Logique & Analyse **189–192** (2005), 15–32

REMARKS ON THE CONTENT AND EXTENSION OF THE NOTION OF PROVABILITY†

LEON HORSTEN

> "In June 1972 Gödel asked in a meeting devoted to John von Neumann: is there anything paradoxical in the idea of a machine that knows its own program completely?" (Wang, 1993, p. 121).

In his article 'Some remarks on the notion of proof',[1] Myhill was occupied with the question of the *content* of the informal notion of proof. He argued that there *is* such a thing as an informal or intuitive notion of proof and provability, and wondered how we might best investigate it. Soon afterwards, Lucas took a bold stance on the question of the *extension* of the informal notion of provability.[2] These matters are obviously related.

During subsequent decades, both conceptual (philosophical) and technical (logical) work bearing on these subjects has been carried out. Especially the philosophical literature is often repetitive and bewildering. In any case, it is undeniable that it is poorly connected to the logical work that was carried out during the past decades. The aim of this paper is to pull the logical and philosophical work closer together. I am not under the illusion to be able to fully integrate them. But I do want to see where they can be made to (almost) touch each other.

This paper has a somewhat dejected, sombre tone — as you will soon find out. It seems to me, for reasons that I will explain, that not much progress has

†I am indebted to Hannes Leitgeb, Torkel Franzen, and Lieven Decock for helpful remarks on an earlier version of this paper. Thanks also to an anonymous referee for helpful suggestions for improvement.

[1] Myhill, 1960.

[2] See Lucas, 1961. One already finds important discussions of the intension and extension of the informal notion of provability in Gödel's work. An important source is Gödel, 1951. But there are important remarks about it already in Gödel, 1933.

been made on the central issues involved in the debate. Nevertheless, some philosophers-logicians have made extremely perceptive and insufficiently appreciated comments on these issues. Their remarks offer us glimpses of insight that was not there before. These mostly take the form of observations that in themselves seem elementary, but which are often overlooked and which are bound to play an important role in any eventual resolution of the perplexities that we are confronted with.

I

Let us start with the *content* of the notion of intuitive provability. What do we mean by informal or intuitive provability?

Provable *by who*? We must fix an agent. Several options are open and have been proposed: provable by a fixed person; provable by some restricted group of people — perhaps the mathematicians; provable by humanity...[3] Not much will hinge on which of these options we take — as long as the agent is not *superhuman* in some strong sense, so provable by God will not do. Other than that, this is not where the *immediate* philosophical difficulties lie.[4] So we will be nonspecific about the options that we have mentioned and just call the agent Peter, without committing ourselves to whether he is male or female (or has a determinate gender at all), whether he (she / it) is an individual or a group.

Now we can consider the notion of (Peter) *having* an informal proof. We can make this notion a little more precise by taking this to mean having an a priori demonstration of a proposition on any subject matter.[5] As a degenerate case, axioms count as one-line proofs. So if Peter has a sentence ϕ as an axiom, then I will also say that (by that very fact) Peter has a proof of φ. It then seems plausible that the collection of Peter's actual proofs is at every moment and in every possible situation *finite*.

We now move on to the intuitive notion of prova*bility*. Here things become less clear. An attempt to clarify this notion is the following:

[3] For instance, a main difference between the position of Lucas and that of Penrose is that, like Gödel (Wang, 1993, p. 121), Lucas takes the agent to be an ideal individual mathematician, whereas Penrose takes the agent to be the mathematical community (Penrose, 1990, p. 696). Such differences should be immaterial for the argumentation of this paper.

[4] But we will see further on that there are philosophical problems connected with the *idealisations* involved in these notions.

[5] See Anderson, 1997. But see also Myhill, 1960.

> φ is provable $\approx$ Peter has the *ability* to prove φ.

But now we have abilities on our hands. Abilities seem as bad and problematic as, and related to, capacities and dispositions. What is meant here?

I will assume in the sequel, as is usually done, that proofs are structured as finite rooted trees, with axioms as leaves, and with lower nodes obtained by rules of inference from higher nodes.[6] Then one way to make the notion 'ability to prove' more precise is to say:

> Peter has the *ability* to prove $\varphi \approx$
> Peter knows axioms and rules of inference from which φ can be finitely derived.

