The Henkin Sentence

Volker Halbach and Albert Visser

Abstract In this paper we discuss Henkin's question concerning a formula that has been described as expressing its own provability. We analyze Henkin's formulation of the question and the early responses by Kreisel and Löb and sketch how this discussion led to the development of provability logic. We argue that, in addition to that, the question has philosophical aspects that are still interesting.

Keywords Self-reference · Fixed Points · Second Incompleteness Theorem · Provability Logic

1 Henkin's Question

In the problem section of the Journal of Symbolic Logic, Vol. 17, No. 2 (June 1952), on p. 160, Leon Henkin posed the following problem

> *A problem concerning provability*. If Σ is any standard formal system adequate for recursive number theory, a formula (having a certain integer q as its Gödel number) can be constructed which expresses the proposition that the formula with Gödel number q is provable in Σ. Is this formula provable or independent in Σ? (Received February 28, 1952.)

There seems to be a certain *naiveté* about the question. There are many formal systems adequate for recursive number theory. It is not just that we can vary the axioms, but we can vary the proof system; we can even vary the setup of the syntax. Moreover, the construction of Henkin's sentence involves arithmetization, which can be implemented in any number of ways. We can choose different Gödel numberings, but even with a given Gödel numbering, we can encode the syntax in different ways as linear strings of symbols, trees, or still something else. Once all these decisions have been made, one needs to single out a formula for expressing provability. Once such a formula is fixed, one can think of many constructions for obtaining a sentence with code q that says that the sentence with Gödel number q is provable. Different choices at any stage will produce a different sentence. Why should we expect that all these formulae share the same status with respect to provability or independence?

Of course, one could insist that we simply fix one specific set of choices for all these "parameters," a set of choices that we intuitively recognize as being correct, straightforward or "natural." But nobody seriously thinks that there is just one admissible way of arithmetizing syntax, of picking a provability predicate, and so on. When contemplating arithmetization, we always understand that a chosen implementation is just one of the many ways, but that our results should be robust with respect to particular choices as long

M. Manzano et al. (eds.), *The Life and Work of Leon Henkin*, Studies in Universal Logic,
DOI 10.1007/978-3-319-09719-0_17

as they are "reasonable" or "natural" because we do not make use of "accidental" features of our choices. This is by no means trivial. But in the case of Henkin's problem and similar problems, it was not to be expected in 1952 that a solution would be robust.[1]

Henkin's way of posing the question differs significantly from those found in the more recent literature. In the modern literature, Henkin's question or Henkin's problem is usually described as the question *whether or not the sentence expressing its own provability is provable* [12, p. 148] or even as the question *whether the sentence expressing its own provability* [...] *is true or false, and provable or not* [37]. It is far from clear whether these abbreviated forms are adequate renderings of Henkin's original question or whether they capture Henkin's intention. Today *Henkin's Problem* is used more like a proper name for a family of logical questions and less as a description of the question asked by Henkin in 1952. Here we would like to take a closer look at what Henkin's did ask—and what he did not ask.

In contrast to many modern accounts, Henkin did not make use of the notion of self-reference in the formulation of his question. He did not describe the sentence as one that *says of itself that it is provable* or the like. The usual catchphrases like *self-reference* are strangely absent from Henkin's question and also from the immediate replies and discussions, even though Gödel had already described his own sentence as a sentence stating its own unprovability [19, p. 175].

In the proof of Gödel's First Incompleteness Theorem and many other results, the notion of self-reference is not needed; The Gödel sentence γ only needs to be a fixed point of nonprovability, that is, it must satisfy $\Sigma \vdash \gamma \leftrightarrow \neg \mathsf{Bew}(\ulcorner \gamma \urcorner)$; whether γ says something about itself and what it says is irrelevant for the proof. But Henkin also did not ask whether a fixed point of the provability predicate, that is, a sentence η with $\Sigma \vdash \eta \leftrightarrow \mathsf{Bew}(\ulcorner \eta \urcorner)$ is provable or not. Any provable formula such as $0 = 0$ clearly is a fixed point of any formula that may be called a provability predicate; and these trivial fixed points are clearly not what Henkin was after.

Henkin also did not ask whether the sentence obtained by applying a certain canonical diagonal construction to the provability predicate is provable or not. As we shall see in our discussion of Kreisel's answer, this is not equivalent to Henkin's requirement that the formula with Gödel number q should "express[es] the proposition that the formula with Gödel number q is provable," as this may be achievable without applying the standard Gödel diagonal construction to a given provability predicate.

Moreover, Henkin did not ask whether his sentence is *refutable*. He probably noted that if Σ refutes this sentence, then Σ ipso facto proves its nonprovability. This, in its turn, implies that Σ proves the consistency of Σ, thus contradicting the Second Incompleteness Theorem—if we assume that Σ is consistent. To make such reasoning valid, the provability predicate involved must have some of the properties usually ascribed to it.

Henkin employed intensional language in the question: the formula in question is supposed to *express* the provability of a formula with a certain Gödel number. Is this use

[1] One may compare this with the truth-teller sentence that states its own Σ_1-truth. The answer to the question whether this sentence is provable, refutable, or independent depends on assumptions on the coding, the diagonalization method, and so on [27]. So Henkin's question for Σ_1-truth instead of provability only admits an answer that is far less robust than Löb's answer to Henkin's original question, which is extremely robust. Among the "Henkin-like" problems, the robustness of the answer to Henkin's original problem may be more the exception than the rule.

of intensionality to be viewed as merely a *façon de parler* or does it carry some serious weight? As we will see, Henkin's review of Kreisel's paper contains some evidence that Henkin did indeed take that business of *expressing something* seriously.

