff] showed that even axiomatic theories of truth are sensitive to the chosen coding (and the language). Again, one has to be very careful about what an axiomatic truth theory is, independently of a fixed coding. In semantic, non-classical theories of truth, [2] had already observed sensitivities to the codings. See also [22, Section 2.2] and [8, Section 9] for more recent examples of intensionality with respect to truth theories which result from numberings which have built-in diagonalization. Whether all numberings with built-in diagonalization yield accidental fixed-points is a delicate question which we will briefly address in Section 8.2.

[5] Here we could choose $\overline{n}$ instead of $\widetilde{m}$, for any odd n. The particular choice of $\widetilde{m}$ will only be relevant in Section 8.2.

[6] This follows from Löb's theorem, which also holds for the contrived numbering α [7, section 4.2].

Plan of the Paper The main goal of this paper is to make the distinction between accidental and non-accidental fixed-points precise. The basic idea is that non-accidental fixed-points are constructed in a uniform way. A precise notion of uniformity is introduced in Section 3, where we also show that the canonical fixed-point constructions found in the literature are uniform in our sense. In the remainder of the paper, we examine the extent to which the uniformity constraint rules out accidental fixed-points.

We start with Kreisel's construction as a paradigmatic case of an accidental fixed-point. According to our analysis, this fixed-point construction is accidental since Kreisel's provability predicate $\mathsf{Bew}_K(x)$ already contains its own fixed-point term. We ask whether, or more generally, under which additional assumptions, the uniformity requirement rules out the possibility that a predicate contains its own fixed-point term (Question 3.6).

In Section 4, we show that uniformity alone is not sufficient to exclude refutable Henkin sentences which contain their own fixed-point terms. Rather, accidental fixed-points can also result from contrived choices of the numbering or the numeral function. However, in Section 5 we show that Kreisel-like constructions can be successfully excluded by 1) requiring uniformity of the diagonal operator and 2) requiring the numbering and the numeral function to induce a non-circular *weak naming relation*. As we argue in Section 6, the constraint of well-foundedness, which implies non-circularity, is natural and well-motivated. This provides a satisfactory answer to Question 3.6 and completes the main part of the paper.

In Section 7, we introduce a new construction of refutable Henkin sentences which are accidental, but do not contain their own fixed-point terms. We show that the constraints of uniformity and the non-circularity of the weak naming relation taken together do not rule out this construction, but uniformity plus the well-foundedness of the weak naming relation do. In Section 8, we provide a different metamathematical context in which the uniformity constraint successfully singles out non-accidental fixed-points. Moreover, we briefly address the question whether all numberings with built-in diagonalization yield accidental fixed-points. Finally, in Section 9 we extract some conclusions.

3 Uniformity

In this section the distinction between accidental and non-accidental is made precise. The non-accidental fixed-points are obtained in a uniform and systematic way. These systematic methods can be extracted from the usual textbook proofs of the diagonal lemma. These methods apply uniformly to all formulas. The precise definition of a *uniform* diagonal construction allows us to distinguish the non-accidental fixed-point constructions, including the usual Gödel's diagonal method and alike, and the more accidental ones provided by Kreisel. The use of uniformity to distinguish the two kinds of fixed-point constructions can be traced back to [11]. But the definition of uniformity given there is defective, since, according to the definition there, Gödel's diagonal construction would not be uniform.

It is our task here to give a more adequate definition of uniformity and provide an extensive study of its implications for self-reference. There are several considerations that motivate and shape our formulation for uniformity below. First of all, as already mentioned, the intuition for uniformity is that a uniform construction should not result in a fixed-point depending on very specific syntactical features of the formula it diagonalizes; it should diagonalize all the formulas with a designated free variable by similar means.

Second of all, since the notion of uniformity is essentially restricting the class of constructions we are allowed to perform on the syntactical objects, it is natural to define it in a recursive way: we specify basic operations that are uniform in a very intuitive sense, and a uniform construction is then a finite composition of all the basic uniform constructions. Except requiring the basic operations to be intuitively uniform, we also want them to be possibly carried out in a syntax theory, e.g. in the sense of Halbach and Leigh [9]. This reflects our general contemplation on the subject: if an arithmetic sentence could refer to a syntactical object at all, then when constructing such a sentence we must mimic what we can do in syntax theory. The operations provided in a syntax theory include substitution, quotation, and concatenation. As you will soon see, these are indeed the basic constructions we allow, except for a modification for concatenation, which links to our final motivation.

The final consideration is that we want our defined constructions to always yield well-defined syntactical objects. This leads us to a typed approach to define uniformity. The unrestricted form of concatenation will not always result in well-formed formulas or terms, thus we have replaced concatenation with three collections of well-typed operations, associated with logical connectives, function symbols, and predicates. We also distinguish operations which only differ with respect to their domain or codomain, such as the substitution or naming function. As we will see below, these distinctions will permit a conceptually more refined introduction of the basic constructions and will lead to technically important results (see Lemma 5.9).

After spelling out all the motives, we now provide the precise formulation of the notion of uniformity. Since our aim is to analyse sentences that self-ascribe properties which can be expressed by unary predicates, we restrict ourselves to terms and formulas with a single (designated) free variable x. Let Fml_x and Term_x be the set of all $\mathcal{L}$-formulas and $\mathcal{L}$-terms respectively which at most contain x as a free variable. As usual, for any set A and $n \geq 1$ let A^n denote the Cartesian product $A \times \cdots \times A$ which consists of n factors; A^1 will simply be A. We set $A^0 := 1$, where 1 denotes a designated singleton set, which remains fixed throughout this paper. Let $*$ denote the unique element of 1, i.e., $1 = \{*\}$. To be precise, the binary product is not strictly associative, though $(A \times B) \times C$ is canonically isomorphic to $A \times (B \times C)$.

The fact that we have these canonical isomorphisms, and that these isomorphisms interact in a coherent way, justifies our usual sloppy way of writing $A \times \cdots \times A$, and permits us to freely view A^n as $A^m \times A^l$, whenever $m + l = n$. However, precisely speaking, we will assume A^n to be the product $A \times (\cdots (A \times (A \times A)) \cdots)$. This level of precision will only affect the precise form of Definition 3.1 below and the materials in Section 5 where a more careful treatment is needed. In other places in this paper, however, we will suppress this level of precision as usual.

The following meta-linguistic operations will serve as the basic constituents of uniform constructions:[7]

(1) Two meta-linguistic substitution functions:

$$\begin{aligned} \mathrm{Sub}_f &: \mathrm{Fml}_x \times \mathrm{Term}_x \to \mathrm{Fml}_x, \\ \mathrm{Sub}_t &: \mathrm{Term}_x \times \mathrm{Term}_x \to \mathrm{Term}_x . \end{aligned}$$

Given any formula $\varphi(x) \in \mathrm{Fml}_x$ and any term $t(x) \in \mathrm{Term}_x$, application of Sub_f yields the result of substituting the term $t(x)$ for x in $\varphi(x)$, i.e.,

$$\mathrm{Sub}_f(\varphi(x), t(x)) \equiv \varphi(t(x)).$$

Similarly, for any two terms $t(x)$ and $s(x)$, we have

$$\mathrm{Sub}_t(t(x), s(x)) \equiv t(s(x)).$$

(2) The naming functions for formulas and terms:[8]

$$\begin{aligned} \ulcorner - \urcorner_f &: \mathrm{Fml}_x \to \mathrm{Term}_x, \\ \ulcorner - \urcorner_t &: \mathrm{Term}_x \to \mathrm{Term}_x . \end{aligned}$$

(3) Given any n-ary logical connective $\star$ (for quantifiers we only consider ones that bind x), the meta-linguistic function

$$\overline{\star}\colon \mathrm{Fml}_x^n \to \mathrm{Fml}_x,$$

given by

$$\overline{\star}(\varphi_1(x), \cdots, \varphi_n(x)) \equiv \star(\varphi_1(x), \cdots, \varphi_n(x)).$$

In our language $\mathcal{L}$, $\star$ ranges over $\{\neg, \wedge, \forall x\}$.

(4) Given any n-ary function symbol f of $\mathcal{L}$,[9] the meta-linguistic function

$$\overline{f}\colon \mathrm{Term}_x^n \to \mathrm{Term}_x,$$

given by

$$\overline{f}(t_1(x), \cdots, t_n(x)) \equiv f(t_1(x), \cdots, t_n(x)).$$

(5) Given any n-ary predicate symbol R of $\mathcal{L}$, the meta-linguistic function

$$\overline{R}\colon \mathrm{Term}_x^n \to \mathrm{Fml}_x,$$

given by

$$\overline{R}(t_1(x), \cdots, t_n(x)) \equiv R(t_1(x), \cdots, t_n(x)).$$

(6) The function which introduces the variable term x,

$$\overline{x}\colon 1 \to \mathrm{Term}_x,$$

where $\overline{x}$ sends the unique element $*$ in 1 to the term x.

[7] It will be evident from the definition below that uniformity applies more generally to a wider range of languages, which we do not consider in this paper.

[8] Recall from Section 1.2 that the naming function $\ulcorner - \urcorner$ are induced by a numbering and a numeral function and can be applied to any string. However, for reasons which will become clear at a later stage of this paper, it is useful to distinguish naming functions for well-formed formulas and terms respectively (cf. Section 5). When this distinction is not important or when we consider a naming function for all strings, we also write $\ulcorner - \urcorner$ as usual, without a subscript.

[9] We identify constants with 0-ary function symbols.

While the functions given in (1) and (3)– (6) are fixed, we treat the naming function $\ulcorner - \urcorner$ as a parameter which has to be specified. To be fully precise and explicit, we call the above functions *basic operations containing* $\ulcorner - \urcorner$. We define uniform functions based on such a class as follows:

Definition 3.1 Let D be the smallest collection of sets containing 1, Fml_x, Term_x which is closed under binary products. Let A, B be sets in D. A function $f\colon A \to B$ is called *uniform for* $\ulcorner - \urcorner$ if it is contained in the smallest class of functions that includes the following *C-basic functions containing* $\ulcorner - \urcorner$:

- each basic operation containing $\ulcorner - \urcorner$;
- identity functions $\mathrm{id}_{\mathrm{Fml}_x}$, $\mathrm{id}_{\mathrm{Term}_x}$ on Fml_x and Term_x;
- projection maps $\pi_1^{A,B}$, $\pi_2^{A,B}$ from $A \times B$ to A, B, with $A, B \in \mathsf{D}$;[10]
- a uniquely determined function $!_A\colon A \to 1$, for each $A \in \mathsf{D}$;

and which is closed under composition and maps canonically induced by the Cartesian product structure; that is, we have

- composition of two uniform functions for $\ulcorner - \urcorner$ is uniform for $\ulcorner - \urcorner$;
- if $f\colon A \to B$, $g\colon A \to C$ are uniform for $\ulcorner - \urcorner$, then so is

$$\langle f, g\rangle\colon A \to B \times C,$$

which maps $a \in A$ to $(f(a), g(a)) \in B \times C$.

If we want to be explicit about both the numbering function α and the numeral function ν that constitute the naming function $\ulcorner - \urcorner$, we also say a function is *uniform for* $\nu \circ \alpha$. If ν is the standard numeral function that takes n to its numeral $\overline{n}$ for any $n \in \omega$, then we also omit mentioning it explicitly and say a function is *uniform for* α. Of course, if the naming function is implicitly understood as determined by the context we will often suppress this parameter. In fact, an explicit version will only play a major role in Section 8.2. Now suppose we have fixed a naming function $\ulcorner - \urcorner$. Note that all the other canonically induced functions associated to the Cartesian product structure are uniform in the above sense. Given any two uniform functions $f\colon A \to B$ and $g\colon C \to D$, there is an induced function

$$f \times g = \langle f \circ \pi_1, g \circ \pi_2\rangle\colon A \times C \to B \times D,$$

which is uniform by definition. All identity functions on sets in D are also uniform. A simple induction on D shows this fact: The identity functions on 1, Fml_x and Term_x are uniform (the identity function on 1 is $!_1$). Let A, B in D be given such that id_A, id_B are uniform. The identity function on $A \times B$ can be expressed as

$$\mathrm{id}_{A\times B} = \mathrm{id}_A \times \mathrm{id}_B\,.$$

[10] When there is no possible confusion about the domain of the considered projection map, we will simply write π_1, π_2 without explicitly mentioning the domain.

Hence, $\mathrm{id}_{A\times B}$ is uniform. The *associators*, or the canonically induced isomorphisms from $(A \times B) \times C$ to $A \times (B \times C)$ are uniform:

$$a_{A,B,C} = \langle \pi_1^{A,B} \circ \pi_1^{A\times B,C}, \langle \pi_2^{A,B} \circ \pi_1^{A\times B,C}, \pi_2^{A\times B,C} \rangle \rangle.$$

So are the canonical isomorphisms between A, $A \times 1$ and $1 \times A$:

$$r_A = \pi_1 \colon A \times 1 \to A, \quad l_A = \pi_2 \colon 1 \times A \to A.$$

The inverses of these canonical isomorphisms are uniform as well, which we leave for the readers to check. Since all these canonical maps are uniform, we are free to use them in the remaining parts of this paper, and we will usually not mention them explicitly as we usually do when dealing with Cartesian products, except in Section 5 where more careful treatment is needed.

Importantly, all functions belonging to the C-basic class do not make any distinction on the initial input, and obtain results in an intuitively uniform way. The recursive definition of uniformity above then captures this intuitive sense of uniformity. This finally leads us to the definition of a uniform diagonal operator. Given a function u with codomain Fml_x or Term_x, we say u is *closed* if $\mathrm{im}\, u \subseteq \mathrm{Sent}$ or $\mathrm{im}\, u \subseteq \mathrm{ClTerm}$, where $\mathrm{im}\, u$ denotes the image of u and Sent and ClTerm denote the set of sentences and closed terms respectively. For example, both $\ulcorner - \urcorner_f$ and $\ulcorner - \urcorner_t$ are closed.

Definition 3.2 A *diagonal operator* d is a *closed* meta-linguistic function of type

$$d \colon \mathrm{Fml}_x \to \mathrm{Term}_x,$$

such that for every formula $\varphi \in \mathrm{Fml}_x$, the closed term $d\varphi$ satisfies the Kreisel–Henkin criterion with respect to φ, i.e.,

$$\Sigma \vdash d\varphi = \ulcorner \varphi(d\varphi) \urcorner.$$

A uniform diagonal operator is a diagonal operator which is uniform (for some naming function) in the sense of Definition 3.1.

Note that the definition of a diagonal operator depends on the chosen coding and the interpretation of the language. Since we only consider theories which prove all true identity statements of closed terms, instead of requiring that the the identity $d\varphi = \ulcorner \varphi(d\varphi) \urcorner$ is provable in Σ, we could equivalently require that $d\varphi = \ulcorner \varphi(d\varphi) \urcorner$ is true with respect to the given interpretation.

Remark 3.3 In this paper we focus on self-reference via a term, but the definition of uniformity we gave can also be applied to study "weak" diagonal sentences that do not satisfy the Kreisel–Henkin condition and thus to languages lacking the required (or indeed any) function symbols. In particular, we can define the uniformity of such a *weak* diagonal operator $d' : \mathrm{Fml}_x \to \mathrm{Fml}_x$ along similar lines. For instance, the diagonal operator which underlies the diagonal lemma introduced in [3, §35] will of course be uniform.

We now show that the canonical diagonalization methods found in the literature are uniform. More specifically, we show that Gödel's standard diagonal construction (which can be found e.g. in Smoryński [23]), Jeroslow's [15] diagonal operator and some further methods of diagonalization are all uniform. We thereby hope to convince the reader that our definition sufficiently captures the intuitive sense of a uniform diagonal construction.

3.1 Gödel's Construction

Let sub_G be the function symbol representing the primitive recursive function that takes (the code of) a formula $\varphi(x)$ and (the code of) an expression e, and outputs (the code of) the formula obtained by substituting $\ulcorner e \urcorner$ for x in $\varphi(x)$, i.e.,

$$\Sigma \vdash \mathsf{sub}_G(\ulcorner \varphi(x) \urcorner, \ulcorner e \urcorner) = \ulcorner \varphi(\ulcorner e \urcorner) \urcorner.$$

By definition we can prove

$$\begin{aligned}\Sigma \vdash\ & \mathsf{sub}_G(\ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner, \ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner) = \\ & \ulcorner \varphi(\mathsf{sub}_G(\ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner, \ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner)) \urcorner.\end{aligned}$$

This shows that $\mathsf{sub}_G(\ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner, \ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner)$ is a term satisfying the Kreisel–Henkin criterion with respect to $\varphi(x)$. The function d_G that when applied to $\varphi(x)$ outputs $\mathsf{sub}_G(\ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner, \ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner)$ is uniform. Here is an explicit construction: First, note that by using the meta-linguistic function $\overline{\mathsf{sub}_G}$, we can construct the term $\mathsf{sub}_G(x,x)$:

$$1 \xrightarrow{\langle \overline{x}, \overline{x} \rangle} \mathrm{Term}_x \times \mathrm{Term}_x \xrightarrow{\overline{\mathsf{sub}_G}} \mathrm{Term}_x\,.$$

Let $\overline{\mathsf{sub}_G(x,x)}$ now denote this composition, i.e., set

$$\overline{\mathsf{sub}_G(x,x)} := \overline{\mathsf{sub}_G} \circ \langle \overline{x}, \overline{x} \rangle \colon 1 \to \mathrm{Term}_x\,.$$

The following diagram provides a composite map that gives us d_G:

$$\begin{array}{l} \mathrm{Fml}_x \xrightarrow{r^{-1}_{\mathrm{Fml}_x}} \mathrm{Fml}_x \times 1 \xrightarrow{\mathrm{id} \times \overline{\mathsf{sub}_G(x,x)}} \mathrm{Fml}_x \times \mathrm{Term}_x \xrightarrow{\mathrm{Sub}_f} \mathrm{Fml}_x \xrightarrow{\ulcorner - \urcorner_f} \\ \quad \mathrm{Term}_x \xrightarrow{\langle \mathrm{id}, \mathrm{id} \rangle} \mathrm{Term}_x \times \mathrm{Term}_x \xrightarrow{\overline{\mathsf{sub}_G}} \mathrm{Term}_x \end{array}$$

If we unwrap the definition and follow the arrows of the diagram, we obtain for any input $\varphi(x) \in \mathrm{Fml}_x$ the following sequence of constructions

$$\begin{aligned}\varphi(x) &\mapsto (\varphi(x), *) \mapsto (\varphi(x), \mathsf{sub}_G(x,x)) \mapsto \varphi(\mathsf{sub}_G(x,x)) \\ &\mapsto \ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner \mapsto (\ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner, \ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner) \\ &\mapsto \mathsf{sub}_G(\ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner, \ulcorner \varphi(\mathsf{sub}_G(x,x)) \urcorner).\end{aligned}$$

The last step of this sequence delivers the desired fixed-point term $d_G(\varphi(x))$ for $\varphi(x)$. This shows that the usual Gödel construction is uniform.