In this proposal, the *prima facie* modal notion 'can' is reduced to an existential quantifier over (finite) proofs. If this proposal is correct, then the informal concept of provability is *recursively enumerable*. This in turn implies that the informal notion of provability can be logically modeled as provability in some formal theory S.[7] Therefore, if S is consistent, Peter cannot prove $\neg\text{Bew}_S[0{=}1]$, where Bew_S is the standard provability predicate for S and [...] stands for 'the gödel code of ...' . We have here of course a version of Gödel's second incompleteness theorem.

Gödel himself was unsatisfied with the foregoing explication of the concept 'ability to prove'. He was of the opinion that the notion of ability to prove should be interpreted in a dynamical way, a time dimension must be taken into account:

> "the mind, in its use, is not static, but constantly developing" (Gödel, 1972, p. 306).

By *reflecting* on the axioms and rules which he knows already, and the process that has generated them, Peter indefinitely generates new proof principles. This gives rise to a second proposal for making the notion of 'ability to prove' more precise:

[6] I will not take a stance here on whether propositions are taken to be interpreted sentences or contents of sentences or yet something else.

[7] For according to Craig's Lemma (Craig, 1953), every recursively enumerable set of sentences is recursively axiomatizable.

> Peter has the *ability* to prove $\varphi \approx$
> Eventually (perhaps by reflecting on his proof principles),
> Peter *will* have axioms and rules of inference from which φ can be finitely derived.

We know very little about the nature of these reflective procedures. If these procedures are sufficiently systematic and Peter reflectively acquires new proof principles in a linear process of at most ω stages, then the extension of informal provability might still be recursively enumerable; if not, bets are off. I will have to come back to this later.

There is a third explication of the notion of 'ability to prove', on which our grip is even more feeble:

> Peter has the *ability* to prove $\varphi \approx$ Peter is *or might have been* (now) able to prove φ.

This seems to amount to more or less the same thing as saying that in some possible world,[8] Peter has a proof of φ. The proposal invites us to consider counterfactual situations in the determination of the extension of the concept of informal provability.

The existential quantification over possible worlds that is involved in this proposal is much less clear than the quantification over finite derivations from a recursive list of axioms that is involved in proposal one, for we know very little about how far Peter's abilities can be extended while still remaining *human* abilities.[9] In any case, with respect to the question of the extension of the intuitive notion of provability, on this explication all bets are definitely off.

Time (proposal 2) and modality (proposal 3) can of course also *both* be taken to be implicitly present in an interactive way in the notion of ability to prove. This then gives rise to at least a fourth proposal:

> Peter has the *ability* to prove $\varphi \approx$
> There is a possible world in which Peter has, at some moment in time, proved φ.

But the modal dimension tends to eclipse the temporal dimension: the space of possibilities is much vaster than the temporal dimension. So this fourth

[8] I assume here and in the sequel the 'light', Kripkean way of interpreting possible worlds talk. See Kripke, 1980, pp. 16–20.

[9] As Kripke points out (Kripke, 1980, pp. 34–35).

proposal seems to collapse into the third proposal, and I will disregard it in the sequel.

In discussions about the content and extension of the notion of provability, it is not always made explicit which of the above three ways of fleshing out the notion is assumed. The first proposal seems too restrictive. On the other hand, the third proposal is too rich: our grip on what is involved in it is too weak. For the purposes of the present paper I will therefore assume the Gödelian explication as spelled out above. I will countenance an infinite time dimension, but not a modal dimension.

II

Now let us retrace our steps. Much of the perceived philosophical relevance of the investigation, in proof theory, of provability in formal systems is based on the assumption that the extension of the informal concept of provability is recursively enumerable. From this assumption it is inferred that the informal notion of provability can be modeled as provability in some formal system S.

Let us focus on the modeling procedure of informal provability as provability in a formal system. Wang has pointed out that even granting the truth of the assumption that the informal concept of provability is recursively enumerable, there is a defect in the modeling proposal. He concisely puts it this way:

> “The unclarities of the meaning of [’I am consistent’] tend to obscure the exact strength of the premise that I am a [Turing] machine. For example, in some sense I can prove [that I am consistent], but the sense of ‘prove’ and of [’I am consistent’] need not be the formal sense [...]. Thus, if we accept the belief that I have also an informal way of knowing things, then I am not a machine for that reason already.” (Wang, 1974, p. 319)

What Wang means can be expressed in our terms as follows. First, if S is consistent, then S cannot prove $\neg \mathrm{Bew}_{\mathrm{S}}[0{=}1]$. But it seems eminently plausible that Peter *can* prove — in the informal sense of the word ‘prove’ — that he can prove no absurdities such as 0=1. So provability by Peter cannot be modeled as provability in S.