One should view Henkin's question as including the challenge to give a definite mathematical extensional meaning to Henkin's intensional description of his sentence. Kreisel and others took up the challenge by breaking it into two problems, once a formal system Σ and a coding are fixed:[2] First one needs to provide conditions that must be satisfied by a formula to express provability. Then, in the second step, from that formula a sentence with code q that ascribes to q the property of being provable must be constructed.

As we will see, both Kreisel and Löb developed criteria, albeit entirely different ones, that must be satisfied by a formula to qualify as a provability predicate. Kreisel argued that the answer to Henkin's question depends on which provability is used and that different provability predicates can be employed to obtain provable and even refutable Henkin sentences. Hence, the burden on his conditions for expressing provability was heavier. The provability predicates used in his examples need to be recognized as correct arithmetizations of provability. As we shall see, neither Henkin nor Löb agreed that his examples were good examples of arithmetizations of provability.

Löb's answer was positive: Henkin's sentence is provable; and his answer was definitive: all sentences of this kind are provable. Consequently, Löb needed only necessary conditions on formulae for expressing provability. The set of formulae satisfying his conditions only needs to *include* all predicates that we would recognize as good arithmetizations of provability. It is perfectly all right if the conditions admit cases that we would not recognize as good arithmetizations (as in fact they do). The same applies to the second step. Löb's result holds for all diagonal sentences, that is, sentences η satisfying $\Sigma \vdash \eta \leftrightarrow \mathsf{Bew}(\ulcorner \eta \urcorner)$ for the chosen provability predicate Bew; and all formulae with code q that express that the formula with q via the predicate Bew will be diagonal sentences. That there are also other diagonal sentences merely shows that Löb's result is stronger than needed to answer Henkin's question.

Henkin published his question in 1952. This is 21 years after the appearance of Gödel's paper *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I* [19] and 13 years after the first publication of part II of *Grundlagen der Mathematik* by Hilbert and Bernays [23], which contained a careful analysis of the proof of the Second Incompleteness Theorem. It seems to us that all ingredients for Henkin's question were already present in Gödel's 1931 paper. Even if you want to insist that a careful analysis of the proof of Gödel's second theorem is needed as the background for the question, then these were certainly available in the 1939 book by Hilbert and Bernays. So, why was the question not asked earlier? A first point is that one should never underestimate the size of the search space even for elementary questions. After all, the other question (received 19 March 1952) asked by Henkin in the same list of questions is whether the Ordering Theorem is equivalent to the Axiom of Choice for classes of finite sets. This other question also only involves concepts that were around for quite some time. Secondly, the former question clearly was off the beaten track. It was not part of existing research lines. It is an almost whimsical question posed in a playful mood. The question gives the reader a feeling of contingency. It could as well never have been posed. Seeing all the exciting developments that followed it, we may be glad that it actually was asked.

[2] In what follows, we somewhat neglect the problems involving the choice of the formal system Σ and a coding of syntax. See [27] for some additional remarks.

2 Kreisel's Solution

In his 1953 paper [31], Georg Kreisel summarized his reply to Henkin as follows:

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

Thus, Kreisel's reply precisely concerns the point where we said that Henkin's question had a certain *naiveté*. Note the scare quotes around *express*, which suggest that Kreisel thought that the business of expressing should be taken *cum grano salis*. Kreisel gives the following condition for *expressing provability*:

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

We can generalize Kreisel's condition for provability to a more general condition for expressing a property.

Kreisel's Condition A formula $\varphi(x)$ is said to express a property P in Σ if and only if, for all numbers n, we have $\Sigma \vdash \varphi(\underline{n})$ iff n has property P.

In metamathematics, Kreisel's Condition became the formal notion of weak representability:[3] A formula $\varphi(x)$ is said to *weakly represent* a set S of numbers if and only if $\Sigma \vdash \varphi(\underline{n})$ iff $n \in S$.

The main problem with Kreisel's Condition is that it is counterintuitive on two scores. First, according to it, for example, over Peano Arithmetic (PA), many predicates express provability that intuitively do not. Examples of such predicates are Rosser provability, Feferman provability, and the two notions of Kreisel–Henkin provability discussed below. Secondly, consider some predicate Bew that we recognize as a good arithmetization of provability in, say, $\Theta := \mathsf{PA} + \mathsf{incon}(\mathsf{PA})$. Then, by Kreisel's condition, Bew does *not* express provability-in-Θ. However, by the Friedman–Goldfarb–Harrington Theorem, we may manufacture, by Rosser trickery, a predicate that *does* express provability in Θ by the light of the Kreisel Condition.[4] With some more effort, we may even build such a predicate satisfying the Löb Conditions too.

We note that for Kreisel's *it depends* answer, it is needed that all predicates admitted by his Condition are indeed recognized as expressing provability. It does no harm when some predicates not recognized by the condition do express provability. Thus, Kreisel's condition must be a sufficient condition. Conversely, for Löb, with his positive answer, one wants that whatever is recognized as expressing provability should be admitted by Löb's condition. It does no harm when certain predicates admitted by Löb's condition are recognized as not representing provability. For instance, over PA, the predicate $x = x$ satisfies the Löb conditions. Nobody would think that this predicate expresses provability. But, also, nobody would think this is a problem for Löb's answer to Henkin's question. Thus, Löb's condition needs to be necessary.

[3] Feferman [17] introduced and used the term "numerate" for "weakly represent."

[4] See, for instance, [51] for a discussion.

Kreisel constructed two sentences that are both supposed to satisfy Henkin's Condition; one of them is provable, the other refutable. Let Basic be the Tarski–Mostowski–Robinson theory R extended by the recursion equations for all primitive recursive functions.

Kreisel's Observation Let Σ be a consistent theory that extends Basic.[5] Then the following hold:

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

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

The examples employed by Kreisel in the proof are of some interest. In particular, the example for $\mathsf{Bew}_I(t_1)$ foreshadows Kreisel's [32] proof of Löb's theorem, as was pointed out by [44]. Henkin suggested simpler examples that are mentioned by [31] in footnotes. We will use Henkin's examples and refer the reader to Smoryński's paper for an exposition of Kreisel's original examples.