3.2 Jeroslow's Diagonal Operator

We now reconstruct Jeroslow's [15] diagonalization method as a uniform diagonal operator. To begin with, we observe that there is a binary function symbol sub_J satisfying the following property

$$\Sigma \vdash \mathsf{sub}_J(\ulcorner\varphi(x)\urcorner, \ulcorner t(x)\urcorner) = \ulcorner\varphi(t(\ulcorner t(x)\urcorner))\urcorner,$$

for any formula $\varphi(x)$ and any term $t(x)$ both with free variable x. In other words, sub_J represents the primitive recursive function which maps (the code of) a formula $\varphi(x)$ and (the code of) a term $t(x)$ to (the code of) $\varphi(t(\ulcorner t(x)\urcorner))$. Hence, for any $\varphi(x)$ and $t(x)$ we have

$$\Sigma \vdash \mathsf{sub}_J(\ulcorner\varphi(x)\urcorner, \ulcorner t(x)\urcorner) = \ulcorner\varphi(t(\ulcorner t(x)\urcorner))\urcorner.$$

We conclude

$$\begin{aligned}\Sigma \vdash\ & \mathsf{sub}_J(\ulcorner\varphi(x)\urcorner, \ulcorner\mathsf{sub}_J(\ulcorner\varphi(x)\urcorner, x)\urcorner) = \\ & \ulcorner\varphi(\mathsf{sub}_J(\ulcorner\varphi(x)\urcorner, \ulcorner\mathsf{sub}_J(\ulcorner\varphi(x)\urcorner, x)\urcorner))\urcorner.\end{aligned}$$

This implies that the term $\mathsf{sub}_J(\ulcorner\varphi(x)\urcorner, \ulcorner\mathsf{sub}_J(\ulcorner\varphi(x)\urcorner, x)\urcorner)$ satisfies the Kreisel–Henkin criterion with respect to φ. Let d_J denote Jeroslow's diagonal operator which maps $\varphi(x)$ to this term. The following diagram shows that d_J is uniform:[11]

$$\begin{array}{l}
\mathrm{Fml}_x \xrightarrow{r^{-1}_{\mathrm{Fml}_x}} \mathrm{Fml}_x \times 1 \xrightarrow{\langle \mathrm{id}, \mathrm{id}\rangle \times \bar{x}} \mathrm{Fml}_x \times \mathrm{Fml}_x \times \mathrm{Term}_x \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \ulcorner - \urcorner_f \times \ulcorner - \urcorner_f \times \mathrm{id} \\
\quad \searrow \\
\quad \mathrm{Term}_x^3 \xrightarrow{\mathrm{id} \times \overline{\mathsf{sub}_{G,J}}} \mathrm{Term}_x \times \mathrm{Term}_x \xrightarrow{\mathrm{id} \times \ulcorner - \urcorner_t} \mathrm{Term}_x \times \mathrm{Term}_x \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \overline{\mathsf{sub}_J} \\
\quad \searrow \\
\quad \mathrm{Term}_x
\end{array}$$

Starting with $\varphi(x)$ and chasing the arrows in the diagram above, results in the following sequence of constructions:

$$\begin{aligned}\varphi(x) \mapsto\ & (\varphi(x), *) \mapsto (\varphi(x), \varphi(x), x) \mapsto (\ulcorner\varphi(x)\urcorner, \ulcorner\varphi(x)\urcorner, x) \\ \mapsto\ & (\ulcorner\varphi(x)\urcorner, \mathsf{sub}_J(\ulcorner\varphi(x)\urcorner, x)) \mapsto (\ulcorner\varphi(x)\urcorner, \ulcorner\mathsf{sub}_J(\ulcorner\varphi(x)\urcorner, x)\urcorner) \\ \mapsto\ & \mathsf{sub}_J(\ulcorner\varphi(x)\urcorner, \ulcorner\mathsf{sub}_J(\ulcorner\varphi(x)\urcorner, x)\urcorner).\end{aligned}$$

Once again, the last step of this sequence delivers the desired fixed-point term $d_J(\varphi(x))$ for $\varphi(x)$. Hence, the diagonal operator d_J is uniform.

Note that the construction of d_J is based on the basic functions $\ulcorner - \urcorner_f$, $\ulcorner - \urcorner_t$ and $\overline{\mathsf{sub}_J}$. In particular, we have not used Sub_f or Sub_t, while the construction of Gödel's diagonal operator requires the use of Sub_f. This is reflected in the resulting

[11] We have implicitly used associators in the diagramme to make the composite maps well-defined.

fixed-point terms: $d_J(\varphi)$ only contain names of $\varphi(x)$, but not of expressions of the form $\varphi(s)$ with $s \not\equiv x$, while $d_G(\varphi)$ contains a name of $\varphi(\mathsf{sub}(x, x))$, which requires the substitution of a term in $\varphi(x)$.

3.3 Other Uniform Diagonal Constructions

In addition to the usual canonical diagonal constructions, our framework is sufficiently robust to also accommodate several variants thereof, which intuitively qualify as uniform.

Example 3.4 We can slightly tweak Gödel's construction to obtain a uniform diagonal operator which contains a "dummy" conjunct. Let $\ulcorner - \urcorner$ be based on some standard numbering #. Let $\mathrm{Sub}^\wedge$ be a primitive recursive function that satisfies the following condition:

$$\mathrm{Sub}^\wedge(\#\varphi(x), \#\chi) = \begin{cases} \#\psi(\ulcorner \chi \urcorner) & \text{if } \varphi(x) \equiv \psi(x) \wedge x = x; \\ \#\varphi(\ulcorner \chi \urcorner) & \text{otherwise.} \end{cases}$$

Let $\mathsf{sub}^\wedge$ be a binary function symbol that represents $\mathrm{Sub}^\wedge$. Let d_A be a diagonal operator which maps a formula $\varphi(x)$ to the term

$$\mathsf{sub}^\wedge(\ulcorner \varphi(\mathsf{sub}^\wedge(x, x)) \wedge x = x \urcorner, \mathsf{sub}^\wedge(\ulcorner \varphi(\mathsf{sub}^\wedge(x, x)) \wedge x = x \urcorner)).$$

By definition of $\mathsf{sub}^\wedge$, we have

$$\Sigma \vdash \mathsf{sub}^\wedge(\ulcorner \varphi(\mathsf{sub}^\wedge(x, x)) \wedge x = x \urcorner, \ulcorner \varphi(\mathsf{sub}^\wedge(x, x)) \wedge x = x \urcorner) = \ulcorner \varphi(\mathsf{sub}^\wedge(\ulcorner \varphi(\mathsf{sub}^\wedge(x, x)) \wedge x = x \urcorner, \ulcorner \varphi(\mathsf{sub}^\wedge(x, x)) \wedge x = x \urcorner)) \urcorner.$$

Hence, d_A is a diagonal operator. Moreover, it is easy to see that d_A is uniform. d_A can be given by adding to the construction of d_G the operation $\equiv$ which yields the formula $x = x$, and $\overline{\wedge}$ which yields the conjunction of $\varphi(\mathsf{sub}^\wedge(x, x))$ and the formula $x = x$. Variations of d_A involving other connectives or expressions other than $x = x$ can be introduced along similar lines.

Example 3.5 Our notion of uniformity also subsumes the original definition of uniformity introduced in [11]. This definition relies on a *function symbol* such that

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

for every $\varphi \in \mathrm{Fml}_x$. Let now d_B be a diagonal operator which is uniform in the sense of [11], i.e., for every $\varphi \in \mathrm{Fml}_x$,

$$d_B\varphi \equiv \dot{d}(\ulcorner \varphi \urcorner).$$

If our language contains such a function symbol $\dot{d}$, then d_B is also uniform in the sense of Definition 3.2. We close by showing how such a function symbol can be specified. First, we fix a unary function symbol f (other than S) of our language. Let F denote the primitive recursive function given by

$$F(\#\varphi(x)) := \#\varphi(f(\ulcorner \varphi \urcorner)).$$

We now possibly change our base theory Σ to Σ_f such that F is represented by the function symbol f in our language – at least on all relevant formulae (see Section 4 for a more precise formulation). We then have for every $\varphi \in \mathrm{Fml}_x$,

$$\Sigma_f \vdash f(\ulcorner\varphi\urcorner) = \ulcorner\varphi(f(\ulcorner\varphi\urcorner))\urcorner.$$

This provides us with a uniform diagonal construction for theory Σ_f.

The initial motivation for uniformity, as introduced in [11, pp. 700], is to provide a condition on self-referential sentences with the Kreisel–Henkin property that is satisfied by fixed-point sentences obtained in a systematic, "canonical" way, but not by contrived fixed-points. We have shown above that the canonical fixed-point constructions are uniform. Now we turn to the question to what extent uniformity can rule out deviant fixed-point constructions, such as the refutable Henkin sentences constructed by Kreisel and variations and generalizations thereof as in Observation 2.2 and examples below. In particular, if $\varphi(t)$ satisfies the Kreisel–Henkin criterion and some natural assumptions are made, uniformity should rule out the possibility that the self-referential term t occurs already in the formula $\varphi(x)$, as it does in the refutable Henkin sentence above. Thus, the usefulness of uniformity depends on the answer to the following question:

Question 3.6 Let $\varphi(x)$ be a formula and d be a uniform diagonal operator. Under which assumptions can we rule out the possibility that the term $d(\varphi(x))$ occurs in $\varphi(x)$?

It will be shown that a natural assumption on the naming function is sufficient to eliminate this possibility. However, we first show that extra assumptions are required and that $d(\varphi(x))$ can occur in $\varphi(x)$, even if d is uniform, in a carefully chosen theory w.r.t. some specifically tailored numbering and numeral functions.

4 Uniform Kreisel-Like Constructions

To establish our claim that Question 3.6 cannot be trivially answered and some assumption is required, we construct a provability predicate $\mathsf{Bew}^\star(x)$ such that

$$\mathsf{Bew}^\star(x) \equiv x \neq d(\mathsf{Bew}^\star(x)) \wedge \mathsf{Bew}(x), \tag{1}$$

where d is a uniform diagonal operator and $\mathsf{Bew}(x)$ is a given provability predicate weakly representing provability. The application of d to $\mathsf{Bew}^\star(x)$ results in the term $d(\mathsf{Bew}^\star(x))$, which occurs in $\mathsf{Bew}^\star(x)$ itself.

Clearly, the resulting self-referential sentence $\mathsf{Bew}^\star(d(\mathsf{Bew}^\star(x)))$, i.e.,

$$d(\mathsf{Bew}^\star(x)) \neq d(\mathsf{Bew}^\star(x)) \wedge \mathsf{Bew}(d(\mathsf{Bew}^\star(x))),$$

is refutable. Hence, using the same reasoning towards Observation 2.1, $\mathsf{Bew}^\star(x)$ is a provability predicate and $\mathsf{Bew}^\star(d(\mathsf{Bew}^\star(x)))$ is a refutable Henkin sentence.

The construction of the provability predicate $\mathsf{Bew}^\star(x)$ relies on some self-referential trickery. This is because $\mathsf{Bew}^\star(x)$ contains the term $d(\mathsf{Bew}^\star(x))$ which depends on the definition of $\mathsf{Bew}^\star(x)$ itself. In order to make the definition of

$\mathsf{Bew}^\star(x)$ explicit, let d be a uniform diagonal operator which serves as a parameter. We define a meta-linguistic operator $k_d\colon \mathrm{Fml}_x \to \mathrm{Fml}_x$ which maps a given formula $\varphi(x)$ to the formula

$$x \neq d(\varphi) \wedge \mathsf{Bew}(x).$$

Any meta-linguistic fixed-point of k_d will serve as the desired provability predicate. This is because every fixed-point $\mathsf{Bew}^\star(x)$ of k_d remains unchanged with regard to application of k_d and thus satisfies Eq. 1:

$$\mathsf{Bew}^\star(x) \equiv k_d(\mathsf{Bew}^\star(x)) \equiv x \neq d(\mathsf{Bew}^\star(x)) \wedge \mathsf{Bew}(x).$$

As it turns out, whether or not fixed-points of k_d exist crucially depends on specific features of the naming function and the interpretation of our language.

In what follows we need to be more precise with the exact way our language $\mathcal{L}$ extends $\mathcal{L}_0$. Let $\mathcal{L}$ be the result of adding a $k{+}1$-ary function symbol f_n^k for each $n, k \in \omega$, to $\mathcal{L}_0$. For simplicity, we assume that $\mathcal{L}$ does not contain any further non-logical symbols. Let $\mathfrak{Pr}$ denote the set of primitive recursive functions. We call an interpretation $\mathcal{I}$ of $\mathcal{L}$ *standard*, if

(1) $\mathcal{I}$ interprets the symbols of $\mathcal{L}_0$ as usual;
(2) $\mathcal{I}(f_n^k)$ is a $k{+}1$-ary function in $\mathfrak{Pr}$, for every $k, n \in \omega$;
(3) each $k{+}1$-ary function in $\mathfrak{Pr}$ is represented by some f_n^k in $\mathcal{L}$ by $\mathcal{I}$.

In particular, if $\mathcal{I}$ is standard then it interprets the domain as ω. Thus, standard interpretations differ only with respect to the p.r. functions they assign to a given function symbol.

If $\mathcal{I}$ is standard, we use $\mathsf{Basic}(\mathcal{I})$ to denote the deductive closure of the theory R extended with all $\mathcal{I}$-true identities of the form $t = \overline{n}$, where t is a closed term and $n \in \omega$. In general, different standard interpretations $\mathcal{I}$ yield different theories $\mathsf{Basic}(\mathcal{I})$.

Recall that the definition of a diagonal operator depends on the numbering and the interpretation of the language. Moreover, we defined uniformity relative to a given naming function. The following definition makes this explicit:

Definition 4.1 Let α be a numbering, ν a numeral function and $\mathcal{I}$ be a standard interpretation.

(1) We say that d is a *diagonal operator with respect to α and $\mathcal{I}$* if d is a *closed* meta-linguistic function of type

$$d\colon \mathrm{Fml}_x \to \mathrm{Term}_x,$$

such that for each $\varphi \in \mathrm{Fml}_x$, the $\mathcal{I}$-value of closed term $d\varphi$ is the α-code of $\varphi(d\varphi)$.

(2) We say that d is *uniform diagonal operator with respect to α, ν and $\mathcal{I}$* if d is a diagonal operator with respect to α and $\mathcal{I}$ and d is uniform for the naming function $\nu \circ \alpha$.

According to the next lemma, a fixed-point term $d(\varphi(x))$ can occur in $\varphi(x)$ for some particular numberings and standard interpretations, even if d is uniform.

Lemma 4.2 *There is a numbering α, a numeral function ν, a standard interpretation $\mathcal{I}$, a uniform diagonal operator d with respect to α, ν and $\mathcal{I}$ and there are formulas* $\mathsf{Bew}(x)$, $\mathsf{Bew}^\star(x)$ *which weakly represent* $\mathsf{Basic}(\mathcal{I})$ *such that*

$$\mathsf{Bew}^\star(x) \equiv k_d(\mathsf{Bew}^\star(x)) \equiv x \neq d(\mathsf{Bew}^\star(x)) \wedge \mathsf{Bew}(x).$$

We sketch two straight-forward constructions of provability predicates $\mathsf{Bew}^\star(x)$ and $\mathsf{Bew}^\dagger(x)$ which both satisfy the conditions of Lemma 4.2. Recall that the uniform diagonal operator d_B introduced in Example 3.5 maps each formula $\varphi(x)$ to the fixed point term $\dot{q}(\ulcorner\varphi(x)\urcorner)$.