It is worthwhile to spell out Peter’s argument for the conclusion $\neg \mathrm{Bew}_{\mathrm{S}}[0{=}1]$ in some detail. It is a *conceptual truth* about provability that for all sentences φ, if φ is provable (by Peter), then φ is true. Or, in Reinhardt’s words:

> "we know that what we know is true, but the reason for this is [...] that if it were not so we would not call it knowledge" (Reinhardt, 1986, pp. 468–469).

Admittedly, the fact that if Peter proves φ then φ is true, is not one of Peter's purely *mathematical* theorems, for 'Peter proves that' is not a purely mathematical notion. Nevertheless , it is a conceptual truth which Peter can and should recognise. Hence, since we have taken proofs to be a priori demonstrations on *any* subject matter, this proposition belongs to the extension of the concept 'provable by Peter'. Moreover, Tarski's truth-biconditionals entail that if φ is true, then φ. Peter puts these two facts together and instantiates 0=1 for φ. He thereby obtains the conclusion that no absurdities are provable by him.

One might at this point suspect a whiff of paradox here, since the notion of truth and Tarski's biconditionals enter into Peter's argument. But a closer look reveals that there is no problem here. Peter has only used $T[0{=}1]\rightarrow 0{=}1$ (where T is the truth-predicate). Tarski has shown us that *this* much is unproblematic.[10]

So it seems to me that the only conclusion we can possibly draw at this point is that $\neg\mathrm{Bew}_S[0{=}1]$ *fails* to express that Peter cannot prove absurdities. And that in turn entails that Bew_S does not adequately, fully express 'Peter is able to prove'. Something has been lost in the process of passing from the informal notion of provability to its formal model: provability in a formal system, "standardly" expressed. Specifically, Peter proves — practically has as an axiom, really — that for all φ,[11]

Peter proves $\varphi \rightarrow \varphi$

while S cannot consistently prove for all φ:

$\mathrm{Bew}_S[\varphi] \rightarrow \varphi$

I will from now on call the fact that Peter proves for all φ:

φ is provable (by Peter) $\rightarrow \varphi$

[10] Nevertheless, as we will find out soon enough, the paradoxes are waiting in the wings.

[11] Stricly speaking, I am confusing use and mention in the following formula (as well as at a couple of other places in the sequel). Gödel has of course taught us how to express these propositions correctly using gödel numbering. In the interest of readability, I will abstract from these complications; they do not affect the points that are made in this paper.

the *reflexivity of the informal notion of proof.* This appears to me to be the kernel of truth in the arguments of Lucas and Penrose to the effect that 'the human mind is not a Turing machine'.[12] The informal notion of provability is reflexive, whereas formal notions of provability are not.

III

Having arrived at this point, an old proposal by Gödel seems very acute. He suggested that we attempt to axiomatize the informal notion of provability *directly* instead of finding a mathematical substitute for it.[13] So from a semantic approach ('modeling'), we now shift to an axiomatic approach. Gödel introduced a primitive sentential operator $\Box$, which is to be read as informal (classical) provability. He then postulated the $\mathsf{S4}$ principles of modal logic as reasonable principles governing informal provability. Applying the Necessitation rule to the T-axiom[14] yields a two-line proof in the modal logic T (which is a subsystem of the modal logic $\mathsf{S4}$) of

$$\Box(\Box\varphi \rightarrow \varphi).$$

This shows that Gödel's theory respects what was called earlier the reflexive nature of the informal notion of provability.

Gödel confined himself to a propositional setting. Later, Tarski and McKinsey extended Gödel's approach to a first-order predicate setting.[15] Later still, Shapiro and others extended the approach to a first-order arithmetical setting without a hitch.[16] Nonetheless, extension of the approach to higher-order settings presents substantial problems, which have until now not been completely overcome.[17]

[12] See Lucas, 1961; Penrose, 1994.

[13] See Gödel, 1933. Myhill explicitly defends this proposal in Myhill, 1960.

[14] This axiom is also called the *reflexivity axiom* in the logical literature.

[15] See Tarski & McKinsey, 1948.