Proof We start with a proof for the second part (b). Fix some predicate $\mathsf{Bew}(x)$ that weakly represents Σ-provability in Σ. In case Σ is Σ_1-sound, a standard arithmetization of provability will do. In the unsound case, one uses the theorem that any recursively enumerable set is weakly representable in a consistent recursively enumerable extension of the Tarski–Mostowski–Robinson theory R. This is a direct consequence of the Friedman–Goldfarb–Harrington Theorem.[6] Using the canonical diagonal construction (or any other method), one obtains a term t_2 satisfying the condition

$$\Sigma \vdash t_2 = \ulcorner t_2 \neq t_2 \wedge \mathsf{Bew}(t_2) \urcorner \tag{1}$$

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

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

Condition b(ii), that is, $\Sigma \vdash t_2 = \ulcorner \mathsf{Bew}_{II}(t_2) \urcorner$, is then obviously satisfied by the choice (1) of t_2. Since Σ refutes $t_2 \neq t_2 \wedge \mathsf{Bew}(t_2)$, item b(iii) is satisfied as well.

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

$$\Sigma \vdash \varphi \quad \text{iff} \quad \Sigma \vdash \mathsf{Bew}_{II}\big(\ulcorner \varphi \urcorner\big). \tag{2}$$

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

[6] See, for instance, [51] for a discussion.

If φ is different from $t_2 \neq t_2 \wedge \mathsf{Bew}(t_2)$, then this is obvious from the definition of $\mathsf{Bew}_{II}(x)$, using the fact that Bew weakly represents provability in Σ. In the other case, the left-hand side of the equivalence is refutable, and so is the right-hand side by (1). This concludes the proof of part (b) of Kreisel's Observation.

We turn to case (a). If we assume that our theory is Σ_1-sound and sufficiently strong (e.g., if it extends the arithmetical version of Buss' theory S^1_2), then the canonical provability predicate can be used as $\mathsf{Bew}_I(x)$, and t_1 can be obtained in any way, including the usual Gödel diagonal construction. Claim a(iii) follows then by Löb's theorem. (See [35] or, e.g., [12].)

Since Löb's Theorem was not known, Henkin and Kreisel had to use a different construction.[7] Henkin suggested the following construction. He picked a term t_1 such that

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

and defines $\mathsf{Bew}_I(x)$ as

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

Clearly, the provability predicates Bew_I and Bew_{II} are somewhat peculiar. Although they satisfy Kreisel's Condition, hardly anyone considers them to be proper provability predicates. As we shall see soon, Henkin was the first to reject them and claim that the sentences $\mathsf{Bew}_I(t_1)$ and $\mathsf{Bew}_{II}(t_2)$ do not fit the description in his question.

However, the alleged Henkin sentences $\mathsf{Bew}_I(t_1)$ and $\mathsf{Bew}_{II}(t_2)$ exhibit another peculiarity that is neither discussed by Kreisel nor by Henkin:[8] They are not obtained by applying the usual diagonal construction to the respective provability predicates Bew_I and Bew_{II}. Rather Kreisel finessed the predicates Bew_{II} in such a way that simply substituting the term t_2 for the free variable in Bew_{II} produces a formula with Gödel number q such that the value of t_2 is q. So one can reasonably claim that $\mathsf{Bew}_{II}(t_2)$ is "a formula (having a certain integer q as its Gödel number) [...] which expresses the proposition that the formula with Gödel number q is provable in Σ" [24] if Bew_{II} is taken to express provability. Similar remarks apply to $\mathsf{Bew}_I(t_1)$ of course. So, with the possible exception of the choice of the provability predicates, Kreisel provided a correct answer to Henkin's question.

However, one may wonder whether Kreisel answered the questions that are currently called *Henkin's Problem*. In other words, is $\mathsf{Bew}_{II}(t_2)$ self-referential and does it state its own provability? In particular, does $\mathsf{Bew}_{II}(t_2)$ ascribe to itself the property expressed by $\mathsf{Bew}_{II}(x)$—whether it is a good provability predicate or not? Usually, when one considers a sentence that says about itself that it has the property expressed by a formula $\psi(x)$, one often intends to talk about the sentence that is obtained from $\psi(x)$ by the usual diagonal construction or a variant thereof. What exactly the usual diagonal construction and its variant are may be unclear, but $\mathsf{Bew}_{II}(t_2)$ has not been obtained by anything that resembles such a method.

This sheds a light on the usual reformulation of Henkin's problem: It is often stated as a problem about a formula that states its own provability or that says about itself that it

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

[8] It was first noted by Craig Smoryński in [42].

is provable. Of course, one may speculate that Henkin intended to ask his question about this formula and Kreisel tried to address the question understood in this sense. It also seems that later authors understood Henkin's question as being about sentences that state their own provability. But the equivalence to the original formulation is not obvious.

At any rate, if the usual diagonal construction involving the substitution function is applied to Bew_{II}, one obtains a *provable* sentence. This follows easily from Löb's theorem, which of course was not known at the time Kreisel published his paper. If that sentence is seen as the only sentence saying about itself that it has the property expressed by Bew_{II}, then Kreisel fails to provide a counterexample to the claim that the sentence stating its own provability is provable—irrespective of whether Bew_{II} is a provability predicate or not. So Kreisel was somewhat imprecise in summarizing his result: He had shown 'that the answer to Henkin's question depends on which formula is used to "express" the notion of *provability in* Σ'—but also on how the formula with code q is obtained that ascribes provability to q via this provability predicate.