Our first construction is based on a peculiar choice of the numbering:

First Proof Sketch Let $\mathsf{Bew}^\star(x)$ be the formula

$$x \neq \dot{q}(0) \wedge \mathsf{Bew}(x),$$

where $\mathsf{Bew}(x)$ weakly represents $\mathsf{Basic}(\mathcal{I})$. Given a standard numbering # (such that each code is positive), let α be a new numbering which assigns 0 to $\mathsf{Bew}^\star(x)$ and $\#\varphi$ to every other expression φ. Hence, $0 \equiv \ulcorner\mathsf{Bew}^\star(x)\urcorner^\alpha$. Therefore, the fixed-point term of $\mathsf{Bew}^\star(x)$, based on the operator d_B and the numbering α, simply is $\dot{q}(0)$. That is, $\dot{q}(0) \equiv d_B(\mathsf{Bew}^\star(x))$. Hence,

$$\mathsf{Bew}^\star(x) \equiv x \neq d_B(\mathsf{Bew}^\star(x)) \wedge \mathsf{Bew}(x). \qquad \square$$

Instead of using a contrived numbering, our second construction is based on an peculiar choice of the numeral function:

Second Proof Sketch Let $\mathsf{Bew}^\dagger(x)$ be the formula

$$x \neq \dot{q}(c) \wedge \mathsf{Bew}(x),$$

where c is some fresh constant symbol and $\mathsf{Bew}(x)$ weakly represents $\mathsf{Basic}(\mathcal{I})$. We choose $\mathcal{I}$ such that c denotes (the code of) $\mathsf{Bew}^\dagger(x)$. Moreover, let ν be the numeral function which maps the code of $\mathsf{Bew}^\dagger(x)$ to c and each other number to its standard numeral. According to these choices, c is the ν-name of $\mathsf{Bew}^\dagger(x)$, that is, $c \equiv \ulcorner\mathsf{Bew}^\dagger(x)\urcorner^\nu$. Therefore, the fixed-point term of $\mathsf{Bew}^\dagger(x)$, based on the operator d_B and the numeral function ν, simply is $\dot{q}(c)$. That is, $\dot{q}(c) \equiv d_B(\mathsf{Bew}^\dagger(x))$. Hence,

$$\mathsf{Bew}^\dagger(x) \equiv x \neq d_B(\mathsf{Bew}^\dagger(x)) \wedge \mathsf{Bew}(x). \qquad \square$$

Remark 4.3 The vigilant reader will complain that our constructions contain subtle but persistent circles. In the second construction, we defined the provability predicate $\mathsf{Bew}(x)$, and therefore also $\mathsf{Bew}^\dagger(x)$, in dependency of the interpretation $\mathcal{I}$. But $\mathcal{I}$ in turn depends on the choice of $\mathsf{Bew}^\dagger(x)$. In other words, we assume without proof that such an interpretation and provability predicates exist. Similarly, in the first construction we defined the provability predicate $\mathsf{Bew}(x)$, and therefore also $\mathsf{Bew}^\star(x)$, in dependency of $\mathcal{I}$. But the interpretation $\mathcal{I}$ in turn depends on $\mathsf{Bew}^\star(x)$. This is because the function symbol $\dot{q}$ represents the function mapping the α-code

of $\mathsf{Bew}^\star(x)$ (i.e., the number 0) to the α-code of $\mathsf{Bew}^\star(\dot{d}(\mathsf{Bew}^\star(x)))$, which of course depends on $\mathsf{Bew}^\star(x)$.

As we show in the Appendix B, we can use the recursion theorem to provide the missing details, thereby turning our proof sketches into rigorous arguments. See B.1 and B.2 for explicit and detailed constructions of $\mathsf{Bew}^\star(x)$ and $\mathsf{Bew}^\dagger(x)$ respectively. Inspection of these constructions reveals that they rely on circular features of some of the involved formalisation choices. In particular, both constructions yield provability predicates which contain their own names: The predicate $\mathsf{Bew}^\star(x)$ contains its own α-name $\mathsf{0}$, while $\mathsf{Bew}^\dagger(x)$ contains its own ν-name c.

5 Circularity of Naming

In the previous section we have given examples of formulas $\varphi(x)$ that already contain $d(\varphi(x))$, even if d is a uniform diagonal operator. However, our examples rely either on contrived numeral functions or codings. In Question 3.6 we asked which additional assumptions can be made to rule out the possibility that the diagonal term $d(\varphi(x))$ occurs already in the formula $\varphi(x)$ which is diagonalized. Remark 4.3 hints at a possible answer: If we rule out the deviant numeral functions and codings, or more specifically, if the naming function does not exhibit any *circular features*, we may hope it will provide us with the additional natural assumptions we are seeking. To make this precise, we define a binary relation on the set of expressions. This relation will play an essential role in the formulation of an answer to Question 3.6.

Definition 5.1 Let $\preceq$ denote the subexpression relation and let $\ulcorner - \urcorner$ be a naming function. Let $\lhd$ be the binary relation on $\mathcal{L}$-expressions given by $e \lhd e'$ iff there exists an expression e'' such that

$$e \preceq e'' \;\&\; \ulcorner e'' \urcorner \preceq e'.$$

We call $\lhd$ the *weak naming relation for* $\ulcorner - \urcorner$ and say that an expression e is *weakly named* in e' if $e \lhd e'$.[12] In order to make the dependency of $\lhd$ on the underlying naming function explicit, we sometimes write $\lhd^{\ulcorner - \urcorner}$ or $\lhd^{\alpha,\nu}$ for the weak naming relation for $\ulcorner - \urcorner = \nu \circ \alpha$. If ν is the standard numeral function we also simply write $\lhd^\alpha$ instead of $\lhd^{\alpha,\nu}$. Finally, let $\lhd_*$ denote the transitive closure of $\lhd$.

The following useful facts follow immediately from Definition 5.1

Fact 5.2 The weak naming relation is both left-downward and right-upward closed with respect to the subexpression relation. That is, if $e \lhd e'$, then $e' \preceq e''$ implies $e \lhd e''$, and $e'' \preceq e$ implies $e'' \lhd e'$.

Fact 5.3 If $e \lhd_* e'$, then there is a subexpression $t \preceq e'$ such that t is a closed term and $e \lhd_* t$.

[12] The symbol $\lhd$ denotes a slightly different relation in [9]. There, $\lhd$ is defined by setting $e \lhd e'$ iff $\ulcorner e \urcorner \preceq e'$. In our terminology, this may be called a *strong naming relation*. Caveat lector!

The $\lhd$ relation allows us to formalise what we mean by circularity of naming functions. What we will show is the following: If $\lhd$ does not exhibits any loops, viz. if its transitive closure $\lhd_*$ is *irreflexive*, then it suffices to yield a possitive answer to Question 3.6.

To prove this, our strategy is to first inspect the meta-linguistic properties of a fixed-point term obtained from uniform constructions, and our strategy is to first provide a more systematic study of the structure of uniform functions. It is evident from Definition 3.1 that every uniform operation can be constructed by successively composing C-basic functions together with canonical maps of the Cartesian product structure. To make this intuition precise, we introduce a representation system for constructions of uniform functions. Such development will allow us to prove, by induction and case distinction, Lemma 5.11, according to which all diagonal terms obtained uniformly share a particular meta-linguistic feature. This lemma directly implies our main result, viz. Proposition 5.12.

We start by using a term algebra to represent uniform functions. Let $(B_n)_{n\in\omega}$ be a fixed bijective (and effective, if you prefer) enumeration of all C-basic functions, and let UniFct be the set of all uniform functions.

Definition 5.4 Let Ω be the signature that contains constant symbols b_n for each $n \in \omega$, and two binary function symbols $\odot$ and $\fatsemi$. Let T_Ω denote the term algebra generated over Ω (with no variables). We recursively define a subset $\mathcal{R} \subset T_\Omega$ and an evaluation function $\mathsf{ev}\colon \mathcal{R} \to \mathsf{UniFct}$:

- $b_n \in \mathcal{R}$ and $\mathsf{ev}(b_n) = B_n$, for any $n \in \omega$;
- If $p, q \in \mathcal{R}$ and $\mathrm{dom}(\mathsf{ev}(p)) = \mathrm{dom}(\mathsf{ev}(q))$, then $\fatsemi(p, q) \in \mathcal{R}$ and

$$\mathsf{ev}(\fatsemi(p, q)) = \langle \mathsf{ev}(p), \mathsf{ev}(q)\rangle;$$

- If $\mathrm{dom}(\mathsf{ev}(p)) = \mathrm{cod}(\mathsf{ev}(q))$, then $\odot(p, q) \in \mathcal{R}$ and

$$\mathsf{ev}(\odot(p, q)) = \mathsf{ev}(p) \circ \mathsf{ev}(q);$$

where $\mathrm{dom}(f)$ and $\mathrm{cod}(f)$ denotes the domain and codomain of a function, respectively. We call a term $r \in \mathcal{R}$ a *representation* of a uniform function u if $\mathsf{ev}(r) = u$.

Obviously, terms in $\mathcal{R}$ are well-typed, and in what follows we simply use $\mathrm{dom}(r)$, $\mathrm{cod}(r)$ to denote $\mathrm{dom}(\mathsf{ev}(r))$, $\mathrm{cod}(\mathsf{ev}(r))$, respectively. We also call a term $r \in \mathcal{R}$ closed (resp. basic) if $\mathsf{ev}(r)$ is closed (resp. basic).

Since the codomain of all our basic operations is either Fml_x or Term_x, the following fact is immediate:

Fact 5.5 If $r \in \mathcal{R}$ and $\mathrm{cod}(r) = A \times B$, then r must be of the form $\fatsemi(p, q)$.

From the definition of uniformity and representation of uniform functions, it is easy to see that $\mathsf{ev}\colon \mathcal{R} \to \mathsf{UniFct}$ is surjective, which means every uniform function

has some representation. But ev is not injective. Suppose that $\mathsf{ev}(b_m) = \mathrm{id}_{\mathrm{Fml}_x}$ and $\mathsf{ev}(b_n) = \pi_1^{\mathrm{Fml}_x,\mathrm{Fml}_x}$. Then

$$\mathsf{ev}(b_m) = \mathrm{id}_{\mathrm{Fml}_x} = \mathsf{ev}(\odot(b_n, \fatsemi(b_m, b_m))),$$

which shows that both b_m and $\odot(b_n, \fatsemi(b_m, b_m))$ are representations of $\mathrm{id}_{\mathrm{Fml}_x}$. This example shows that some representations contain redundant information which is irrelevant to the actual uniform function it represents. To reduce such redundancies, we define a reduction process for terms in $\mathcal{R}$:

Definition 5.6 The reduction relation is the smallest binary relation $\longrightarrow \subseteq \mathcal{R} \times \mathcal{R}$ satisfying the following clauses: for every $p, q, r \in \mathcal{R}$,

(1) if $\mathsf{ev}(r) = \mathrm{id}_A$ and $\odot(r, p), \odot(q, r) \in \mathcal{R}$ then

$$\odot(r, p) \longrightarrow p, \quad \odot(q, r) \longrightarrow q;$$

(2) if $\mathsf{ev}(b_m) = !_A$ and $\odot(b_m, p) \in \mathcal{R}$, then

$$\odot(b_m, p) \longrightarrow b_n,$$

where b_n is the unique constant that $\mathsf{ev}(b_n) = !_A \circ \mathsf{ev}(p) = !_{\mathrm{dom}(p)}$;

(3) if $\mathsf{ev}(b_m)$, $\mathsf{ev}(b_n)$ are projections maps from $A \times B$ to A, B respectively and if $\odot(b_m, \fatsemi(p, q)), \odot(b_n, \fatsemi(p, q)), \fatsemi(\odot(b_m, p), \odot(b_n, p)) \in \mathcal{R}$, then

$$\odot(b_m, \fatsemi(p, q)) \longrightarrow p, \quad \odot(b_n, \fatsemi(p, q)) \longrightarrow q,$$
$$\fatsemi(\odot(b_m, p), \odot(b_n, p)) \longrightarrow p;$$

(4) if $r \in \mathcal{R}$ is closed, b_i represents either Sub_f or Sub_t, $\odot(b_i, \fatsemi(r, p)) \in \mathcal{R}$, then

$$\odot(b_i, \fatsemi(r, p)) \longrightarrow r;$$

(5) if $\odot(\odot(p, q), r) \in \mathcal{R}$, then

$$\odot(\odot(p, q), r) \longrightarrow \odot(p, \odot(q, r));$$

(6) if r contains a subterm p and $p \longrightarrow q$, then $r \longrightarrow r'$ where r' is the result of substituting this particular occurrence of p with q in r.

It is easy to verify that we have a well-defined notion of reduction among terms in $\mathcal{R}$, i.e. if $p \in \mathcal{R}$ and $p \longrightarrow q$, then q must also be a term in $\mathcal{R}$. Let $\longrightarrow_*$ denote the transitive closure of $\longrightarrow$. Clearly, the function that a term $p \in \mathcal{R}$ represents remains invariant under the reduction process. Thus, if $p \longrightarrow_* q$ then $\mathsf{ev}(p) = \mathsf{ev}(q)$.

We say a representation p is *reduced* if there is no other $q \in \mathcal{R}$ such that $p \longrightarrow q$. By clause (6), if $p \in \mathcal{R}$ is reduced then so is every subterm of p. The following holds for the reduction process we have described above:

Fact 5.7 Given any term $r \in \mathcal{R}$, successive application of reduction to r will always yield a reduced representation after finitely many steps.

Proof All the clauses of the reduction relation do not increase the length of terms, where only (5) does not strictly decrease the length. Hence, we only need to verify that (5) does not generate an infinite chain of reductions, which is clearly the case. □

Hence, without any loss of generality, we can work only with *reduced* representations of a uniform function. Note that even though the reduction process is terminating, it does not necessarily enjoy unique normalisation. That is, a representation may give rise to different reduced representations, and the same uniform function may have different reduced representations.[13]

With reduced representations, we may commence studying the behavior of uniform functions. For our purpose, we are mainly concerned with those whose codomain is Fml_x or Term_x. The following is a simple observation which will be used later:

Fact 5.8 Given a reduced $r \in \mathcal{R}$ such that $\mathrm{cod}(r)$ is either Fml_x or Term_x, then it is either a constant b_n for some $n \in \omega$, or of the form $\odot(b_m, q)$ for some $m \in \omega$ such that b_m is basic.

Proof Suppose r is a composite term. Since r is reduced and $\mathrm{cod}(r)$ is not a binary product, we have $r = \odot(b_m, q)$ for some $m \in \omega$. Clearly, $\mathsf{ev}(b_m)$ cannot be $\mathrm{id}_{\mathrm{Fml}_x}$, $\mathrm{id}_{\mathrm{Term}_x}$ or $!_A$ for any $A \in \mathsf{D}$. Moreover, $\mathsf{ev}(b_m)$ cannot be a projection; otherwise, $\mathrm{cod}(q)$ would be a binary product and by Fact 5.5 it would be of the form $\wp(q_1, q_2)$. This would imply that $\odot(b_m, q)$ is not reduced. Hence, r has the form $\odot(b_m, q)$ with b_m basic. □

This fact holds essentially because the codomains of all our basic functions are *not* multiple products of Fml_x or Term_x, hence if the codomain of a reduced representation is a single multiple of Fml_x or Term_x, the final composite cannot be a projection.

Uniform diagonalization operators are of type $\mathrm{Fml}_x \to \mathrm{Term}_x$. The following lemma shows that any such operator must involve an essential use of the function $\ulcorner - \urcorner_f$. This is simply because the uniformity constraint does not allow any other way to obtain a function of such type.

Lemma 5.9 *Let* $u\colon \mathrm{Fml}_x \to \mathrm{Term}_x$ *be a uniform function, and suppose* b_n *represents* $\ulcorner - \urcorner_f$. *If* u *is not a constant function then* b_n *appears in every reduced representation of* u.

Proof Suppose $r \in \mathcal{R}$ is a reduced term representing u. We show by induction on the complexity of terms that if r does not contain b_n then $\mathsf{ev}(r)$ is a constant function.

[13] It is possible to extend the reduction rules such that we obtain unique normalisation. The current rules suffice for our purpose in this paper.

Suppose r is b_m for some $m \in \omega$. There is no C-basic function other than $\ulcorner - \urcorner_f$ that is of type $\mathrm{Fml}_x \to \mathrm{Term}_x$, hence the base case is closed. For the inductive step suppose r is a composite term. By Fact 5.8 r is of the form $\odot(b_m, q)$ with b_m basic. From the codomain of $\mathsf{ev}(b_m)$ and the fact that $m \neq n$ we conclude that b_m must represent Sub_t, $\ulcorner - \urcorner_t$, $\overline{f}$ or $\overline{x}$. Also, q is reduced and does not contain b_n. The domains of these basic functions are all of the form Term_x^k, for $k \in \omega$. If $k = 0$, then $\mathsf{ev}(r)$ is obviously constant. If $k = 1$, then $\mathsf{ev}(q)\colon \mathrm{Fml}_x \to \mathrm{Term}_x$. By induction hypothesis $\mathsf{ev}(q)$ is constant, and so is $\mathsf{ev}(r)$. Finally, if $k \geq 2$, then according to our choice of the product structure and by Fact 5.5, q must be of the form $\fatsemi(q_1, \fatsemi(q_2, \cdots \fatsemi(q_{k-1}, q_k) \cdots))$, where every q_i is reduced, does not contain b_n and $\mathsf{ev}(q_i)\colon \mathrm{Fml}_x \to \mathrm{Term}_x$. By induction hypothesis again, every $\mathsf{ev}(q_i)$ is constant, and so is $\mathsf{ev}(r)$. □

Note that diagonal operators are not constant functions. Hence, according to this lemma, all the uniform diagonal constructions provided in Section 3 include the function $\ulcorner - \urcorner_f$. Note that the constructions of Gödel's diagonal operator and the two examples presented in Example 3.4 and Example 3.5 do not employ the function $\ulcorner - \urcorner_t$; while Jeroslow's operator does. Hence, Lemma 5.9 in particular shows that, at least in the context of uniform diagonalization, the function $\ulcorner - \urcorner_f$ is more fundamental than the function $\ulcorner - \urcorner_t$. This is one of the reasons why we explicitly distinguish these two naming functions.

Uniform functions also preserve the occurrence of free variables. The following lemma shows that, if a uniform function u takes *some* open formula $\varphi(x)$ to a closed (resp. open) expression, then the value of *every* open formula is a closed (resp. open) expression:

Lemma 5.10 *For each uniform function u of type $\mathrm{Fml}_x \to \mathrm{Fml}_x$ or $\mathrm{Fml}_x \to \mathrm{Term}_x$, if u is not closed then for every formula $\varphi(x)$ with x freely occurring in φ, also $u(\varphi)$ contains x as a free variable.*

Proof Let $r \in \mathcal{R}$ be a reduced representation of u. We prove the claim by induction on the complexity of r. For the base case suppose r is some constant b_m. If u has type $\mathrm{Fml}_x \to \mathrm{Fml}_x$ then $\mathsf{ev}(b_m) = \mathrm{id}_{\mathrm{Fml}_x}$, which satisfies the condition. If u has type $\mathrm{Fml}_x \to \mathrm{Term}_x$, the only C-basic function with the right type is $\ulcorner - \urcorner_f$, which is closed. For the inductive step we can assume by Fact 5.8 that r is of the form $\odot(b_m, q)$ with b_m being basic.