[16] See Shapiro, S. 1985. W. Reinhardt independently arrived at a theory that is very similar to that of Shapiro and his co-workers.

[17] These problems are to some extent discussed in Horsten, 1998. See also Horsten, 2005.

Shapiro's theory of Epistemic Arithmetic (EA) is easily defined. Take the language of first-order arithmetic plus Gödel's operator $\Box$. EA is formulated in this language, and consists of the Peano-axioms plus the S4-principles governing the informal provability-operator. Myhill has shown that the model M which interprets the arithmetical vocabulary in the standard way and takes as the extension of $\Box$ the class of theorems of EA, makes all of EA true.[18] So the extension of EA's notion of informal provability *can* be taken to be provability in EA itself. If this line of research is taken further, a connection with the Lucas-Penrose arguments becomes visible.

Benacerraf pointed out early on in the discussion about the Lucas-argument that two propositions must be clearly distinguished:

(1) There is a Turing machine e such that Peter can prove that he is e.[19]

(2) Peter can prove that he is some Turing machine, but he does not know which one.

Benacerraf concedes that Lucas' arguments show that (1) is inconsistent. But (2) seems weaker than (1). Benacerraf argues that Lucas' considerations leave the question whether (2) is consistent wide open. Let us call (2) *Benacerraf's contingency*.

Reinhardt observed that in a languages very much like the language of Epistemic Arithmetic, sentences (1) and (2) can be formalised (Reinhardt, 1986, pp. 435–438). (1) and (2) quantify over an *infinite* number of sentences. So the first-order language of EA itself is not expressive enough to formalise them. For this reason, Reinhardt extends the language of EA with a truth predicate T. *Reinhardt Arithmetic* (RA) can then be defined as EA augmented with Tarski's compositional truth axioms for T applied to the language of EA. In the language of RA, (1) and (2) can be expressed, roughly as:

(3) There is a Turing machine *of which* it can be proved (by Peter) that the collection of sentences that it enumerates coincides with the collection of sentences that are provable (by Peter).

(4) Peter is able to prove that the collection of sentences that are informally provable (by Peter) coincides with the collection of sentences enumerated by *some* Turing-machine.

[18] See Myhill, 1985.

[19] I assume from now on some standard way of coding Turing machines as natural numbers.

Now we can ask whether RA + (3) and RA + (4) are consistent. Reinhardt proved that RA + (3) is inconsistent (Reinhardt, 1986, pp. 439–440). The proof is straightforward. At the same time, Reinhardt conjectured that RA plus (4), i.e. Benacerraf's contingency, *is* consistent (Reinhardt, 1986, p. 436). Proving this is not nearly as straightforward. Nevertheless, Carlson was able to verify Reinhardt's conjecture.[20]

In the literature on the Lucas-Penrose arguments, Reinhardt and Carlson's work has, as far as I know, scarcely been mentioned — and this is deplorable: philosophers *ought* to be aware of it. But Lucas and Penrose have attempted to dismiss Benacerraf's contingency in its unformalised form. Lucas at one point argued that while Benacerraf's contingency might not be outright inconsistent, it at least leads to ω-inconsistency:

> "But to maintain that there is a programme number j such that the corresponding programme W_j represents me, while knowing that for each particular programme number j there is an argument, different in each case, showing that W_j does not represent me, is to be omega-inconsistent. Benacerraf is claiming that man is a machine, although for each particular machine he could be we can show that he is not that one." (Lucas, 1968, p. 152).

But this rests on an elementary misunderstanding. Benacerraf's contingency does not entail that for every Turing machine e, Peter can *refute* that he is e (as Lucas claims), but merely that he *cannot prove* that he is e. Penrose concedes that Benacerraf's contingency is a possibility, but judges it to be a remote one:

> "I never claimed that [Benacerraf's contingency] is a mathematical impossibility, but it would seem exceedingly unlikely. [...] As I said in *Emperor*, mathematics is 'built up from simple and obvious ingredients'... [Mathematical arguments] are sometimes exceedingly complicated. It is just that such arguments are, in principle, built up from such 'obvious' ingredients." (Penrose, 1990, p. 696).