However, after all, it can be shown that, if Kreisel's Condition is adopted, it *only* depends which provability predicate is chosen whether the Henkin sentence is provable or not. We can even use the standard diagonal construction to obtain a refutable Henkin sentence from the given provability predicate. For in [27] it has been shown that there still another provability that yields a refutable Henkin sentence if the standard diagonal construction is applied to it. Such a provability predicate can be obtained by tinkering with the Kreisel–Henkin construction.

3 Henkin's Review

In 1954 Leon Henkin responds to Kreisel's paper in a review [25] in the Journal of Symbolic Logic. Henkin's main critical point is the following.

> A clear explication of the concept of *that which is expressed by a formula* must be based on an axiomatic treatment of this notion (perhaps along the lines of Church XVII 133). However, it seems fair to say that in one sense, at least, neither formula $P_1(a)$ nor $P_2(a)$ expresses the propositional function *a is provable*; but the former, for example, expresses the proposition *a is provable or is equal to q*, which is a different proposition even though it has the same extension. The direct way to express *a is provable* is, of course, by the formula $(\exists x)B(x, a)$. But the methods of the present paper give no indication as to whether the formula $(\exists x)B(x, q)$ whose Gödel number is denoted by q is provable.

The reference Church XVII 133 is to a review in JSL by Rulon Wells [53] of Church's paper *A formulation of the logic of sense and denotation* [15]. Regrettably, the desired axiomatic explication of *that which is expressed by a formula* never materialized. The remark about *expressing* underscores the fact that Henkin took the philosophical problem of intensionality quite seriously—no scare quotes for him. The subsequent remarks about P_1 and P_2 show that Henkin rejected Kreisel's Condition as a sufficient condition for *expressing provability*.

Finally, Henkin insisted that Kreisel did not solve the problem for the intended predicate $(\exists x)B(x, q)$, where $B(x, y)$ is, as Henkin put it in [25], the "standard formula such that $B(m, n)$ or its negation is provable according as m does or does not denote the number of a formal proof of a formula whose Gödel number is denoted by n." It is not completely clear what the standard formula is, given that Henkin did not fix a formal system Σ;

but for systems like PA, the standard formulas can be thought of as those found in the literature.

Henkin's insistence on a less contrived provability predicate is at least consistent with the fact that he asked in his original 1952 question whether his sentence is provable or independent. As remarked above, he may have reasoned that the refutability of the Henkin sentence would imply consistency contradicting Gödel's Second Incompleteness Theorem. But this applies only if the consistency statement and the Henkin sentence are formulated with a well-behaved provability predicate. The Second Incompleteness Theorem fails for the Rosser provability predicate, for instance. For Kreisel's predicate Bew_{II}, the second incompleteness theorem holds; after all, it agrees with the standard one on all sentences except for $\neg\mathsf{Bew}_{II}(t_2)$. Kreisel's provability predicate Bew_{II}, however, does not satisfy Löb's second derivability condition LC2 below. Of course, Henkin had published his question and the review of Kreisel's reply before Löb's derivability conditions were formulated, so Henkin could resort only to intensional properties of the provability predicates and what they do or do not "express."

At any rate concentration on the "standard" provability predicate led to the breakthrough and the commonly accepted answer to Henkin's question.

4 Löb's Paper

We are again one year later. In his celebrated JSL paper [35], Martin Löb starts by echoing Henkin's review:

> One approach to this problem is discussed by Kreisel in [4]. However, he still leaves open the question whether the formula $(\exists x)\mathfrak{B}(x, a)$, with Gödel number a, is provable or not. Here $\mathfrak{B}(x, y)$ is the number-theoretic predicate which expresses the proposition that x is the number of a formal proof of the formula with Gödel-number y.

So we see that Löb adhered to Henkin's intensional phrasing of the question.

Let us write $\vdash$ for $\Sigma \vdash$ and $\Box\varphi$ for $\mathsf{Bew}(\ulcorner\varphi\urcorner)$. Then, we can state Löb's conditions like this:[9]

LC1 $\vdash \varphi \;\Rightarrow\; \vdash \Box\varphi$.
LC2 $\vdash \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$.
LC3 $\vdash \Box\varphi \to \Box\Box\varphi$.

Löb derives what is known as Löb's Rule from his conditions. We have: if $\vdash \Box\varphi \to \varphi$, then $\vdash \varphi$.

The reasoning is as follows. Suppose (a) $\vdash \Box\varphi \to \varphi$. By Gödel's Fixed Point Lemma, we can find a sentence λ such that (b) $\vdash \lambda \leftrightarrow (\Box\lambda \to \varphi)$. Now reason in Σ. Suppose (c) $\Box\lambda$. Then, by LC3, (d) $\Box\Box\lambda$. By (b), (c), LC1, and LC2, we find: (e) $\Box(\Box\lambda \to \varphi)$. Combining (d) and (e) using LC2, we may conclude $\Box\varphi$, and, hence, by (a): φ. Thus, by canceling assumption (c), we have found (f) $\Box\lambda \to \varphi$. By (b) and (f), we have (g) λ. We have derived λ without assumptions, hence, by LC1, (h) $\Box\lambda$. Combining this with (f), we may conclude φ.

[9] Actually, Löb mentions more conditions in his paper. However, upon analysis, we only need the ones given here.

Löb's solution of Henkin's Problem now follows immediately. Suppose $\vdash \eta \leftrightarrow \Box\eta$. Then, a fortiori, $\vdash \Box\eta \rightarrow \eta$, and, hence, by Löb's Rule, $\vdash \eta$.

In footnote 2, Löb states:

> In a previous version of this note the method of proof was applied specifically to Henkin's problem. The present more general formulation of our result was suggested by the referee.

Albert Visser asked George Kreisel who he thought was the referee of Löb's paper. Kreisel answered that, of course, this must have been Henkin. Later Albert Visser asked Henkin whether he was the referee, and Henkin confirmed that this was indeed the case.