If $u\colon \mathrm{Fml}_x \to \mathrm{Fml}_x$ is not closed, then $\mathsf{ev}(b_m)$ is Sub_f, $\overline{\star}$, where $\star$ is not $\forall x$, or $\overline{R}$ with $\mathsf{ar}(R) \geq 1$. We check the claim for each of these cases:

(1) If $\mathsf{ev}(b_m) = \mathrm{Sub}_f$, then by Fact 5.5 q must be of the form $\fatsemi(q_1, q_2)$, with $\mathsf{ev}(q_1)\colon \mathrm{Fml}_x \to \mathrm{Fml}_x$, $\mathsf{ev}(q_2)\colon \mathrm{Fml}_x \to \mathrm{Term}_x$. Since $\mathsf{ev}(r)$ is not closed, both $\mathsf{ev}(q_1)$ and $\mathsf{ev}(q_2)$ are not closed. By the induction hypothesis, $\mathsf{ev}(q_1)(\varphi)$ and $\mathsf{ev}(q_2)(\varphi)$ are a formula $\varphi_1(x)$ and a term $t_2(x)$ respectively, both with a free variable x. Hence, $\varphi_1(t_2(x))$ contains x as a free variable.

(2) If $\mathsf{ev}(b_m)$ is $\overline{\star}$ for $\star$ either being $\neg$ or $\wedge$, then either $\mathsf{ev}(q)\colon \mathrm{Fml}_x \to \mathrm{Fml}_x$ or $q = \fatsemi(q_1, q_2)$, where $\mathsf{ev}(q_1), \mathsf{ev}(q_2)\colon \mathrm{Fml}_x \to \mathrm{Fml}_x$. For the former case, $\mathsf{ev}(q)$ cannot be closed. For the latter case, at least one of $\mathsf{ev}(q_i)$ is not closed.

By induction hypothesis, $\mathsf{ev}(q)(\varphi)$ or $\mathsf{ev}(q_i)(\varphi)$ contains x as a free variable, and hence so does $\mathsf{ev}(r)(\varphi)$.

(3) The case for $\mathsf{ev}(b_m) = \overline{R}$ with $\mathsf{ar}(R) \geq 1$ proceeds similarly to (2), with the minor difference that here q is of the form $\mathfrak{s}(q_1, \cdots \mathfrak{s}(q_{k-1}, q_k) \cdots)$, if $\mathsf{ar}(R) = k$.

The case where $u\colon \mathrm{Fml}_x \to \mathrm{Term}_x$ is completely similar. Here, $\mathsf{ev}(b_m)$ is Sub_t, $\overline{f}$ with $\mathsf{ar}(f) \geq 1$, or $\overline{x}$, and a proof by case distinction proceeds in almost the same manner as shown above. □

The above proof is a bit tedious, since it relies on an induction with several case distinctions. But the statement of Lemma 5.10 should be expected, since each C-basic function either preserves the existence of free variables, or it maps anything to closed expressions; composition does not change this fact.

With all these preliminary work, we can finally show the following result about uniform diagonalization: If $d(\varphi(x))$ is the result of uniform diagonalization, then it weakly names an expression, which weakly names an expression ..., which weakly names an expression of the form $\varphi(s)$:

Lemma 5.11 *Let d be a uniform diagonal operator. Then for every formula $\varphi(x) \in \mathrm{Fml}_x$ that contains a free variable x, there is a term s such that*

$$\varphi(s) \lhd_* d(\varphi(x)).$$

The idea behind Lemma 5.11 is very simple. Intuitively, we may view a reduced representation r of a uniform diagonal operator d as an instruction for carrying out the diagonalization process for each formula $\varphi(x)$. We have shown in Lemma 5.9 that d must make an essential use of the function $\ulcorner - \urcorner_f$. Before that use, what we may essentially do is to construct terms and substitute them into $\varphi(x)$, which results in a formula of the form $\varphi(s)$ for some term s. Moreover, we may combine $\varphi(s)$ with other formulas using connectives to form a longer expression, which we temporarily denote as ψ. Note that $\varphi(s) \preceq \psi$, hence, after applying $\ulcorner - \urcorner_f$, we have $\varphi(s) \lhd_* \ulcorner \psi \urcorner$. Note that $\ulcorner \psi \urcorner$ is a *closed* term, which means that we can only substitute it into other expressions, but not the other way around. This implies that further applications of other basic functions to $\ulcorner \psi \urcorner$ would retain the $\lhd_*$-relation with $\varphi(s)$, which is exactly what we want.

Of course, to make this rough proof sketch precise, it requires a rigorous argument. Since the detailed proof is quite tedious and technical, we omit is here. Its structure resembles the proof of Lemma 5.10, where we also need an induction together with several case distinctions. The enthusiastic reader can find the proof in full detail in Appendix A.

We can now formulate an answer to our Question 3.6: To make it impossible for $d(\varphi(x))$ to occur in $\varphi(x)$ for a uniform diagonal operator d, it is sufficient to rule out loops in the weak naming relation or, equivalently to demand that $\lhd_*$ is irreflexive. We maintain that this assumption on $\lhd_*$ is natural. The usual Gödel codings and numeral functions make $\lhd_*$ irreflexive.

Proposition 5.12 *Let $\lhd_*$ be irreflexive. Then for every uniform diagonal operator d and formula $\varphi(x)$, the fixed-point term $d(\varphi(x))$ cannot occur in $\varphi(x)$.*

Proof Assume that there is a uniform diagonal operator d such that $d(\varphi(x))$ occurs in $\varphi(x)$. By Lemma 5.11, there exists a term s such that $\varphi(s) \lhd_* d(\varphi(x))$. Since $d(\varphi(x))$ is a subterm of $\varphi(s)$, we obtain $d(\varphi(x)) \lhd_* d(\varphi(x))$ by Fact 5.2. Hence, $\lhd_*$ is not irreflexive. □

Recall that the deviant Henkin sentences introduced in Section 4 are based on provability predicates $\mathsf{Bew}^\star(x)$ which satisfy condition Eq. 1, i.e.,

$$\mathsf{Bew}^\star(x) \equiv x \neq d(\mathsf{Bew}^\star(x)) \wedge \mathsf{Bew}(x).$$

It follows immediately from Proposition 5.12 that every predicate $\mathsf{Bew}^\star(x)$ satisfying this condition involves circular weak naming relations.

Remark 5.13 While the construction of $\mathsf{Bew}^\star(x)$ in B.1 employs a canonical numeral function, namely, standard numerals, the circularity of the weak naming relation results from a contrived choice of the numbering. To further analyse this situation, we say that a numbering α is *monotonic* if for any expressions e, e', $e \preceq e'$ implies $\alpha(e) \leq \alpha(e')$. Clearly, the numbering function α used to construct $\mathsf{Bew}^\star(x)$ in B.1 is not monotonic. Can we do better and base this construction on a monotonic numbering instead of α? We note that this is not possible. In order to see this, we call a numbering α *strongly monotonic for ν-numerals*, if

$$\alpha(e) < \alpha(\ulcorner e \urcorner^{\alpha,\nu}), \text{ for all expressions } e.$$

We observe that every monotonic numbering is strongly monotonic for standard numerals (see also [8, Section 6]). Moreover, if α is strongly monotonic for ν-numerals, then the weak naming relation $\lhd^{\alpha,\nu}$ induced by α and ν is *well-founded*. In particular, the relation $\lhd_*^{\alpha,\nu}$ cannot be circular. Hence, by Proposition 5.12 we cannot construct a uniform diagonal operator d and a provability predicate $\mathsf{Bew}^\star(x)$ satisfying Eq. 1, whenever we use a numbering α and a numeral function ν such that α is strongly monotonic for ν-numerals.

As we have seen in B.2, the circularity of the weak naming relation can also result from a non-standard numeral function, even if we fix a standard monotonic numbering. Hence, an answer to Question 3.6 involves a constraint on the numbering *and* the numeral function. Of course, this is precisely what we do in Proposition 5.12 when we require the irreflexivity of $\lhd_*$.

We have said that the constraint of non-circularity is natural. The next section will provide more detailed conceptual and philosophical grounds on which this requirement can be based.

6 Quotation and the Well-Foundedness of Naming

A naming function maps every expression e to a closed term $\ulcorner e \urcorner$, which serves as its name. Since we work in an arithmetical framework, a naming function consists

of a Gödel numbering and a numeral function. Except for requiring effectiveness, thus far we have not placed any constraints on numberings and numeral functions. Gödel numerals are often conceived of as arithmetical counterparts of quotational names. However, there are codings and numeral functions that make it very implausible to think of these numerals as quotations. In this section, we introduce precise constraints that single out certain coding schemata and numeral functions as adequate counterparts of quotation devices.

In order to do so, we conceive of arithmetical naming functions as particular instances of string-theoretical naming devices. Let $\mathcal{A}$ be an alphabet and let $\mathcal{A}^*$ denote the set of finite strings over $\mathcal{A}$ including the empty string ϵ. For strings e, f in $\mathcal{A}^*$, let ef denote the result of concatenating e with f. Let $\mathcal{E} \subseteq \mathcal{A}^*$ be a set of expressions. We call any injective function $N\colon \mathcal{E} \to \mathcal{E}$ a *string-theoretical naming function for* $\mathcal{E}$. For example, let $\mathcal{A}$ consist of English letters together with a pair of single quotation marks. The function Q which maps each string $s \in \mathcal{A}^*$ to its proper quotation 's' is a canonical example of a string-theoretical naming function for A^*.[14] Let $\ulcorner - \urcorner\colon \mathrm{Term}_x \cup \mathrm{Fml}_x \to \mathrm{ClTerm}$ be a naming function as introduced in Section 3, i.e., $\ulcorner - \urcorner$ is the composition of a numbering and a numeral function. Then $\ulcorner - \urcorner$ can be also conceived of as a string-theoretical naming function. For example, Q and $\ulcorner - \urcorner$ name the letter "x" by the strings " 'x' " and "$\ulcorner x \urcorner$" respectively.

In the philosophical literature, $\ulcorner - \urcorner$ is often viewed as an arithmetical proxy of the quotation function Q. Heck [12, pp. 27], for example, takes the "disquotation" schema $T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$ to be an arithmetical formalisation of the informal schema 'S' is true iff S (see also [24, pp. 156]). On this view, it is plausible to require that $\ulcorner - \urcorner$ satisfies certain quotation-like features. In particular, we require that $\lhd^{\ulcorner - \urcorner}$ behaves similarly to the weak naming relation induced by quotation. In order to make this precise, we first generalize Definition 5.1 to string-theoretical naming functions.

Definition 6.1 Let N be a string-theoretical naming function for $\mathcal{E}$. We say that an expression $e \in \mathcal{E}$ is *weakly named in* e' *by* N, in symbols: $e \lhd^N e'$, if there exists another expression $e'' \in \mathcal{E}$ such that $e \preceq e''$ and $N(e'') \preceq e'$. We also call $\lhd^N$ the *weak naming relation for* N. Let $\lhd^N_*$ denote the transitive closure of $\lhd^N$.

Clearly, Fact 5.2 also holds in the more general setting. The following useful observation follows from the above definition and Fact 5.2.

Fact 6.2 Let N be a string-theoretical naming function for $\mathcal{E}$. If $\lhd^N$ is ill-founded, then there is a sequence $(f_i)_{i\in\omega}$ of elements in $N(\mathcal{E})$ such that $f_{i+1} \lhd^N f_i$, for each $i \in \omega$.

[14] For the sake of better readability, we usually omit (meta-linguistic) quotation symbols when there is no confusion. That is, we write " 's' " instead of " 's' " (or, more precisely, " ' "s" ' "). In order to avoid confusion, we use the convention that single quotation marks " ' " and " ' " are part of the object language and double quotation marks " " " and " " " are part of the metalanguage.

Since each quotation properly contains its named expression, no quotation q can denote an expression containing q itself. More generally, the weak naming relation $\lhd^{Q}$ is well-founded. From this observation, we can extract the following necessary condition for a naming function to mimic or resemble quotation:

Well-Foundedness: Every naming function which resembles quotation induces a well-founded weak naming relation.

Thus, we can justify the assumption of $\lhd_*$'s irreflexivity in our answer to Question 3.6 with Proposition 5.12 by drawing on the conception of $\ulcorner - \urcorner$ as resembling quotation.

While proper quotation is perhaps the most common naming function, the specific method of enclosing expressions by quotation marks is by no means theoretically essential.[15] Alternatively, we may name strings by describing their constituent symbols, e.g. using Tarski's [24] structural-descriptive names, or by a Kripkean act of baptism [17, pp. 693]. The reader may wonder to what extend the well-foundedness requirement depends on the specifics of the quotation function. In other words, can we maintain the requirement of $\lhd$'s well-foundedness if we conceive of $\ulcorner - \urcorner$ as resembling other naming functions different to proper quotation? In the remainder of this section we show that the well-foundedness criterion can be based on a broad conception of quotation which encompasses several canonical naming devices found in the literature.

To delineate this broad conception, consider the expressions " 'snow' " , "the word which consists of the following letters: es, en, o, double-u, following one another", "the 4354th word of *Chants Democratic*" and "Jack". There is an important difference in the way these expressions serve as names of strings. The first two preserve the literal information of the named string "snow", i.e., for each letter of "snow", they contain a designated corresponding string. For example, "s" and "es" correspond to the first letter of "snow" respectively. This preservation of literal information enables us to read off the designated words from their names. The last two expressions do not preserve literal information. As opposed to the situation above, their referents can only be determined by reference to an external source of information or act of baptism.

In what follows, we confine ourselves to naming devices which preserve literal information. The following definitions are an attempt to make this precise.

Definition 6.3 Let $e, f, g \in \mathcal{A}^*$. We write $(e, f) \sqsubset g$ if g contains non-overlapping occurrences of e and f. More precisely, $(e, f) \sqsubset g$ iff there are (possibly empty) $a, b, c \in \mathcal{A}^*$ such that $g = aebfc$ or $g = afbec$. We call a function $G\colon S \times S \to S$ *weakly $\sqsubset$-increasing*, if $(e, f) \sqsubset G(e, f)$, for all $e, f \in S$.

Definition 6.4 We call a function $N'\colon \mathcal{A}^* \to \mathcal{A}^*$ a *literal pre-naming function for* $\mathcal{A}^*$, if N' can be recursively defined by

- $N'(\epsilon) = L(\epsilon, \epsilon)$;

[15] This has been already observed by Tarski [24, pp. 156] and Quine [21, pp. 26].

- $N'(s) = G(N'(a_{\pi_s(1)} \cdots a_{\pi_s(n-1)}), L(a_{\pi_s(n)}, s))$, where $s \equiv a_1 \cdots a_n$ such that $n > 0$ and $a_i \in \mathcal{A}$ for each $i \leq n$;

for some function $L\colon (\mathcal{A} \cup \{\epsilon\}) \times \mathcal{A}^* \to \mathcal{A}^* \setminus \{\epsilon\}$, a weakly $\sqsubset$-increasing function $G\colon \mathcal{A}^* \times \mathcal{A}^* \to \mathcal{A}^*$ and a function π_- which maps each string $s \in \mathcal{A}^*$ to a permutation π_s of the set $\{1, \ldots, \mathrm{lh}(s)\}$. We also call L a *literal function*. The second argument of L serves as a parameter, permitting alphabetical symbols to be named by L in dependency of the full string in which they occur. If L is defined without parameters, we sometimes suppress the second argument of L for better readability.

Let $\mathcal{E} \subseteq \mathcal{A}^*$ be a set of expressions. We call a function $N\colon \mathcal{E} \to \mathcal{E}$ a *literal naming function for* $\mathcal{E}$, if there is a literal pre-naming function N' for $\mathcal{A}^*$ and functions $B, E\colon \mathcal{E} \to \mathcal{A}^*$ such that

$$N(e) = B(e)N'(e)E(e), \text{ for each } e \in \mathcal{E}.$$

We also call $B(s)$ and $E(s)$ the *begin marker* and the *end marker* of the name $N(e)$ of e respectively.

This definition accommodates a large class of naming devices found in the literature:

A. Let $\mathcal{A}$ consist of English lower case letters together with a pair of single quotation marks. Let $\mathcal{E}$ be given by

$$\alpha ::= \mathsf{a} \mid \mathsf{b} \mid \mathsf{c} \mid \cdots \mid \mathsf{z} \mid \text{`}\alpha\text{'} \mid \alpha\alpha$$

The functions $\mathrm{Q}_1, \mathrm{Q}_2, \mathrm{Q}_3\colon \mathcal{E} \to \mathcal{E}$, given by

$$\begin{aligned} \mathrm{Q}_1(a_1a_2 \cdots a_n) &:\equiv \text{`}a_1a_2 \cdots a_n\text{'} \\ \mathrm{Q}_2(a_1a_2 \cdots a_n) &:\equiv \text{`}a_n \cdots a_2a_1\text{'} \\ \mathrm{Q}_3(a_1a_2 \cdots a_n) &:\equiv \text{`}a_1a_1a_2a_2 \cdots a_na_n\text{'} \end{aligned}$$

are literal naming functions. The literal function of Q_3 duplicates the symbol a to a string aa. For Q_2, the family of permutations is non-trivial: for each s with length $k \geq 2$, $\pi_s(j) = k + 1 - j$, for any $1 \leq j \leq k$. We can also duplicate or permute in dependency of whether or not the input string contains a designated marker:

$$\mathrm{Q}_4(a_1a_2 \cdots a_n) :\equiv \begin{cases} \text{`}a_1a_1a_2a_2 \cdots a_na_n\text{'} & \text{if } a_i \equiv \mathsf{d}, \text{ for some } i \leq n; \\ \text{`}a_1a_2 \cdots a_n\text{'} & \text{otherwise.} \end{cases}$$

$$\mathrm{Q}_5(a_1a_2 \cdots a_n) :\equiv \begin{cases} \text{`}a_n \cdots a_2a_1\text{'} & \text{if } a_i \equiv \mathsf{d}, \text{ for some } i \leq n; \\ \text{`}a_1a_2 \cdots a_n\text{'} & \text{otherwise.} \end{cases}$$

where $a_1, a_2, \ldots, a_n \in \mathcal{A}$, are literal naming functions. Also Q_4 and Q_5 are literal naming functions. Note that the definitions of Q_4 and Q_5 essentially rely on parameters for L and π_- respectively.