In other words, Penrose argues that our informal notion of provability must be *effective* in the sense that Peter knows what his proofs are. But it is not at all clear to me that provability is effective in this sense. Might we not see

[20] See Carlson, 1997, 1999, and Carlson, 2000. Actually, Carlson proves the consistency of the system EA+SMT ("Strong Mechanistic Thesis"), where the schema SMT expresses Benacerraf's contingency as closely as is possible without introducing a truth predicate in the formal language. See Carlson, 2000, p. 54.

it in the following way? We know our capacities — in a sense. But we do not know of all of them that they are *real* capacities. We must respect the condition that if the conclusion of an argument is not true, then the argument cannot be a proof. But we do not have the resources to effectively check this. So we do know $\neg\Box(0{=}1)$. But we do not know whether everything that we take as an axiom is true. We *hope* that all of our mathematical arguments are genuine proofs — but we cannot know for sure. In this sense, our notion of proof is noneffective. In other words, might not Peter's situation be the following? He knows *of* some Turing machine e *that* it enumerates the sentences which he uses as axioms in his a priori demonstrations. But he does not and cannot know that all sentences which he takes to be axioms really *are* axioms or even real theorems (and therefore true).

At any rate, it seemed to Carlson and Reinhardt that progress had been made with respect to the Lucas-Penrose arguments, especially with Benacerraf's contingency. Their results seem to show that while in some sense, it is inconsistent for Peter to know *of* any particular Turing machine that he "is" that very Turing machine, it is consistent for Peter to know *that* he "is" some Turing machine without knowing which.

But there is a difficulty that cuts deeper than Lucas' and Penrose's objections. A point which is elementary but of crucial importance in this connection is that there are strong reasons to think that informal provability should really be formalized as a *predicate* (P) of (codes of) sentences, rather than as a sentential operator. After all, we want to be able to quantify into the informal provability-context. We want to be able to say things like "There are some sentences which are not informally provable (by Peter)". If we opt for the sentential operator approach, then this is not possible. We need the predicate approach.[21]

In the context of arithmetic, the T-axiom and the Necessitation rule for P give us an explosive mix: it leads to a fairly immediate contradiction. This result is generally credited to Kaplan and Montague,[22] but Myhill discovered it around the same time and apparently independently.[23] It was recognised from the start that this 'Paradox of the Knower', as it is called, is deeply related to the semantical paradoxes; indeed, it can be seen as a *strengthening* of the liar paradox, for its premises are weaker. So we have the paradoxes on

[21] Carlson explicitly opts for the operator approach and against the predicate approach (Carlson 1997a, p. 2). But he does not dwell on the consequences of this choice for the relevance of his results for the *philosophical* question whether Benacerraf's contingency is consistent.

[22] See Kaplan & Montague, 1960.

[23] See Myhill, 1960, pp. 469–470.

our hands.[24] It comes as no real surprise that this must somehow also have been clear to Gödel:

> "Regarding the discussion [about the Lucas-Penrose argument], Gödel thinks that because of the unsolved intentional paradoxes for concepts, like 'concept', 'proposition', 'proof', etc., in their most general sense, no proof using the self-reflexivity of these concepts can be regarded as conclusive in the present stage of the development of logic, although, after a satisfactory solution of these paradoxes, such argument may turn out to be conclusive." (Wang, 1974, fn 14, p. 328).[25]

This gives a man a sinking feeling. If we have to solve the paradoxes *before* we can arrive at definite conclusions about the extension of the informal notion of provability, then the situation looks gloomy indeed! But let us not give up hope completely. Let us see if there is anything we can take home from the battlefield.

The result that RA + (3) is inconsistent is probably not going to be affected by the shift from an operator approach to a predicate approach. The reason is roughly the following. Consider the translation τ which transforms every formula ϕ of the language of EA by uniformly replacing $\Box$(...) by P[...], starting from the atomic constituents of the formula and working systematically outwards in the formula. Then the translation τ(EA) of EA is a consistent and presumably even sound theory in the predicate approach.[26] Intuitively one can see that this *must* be so, for no axioms of EA will be translated by τ into 'viciously self-referring' sentences. Therefore τ will also translate the theorem that RA + (3)$\vdash \perp$ (where $\perp$ is the *falsum* symbol) into a theorem of a sound and unproblematic formal theory which treats informal provability as a predicate. So Reinhardt's theorem that RA + (3) is inconsistent is *persistent.*

[24] In order to obtain the power to quantify into the context of the informal notion of provability, we could perhaps work with propositional quantifiers. But Grim has shown that, given enough expressive power in the propositional quantifier-setting, the paradox of the knower reappears. See Grim 1993.