Löb's Principle is the formalized form of Löb's Rule: $\vdash \Box(\Box\varphi \rightarrow \varphi) \rightarrow \Box\varphi$. We can derive Löb's Principle by formalizing the reasoning leading to Löb's Rule. However, we can also derive Löb's Principle from Löb's Rule. Reason in Σ. We suppose that (i) $\Box(\Box(\Box\varphi \rightarrow \varphi) \rightarrow \Box\varphi)$ and (ii) $\Box(\Box\varphi \rightarrow \varphi)$. From (ii) and LC3, we have (iii) $\Box\Box(\Box\varphi \rightarrow \varphi)$. Combining (i) and (iii) using LC2, we find (iv) $\Box\Box\varphi$. From (ii) and (iv), using LC2, we get: (v) $\Box\varphi$. By canceling assumptions (ii) and (i), we find:

$$\text{(vi)} \quad \Box\big(\Box(\Box\varphi \rightarrow \varphi) \rightarrow \Box\varphi\big) \rightarrow \big(\Box(\Box\varphi \rightarrow \varphi) \rightarrow \Box\varphi\big).$$

Then, using Löb's Rule, we find: $\Box(\Box\varphi \rightarrow \varphi) \rightarrow \Box\varphi$, as desired.

Conversely, both Löb's Rule and LC3 follow from Löb's Principle using only LC1 and LC2.

To derive Löb's Rule from Löb's Principle, suppose that (A) $\vdash \Box\varphi \rightarrow \varphi$. Then, by LC1, we find (B) $\vdash \Box(\Box\varphi \rightarrow \varphi)$. Löb's Principle then gives us (C) $\vdash \Box\varphi$. By (A) and (C), we have $\vdash \varphi$.

The derivation of LC3 from Löb's Principle is due to Dick de Jongh. It works as follows. Reason in Σ. Suppose (α) $\Box\varphi$. Then, we find (β) $\Box(\Box(\varphi \wedge \Box\varphi) \rightarrow (\varphi \wedge \Box\varphi))$, by LC1 and LC2. Hence, by Löb's Principle, (γ) $\Box(\varphi \wedge \Box\varphi)$. We may conclude by LC1 and LC2 that $\Box\Box\varphi$.

What do Löb's answer to Henkin's question and Gödel's Second Incompleteness Theorem have in common? Löb showed that any fixed point of $\mathsf{Bew}(x)$ is provably equivalent to $0 = 0$. Gödel showed that any fixed point of $\neg\,\mathsf{Bew}(x)$ is provably equivalent to $\mathsf{con}(\Sigma)$. Thus, both fixed point equations have, modulo provable equivalence, a unique solution with a self-reference free formulation. As we will see in the next section, the proper framework to formulate, study, and generalize this insight is *Provability Logic*.

5 Provability Logic

The first to note the possibility of reading formal provability in a theory as a modal operator was Kurt Gödel in his paper [20]. The main result of the paper is the translation of intuitionistic propositional logic IPC in the modal system S4. At the end of the paper, Gödel remarks:

> Es ist zu bemerken, daß für den Begriff "beweisbar in einem bestimmten formalen System S" die aus $\mathfrak{S}$ beweisbaren Formeln nicht alle gelten. Es gilt z.B. für ihn $B(Bp \rightarrow p)$ niemals, d.h. für kein System S, das die Arithmetik enthalt. Denn andernfalls wäre beispielsweise $B(0 \neq 0) \rightarrow 0 \neq 0$ und daher auch $\sim B(0 \neq 0)$ in S beweisbar, d.h. die Widerspruchsfreiheit von S wäre in S beweisbar.

Here $\mathfrak{S}$ is S4. In the English translation of the Collected Works, Gödel's text becomes:

> It is to be noted that for the notion "provable in a certain formal system S" not all of the formulas provable in $\mathfrak{S}$ hold. For example, $B(Bp \to p)$ never holds for that notion, that is, it holds for no system S that contains arithmetic. For, otherwise, for example, $B(0 \neq 0) \to 0 \neq 0$ and therefore also $\sim B(0 \neq 0)$ would be provable in S, that is, the consistency of S would be provable in S.

In this paper, we will just look at that part of provability logic that is directly connected to Henkin's question: the study of fixed points. For a treatment of Gödel's remarks that is closely connected to Provability Logic, see [4].

Löb's logic GL is a modal propositional logic that has, in addition to the axioms and rules of propositional logic, the following principles:[10]

L1 $\vdash \varphi \;\Rightarrow\; \vdash \Box\varphi$.
L2 $\vdash \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$.
L3 $\vdash \Box\varphi \to \Box\Box\varphi$.
L4 $\vdash \Box(\Box\varphi \to \varphi) \to \Box\varphi$.

The principles of GL correspond to the Löb Conditions plus Löb's Theorem. Thus, these principles are schematically valid for arithmetical interpretations where the $\Box$ is interpreted by Bew. We note that by the remarks of Sect. 4 the axiom L3 is superfluous. We can prove that the following strengthened version of Löb's Rule is admissible over GL. Let $\boxdot\chi$ stand for $\chi \wedge \Box\chi$. We have:

$$\text{SLR} \quad \text{if } \Box\psi_0, \ldots, \Box\psi_{n-1}, \boxdot\chi_0, \ldots, \boxdot\chi_{k-1}, \Box\varphi \vdash \varphi, \qquad \text{then}$$
$$\Box\psi_0, \ldots, \Box\psi_{n-1}, \boxdot\chi_0, \ldots, \boxdot\chi_{k-1} \vdash \varphi$$

We have seen in Sect. 4 that both Gödel's fixed point equation and Henkin's fixed point equation have unique self-reference free solutions. Provability Logic gives us the proper context both to formulate and to generalize these results.