B. Let $\mathcal{A}$ consist of English lower case letters together with the symbols $\prime$ and $\circ$. For any $\alpha \in \mathcal{A}^*$, let $B(\alpha)$ be the shortest string of the form $\prime \cdots \prime^\circ$ which does not occur in α. Boolos' quotation function $\mathrm{Q_B} : \mathcal{A}^* \to \mathcal{A}^*$ presented in [1], given by

$$\mathrm{Q_B}(\alpha) :\equiv B(\alpha)\alpha B(\alpha),$$

is a literal naming function. Here the non-trivial bit lies in the begin markers and end markers. A variant of Boolos' construction was communicated to us by Albert Visser and is given as follows. For any $\alpha \in \mathcal{A}^*$, let $B(\alpha) :\equiv {}^\circ\!/\!/\cdots\!/^\circ$, where the length of $/\!/\cdots\!/$ equals the length of α. The quotation function $\mathrm{Q_V} : \mathcal{A}^* \to \mathcal{A}^*$, given by

$$\mathrm{Q_V}(\alpha) :\equiv B(\alpha)\alpha,$$

is a literal naming function.

C. We now show that the quotation device introduced by Halbach and Leigh [9, Chapter 8] can be accommodated in our framework. Let the alphabet $\mathcal{A}_{\mathrm{HL}}$ consist of the following symbols (see [9, Definition 8.1]):

(a) variable symbols $\mathsf{v}_1, \mathsf{v}_2, \mathsf{v}_3, \ldots$;
(b) logical connectives and quantifiers;
(c) non-logical symbols, including a unary function symbol q and a binary function symbol $\frown$;
(d) auxiliary symbols (and);
(e) a quotation constant $\overline{a}$, for each symbol introduced in (a)–(d);
(f) a quotation constant $\underline{0}$ for the empty string ϵ.

Let L be given as follows (without parameters):

- $L(\epsilon) :\equiv \underline{0}$;
- $L(a) :\equiv \overline{a}$, if a was introduced in (a)–(d);
- $L(\overline{a}) :\equiv \mathsf{q}\overline{a}$, if $\overline{a}$ is a quotation constant.

Halbach and Leigh's quotation function $\mathrm{Q_{HL}} : \mathcal{A}_{\mathrm{HL}}^* \to \mathcal{A}_{\mathrm{HL}}^*$ can now be recursively defined as follows:

- $\mathrm{Q_{HL}}(e) :\equiv L(e)$, if $\mathrm{lh}(e) \leq 1$;
- $\mathrm{Q_{HL}}(a_1 \cdots a_n) :\equiv \frown\mathrm{Q_{HL}}(a_1 \cdots a_{n-1})L(a_n)$, where $n > 1$ and $a_i \in \mathcal{A}_{\mathrm{HL}}$ for each $i \leq n$;

Hence, $\mathrm{Q_{HL}}$ is a literal naming function.

D. Let $\mathcal{A}$ consist of English lower case letters together with the space character " ". Let $\mathcal{E}$ be the set of $\mathcal{A}$-strings without the empty string ϵ. Let L map ϵ to the string "empty" and each letter of $\mathcal{A}$ to its ICAO spelling name. For instance, L maps the letter "b" to the string "bravo" and the space character to the string "space". Let $G : \mathcal{A}^* \times \mathcal{A}^* \to \mathcal{A}^*$ be given by

$$G(e, f) :\equiv \begin{cases} f & \text{if } e \equiv \text{empty}; \\ e \text{ " concatenated with " } f & \text{otherwise.} \end{cases}$$

Let $B(s)$ and $E(s)$ be empty. The resulting structural-descriptive naming device SD_1 which maps any expression $a_1a_2\cdots a_n$ to the $\mathcal{A}$-string

$$L(a_1)\text{“ concatenated with” }\cdots\text{“concatenated with”}L(a_n),$$

where $a_1, a_2, \ldots, a_n \in \mathcal{A}$, is a literal naming function.

E. Let the alphabet $\mathcal{A}$ be given by

$$\alpha ::= \mathsf{a} \mid \mathsf{b} \mid \mathsf{c} \mid \cdots \mid \mathsf{z} \mid \overline{\alpha}$$

Here, $\overline{\alpha}$ is conceived of as an alphabetical symbol of length 1. The structural-descriptive naming device $\mathrm{SD}_2\colon \mathcal{A}^* \to \mathcal{A}^*$, given by

$$\mathrm{SD}_2(a_1a_2\cdots a_n) :\equiv \overline{a}_1\overline{a}_2\cdots\overline{a}_n,$$

where $a_1, a_2, \ldots, a_n \in \mathcal{A}$, is a literal naming function.

F. Let $\mathcal{A} = 0, \ldots, 9, \mathsf{a}, \mathsf{b}, \ldots, \mathsf{z}$,“ ” be an ordered alphabet containing the Arabic numerals, the English lower case letters and the space character “ ”. We specify a base 37 notation system for ω by using the k-th alphabetical symbol of $\mathcal{A}$ as the base 37 digit for k (with $0 \leq k < 37$). We write $(a_1\cdots a_n)_{37}$ for the number with base 37 notation $a_1\cdots a_n$. For example, since 2 and b are the 2nd and the 11th symbol of $\mathcal{A}$ respectively, we have

$$(2\mathsf{b})_{37} = 11 + (2 \cdot 37) = 85.$$

We now order the strings of $\mathcal{A}^*$ using the length-first ordering $(\alpha_i)_{i\in\omega}$ in which we enumerate the strings according to increasing length, where the strings of same length are ordered alphabetically. We have $\alpha_m \equiv a_1\cdots a_n$ iff $m = (a_1\cdots a_n)_{37}$. The list $(\alpha_i)_{i\in\omega}$ can be seen as a lexicon for strings over $\mathcal{A}$. We define a naming function $D\colon \mathcal{A}^* \to \mathcal{A}^*$ by mapping each string $a_1\cdots a_n$ to its descriptive name

$$\text{“the word in the lexicon whose index is ”}$$
$$a_1\cdots a_n\text{“ in base 37 notation”}$$

Clearly, D is a literal naming function.

G. We now transfer the descriptive device D from the previous example to an arithmetical setting. Let $\mathcal{A} = \mathsf{a}_1, \ldots, \mathsf{a}_k$ be an alphabet for our arithmetical language $\mathcal{L}$ (including parentheses) and let $(\alpha_i)_{i\in\omega}$ be a length-first ordering of $\mathcal{A}^*$. We now define for each $i \leq k$

$$\mathsf{S}^{\mathsf{a}_i} :\equiv \underbrace{\mathsf{S}\cdots\mathsf{S}}_{i\text{-times}}$$
$$\mathsf{S}_{\mathsf{a}_i}(x) :\equiv \underbrace{\mathsf{S}\cdots\mathsf{S}}_{i\text{-times}}(\overline{k} \times x)$$

We now define the “efficient” naming function $E\colon \mathcal{A}^* \to \mathrm{ClTerm}$ by setting $E(\epsilon) :\equiv \mathsf{0}$ and

$$E(a_1\cdots a_n) :\equiv \mathsf{S}_{a_n}(\cdots\mathsf{S}_{a_1}(\mathsf{0})\cdots).$$

Note that the value of $E(a_1\cdots a_n)$ is the number m with $a_1\cdots a_n \equiv \alpha_m$.

In order to show that E is a literal naming function, we define L (without parameters) by setting $L(\epsilon) :\equiv \mathsf{0}$ and $L(a) :\equiv \mathsf{S}^a$, for $a \in \mathcal{A}$. We set $G(e, f) :\equiv f(\overline{k} \times e)$, for $e, f \in \mathcal{A}^*$. We then have

$$\begin{aligned} E(\epsilon) &\equiv L(\epsilon) \equiv \mathsf{0}; \\ E(a_1 \cdots a_n) &\equiv L(a_n)(\overline{k} \times E(a_1 \cdots a_{n-1})) \\ &\equiv G(E(a_1 \cdots a_{n-1}), L(a_n, a_1 \cdots a_n)), \end{aligned}$$

where $n > 0$ and $a_i \in \mathcal{A}$ for each $i \leq n$. Hence, E is a literal naming function.

H. Let $\mathcal{A}$ be an alphabet for our arithmetical language $\mathcal{L}$ and let $\#\colon \mathcal{A}^* \to \omega$ be a monotonic numbering, i.e., $\#e \leq \#e'$, for all $e, e' \in \mathcal{A}^*$ with $e \preceq e'$. We set $G(e, f) :\equiv fe$, for $e, f \in \mathcal{A}^*$. We define the literal function L by setting

$$L(a, e) :\equiv \begin{cases} \ulcorner a \urcorner^{\#} & \text{if } \mathrm{lh}(e) \leq 1; \\ \underbrace{\mathsf{S} \cdots \mathsf{S}}_{\Delta\text{-times}} & \text{if } e \equiv a_1 a_2 \cdots a_n a \text{ for some } n \geq 1 \ \& \ \forall i \leq n \ a_i \in \mathcal{A} \\ & \text{and } \Delta = \#(a_1 a_2 \cdots a_n a) - \#(a_1 a_2 \cdots a_n); \\ \mathsf{0} & \text{otherwise.} \end{cases}$$

Note that $\Delta \in \omega$, since $\#$ is monotonic. We then have

$$\begin{aligned} \ulcorner e \urcorner^{\#} &\equiv L(e, e), \text{ if } \mathrm{lh}(e) \leq 1; \\ \ulcorner a_1 \cdots a_n \urcorner^{\#} &\equiv L(a_n, a_1 \cdots a_n) \ulcorner a_1 \cdots a_{n-1} \urcorner^{\#} \\ &\equiv G(\ulcorner a_1 \cdots a_{n-1} \urcorner^{\#}, L(a_n, a_1 \cdots a_n)), \end{aligned}$$

where $n > 1$ and $a_i \in \mathcal{A}$ for each $i \leq n$. Hence, $\ulcorner \cdot \urcorner^{\#}$ is a literal naming function.[16] This shows that every naming function which is based on a monotonic numbering and standard numerals is a literal naming function.

Note that unique readability is not satisfied by every literal naming function. However, all of the above examples permit unique readability. We now provide sufficient conditions for the well-foundedness of weak naming relations.

Lemma 6.5 *Let N be a literal naming function for $\mathcal{E} \subseteq \mathcal{A}^*$ which satisfies at least one of the following conditions:*

(1) *N can be defined using markers B and E such that $B(e) \not\equiv \epsilon$ or $E(e) \not\equiv \epsilon$, for each $e \in \mathcal{E}$;*
(2) *N can be defined using a literal function L such that $L(a, e) \notin \mathcal{A}$, for each $a \in \mathcal{A}$ and $e \in \mathcal{A}^*$;*
(3) *N can be defined using a well-founded literal function L, i.e., there are no sequences $(a_i)_{i \in \omega}$ and $(e_i)_{i \in \omega}$ of alphabetical symbols and $\mathcal{A}$-strings respectively such that $L(a_{i+1}, e_{i+1}) \equiv a_i$ for every $i \in \omega$.*
(4) *N satisfies the following two conditions:*

[16] Inspection of the constructions of E (see 6) and $\ulcorner - \urcorner^{\#}$ suggests that the naming function based on efficient numerals preserves literal information in a more strict sense than in the case of standard numerals. In particular, the literal function for E can be defined without parameters, while the literal function for $\ulcorner - \urcorner^{\#}$ essentially relies on the surrounding string as a parameter.

(i) *the relation* $\lhd^N$ *restricted to* $\mathcal{A} \cap \mathrm{im}(N)$ *is well-founded;*
(ii) *for every two* N*-names* e, $f \in \mathrm{im}(N)$, *if* $e \prec f$ *then* $\mathrm{lh}(e) + 1 < \mathrm{lh}(f)$.

Then $\lhd^N$ *is well-founded.*

Proof (1) & (2) follow from the fact that in each case we have $\mathrm{lh}(N(e)) > \mathrm{lh}(e)$, for every $e \in \mathcal{E}$.

We now show (3). Assume that there is a sequence $(e_i)_{i\in\omega}$ of expressions of $\mathcal{E}$ such that $e_{i+1} \lhd^N e_i$, for each $i \in \omega$. Since $e_{i+1} \lhd^N e_i$ implies $\mathrm{lh}(e_{i+1}) \leq \mathrm{lh}(e_i)$, there is a number k such that $\mathrm{lh}(e_{i+1}) = \mathrm{lh}(e_i)$, for each $i \geq k$. Hence, we have $N(e_{i+1}) \equiv e_i$, for each $i \geq k$. Let N be defined by means of a literal pre-naming function N'. We then have

$$\mathrm{lh}(e_{i+1}) \leq \mathrm{lh}(N'(e_{i+1})) \leq \mathrm{lh}(N(e_{i+1})) = \mathrm{lh}(e_{i+1}),$$

for each $i \geq k$. Hence, $N'(e_{i+1}) \equiv N(e_{i+1})$, for each $i \geq k$. We therefore obtain an infinite sequence $(a_i)_{i\in\omega}$ of alphabetical symbols of $\mathcal{A}$ such that $L(a_{i+1}, e_{i+1}) \equiv a_i$. Thus, L is ill-founded.

We now show (4). Let N be defined by some literal pre-naming function N'. We first show that $\mathrm{lh}(s) < \mathrm{lh}(N'(s))$ for all $s \in \mathcal{E}$ with $\mathrm{lh}(s) > 1$. Let $s \equiv a_1 \cdots a_n$, for some $n > 1$ and $a_1, \ldots, a_n \in \mathcal{A}$. Let

$$N'(s) \equiv G(N'(a_{\pi_s(1)} \cdots a_{\pi_s(n-1)}), L(a_{\pi_s(n)}, s)),$$

where G, L and π_- are given as in Definition 6.4. Since G is weakly $\sqsubset$-increasing, we have $N'(a_{\pi_s(1)} \cdots a_{\pi_s(n-1)}) \prec N'(s)$. Hence, using (ii), we have

$$\mathrm{lh}(s) = n \leq \mathrm{lh}(N'(a_{\pi_s(1)} \cdots a_{\pi_s(n-1)})) + 1 < \mathrm{lh}(N'(s)).$$

Now, assume that there is a sequence $(e_i)_{i\in\omega}$ of expressions of $\mathcal{E}$ such that $e_{i+1} \lhd^N e_i$, for each $i \in \omega$. As we have seen in the proof of clause (3), there is a number k such that $\mathrm{lh}(e_{i+1}) = \mathrm{lh}(e_i)$ and $N'(e_{i+1}) \equiv e_i$, for each $i \geq k$. Since $\mathrm{lh}(e) < \mathrm{lh}(N'(e))$ for all composite expressions e, we have $\mathrm{lh}(e_i) = 1$ for each $i \geq k$. Hence, there is a sequence $(a_i)_{i\in\omega}$ of alphabetical symbols such that $a_{i+1} \lhd^N a_i$ for each $i \in \omega$. But this contradicts (i). □

We observe that all literal naming functions introduced above give rise to well-founded weak naming relations.

Corollary 6.6 *The literal naming functions given in (A)-(H) induce well-founded weak naming relations.*

Proof The functions Q_1, Q_2, Q_3, Q_4, Q_5, Q_B, Q_V and D satisfy clause (1) of Lemma 6.5. The function SD_1 satisfies clause (2) of Lemma 6.5. Finally, Q_HL, SD_2, E and $\ulcorner - \urcorner^\#$ satisfy clause (3) of Lemma 6.5 (note that $0 \leq \#\epsilon < \#0$, since $\#$ is monotonic). □

These examples suggest that the well-foundedness principle is grounded in a rather robust and general conception of quotation.

Remark 6.7 Instead of requiring well-foundedness, one may require naming functions which resemble quotation to be strongly monotonic (cf. Section 5). Here, the essential assumption is that each quotation properly contains its quoted expression (as strings). It can then be argued that numberings which mimic quotations are required to code the Gödel numeral of an expression e by a larger number than the expression e itself (see [8, Section 6] for an elaboration of this view). Note that only the naming functions Q_1, Q_B, Q_V and D satisfy this assumption. Hence, the justification of strong monotonicity seems to require a much more narrow conception of quotation than in the case of well-foundedness.

Moreover, we immediately obtain the following result from clause (4) of Lemma 6.5.

Corollary 6.8 *Let* ClTerm *be the set of closed $\mathcal{L}$-terms such that each complex term is of the form $f(u_1 \dots u_k)$, where f is a k-ary function symbol. Let $\mathcal{E} \subseteq \mathcal{A}^*$ be such that* $\mathrm{Term}_x \cup \mathrm{Fml}_x \subseteq \mathcal{E}$. *Let $N\colon \mathcal{E} \to$* ClTerm *be a literal naming function for $\mathcal{E}$. Then $\lhd^N$ is well-founded iff $N(\mathsf{0}) \not\equiv \mathsf{0}$.*

Remark 6.9 The philosophical significance of the above corollary is limited by the fact that it depends on subtleties regarding the employed notation system for arithmetical terms. For example, by Lemma 6.5.4 the corollary also holds if each complex term is enclosed by parenthesis, or if each function symbol consists of a composite string. However, we can easily construct a counterexample to Corollary 6.8 if complex terms are of the form $f u_1 \dots u_k$, where f is an alphabetical symbol. These considerations suggest that the choice of the notation system is yet another source of intensionality in the context of self-reference.