[25] Anderson comes to a similar conclusion regarding the discussion of Fitch's paradox (Anderson, 2001). An awareness of the fact that the paradoxes are close is also found in Chalmers' discussion of Penrose's 'second argument' (Chalmers, 1995, section 3). Chalmers' argument is clearly a version of the paradox of the knower. (Oddly enough he does not mention Kaplan and Montague's paper.)

[26] This is worked out in detail in Schweizer, 1992.

This much cannot be said of Carlson's theorem. As far as I can see, it is completely open whether Carlson's consistency theorem carries over to future predicate treatments of informal provability. One might at this point say: "Well, just repeat the whole exercise for informal provability treated as a predicate. Construct an adequate formal theory, and check if Benacerraf's contingency is consistent with it." But unfortunately we do not, at present, have even the foggiest idea of how a natural, elegant, powerful axiomatic theory of informal provability (treated as a predicate) would look like,[27] let alone whether Benacerraf's contingency is consistent with it.

IV

I have been looking at the informal concept of provability in the Gödelian way: a temporal dimension, but no modal dimension is taken into account. As time goes on, Peter somehow accepts new proof principles as basic axioms. And I have assumed that time is infinite in the future-direction: this seems to be the sort of idealisation that is allowed.[28] It follows from these assumptions that Peter has sufficient time to adopt infinitely many new axioms in the fullness of time.

It was noted in section I that if the process of introduction of new axioms is sufficiently nonsystematic, then it seems hardly possible to place an upper bound on the complexity of the extension of informal provability. To be a little more concrete, suppose that the only thing that we know is that in the course of time, there are ω instances of introduction of a new large cardinal axiom by Peter. Then the union of these axioms *may* be recursively axiomatizable; but it may as well be horribly nonrecursive. It therefore is more fruitful to consider *systematic* ways of introducing new axioms. The most significant work that has been done in this area is of course that by Feferman.[29]

The idea is roughly the following. Peter starts at a moment in time with some axiomatic theory — say Peano Arithmetic (PA).[30] Then *reflection* on

[27] The sad current state of affairs is reviewed in Horsten, 2003.

[28] For instance, the intuitionists make this assumption in their discussion of the 'creative subject'. Gödel also assumes the ideal mathematician's lifespan to be infinite (Wang, 1993: p. 121).

[29] See Feferman, 1962. Franzen, 2002, ch. 13–15 gives a good inroad to Feferman's work on progressions of formal theories.

[30] But the same will apply, *mutatis mutandis*, if Peter starts out with a strong mathematical theory, such as ZFC.

what he has accepted so far allows Peter to pass to a stronger axiomatic theory. We focus here on two insights which can result from such a reflective process:

(1) Insight in the *consistency* of the axiomatic system one has obtained so far.
(2) Insight in the *soundness* of the axiomatic system that one has obtained so far.

Intuitively, a consistency insight is weaker than a soundness insight.

By repeating these procedures into the transfinite, *transfinite progressions* of axiomatic theories can be constructed. Two such progressions can be defined *roughly* as follows:

$$\begin{aligned}
&\mathrm{T}^{\mathrm{C}}_0 = \mathrm{PA}\\
&\mathrm{T}^{\mathrm{C}}_{\alpha+1} = \mathrm{T}^{\mathrm{C}}_\alpha \cup \mathrm{Consis}(\mathrm{T}^{\mathrm{C}}_\alpha)\\
&\mathrm{T}^{\mathrm{C}}_\lambda = \textstyle\bigcup_{\kappa<\lambda} \mathrm{T}^{\mathrm{C}}_\kappa \text{ for } \lambda \text{ a limit ordinal}
\end{aligned}$$

$$\begin{aligned}
&\mathrm{T}^{\mathrm{UR}}_0 = \mathrm{PA}\\
&\mathrm{T}^{\mathrm{UR}}_{\alpha+1} = \mathrm{T}^{\mathrm{UR}}_\alpha \cup \{\mathrm{Bew}^{\mathrm{UR}}_{\mathrm{T}_\alpha}[\varphi] \rightarrow \varphi \mid \varphi \text{ an arithmetical formula possibly containing free variables}\}\\
&\mathrm{T}^{\mathrm{UR}}_\lambda = \textstyle\bigcup_{\kappa<\lambda} \mathrm{T}^{\mathrm{UR}}_\kappa \text{ for } \lambda \text{ a limit ordinal}
\end{aligned}$$