Let us say that φ is *modalized* in p if all occurrences of p in φ are in the scope of $\Box$. A first observation is that, if φ is modalized in p, then φ has a unique fixed point (modulo provable equivalence) w.r.t. p. The uniqueness of fixed points was proved independently by Dick de Jongh (unpublished), Giovanni Sambin [39] and Claudio Bernardi in 1974 [10].

Let φp be modalized in p, and let q be a fresh propositional variable. Then we have:

$$\text{GL} \vdash \big(\boxdot(p \leftrightarrow \psi p) \wedge \boxdot(q \leftrightarrow \psi q)\big) \to (p \leftrightarrow q).$$

To see this, reason in GL. Suppose (a) $\boxdot(p \leftrightarrow \psi p)$, (b) $\boxdot(q \leftrightarrow \psi q)$, and (c) $\Box(p \leftrightarrow q)$. Since φp is modalized in p and φq is modalized in q, we find from (c): (d) $\varphi p \leftrightarrow \varphi q$.[11] Ergo, by (a) and (b), (e) $p \leftrightarrow q$. By SLR, we may conclude $p \leftrightarrow q$ without assumption (c).

We turn to the existence of explicit fixed points in the modal language. Suppose φp is modalized in p. Then there is a formula χ, where the free variables of χ are included in the free variables of φ minus p, such that $\text{GL} \vdash \chi \leftrightarrow \varphi\chi$.

[10] It would be more appropriate to call this logic simply L. Unfortunately, L also suggests *language*, so the designation GL was preferred.

[11] The substitution principle used here can be proved by induction of φ.

As is proper for great results, the theorem has many proofs. The existence of explicit fixed points was first proved by Dick de Jongh in 1974 (unpublished). Dick provided both a semantical and a syntactical proof. In 1976 another proof was given by Giovanni Sambin [39]. Also in 1976, George Boolos found a proof of explicit definability using characteristic formulas. In 1978, Craig Smoryński [41] proved explicit definability via Beth's Theorem.[12] There is an improved version of Sambin's approach by Giovanni Sambin and Silvio Valentini [46] in 1982 and an improved version of Boolos' proof by Zachary Gleit and Warren Goldfarb [18] in 1990. Finally, in 1990 there is a proof by Lisa Reidhaar-Olson [38] that is close to the proof of Sambin–Valentini. In 2009, Luca Alberucci and Alessandro Facchini [3] provide a proof using the modal μ-calculus.

We give the proof that is due to Craig Smoryński. In the proof we will assume the interpolation theorem for GL that can be proved both by semantic methods and by proof-theoretical methods. We assume that φp is modalized in p and that q does not occur in φp. We note that the uniqueness theorem gives us:

$$\begin{aligned}\Box(p \leftrightarrow \varphi p) \wedge \Box(q \leftrightarrow \varphi q) &\vdash_{\mathsf{GL}} \Box\boxdot(p \leftrightarrow \varphi p) \wedge \Box\boxdot(q \leftrightarrow \varphi q)\\ &\vdash_{\mathsf{GL}} \Box(p \leftrightarrow q)\\ &\vdash_{\mathsf{GL}} \varphi p \leftrightarrow \varphi q\end{aligned}$$

It follows that

$$\Box(p \leftrightarrow \varphi p) \wedge \varphi p \vdash_{\mathsf{GL}} \Box(q \leftrightarrow \varphi q) \rightarrow \varphi q.$$

Let χ be an interpolant between $\Box(p \leftrightarrow \varphi p) \wedge \varphi p$ and $\Box(q \leftrightarrow \varphi q) \rightarrow \varphi q$. By substituting χ for p and q we obtain:

$$\Box(\chi \leftrightarrow \varphi\chi) \wedge \varphi\chi \vdash_{\mathsf{GL}} \chi \vdash_{\mathsf{GL}} \Box(\chi \leftrightarrow \varphi\chi) \rightarrow \varphi\chi.$$

We may conclude that $\mathsf{GL} \vdash \Box(\chi \leftrightarrow \varphi\chi) \rightarrow (\chi \leftrightarrow \varphi\chi)$ and, hence, $\mathsf{GL} \vdash \chi \leftrightarrow \varphi\chi$.

The story of the fixed points is not finished here. First, the uniqueness and explicit definability results extend to interpretability logic as was shown in [16]. A Smoryński-style proof of this result is provided in [2]. See also [26]. Secondly, the fixed points of provability logic connect it to that other great modal logic of fixed points, the modal μ-calculus. See [48, 52], and [3]. See also Giacomo Lenzi's survey [33] of results concerning the μ-calculus.

The world looks different when the Löb conditions fail. The uniqueness/non-uniqueness of the Rosser fixed points was studied by David Guaspari and Robert Solovay [22]. Their answer is a laconic *it depends*. Interestingly, Kreisel never found their work convincing. His experience with his own *it depends* led him to hope for a missing natural condition ....

What happens in case of the Feferman provability predicate was studied by Albert Visser [50], by Craig Smoryński [43], and by Volodya Shavrukov [40]. Visser shows that there are infinitely many pair-wise nonequivalent Henkin fixed points for Feferman provability. Smoryński shows that under rather natural assumptions the Gödel fixed point

[12] In her paper [36], Larisa Maksimova shows that, conversely, Beth's theorem follows from the existence of explicit fixed points. See also [26].

for the Feferman predicate is unique. Shavrukov gives a beautiful modal derivation of the same result.

Of course, there is much more to provability logic than the treatment of fixed points. For example, there is Solovay's celebrated arithmetical completeness theorem ([45]). However, this further story is outside of the scope of this paper. We just provide some pointers to the further literature.

The history of the mathematical modality *provability in a certain formal system* has been described in the paper [13]. This paper is warmly recommended to the reader. The systematic content of provability logic can be found in the textbooks [12] and [42]. It is worth looking at both books since they offer a somewhat different perspective. For more recent treatments, containing also new material, see also [1, 14, 30, 34, 47].