We now return to our study of Kreisel-like constructions of refutable Henkin sentences. In Section 5, we have seen that Kreisel-like fixed-points can be uniformly constructed with respect to circular naming relations. In the next section we will introduce another variant of Kreisel-like constructions of refutable Henkin sentences which are based on an ill-founded but non-circular naming relation. By slightly generalising our results of Section 5, we will show that the requirements of well-foundedness of the naming relation, together with uniformity, also rule out this new variety of deviant fixed-point constructions.

7 Ill-Foundedness Without Circles

Recall that the Kreisel-like Henkin sentences introduced in Section 4 consist of a provability predicate $\mathsf{Bew}^\star(x)$ of the form

$$x \neq d(\mathsf{Bew}^\star(x)) \wedge \mathsf{Bew}(x),$$

where d is a diagonal operator. That is, $\mathsf{Bew}^\star(x)$ contains its own fixed-point term $d(\mathsf{Bew}^\star(x))$. As we have seen in Section 5, if d is uniform then $d(\mathsf{Bew}^\star(x))$ induces a circle with regard to the underlying naming relation (Proposition 5.12). Hence, the

constraints of uniformity and non-circularity are sufficient to rule out the refutable Henkin sentences considered thus far.

However, we can tweak the construction of $\mathsf{Bew}^\star(x)$ such that it no longer contains its own fixed-point term but still yields refutable Henkin sentences. In order to do so, we construct an ω-chain of formulas $(\mathsf{Bew}_n(x))_{n\in\omega}$ such that $\mathsf{Bew}_n(x)$ is of the form

$$\mathsf{jump}(x) \neq d(\mathsf{Bew}_{n+1}(x)) \wedge \mathsf{Bew}(x),$$

where $\mathsf{Bew}(x)$ is some fixed provability predicate and jump represents a function mapping $d(\mathsf{Bew}_n(x))$ to $d(\mathsf{Bew}_{n+1}(x))$, for every $n \in \omega$. As in the case of $\mathsf{Bew}^\star(x)$ above, the Henkin sentence of each provability predicate $\mathsf{Bew}_n(x)$ is refutable. As opposed to $\mathsf{Bew}^\star(x)$, however, $d(\mathsf{Bew}_n(x))$ is not contained in $\mathsf{Bew}_n(x)$ itself. Hence, there are refutable Henkin sentences which are based on provability predicates which do not contain their own fixed-point terms and thus evade Proposition 5.12:

Lemma 7.1 *There is a numbering α, a numeral function ν, a standard interpretation $\mathcal{I}$, a uniform diagonal operator d with respect to α, ν and $\mathcal{I}$ and for each $n \in \omega$ there is a formula $\mathsf{Bew}_n(x)$ such that*

(1) *$\mathsf{Bew}_n(x)$ weakly represents $\mathsf{Basic}(\mathcal{I})$;*
(2) *$\mathsf{Basic}(\mathcal{I}) \vdash \neg\mathsf{Bew}_n(t_n)$;*
(3) *t_n does not occur in $\mathsf{Bew}_n(x)$;*

where t_n is the fixed-point term $d(\mathsf{Bew}_n(x))$ of $\mathsf{Bew}_n(x)$ w.r.t. α and $\mathcal{I}$.

Proof The details of the construction sketched above can be found in B.3. □

Inspection of the ω-chain $(\mathsf{Bew}_n(x))_{n\in\omega}$ constructed in B.3 shows that the fixed-point term t_{n+1} of $\mathsf{Bew}_{n+1}(x)$ occurs in $\mathsf{Bew}_n(x)$, for each $n \in \omega$. We now ask under which additional assumptions we can rule out both Kreisel's original construction and its variant based on ω-chains as given above:

Question 7.2 Let d be a uniform diagonal operator and $(\varphi_n(x))_{n\in\omega}$ and $(t_n)_{n\in\omega}$ sequences of formulas and closed terms respectively. Under which assumptions can we rule out the possibility that for every $n \in \omega$ we have $d(\varphi_n(x)) \equiv t_n$, where $\varphi_n(x)$ contains t_m for some $m \geq n$.

An answer to Question 7.2 yields also an answer to Question 3.6 by setting $\varphi_n(x) :\equiv \varphi(x)$ and $t_n :\equiv t$, for each $n \in \omega$.

We first observe that requiring uniformity together with the non-circularity of the weak naming relation is not sufficient to rule out the construction of ω-chains of refutable Henkin sentences as given above:

Lemma 7.3 *We can choose the numbering α and the numeral function ν in Lemma 7.1 such that they induce a non-circular weak naming relation $\lhd^{\alpha,\nu}$.*

Proof See B.4. □

However, once we require uniformity together the well-foundedness of the weak naming relation, deviant Henkin sentences such as constructed above can be successfully excluded. More generally, we obtain the following answer to Question 7.2:

Proposition 7.4 *Let $(\varphi_n(x))_{n\in\omega}$ and $(t_n)_{n\in\omega}$ be sequences of formulas and closed terms respectively. If the $\lhd_*$ relation induced by the naming function $\ulcorner - \urcorner$ is well-founded, there is no uniform diagonal operator d such that for every $n \in \omega$ we have $d(\varphi_n(x)) \equiv t_n$, where $\varphi_n(x)$ contains t_m for some $m \geq n$.*

Proof Let $n \in \omega$. We have $d(\varphi_n(x)) \equiv t_n$, where $\varphi_n(x)$ contains t_m for some $m \geq n$. By Lemma 5.11, there exists a term s such that $\varphi_n(s) \lhd_* t_n$. Since t_m is a subterm of the formula $\varphi_n(s)$, we obtain $t_m \lhd_* t_n$ by Fact 5.2. Hence, there is an infinite subsequence $(u_n)_{n\in\omega}$ of $(t_n)_{n\in\omega}$ such that

$$u_{n+1} \lhd_* u_n, \text{ for each } n \in \omega.$$

This contradicts our assumption that $\lhd_*$ is well-founded. □

7.1 Limitations

At this point we should stress that the constraints of uniformity and well-foundedness by no means rule out *every* deviant construction of a Henkin sentence. After all, in this paper we have only investigated constraints on the fixed-point operator and the naming function, while we impose no constraints whatsoever on provability predicates, except that they should weakly represent the set of theorems of the theory. It is therefore hardly surprising that there exists a contrived provability predicate whose canonical diagonalization, say via Gödel's method, yields a refutable Henkin sentence (see [10, Section 5] for an example).

We conclude this section by providing another concrete example showing that uniformity and well-foundedness are not be sufficient to rule out every accidental diagonal sentence. Recall that the Kreisel-like constructions considered in this paper consist of a provability predicate $\mathsf{Bew}^\circ(x)$ of the form

$$\chi(x) \wedge \mathsf{Bew}(x),$$

where $\mathsf{Bew}(x)$ is a provability predicate weakly representing provability and $\chi(x)$ is a formula such that $\chi(d(\mathsf{Bew}^\circ(x)))$ is refutable and $\chi(\ulcorner\varphi\urcorner)$ is provable for all sentences φ which are distinct to $\mathsf{Bew}^\circ(d(\mathsf{Bew}^\circ(x)))$. The conjunct χ of $\mathsf{Bew}^\star(x)$ (see Section 4) contains its fixed-point term $d(\mathsf{Bew}^\star(x))$, and the conjunct χ of $\mathsf{Bew}_n(x)$ (cf. Section 7) contains the fixed-point term $d(\mathsf{Bew}_{n+1}(x))$ of its subsequent provability predicate in the given ω-chain. As we have seen, this is precisely the reason why uniform versions of these constructions force the naming relation to be circular and ill-founded respectively. We now provide an example of a Kreisel-like construction which is based on a provability predicate whose conjunct χ does not contain any fixed-point term. Hence, this construction can be given with respect to a uniform

diagonal operator d and a well-founded naming relation. To do so, let $\mathsf{Bew}^\circ(x)$ be of the form

$$f_0^0(x) \neq 0 \wedge \mathsf{Bew}(x),$$

where $\mathsf{Bew}(x)$ is a provability predicate weakly representing provability and f_0^0 represents the function which maps the code of $d(\mathsf{Bew}^\circ(x))$ to 0 and each other number to itself. Clearly, $f_0^0(\overline{n}) \neq 0$ is satisfied by all positive numbers n which are not the code of $d(\mathsf{Bew}^\circ(x))$. Assuming that sentences all have positive codes, we therefore obtain a refutable Henkin sentence $d(\mathsf{Bew}^\circ(x))$ with respect to the uniform diagonal operator d and a standard naming function.

Finally, we can even construct refutable Henkin sentences without any additional conjunct (see B.5 for a detailed construction).

8 Applications

In this section, we present various applications of uniformity in distinguishing and identifying accidental fixed-points constructed by various means in the literature.

8.1 Logical Derivability

We first provide an example of accidental self-reference from a setting closer to natural language. For a given English sentence φ, we may ask about the status of the sentence that says of itself that it is logically derivable from φ. The status of such sentences depends on how self-reference is obtained:

(1) The sentence (1) is logically derivable from (1).
(2) The sentence (2) is logically derivable from (1).

Clearly, the sentence (1) is true, while (2) is false. In the metamathematical study of self-reference we would like to rule out diagonal operators which mirror the accidental self-referential feature of sentence (1). While the KH-property is not sufficient to rule out such diagonal operators, the requirement of uniformity successfully excludes metamathematical counterparts of sentence (1).

Let $\mathsf{Pr}_{\alpha(v)}(x)$ denote Feferman's [5] provability predicate. For any closed term t which denotes a sentence, set $\mathsf{Bew}_t(x) :\equiv \mathsf{Pr}_{v=t}(x)$. That is, $\mathsf{Bew}_t(x)$ is a standard provability predicate for the $\mathcal{L}$-theory whose only non-logical axiom is the sentence whose code is denoted by t.

Lemma 8.1 *Let $\ulcorner - \urcorner$ be a well-founded naming function. There is a KH-diagonal operator d_0 and a closed term t such that*

(1) $t \equiv d_0(\mathsf{Bew}_t(x))$;
(2) $\mathsf{PA} \vdash t = \ulcorner \mathsf{Bew}_t(t) \urcorner$
(3) $\mathsf{PA} \vdash \mathsf{Bew}_t(d_0(\mathsf{Bew}_t(x)))$;
(4) $\mathsf{PA} \vdash \neg\mathsf{Bew}_t(d(\mathsf{Bew}_t(x)))$, *for every uniform diagonal operator d.*

Proof We first show (1)-(3). Let $t :\equiv d_J(\mathsf{Bew}_x(x))$. Hence, $\Sigma \vdash t = \ulcorner \mathsf{Bew}_t(t) \urcorner$. Let d_0 be a diagonal operator which maps $\mathsf{Bew}_t(x)$ to t and any other formula of the form $\varphi(x)$ to $d_J(\varphi(x))$. Clearly, d_0 satisfies the KH-property. Moreover, we have $\mathsf{Bew}_t(t) \vdash \mathsf{Bew}_t(t)$ and hence $\mathsf{PA} \vdash \mathsf{Bew}_t(d_0(\mathsf{Bew}_t))$.

In order to show (4), let d be any uniform diagonal operator. By Proposition 5.12, we have $t \not\equiv d(\mathsf{Bew}_t(x))$. Hence, $\mathsf{Bew}_t(t) \not\vdash \mathsf{Bew}_t(d(\mathsf{Bew}_t(x)))$. In order to show that $\mathsf{PA} \vdash \neg\mathsf{Bew}_t(d(\mathsf{Bew}_t(x)))$, we observe that

$$\mathsf{PA} \vdash \neg\mathsf{Bew}_\emptyset(t \dot{\to} d(\mathsf{Bew}_t(x)))$$

and

$$\mathsf{PA} \vdash \mathsf{Bew}_t(d(\mathsf{Bew}_t(x))) \leftrightarrow \mathsf{Bew}_\emptyset(t \dot{\to} d(\mathsf{Bew}_t(x))).$$ □

8.2 Codings with Built-in Diagonalization

We now turn to the question to what extent uniformity excludes fixed-points which are obtained by codings with built-in diagonalization. Recall the construction of the refutable Henkin sentence $\mathsf{Bew}(\widetilde{m})$ in Observation 2.3. Intuitively, $\mathsf{Bew}(\widetilde{m})$ is an accidental diagonal sentence since it relies on a numbering which is constructed in a highly ad hoc fashion. This intuition can be grounded in mathematical facts as follows. While the employed numbering α together with standard numerals induce a well-founded naming relation, the fixed-point term $\widetilde{m}$ cannot be constructed uniformly. Recall that if we do not mention the numeral function in a parameter of naming function, it means we take the standard numerals.

Lemma 8.2 *Let* $\mathsf{Bew}(x)$*,* α *and* $\widetilde{m}$ *be given as in Observation 2.3.*

(1) *The naming relation* $\lhd^\alpha$ *is well-founded;*
(2) *No diagonal operator which maps* $\mathsf{Bew}(x)$ *to* $\widetilde{m}$ *is uniform for* α*.*

Proof The relation $\lhd^\#$ is well-founded by Corollary 6.6. Using Fact 6.2, it is therefore sufficient to show that $\overline{m} \not\lhd^\alpha \overline{m}$. Assume that $\overline{m} \lhd^\alpha \overline{m}$. Then there is a string e such that $\overline{m} \preceq e$ and $\alpha(e) \leq m$. We have $e \not\equiv \mathsf{Bew}(\widetilde{m})$ by definition of m. Hence, $\alpha(e) = \#e$. But since $\#$ is monotonic, we have

$$m < \#\overline{m} \leq \#e \leq m,$$

a contradiction.

Assume now that d is a uniform diagonal operator which maps $\mathsf{Bew}(x)$ to $\widetilde{m}$. By Lemma 5.11 there is a term s such that $\mathsf{Bew}(s) \lhd^\alpha_* d(\mathsf{Bew}(x)) \equiv \widetilde{m}$. Since 0 and 1 are not $\#$-codes, there is no string e with $\ulcorner e \urcorner^\alpha \preceq \widetilde{m}$. Hence $\mathsf{Bew}(s) \lhd^\alpha_* \widetilde{m}$ cannot be true. □

We now turn to other codings with built-in diagonalization. Let β be a monotonic numbering of strings such that for each formula $\varphi(x)$ with x free there is a number n^φ such that $n^\varphi = \beta(\varphi(\underline{n^\varphi}))$, where $\underline{n^\varphi}$ denotes the efficient numeral of n^φ (e.g., take β to be the numbering gn_1 constructed in [8, Section 5]).

Clearly, β gives rise to an ill-founded naming relation if we use efficient numerals. This is because the diagonal sentence $\varphi(\underline{n^\varphi})$ contains its own name $\underline{n^\varphi}$. However, β together with standard numerals induce a well-founded relation $\lhd^\beta$. Yet, no diagonal operator which maps $\varphi(x)$ to $\underline{n^\varphi}$ can be uniform for β.

Lemma 8.3

(1) *The naming relation* $\lhd^\beta$ *is well-founded;*
(2) *The diagonal operator* d *given by* $d(\varphi(x)) :\equiv \underline{n^\varphi}$ *is not uniform for* β.

Proof By Corollary 6.6, $\lhd^\beta$ is well-founded. Assume that d is uniform. By Lemma 5.11 there is a term s such that $\varphi(s) \lhd^\beta_* \underline{n^\varphi}$. But $\underline{n^\varphi}$ does not contain any standard numeral which is the β-code of any expression. Hence, $\varphi(s) \lhd^\beta_* \underline{n^\varphi}$ cannot be true. □

We close by showing that there is a coding with built-in diagonalization which induces a well-founded naming relation and yields a uniform diagonal operator. Let δ be a numbering of the well-formed expressions of $\mathcal{L}$ such that for any given $\varphi(x)$ with x free there is a number n^φ with $n^\varphi = \delta(\varphi(x))$ and $(n^\varphi)^2 = \delta(\varphi(\underline{n^\varphi} \times \underline{n^\varphi}))$.[17] Set $\ulcorner - \urcorner = \dot{} \circ \delta$.

Lemma 8.4

(1) *The naming relation* $\lhd^{\ulcorner - \urcorner}$ *is well-founded;*
(2) *The diagonal operator* d *given by* $d(\varphi(x)) :\equiv \underline{n^\varphi} \times \underline{n^\varphi}$ *is uniform for* $\ulcorner - \urcorner$.

Proof In order to show that $\lhd^{\ulcorner - \urcorner}$ is well-founded, it is sufficient to show that for every expression e and numbers m, n:

$$e \lhd^{\ulcorner - \urcorner} \underline{n} \text{ and } \underline{m} \preceq e \text{ implies } m < n.$$

If the antecedent holds, then there is an expression f such that $e \preceq f$ and $\ulcorner f \urcorner \preceq \underline{n}$. By [8, Lemma 6.10] we have $m < \delta(\underline{m})$. Since δ is monotonic, we have

$$m < \delta(\underline{m}) \leq \delta(e) \leq \delta(f) \leq n.$$

Let now any formula $\varphi(x)$ with x free be given. By definition of δ there is a number n^φ with $n^\varphi = \delta(\varphi(x))$ and $(n^\varphi)^2 = \delta(\varphi(\underline{n^\varphi} \times \underline{n^\varphi}))$. Let $\overline{\times}$ be the basic meta-linguistic operation which maps two terms s, t to $s \times t$. We then have

$$d(\varphi(x)) \equiv \underline{n^\varphi} \times \underline{n^\varphi} \equiv \overline{\times}(\ulcorner \varphi(x) \urcorner, \ulcorner \varphi(x) \urcorner).$$

Hence, d is uniform. □

[17] δ can be obtained by slightly tweaking the construction of the numbering in [8, Section 6.2].