The superscript C here stands for *consistency*; the superscript UR stands for *uniform reflection*. $\mathrm{Bew}^{\mathrm{UR}}_{\mathrm{T}_\alpha}$ is an arithmetical proof predicate for $\mathrm{T}^{\mathrm{UR}}_\alpha$; Consis($\mathrm{T}^{\mathrm{C}}_\alpha$) expresses in some obvious form the consistency of $\mathrm{T}^{\mathrm{C}}_\alpha$, and is formed using an arithmetical proof predicate $\mathrm{Bew}^{\mathrm{C}}_{\mathrm{T}_\alpha}$ for $\mathrm{T}^{\mathrm{C}}_\alpha$.

But there are caveats. First, we must use somehow a *standard* arithmetical proof predicate for the $\mathrm{T}^{\mathrm{C}}_\alpha$'s and the $\mathrm{T}^{\mathrm{UR}}_\alpha$'s. For with the aid of nonstandard proof predicates, we can hide inside "consistency" assertions information which is not given by the consistency insight alone.

Second, in order to formulate $\mathrm{Bew}^{\mathrm{C}}_{\mathrm{T}_\alpha}$, we need an *arithmetical name* for the ordinal α. To this end, Feferman uses the systematic arithmetic notation system for countable ordinals that was developed by Kleene. This is the so-called system O of *constructive ordinal notations*. This system of arithmetical notations for ordinals is defined as follows (we define the coding scheme together with an induced ordering relation $<_{\mathrm{O}}$ on the numerical codes):[31]

[31] See Rogers, 1967, p. 208.

(a) 1 is a notation for the ordinal 0;

Assume that all ordinals $<\gamma$ have received their notations, and that $<_{\mathrm{O}}$ has been defined for these arithmetical notations:

(b) Suppose $\gamma = \beta + 1$. Then for each x such that x is a notation of β, 2^{x} is a notation of γ, and we put $2^{\mathrm{x}} <_{\mathrm{O}} \mathrm{z}$ for each z which is a notation for an ordinal $\leq \beta$.

(c) Suppose γ is a limit ordinal. Then for each code e of a Turing machine $\mathrm{T_e}$ such that $< \mathrm{T_e(n)} >_{\mathrm{n=1}}^{\mathrm{n=\infty}}$ is a sequence of notations for an increasing sequence of ordinals with limit γ and such that for all i,j: if i<j, then $\mathrm{T_e(i)} <_{\mathrm{O}} \mathrm{T_e(j)}$, γ receives $3 \times 5^{\mathrm{e}}$ as a notation.

The details of the definition and of the theory of this notation system O do not really matter here. I just note some of its properties. First, this system of notations has the property that only at limit stages, notations for a given ordinal multiply. The reason for this multiplication at limit stages is that there are generally *many* ways of recursively enumerating an infinite sequence of natural numbers. So the system O has the form of a transfinite tree which (infinitely) branches only at limit points. Second, Kleene's O is a *maximal* notation system with certain canonicity properties: in a sense, all ordinal notation systems which have certain niceness properties reduce to this system.[32] Third, O is a very complex set of natural numbers. It is Π^1_1-complete, i.e., more complex than the collection of all first-order arithmetical truths.[33] Fourth, a relatively large number of countable ordinals has notations in O.

Since O gives us arithmetical names for many ordinals, one can construct progressions of theories that reach far into the (countably) transfinite. And one can ask precise proof-theoretic questions about such progressions. For instance, one can ask for a characterization of $\bigcup_{\alpha\in\mathrm{O}} \mathrm{T}^{\mathrm{C}}_{\infty}$. Feferman was able to show that this is a Π_1-complete set: it coïncides with the collection of all arithmetical sentences derivable from Π_1 arithmetical truths. Feferman also proved that $\bigcup_{\alpha\in\mathrm{O}} \mathrm{T}^{\mathrm{UR}}_{\alpha}$ is much more complex: it is the collection of *all* first-order arithmetical truths. There even exists a "path through O", i.e. a path P through the transfinite tree O, such that already $\bigcup_{\alpha\in\mathrm{P}} \mathrm{T}^{\mathrm{UR}}_{\alpha}$ proves all arithmetical truths. Can we not see this path P through O as a description of the evolution of Peter's mind?