We have seen that from Henkin's question and Löb's work, the field of Provability Logic emerged. Provability Logic, apart from being a beautiful subject, has some application outside of its own domain.

- Michael Beeson employed fp-realizability, a form of realizability based on provability to prove the independence of the Myhill–Shepherdson theorem and of the Kreisel–Lacombe–Schoenfield theorem from Heyting arithmetic. See [5] and [49]. Beeson's result uses Löb's principle.
- Research on the Provability Logic of Heyting Arithmetic inspired an axiomatization of the admissible rules of the intuitionistic propositional calculus. See [28].
- S.N. Goryachev connects in his work [21] Provability Logic and reflection principles. His results are used by Lev Beklemishev for the analysis of reflection principles. See [6] and [8].
- Japaridze's polymodal logic [29] is used by Lev Beklemishev to study ordinal notations. See [7, 9, 11].

6 Concluding Remarks

The question of what it means for a formula to *express* a property like provability has fallen from grace since the success of Löb's work. First, the question seems rather hopeless, and second, Löb showed that, at least for some important results, we can successfully get by without an answer to the question. Does this mean that the question has for once and for all been laid to rest? We do not think so. First, even if, perhaps, the question is mathematically less important, then it is still relevant philosophically. If we say, for example, that Peano arithmetic does not prove its own consistency, is this merely a *façon de parler*—to be paraphrased away by a more mathematical pronouncement, or does it really mean what it seems to mean?

Can we tell a better story about arithmetical *ventriloquism*? Such a story should at least take into account that content ascriptions like *the formula* Bew *expresses provability in* Σ are heavily contextual. For instance, the ascription only makes sense against the background of some Gödel numbering. Perhaps, we need, as Henkin thought, to have a theory of content as a *prolegomemon*. But, maybe, the task is rather to describe on the basis of our everyday understanding of how formulas express properties, to explain how formulas that are obviously about something else (like numbers), still manage to express properties of sentences against the background of conventional choices relating, for example, numbers and formulas.

We think that there is reason to have hope for progress. In a sense, we have all the needed information concerning what is going on. After all, the good cases where we think that we really construct a formula Bew expressing provability are open for detailed inspection. *All there is, is here.* Also we have lots of deviant examples where we could have doubts like the nonstandard Gödel numberings with in-built self-reference. What is lacking is an articulated analysis bringing to the fore what is good and what is bad.

As we have seen, Henkin's playful question led to the development of Provability Logic. Moreover, it touches immediately upon philosophical questions concerning intensionality in mathematics. Voltaire said "Il est encore plus facile de juger de l'esprit d'un homme par ses questions que par ses réponses" (it is easier to judge a man by his questions than by his answers). Clearly, Leon Henkin is doing very well on Voltaire's criterion.

Acknowledgements We thank Volodya Shavrukov for his comments on the penultimate version.

References

1. Artemov, S.N., Beklemishev, L.D.: Provability logic. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, 2nd edn. vol. 13, pp. 229–403. Springer, Dordrecht (2004)
2. Areces, C., de Jongh, D., Hoogland, E.: The interpolation theorem for IL and ILP. In: Proceedings of AiML98. Advances in Modal Logic, Uppsala, Sweden (1998)
3. Alberucci, L., Facchini, A.: On modal μ-calculus and Gödel–Löb logic. Stud. Log. **91**(2), 145–169 (2009)
4. Artemov, S.: Logic of proofs. Ann. Pure Appl. Log. **67**(1), 29–59 (1994)
5. Beeson, M.: The nonderivability in intuitionistic formal systems of theorems on the continuity of effective operations. J. Symb. Log. **40**, 321–346 (1975)
6. Beklemishev, L.D.: Proof-theoretic analysis by iterated reflection. Archive **42**, 515–552 (2003). doi:10.1007/s00153-002-0158-7
7. Beklemishev, L.D.: Provability algebras and proof-theoretic ordinals. Ann. Pure Appl. Log. **128**, 103–124 (2004)
8. Beklemishev, L.D.: Reflection principles and provability algebras in formal arithmetic. Russ. Math. Surv. **60**(2), 197–268 (2005)
9. Beklemishev, L.D.: The Worm principle. Logic Colloquium'02. Lect. Notes Log. **27**, 75–95 (2006)
10. Bernardi, C.: The uniqueness of the fixed-point in every diagonalizable algebra. Stud. Log. **35**(4), 335–343 (1976)
11. Beklemishev, L.D., Joosten, J.J., Vervoort, M.: A finitary treatment of the closed fragment of Japaridze's provability logic. J. Log. Comput. **15**(4), 447–463 (2005)
12. Boolos, G.: The Logic of Provability. Cambridge University Press, Cambridge (1993)
13. Boolos, G., Sambin, G.: Provability: the emergence of a mathematical modality. Stud. Log. **50**, 1–23 (1991)
14. Beklemishev, L.D., Visser, A.: Problems in the logic of provability. In: Gabbay, D.M., Concharov, S.S., Zakharyashev, M. (eds.) Mathematical Problems from Applied Logic I. Logics for the XXIst Century. International Mathematical Series, vol. 4, pp. 77–136. Springer, New York (2006)
15. Church, A.: A formulation of the logic of sense and denotation. In: Structure, Method, and Meaning, pp. 3–24 (1951)
16. de Jongh, D.H.J., Visser, A.: Explicit fixed points in interpretability logic. Stud. Log. **50**, 39–50 (1991)
17. Feferman, S.: Arithmetization of metamathematics in a general setting. Fundam. Math. **49**, 35–92 (1960)
18. Gleit, Z., Goldfarb, W.: Characters and fixed points in provability logic. Notre Dame J. Form. Log. **31**, 26–36 (1990)