9 Conclusion

If not all diagonal sentences for a formula expressing a property behave in the same way, we can exclude those diagonal sentences that are not self-referential by the Kreisel–Henkin criterion, assuming we are interested in *the* sentence ascribing P to themselves. Maybe there is no single such sentence, but any diagonal sentences ascribing P to itself must be self-referential.

However, self-referential diagonal sentences ascribing provability to themselves via Kreisel provability predicates still vary in their properties, as they may be provable, refutable, or independent. Refutable Henkin sentences are obtained by plugging in a specific term into the formula that happens to be a self-referential Henkin sentence in virtue of the cunning construction. If the usual diagonal constructions are applied to the provability predicate, provable sentences are obtained. If we are interested in *the* sentence ascribing provability to itself via this provability predicate it must be among the provable ones. The refutable ones can only be obtained via a trick very specific to the provability predicate in question. Thus, we single out those self-referential diagonal sentences that have been obtained in a *uniform* way. This is sufficient to eliminate refutable Henkin sentences, as long as we employ a canonical coding and numeral function.

Of course, appealing to "canonical" codings and numeral functions is as unsatisfactory as appealing to "canonical" diagonal sentences. Hence, we replace this vague condition with a precise condition on the naming relation: Ruling out illfounded naming relations is then sufficient to obtain only provable Henkin sentences from Kreisel-style provability predicates. Generally, the well-foundedness of the naming relation is another constraint for narrowing down the class of diagonal sentences.

We do not not maintain that these constraints are the final word. Section 7.1 contains an example hinting at the need for further constraints. However, in some cases our constraints suffice to answer the question about *the* sentence ascribing some property to itself. Of course, our constrains on diagonal sentences interact with other constraints on the language, the coding, the axiomatization of the theory, and the formula expressing the property. All these interrelate and there is much scope for future work.

Appendix A. Proof of Lemma 5.11

We first prove the more general result that for any non-constant uniform function $u\colon \mathrm{Fml}_x \to \mathrm{Term}_x$ the claim holds. Let $r \in \mathcal{R}$ be a reduced representation of u. By Lemma 5.9, it contains b_n as a subterm, where $\mathsf{ev}(b_n) = \ulcorner - \urcorner_f$. We prove by induction over the complexity of r that $\varphi(s) \lhd_* \mathsf{ev}(r)(\varphi)$, for some term s. For the base case it is sufficent to check the claim for $r = b_n$, since no other C-basic function represents a function of the right type. Clearly, we have $\varphi(x) \lhd \ulcorner \varphi(x) \urcorner$. For the induction step, again by Fact 5.8 and the fact that $\mathsf{ev}(r)$ is not constant, r must be of the form $r = \odot(b_m, q)$, where $\mathsf{ev}(b_m)$ is one of the basic functions $\mathrm{Sub}_t, \ulcorner - \urcorner_f, \ulcorner - \urcorner_t$

or $\overline{f}$ with $\mathsf{ar}(f) \geq 1$. We now show by the following case distinction that $\varphi(s) \lhd_* \mathsf{ev}(r)(\varphi)$, for some term s.

(1) If $\mathsf{ev}(b_m) = \mathrm{Sub}_t$, then by Fact 5.5 q is of the form $\mathring{\wp}(q_1, q_2)$, where $\mathsf{ev}(q_i) : \mathrm{Fml}_x \to \mathrm{Term}_x$ for $i = 1, 2$. Both q_1 and q_2 are reduced, and at least one of $\mathsf{ev}(q_1)$ and $\mathsf{ev}(q_2)$ is not a constant function. If $\mathsf{ev}(q_1)$ is not constant, then by the induction hypothesis we have $\varphi(s) \lhd_* \mathsf{ev}(q_1)(\varphi)$ for some term s. By Fact 5.3 there is a closed term $t \preceq \mathsf{ev}(q_1)(\varphi)$ such that $\varphi(s) \lhd_* t$; and since $t \preceq \mathsf{ev}(r)(\varphi)$, this shows $\varphi(s) \lhd_* \mathsf{ev}(r)(\varphi)$. Now suppose $\mathsf{ev}(q_1)$ is a constant function. Then $\mathsf{ev}(q_2)$ must not be constant, and thus by the induction hypothesis $\varphi(s) \lhd_* \mathsf{ev}(q_2)(\varphi)$ for some term s. Also, $\mathsf{ev}(q_1)$ cannot be closed, otherwise $\mathsf{ev}(r)$ would be a constant function. Hence, by Lemma 5.10, $\mathsf{ev}(q_1)(\varphi)$ contains a free variable x. This means that $\mathsf{ev}(q_2)(\varphi) \preceq \mathsf{ev}(r)(\varphi)$, and we conclude that $\varphi(s) \lhd_* \mathsf{ev}(r)(\varphi)$ by Fact 5.2.

(2) If $\mathsf{ev}(b_m) = \ulcorner - \urcorner_t$, then $\mathsf{ev}(q) : \mathrm{Fml}_x \to \mathrm{Term}_x$ is a non-constant function. By induction hypothesis there is a term s such that $\varphi(s) \lhd_* \mathsf{ev}(q)(\varphi)$. We also have $\mathsf{ev}(q)(\varphi) \lhd \ulcorner \mathsf{ev}(q)(\varphi) \urcorner$, hence $\varphi(s) \lhd_* \mathsf{ev}(r)(\varphi)$.

(3) If $\mathsf{ev}(b_m) = \overline{f} : \mathrm{Term}_x^n \to \mathrm{Term}_x$, then by Fact 5.5, q must be of the form $\mathring{\wp}(q_1, \mathring{\wp}(q_2, \mathring{\wp}(\cdots, \mathring{\wp}(q_{n-1}, q_n) \cdots)))$ (if $n = 1$ we simply have q), with each q_i of type $\mathrm{Fml}_x \to \mathrm{Term}_x$. There is $i \leq n$ such that q_i is not a constant function. By induction hypothesis, $\varphi(s) \lhd_* \mathsf{ev}(q_i)(\varphi)$, for some term s. Since $\mathsf{ev}(q_i)(\varphi) \preceq \mathsf{ev}(r)(\varphi)$, we obtain $\varphi(s) \lhd_* \mathsf{ev}(r)(\varphi)$ by Fact 5.2.

(4) Finally, if $\mathsf{ev}(b_m) = \ulcorner - \urcorner_f$, then $\mathsf{ev}(q) : \mathrm{Fml}_x \to \mathrm{Fml}_x$ is a non-constant function. We now show that $\varphi(s) \preceq \mathsf{ev}(q)(\varphi)$ or $\varphi(s) \lhd_* \mathsf{ev}(q)(\varphi)$, for some term s. In either case, we obtain the desired result, i.e., $\varphi(s) \lhd_* \mathsf{ev}(r)(\varphi)$. We prove this disjunction by a further local induction over the complexity of q. If q is b_l, for some $l \in \omega$, then we have $\varphi(x) \preceq \mathsf{ev}(b_l)(\varphi)$. This is because the only C-basic function of the right type is $\mathrm{id}_{\mathrm{Fml}_x}$. For the inductive step, it is sufficient to assume by Fact 5.8 and the fact that $\mathsf{ev}(q)$ is not a constant function, that q is of the form $\odot(b_k, q')$ where $\mathsf{ev}(b_k)$ is Sub_f, $\overline{\star}$ or $\overline{R}$ with $\mathsf{ar}(R) \geq 1$. We proceed by considering each of these cases:

(a) If $\mathsf{ev}(b_k) = \mathrm{Sub}_f$, then q' must be $\mathring{\wp}(q'_1, q'_2)$ with $\mathsf{ev}(q'_1) : \mathrm{Fml}_x \to \mathrm{Fml}_x$ and $\mathsf{ev}(q'_2) : \mathrm{Fml}_x \to \mathrm{Term}_x$. At least one of $\mathsf{ev}(q'_1)$, $\mathsf{ev}(q'_2)$ is not a constant function. If $\mathsf{ev}(q'_1)$ is not constant, then by the induction hypothesis, we have $\varphi(s) \preceq \mathsf{ev}(q'_1)(\varphi)$ or $\varphi(s) \lhd_* \mathsf{ev}(q'_1)(\varphi)$ for some term s. In the former case, we have $\varphi(s[x/\mathsf{ev}(q'_2)(\varphi)]) \preceq \mathsf{ev}(r)(\varphi)$. In the latter case we conclude $\varphi(s) \lhd_* \mathsf{ev}(r)(\varphi)$ by a similar argument as in (1). Now if $\mathsf{ev}(q'_1)$ is a constant function, then $\mathsf{ev}(q'_2)$ is not constant and hence $\mathsf{ev}(q'_1)$ cannot be closed. Then $\mathsf{ev}(q'_1)(\varphi)$ contains a free variable by Lemma 5.10, and we obtain $\mathsf{ev}(q'_2)(\varphi) \preceq \mathsf{ev}(r)(\varphi)$. By the *outer* induction hypothesis, $\varphi(s) \lhd_* \mathsf{ev}(q'_2)(\varphi)$ for some term s. We then conclude $\varphi(s) \lhd_* \mathsf{ev}(r)(\varphi)$ by another use of Fact 5.2.

(b) If $\mathsf{ev}(b_k) = \overline{\star}$, then q' must be $\mathring{\wp}(q'_1, \mathring{\wp}(\cdots, \mathring{\wp}(q'_{n-1}, q'_n) \cdots))$ (if $n = 1$ we simply have q'), with each q'_i, $\mathsf{ev}(q'_i) : \mathrm{Fml}_x \to \mathrm{Fml}_x$. At least one q'_i is not

a constant function. Hence by induction hypothesis $\varphi(s) \preceq \mathsf{ev}(q_i')(\varphi)$ or $\varphi(s) \lhd_* \mathsf{ev}(q_i')(\varphi)$. Since $\mathsf{ev}(q_i')(\varphi) \preceq \mathsf{ev}(q')(\varphi)$, we are done.

(c) The case for $\mathsf{ev}(b_k) = \overline{R}$, where $\mathsf{ar}(R) \geq 1$, proceeds similarly to (4.b).

The proof is complete since diagonal operators cannot be constant functions.

Appendix B. Uniform Constructions of Deviant Henkin Sentences

This part of the appendix contains several explicit constructions of deviant provability predicates which yield refutable Henkin sentences. Since all of these constructions will rely on the recursion theorem, we start by briefly introducing this important recursion theoretic result.

We start by assigning indices to p.r. functions.[18] We write F_a for the p.r. function with index a. Note that each p.r. function has infinitely many indices. Moreover, the set of indices is p.r. The following recursion theorem for p.r. functions shows that we can construct p.r. functions in a self-referential way, by using their indices in their own definitions.[19]

Theorem B.1 (Primitive Recursion Theorem) *For every k+1-ary p.r. function $G(\vec{x}, y)$, there is an index a such that $F_a(\vec{x}) = G(\vec{x}, a)$, where $\vec{x}$ is a k-tuple of variables.*

In this paper we only consider standard interpretations which are intuitively "effective" (for a definition of a standard interpretation see Section 4). More precisely, we require that for every standard interpretation $\mathcal{I}$ there is a recursive function which maps each pair (n, k) of numbers to an index a such that $\mathcal{I}(f_n^k) = F_a$.

Before we provide our constructions we fix some more notation. Let $\mathcal{L}$ be given as in Section 4 and let # be some standard elementary numbering of $\mathcal{L}$ such that $\#e > 0$, for all well-formed expressions e in $\mathcal{L}$. Let P_i^k be the projection function which maps a $k+1$-tuple to its $i+1$-th component (where $i \leq k$).

Let the function $\mathcal{I}' \colon \{f_n^k \mid n \neq 0, k \neq 0\} \to \mathfrak{Pr}$ be given by

$$\mathcal{I}'(f_n^k) = \begin{cases} F_n & \text{if } n \text{ is the index of a } k+1\text{-ary function;} \\ P_0^k & \text{otherwise.} \end{cases}$$

Clearly, $\mathcal{I}'$ is surjective. For each index a of a unary function, let $\mathcal{I}^a$ be the extension of $\mathcal{I}'$ by mapping the function symbol f_0^0 to F_a. Each such $\mathcal{I}^a$ is a standard interpretation function of $\mathcal{L}$. We observe:

[18] See [14, p. 34] or [19, pp. 91]. It is crucial that we only assign indices to p.r. functions *via their p.r. constructions*. Note that there are functions which are defined by μ-recursion, but "happen" to be p.r. The set of indices of p.r. functions which are constructed primitive recursively is p.r., while the set of (partial) recursive functions which are p.r. is undecidable by Rice's theorem.

[19] A proof of this theorem can be found in [14, p. 41].

Fact B.2 For each index a of a unary function, there is an $\mathcal{L}_0$-formula $\mathsf{Bew}^a(x)$ which weakly represents $\mathsf{Basic}(\mathcal{I}^a)$ relative to #. Moreover, there is a p.r. function H which maps each such a to the #-code of $\mathsf{Bew}^a(x)$.

We now prove Lemma 4.2 by providing two examples of uniform diagonal operators which satisfy the fixed-point property of Lemma 4.2. The first example employs a canonical numeral function, namely, standard numerals, but is based on a contrived numbering. The second example uses a standard numbering, but relies on an artificial numeral function. These examples can be developed for all uniform diagonal operators introduced in this paper. For the sake of simplicity, however, we will base the exposition on diagonal operators which are particularly suitable.

B.1 First Proof of Lemma 4.2

We base the first construction on the diagonal operator d_B introduced in Example 3.5. Similar but slightly more complicated constructions can be given for d_G and d_J.

Let J be a unary p.r. function which maps the #-code of φ to the #-code of $f_0^0(\ulcorner\varphi\urcorner^{\#})$. Let $\wedge_{\#}$ denote the #-tracking function of $\overline{\wedge}$, i.e., $\wedge_{\#}(\#\varphi, \#\psi) = \#\varphi \wedge \psi$. Similarly, let $\mathrm{Sub}_{\#}$ denote the #-tracking function of $\mathrm{Sub}_f : \mathrm{Fml}_x \times \mathrm{Term}_x \to \mathrm{Fml}_x$, defined in Section 3. Let z be the #-code of the formula

$$f_0^0(0) \neq f_0^0(0).$$

We now define a function $G\colon \omega^2 \to \omega$ by setting

$$G(p,q) := \begin{cases} \mathrm{Sub}_{\#}(p, J(p)) & \text{if } p \in \#(\mathrm{Fml}_x); \\ \wedge_{\#}(z, \mathrm{Sub}_{\#}(H(q), \#f_0^0(0))) & p = 0; \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, G is p.r. Using Theorem B.1, we find an index a such that

$$F_a(p) = G(p,a), \text{ for each } p \in \omega.$$

We now set

$$\mathsf{Bew}^*(x) :\equiv x \neq f_0^0(0) \wedge \mathsf{Bew}^a(x).$$

Using the fixed-point property of a, we get

$$F_a(0) = \#(f_0^0(0) \neq f_0^0(0) \wedge \mathsf{Bew}^a(f_0^0(0))) = \#(\mathsf{Bew}^*(f_0^0(0))). \tag{2}$$

We now define a numbering α as follows:

$$\alpha(\varphi) := \begin{cases} 0 & \text{if } \varphi \equiv \mathsf{Bew}^*(x); \\ \#\varphi & \text{otherwise.} \end{cases}$$

The numbering α is injective and elementary. Moreover, $\mathsf{Bew}^a(x)$ also weakly represents $\mathsf{Basic}(\mathcal{I}^a)$ relative to α. We set $\dot{d} :\equiv f_0^0$. By Eq. 2 and the fact that $\mathcal{I}^a(\dot{d}) = F_a$, we have for each $\varphi(x) \in \mathrm{Fml}_x$ that

$$\mathsf{Basic}(\mathcal{I}^a) \vdash \dot{d}(\ulcorner\varphi\urcorner^{\alpha}) = \varphi(\dot{d}(\ulcorner\varphi\urcorner^{\alpha})).$$

Hence, d_B given by $\varphi(x) \mapsto \dot{d}(\ulcorner \varphi(x) \urcorner^{\alpha})$ is a uniform diagonal operator with respect to α, standard numerals and $\mathcal{I}^a$ (see also Definition 4.1 and Example 3.5). Moreover, $\mathsf{Bew}^*(x)$ is a fixed-point of k_{d_B}, i.e.,

$$\begin{aligned} k_{d_B}(\mathsf{Bew}^*(x)) &\equiv x \neq \dot{d}(\ulcorner \mathsf{Bew}^*(x) \urcorner^{\alpha}) \wedge \mathsf{Bew}^a(x) \\ &\equiv x \neq \dot{d}(0) \wedge \mathsf{Bew}^a(x) \\ &\equiv \mathsf{Bew}^*(x). \end{aligned}$$

This completes our first proof of Lemma 4.2. Note that while our construction employs the standard numeral function, the numbering α is contrived. We now show that if we leave the numeral function unconstrained, we can construct a deviant provability predicate satisfying Lemma 4.2 for any given standard numbering.

B.2 Second Proof of Lemma 4.2

We base the second construction on Jeroslow's operator d_J introduced in Section 3.2. Once again, similar but slightly more complicated constructions can be given for d_B and d_G.