19. Gödel, K.: Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatshefte Math. **38**, 173–198 (1931)
20. Gödel, K.: Ein Interpretation des intuitionistischen Aussagenkalküls. In: Ergebnisse eines mathematischen Kolloquiums, vol. 4, pp. 39–40 (1933). Reprinted as: An interpretation of the intuitionistic propositional calculus. In: Feferman, S. (ed.), Gödel Collected Works I, pp. 300–303, Oxford (1986)
21. Goryachev, S.: On interpretability of some extensions of arithmetic. Mat. Zametki **40**, 561–572 (1986) (in Russian). English translation in Math. Notes **40**
22. Guaspari, D., Solovay, R.M.: Rosser sentences. Ann. Math. Log. **16**, 81–99 (1979)
23. Hilbert, D., Bernays, P.: Grundlagen der Mathematik II, Springer, Berlin (1939). 2nd edn. in (1970)
24. Henkin, L.: A problem concerning provability. J. Symb. Log. **17**, 160 (1952)
25. Henkin, L.: Review of G. Kreisel: On a problem of Henkin's. J. Symb. Log. **19**(3), 219–220 (1954)
26. Hoogland, E.: Definability and Interpolation: Model-Theoretic Investigations. Institute for Logic, Language and Computation, Amsterdam (2001)
27. Halbach, V., Visser, A.: Self-reference in Arithmetic. Logic Group Preprint Series, vol. 316. Faculty of Humanities, Philosophy, Utrecht (2013)
28. Iemhoff, R.: On the admissible rules of intuitionistic propositional logic. J. Symb. Log. **66**(1), 281–294 (2001)
29. Japaridze, G.: The polymodal logic of provability. In: Intensional Logics and Logical Structure of Theories: Material from the Fourth Soviet–Finnish Symposium on Logic, Telavi, pp. 16–48 (1985)
30. Japaridze, G., de Jongh, D.: The logic of provability. In: Handbook of Proof Theory, pp. 475–546. North-Holland, Amsterdam (1998)
31. Kreisel, G.: On a problem of Henkin's. Indag. Math. **15**, 405–406 (1953)
32. Kreisel, G., Takeuti, G.: Formally self-referential propositions for cut free classical analysis and related systems. Diss. Math. **118**, 1–50 (1974)
33. Lenzi, G.: Recent results on the modal μ-calculus: a survey. Rend. Ist. Mat. Univ. Trieste **42**, 235–255 (2010)
34. Lindström, P.: Provability logic—a short introduction. Theoria **62**(1–2), 19–61 (1996)
35. Löb, M.H.: Solution of a problem of Leon Henkin. J. Symb. Log. **20**, 115–118 (1955)
36. Maksimova, L.: Definability theorems in normal extensions of the probability logic. Stud. Log. **48**(4), 495–507 (1989)
37. Raatikainen, P.: Gödel's incompleteness theorems. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2013). Available via http://plato.stanford.edu/archives/win2013/entries/goedel-incompleteness/
38. Reidhaar-Olson, L.: A new proof of the fixed-point theorem of provability logic. Notre Dame J. Form. Log. **31**(1), 37–43 (1990)
39. Sambin, G.: An effective fixed-point theorem in intuitionistic diagonalizable algebras. Stud. Log. **35**, 345–361 (1976)
40. Shavrukov, V.Yu.: A smart child of Peano's. Notre Dame J. Form. Log. **35**, 161–185 (1994)
41. Smoryński, C.: Beth's theorem and self-referential sentences. In: Macintyre, A., Pacholski, L., Paris, J. (eds.) Logic Colloquium '77. Studies in Logic. North-Holland, Amsterdam (1978)
42. Smoryński, C.: Self-reference and Modal Logic. Springer, New York (1985)
43. Smoryński, C.: Arithmetic analogues of McAloon's unique Rosser sentences. Arch. Math. Log. **28**, 1–21 (1989)
44. Smoryński, C.: The development of self-reference: Löb's theorem. In: Drucker, T. (ed.) Perspectives on the History of Mathematical Logic, pp. 110–133. Springer, Berlin (1991)
45. Solovay, R.M.: Provability interpretations of modal logic. Isr. J. Math. **25**, 287–304 (1976)
46. Sambin, G., Valentini, S.: The modal logic of provability. The sequential approach. J. Philos. Log. **11**(3), 311–342 (1982)
47. Švejdar, V.: On provability logic. Nord. J. Philos. Log. **4**(2), 95–116 (2000)
48. Van Benthem, J.: Modal frame correspondences and fixed-points. Stud. Log. **83**(1–3), 133–155 (2006)
49. Visser, A.: On the completeness principle. Ann. Math. Log. **22**, 263–295 (1982)
50. Visser, A.: Peano's smart children: a provability logical study of systems with built-in consistency. Notre Dame J. Form. Log. **30**(2), 161–196 (1989)
51. Visser, A.: Faith & Falsity: a study of faithful interpretations and false Σ^0_1-sentences. Ann. Pure Appl. Log. **131**(1–3), 103–131 (2005)

52. Visser, A.: Löb's logic meets the μ-calculus. In: Middeldorp, A., van Oostrom, V., van Raamsdonk, F., de Vrijer, R. (eds.) Processes, Terms and Cycles, Steps on the Road to Infinity. Essays Dedicated to Jan Willem Klop on the Occasion of His 60th Birthday. LNCS, vol. 3838, pp. 14–25. Springer, Berlin (2005)
53. Wells, R.: Review of: A formulation of the logic of sense and denotation, by Alonzo Church. J. Symb. Log. **17**(2), 133–134 (1952)

V. Halbach
New College, OX1 3BN Oxford, England, UK
e-mail: volker.halbach@new.ox.ac.uk

A. Visser (✉)
Philosophy, Faculty of Humanities, Utrecht University, Janskerkhof 13, 3512BL Utrecht, The Netherlands
e-mail: a.visser@uu.nl