Let sub_J be f^1_n for some n such that F_n maps the #-codes of $\varphi(x)$ and $t(x)$ to the #-code of $\varphi(t(\ulcorner t(x) \urcorner^{\#})$. Let z be the #-code of the formula

$$x \neq \mathsf{sub}_J(f^0_0(0), \ulcorner \mathsf{sub}_J(f^0_0(0), x) \urcorner^{\#}).$$

We now define a function $G\colon \omega^2 \to \omega$ by setting $G(p, q) := \wedge_{\#}(z, H(q))$. By Theorem B.1, there is an index a such that $F_a(p) = G(p, a)$, for all $p \in \omega$. We now define the formula

$$\mathsf{Bew}^{\dagger}(x) :\equiv x \neq \mathsf{sub}_J(f^0_0(0), \ulcorner \mathsf{sub}_J(f^0_0(0), x) \urcorner^{\#}) \wedge \mathsf{Bew}^a(x).$$

Given the fixed-point property of a, we have

$$\mathcal{I}^a(f^0_0)(0) = \#\mathsf{Bew}^{\dagger}(x).$$

Hence, the mapping $\nu\colon \omega \to \mathrm{ClTerm}$ given by

$$\nu(n) :\equiv \begin{cases} f^0_0(0) & \text{if } n = \#\mathsf{Bew}^{\dagger}(x); \\ \overline{n} & \text{otherwise;} \end{cases}$$

is a numeral function (for $\mathcal{I}^a$). We moreover have that

$$\mathsf{Basic}(\mathcal{I}^a) \vdash \mathsf{sub}_J(\ulcorner \varphi(x) \urcorner^{\#,\nu}, \ulcorner t(x) \urcorner^{\#,\nu}) = \ulcorner \varphi(t(\ulcorner t(x) \urcorner^{\#,\nu})) \urcorner^{\#,\nu}.$$

Hence, for each $\varphi \in \mathrm{Fml}_x$

$$\mathsf{Basic}(\mathcal{I}^a) \vdash d_J(\varphi) = \mathsf{sub}_J(\ulcorner \varphi(x) \urcorner^{\#,\nu}, \ulcorner \mathsf{sub}_J(\ulcorner \varphi(x) \urcorner^{\#,\nu}, x)) = \ulcorner \varphi(d_J(\varphi)) \urcorner^{\#,\nu}.$$

In other words, d_J given by $\varphi(x) \mapsto \mathsf{sub}_J(\ulcorner \varphi(x) \urcorner^{\#,\nu}, \ulcorner \mathsf{sub}_J(\ulcorner \varphi(x) \urcorner^{\#,\nu}, x) \urcorner)$ is a uniform diagonal operator with respect to #, ν and $\mathcal{I}^a$.

Moreover, $\mathsf{Bew}^\dagger(x)$ is a fixed-point of k_{d_J}, i.e., we have

$$\begin{aligned} k_{d_J}(\mathsf{Bew}^\dagger(x)) &\equiv x \neq \mathsf{sub}_J(\ulcorner \mathsf{Bew}^\dagger(x)\urcorner^{\#,\nu}, \ulcorner \mathsf{sub}_J(\ulcorner \mathsf{Bew}^\dagger(x)\urcorner^{\#,\nu}, x)\urcorner^{\#,\nu}) \wedge \mathsf{Bew}^a(x) \\ &\equiv x \neq \mathsf{sub}_J(f_0^0(0), \ulcorner \mathsf{sub}_J(f_0^0(0), x)\urcorner^{\#}) \wedge \mathsf{Bew}^a(x) \\ &\equiv \mathsf{Bew}^\dagger(x). \end{aligned}$$

This completes our second proof of Lemma 4.2. As opposed to the contrived numbering α used in B.1, our second proof works for any given standard numbering. However, the uniform diagonal operator d_J is based on the contrived numeral function ν.

B.3 Proof of Lemma 7.1

Let sub be f_n^1 for some n such that F_n is the #-tracking function of the substitution function for formulas. Moreover, let num be f_n^0 for some n such that F_n maps every number to the #-code of its standard numeral. Finally, let $\dot{d}$ be f_n^0 for some n such that F_n maps the #-code of $\varphi(x, y)$ to the #-code of $\varphi(\dot{d}(\ulcorner\varphi(x, y)\urcorner^{\#}), y)$ (here we assume that our indexing of p.r. functions permits this construction). For each index a, consider the formula[20]

$$\varphi^a(x, y) :\equiv f_0^0(\mathsf{sub}(x, \mathsf{num}(y))) \neq \mathsf{sub}(x, \mathsf{num}(\mathsf{S}y)) \wedge \mathsf{Bew}^a(\mathsf{sub}(x, \mathsf{num}(y))).$$

Theorem B.1 yields an index a of a unary p.r. function mapping the #-code of $\mathsf{sub}(\dot{d}(\varphi^a(x, y)), \mathsf{num}(\overline{n}))$ to the #-code of $\mathsf{sub}(\dot{d}(\varphi^a(x, y)), \mathsf{num}(\overline{n+1}))$, for each $n \in \omega$. We now set $t :\equiv \dot{d}(\ulcorner\varphi^a(x, y)\urcorner^{\#})$. Hence, we have

$$\mathsf{Basic}(\mathcal{I}) \vdash t = \ulcorner\varphi^a(t, y)\urcorner^{\#}.$$

Therefore,

$$\mathsf{Basic}(\mathcal{I}) \vdash \forall y\ \mathsf{sub}(t, \mathsf{num}(y)) = \mathsf{sub}(\ulcorner\varphi^a(t, y)\urcorner^{\#}, \mathsf{num}(y)).$$

We now set $t_n :\equiv \mathsf{sub}(t, \mathsf{num}(\overline{n}))$, for each $n \in \omega$. Note that t_n is not contained in t_m, for $n \neq m$. We can then show in $\mathsf{Basic}(\mathcal{I}^a)$ that

$$\begin{aligned} t_n &= \mathsf{sub}(\ulcorner\varphi^a(t, y)\urcorner^{\#}, \mathsf{num}(\overline{n})) \\ &= \mathsf{sub}(\ulcorner f_0^0(\mathsf{sub}(t, \mathsf{num}(y))) \neq \mathsf{sub}(t, \mathsf{num}(\mathsf{S}y)) \wedge \mathsf{Bew}^a(\mathsf{sub}(t, \mathsf{num}(y)))\urcorner^{\#}, \\ &\quad \mathsf{num}(\overline{n})) \\ &= \ulcorner f_0^0(\mathsf{sub}(t, \mathsf{num}(\overline{n}))) \neq \mathsf{sub}(t, \mathsf{num}(\mathsf{S}\overline{n})) \wedge \mathsf{Bew}^a(\mathsf{sub}(t, \mathsf{num}(\overline{n})))\urcorner^{\#} \\ &= \ulcorner f_0^0(t_n) \neq t_{n+1} \wedge \mathsf{Bew}^a(t_n)\urcorner^{\#} \end{aligned}$$

For each $n \in \omega$, we define

$$\mathsf{Bew}_n(x) :\equiv f_0^0(x) \neq t_{n+1} \wedge \mathsf{Bew}^a(x).$$

We observe:

[20] This construction is inspired by Picollo's [20] method of obtaining ω-chains of sentences, each referring to its subsequent expression.

Fact B.3 For each $n \in \omega$,

(1) $\mathsf{Bew}_n(x)$ weakly represents provability in $\mathsf{Basic}(\mathcal{I}^a)$;
(2) $\mathsf{Basic}(\mathcal{I}^a) \vdash t_n = \ulcorner \mathsf{Bew}_n(t_n) \urcorner^{\#}$;
(3) $\mathsf{Basic}(\mathcal{I}^a) \vdash \neg \mathsf{Bew}_n(t_n)$.

Finally, $\mathsf{Bew}_n(x)$ does not contain t_n (at least for sensible choices of $\mathsf{Bew}^a(x)$). This completes our proof of Lemma 7.1.

B.4 Proof of Lemma 7.3

Let sub_J be given as in Section B.2. Let jump be f_n^0 for some n such that F_n maps the #-code of $\mathsf{sub}_J(f_0^0(\overline{m}), \ulcorner \mathsf{sub}_J(f_0^0(\overline{m}), x) \urcorner^{\#})$ to the #-code of $\mathsf{sub}_J(f_0^0(\overline{m+1}), \ulcorner \mathsf{sub}_J(f_0^0(\overline{m+1}), x) \urcorner^{\#})$, for $m \in \omega$. By Theorem B.1, there is an index a of a unary p.r. function which maps n to the #-code of

$$\mathsf{jump}(x) \neq \mathsf{sub}_J(f_0^0(\overline{n}), \ulcorner \mathsf{sub}_J(f_0^0(\overline{n}), x) \urcorner^{\#}) \wedge \mathsf{Bew}^a(x).$$

For each $n \in \omega$, we define the formula

$$\mathsf{Bew}_n(x) :\equiv \mathsf{jump}(x) \neq \mathsf{sub}_J(f_0^0(\overline{n+1}), \ulcorner \mathsf{sub}_J(f_0^0(\overline{n+1}), x) \urcorner^{\#}) \wedge \mathsf{Bew}^a(x).$$

Given the fixed-point property of a, we have

$$\mathcal{I}^a(f_0^0)(n) = \#\mathsf{Bew}_n(x).$$

Hence, the mapping $\nu \colon \omega \to \mathrm{ClTerm}$ given by

$$\nu(n) :\equiv \begin{cases} f_0^0(\overline{m}) & \text{if } n = \#\mathsf{Bew}_m(x); \\ \overline{n} & \text{otherwise;} \end{cases}$$

is a numeral function (for $\mathcal{I}^a$). We moreover have that

$$\mathsf{Basic}(\mathcal{I}^a) \vdash \mathsf{sub}_J(\ulcorner \varphi(x) \urcorner^{\#,\nu}, \ulcorner t(x) \urcorner^{\#,\nu}) = \ulcorner \varphi(t(\ulcorner t(x) \urcorner^{\#,\nu})) \urcorner^{\#,\nu}.$$

For each $n \in \omega$ we set $t_n :\equiv \mathsf{sub}_J(f_0^0(\overline{n}), \ulcorner \mathsf{sub}_J(f_0^0(\overline{n}), x) \urcorner^{\#,\nu})$. Recall that d_J is the uniform diagonal operator which maps any formula $\varphi(x)$ to the term

$$\mathsf{sub}_J(\ulcorner \varphi(x) \urcorner^{\#,\nu}, \ulcorner \mathsf{sub}_J(\ulcorner \varphi(x) \urcorner^{\#,\nu}, x) \urcorner^{\#,\nu}).$$

Fact B.4 We have $t_n \equiv d_J(\mathsf{Bew}_n(x))$, for each $n \in \omega$.

Proof We have

$$\begin{aligned} d_J(\mathsf{Bew}_n(x)) &\equiv \mathsf{sub}_J(\ulcorner \mathsf{Bew}_n(x) \urcorner^{\#,\nu}, \ulcorner \mathsf{sub}_J(\ulcorner \mathsf{Bew}_n(x) \urcorner^{\#,\nu}, x) \urcorner^{\#,\nu}) \\ &\equiv \mathsf{sub}_J(f_0^0(\overline{n}), \ulcorner \mathsf{sub}_J(f_0^0(\overline{n}), x) \urcorner^{\#,\nu}) \\ &\equiv t_n. \end{aligned}$$

□

We observe:

Fact B.5 For each $n \in \omega$,

(1) $\mathsf{Bew}_n(x)$ weakly represents provability in $\mathsf{Basic}(\mathcal{I}^a)$;
(2) $\mathsf{Basic}(\mathcal{I}^a) \vdash t_n = \ulcorner \mathsf{Bew}_n(t_n) \urcorner^{\#}$;
(3) $\mathsf{Basic}(\mathcal{I}^a) \vdash \neg \mathsf{Bew}_n(t_n)$.

Finally, we observe that

$$\cdots \lhd_*^{\#,\nu} t_2 \lhd_*^{\#,\nu} t_1 \lhd_*^{\#,\nu} t_0.$$

Thus, $\lhd_*^{\#,\nu}$ is ill-founded but irreflexive.

B.5 A Henkin Sentence Without an Additional Conjunct

We now construct a Henkin sentence which is not of the form $\chi(x) \wedge \mathsf{Bew}(x)$. Let sub_J be given as in Section B.2. We assume that for each index a of a unary function, $\mathsf{Bew}^a(x)$ also satisfies $\mathsf{Basic}(\mathcal{I}^a) \vdash \neg \mathsf{Bew}^a(\ulcorner \varphi \urcorner^{\#})$, for all non-formulas φ. Let H be a p.r. function which maps each such a to the #-code of $\mathsf{Bew}^a(x)$ (see Fact B.2). Let K be a p.r. function which maps each #-code of φ to the #-code of $\ulcorner \varphi \urcorner^{\#}$. Let $\mathrm{Sub}_{\#}$ denote the #-tracking function of the substitution function Sub_f defined in Section 3. Let $\overline{\mathsf{sub}_J}_{\#}$ denote the #-tracking function of the binary function $\overline{\mathsf{sub}_J}$ introduced in Section 3. Let the function $L\colon \omega \to \omega$ be given by

$$L(q) := \mathrm{Sub}_{\#}(H(q), \#f_0^0(x)).$$

We define $G\colon \omega^2 \to \omega$ by setting

$$G(p,q) := \begin{cases} \#0 & \text{if } p = \mathrm{Sub}_{\#}(L(q), \overline{\mathsf{sub}_J}_{\#}(K(L(q)), \overline{\mathsf{sub}_J}_{\#}(K(L(q)), \#x))); \\ p & \text{otherwise.} \end{cases}$$

Using Theorem B.1, there is an index a such that $F_a(p) = G(p,a)$, for all $p \in \omega$. We now define the formula

$$\mathsf{Bew}^{\diamond}(x) :\equiv \mathsf{Bew}^a(f_0^0(x)).$$

Given the fixed-point property of a, we have

$$\mathcal{I}^a(f_0^0)(\#\mathsf{Bew}^{\diamond}(d_J(\mathsf{Bew}^{\diamond}(x)))) = \#0.$$

We observe that application of Jeroslow's diagonal operator to the provability predicate $\mathsf{Bew}^{\diamond}(x)$ yields a refutable Henkin sentence:

Fact B.6

(1) $\mathsf{Bew}^{\diamond}(x)$ weakly represents provability in $\mathsf{Basic}(\mathcal{I}^a)$;
(2) $\mathsf{Basic}(\mathcal{I}^a) \vdash \neg \mathsf{Bew}^{\diamond}(d_J(\mathsf{Bew}^{\diamond}(x)))$.

To sum up, $d_J(\mathsf{Bew}^{\diamond}(x))$ is a refutable Henkin sentence, where d_J is a uniform diagonal operator and the underlying naming relation is given by a standard numbering and numeral function. Hence, in particular, the corresponding naming relation is well-founded.

References

1. Boolos, G. (1995). Quotational ambiguity. In P. Leonardi, & M. Santambrogio (Eds.) *On Quine: New Essays, pp. 283–296. Cambridge University Press*.
2. Cain, J., & Damnjanovic, Z. (1991). On the weak Kleene scheme in Kripke's theory of truth. *The Journal of Symbolic Logic*, *56*, 1452–1468.
3. Carnap, R. (1934). Logische Syntax der Sprache. Springer. Translated and reprinted in [4].
4. Carnap, R. (2001). Logical syntax of language. Routledge.
5. Feferman, S. (1960). Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, *49*, 35–92.
6. Fraassen, B. C. (1970). Inference and self-reference. *Synthese*, *21*(3-4), 425–438.
7. Grabmayr, B. (2021). On the invariance of Gödel's second theorem with regard to numberings. *Review of Symbolic Logic*, *14*(1), 51–84.
8. Grabmayr, B., & Visser, A. (2021). Self-reference upfront: a study of self-referential Gödel numberings. Review of Symbolic Logic, pp. 1–40. https://doi.org/10.1017/S1755020321000393.
9. Halbach, V., & Leigh, G. (2021). The road to paradox: a guide to syntax, truth, and modality. Cambridge University Press. To be published.
10. Halbach, V., & Visser, A. (2014a). Self-reference in arithmetic I. *Review of Symbolic Logic*, *7*(4), 671–691.
11. Halbach, V., & Visser, A. (2014b). Self-reference in arithmetic II. *Review of Symbolic Logic*, *7*(4), 671–691.
12. Heck, R. K. (2007). Self-reference and the languages of arithmetic. *Philosophia Mathematica*, *15*(1), 1–29. (originally published under the name "Richard G. Heck, Jr").
13. Henkin, L. (1952). A problem concerning provability. *Journal of Symbolic Logic*, *15*, 160.
14. Hinman, P. G. (1978). *Recursion-theoretic hierarchies*. Berlin: Springer.
15. Jeroslow, R. G. (1973). Redundancies in the Hilbert-Bernays derivability conditions for Gödel's second incompleteness theorem. *Journal of Symbolic Logic*, *38*(3), 359–367.
16. Kreisel, G. (1953). On a problem of Henkin's. *Indagationes Mathematicae*, *15*, 405–406.
17. Kripke, S. A. (1975). Outline of a theory of truth. *Journal of Philosophy*, *72*(19), 690–716.
18. Löb, M. H. (1955). Solution of a problem of Leon Henkin. *Journal of Symbolic Logic*, *20*(2), 115–118.
19. Odifreddi, P. (1989). *Classical recursion theory*. Amsterdam: North-Holland.
20. Picollo, L. (2018). Reference in arithmetic. *Review of Symbolic Logic*, *11*(3), 573–603.
21. Quine, W. V. (1940). Mathematical Logic. W. W. Norton, New York, 1 edition.
22. Schindler, T. (2015). *Type-free truth*. Ludwig Maximilians Universität München: PhD thesis.
23. Smoryński, C. (1985). Self-reference and modal logic. Springer.
24. Tarski, A. (1936). Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica*, *1*, 261–405. Reprinted as 'The Concept of Truth in Formalized Languages' in [25] pp. 152–278. Page references are given for the translation.
25. Tarski, A. (1956). *Logic, Semantics, Metamathematics*. Oxford: Clarendon Press.
26. Visser, A. (1989). Semantics and the liar paradox. In D. Gabbay, & F. Guenthner (Eds.) *Handbook of Philosophical Logic,* vol. IV, pp. 617–706. *Reidel, Dordrecht*.