There are two obstacles. First, one would have to make sense of Peter's going through a *transfinite* number of discrete stages in the evolution of his

[32] See Rogers, 1967, p. 210.

[33] See Sacks, 1990, p. 19.

mind. In other words, how does Peter pass the limit stage ω? It seems there isn't time. Second, it follows from the Π^1_1-completeness of O that for many of the ordinal notations o in O, it is not provable in PA or even in much stronger arithmetical theories *that* o is an ordinal notation.

Concerning the first problem, it is not entirely clear to me that allowing Peter to go through a transfinite number of stages is an inadmissible form of idealization — although it certainly also is not clear to me that it *is* admissible.

It seems very unlikely that Peter is *in fact* allowed even ω stages. As a general rule, it seems to me that any natural extension of Peter's capacities which does not obviously collapse into truth merits investigation. Already for this reason alone we should not dismiss the idealisation of Peter going through a transfinite number of stages out of hand. Moreover, in this connection one can refer to an ongoing discussion in the philosophy of science. There it has been shown that there are cosmological models of the *General Theory of Relativity* in which a computor carries out an infinite number of computations in a finite amount of time.[34] But it would take us too far to enter into this discussion.

The concept of *autonomous progressions*, also introduced by Feferman, can be used to remedy the second problem. The idea is the following. In PA, one can prove of a lot of ordinal notations α that they are ordinal notations. For such α, $\mathrm{T}^{\mathrm{UR}}_{\alpha}$ is then an acceptable theory for Peter. But then there will be *new* ordinals β such that in $\mathrm{T}^{\mathrm{UR}}_{\alpha}$, but not in PA, it can be shown that $\beta \in \mathrm{O}$. This makes $\mathrm{T}^{\mathrm{UR}}_{\beta}$ acceptable for Peter, and so on and so forth. Let A be the class of ordinals that can eventually be provable in this way to be in O. Theories $\mathrm{T}^{\mathrm{UR}}_{\alpha}$ with $\alpha \in \mathrm{A}$ are called *autonomously* reachable. It is clear that to the class $\bigcup_{\alpha\in\mathrm{A}} \mathrm{T}^{\mathrm{UR}}_{\alpha}$, the second objection no longer applies. But unfortunately $\bigcup_{\alpha\in\mathrm{A}} \mathrm{T}^{\mathrm{UR}}_{\alpha}$ still is an axiomatic theory, it a fragment of the the so-called theory of *predicative analysis*.[35] So in the process of addressing the second objection, we have lost the conclusion that Peter is not a machine. I suspect that it is for this reason that Feferman, in his review of Penrose's book, seems to attach no special relevance to his own results about transfinite progressions of axiomatic theories.[36] Lindström, in a recent article, goes further than this. He parenthetically remarks that Feferman's work on progressions of formal theories can be used to show that the mind *is* a Turing

[34] See Earman, 1995: chapter 4.

[35] See Feferman, 1964. We have a similar development if we start not from PA, but from a strong mathematical theory, such as ZFC.

[36] See Feferman, 1995.

machine (Lindström, 2001, p. 242). The arguments discussed in this paper do not support this stronger claim.

V

Where does this leave us? Where do we get off?

We are caught in a dilemma. Either we admit very irregular ways of formulating new axioms, or we confine our attention to procedures for finding new axioms that are somehow systematic. Neither strategy appears promising at the moment.

If we take the first option, we lose our bearings. For then we are unable to place any informative bounds on the complexity of the extension of the informal notion of provability. The second option has problems of its own. We are unable to find a systematic procedure which takes us beyond a recursively enumerable collection of theorems, even if we allow ourselves a transfinite number of stages for extending our axioms.

We can try to approach the matter from the other end by trying to show that it is at least *consistent* with strong logical theories of informal provability that its extension is recursively enumerable. This would at least soften the dilemma. But here we find our attempts blocked by the paradoxes. All we have is the relatively weak proposition that we cannot be in a position where we have an axiom system which we *know* to be the extension of the informal concept of provability. And as we have seen, even this weak result is not as straightforward as it seems at first sight.

Hence my conclusion that the field is in a sorry state. None of the avenues that have been pursued in recent decades appear to be very promising to date. One would have to be a staunch optimist to predict that we will soon see a way out of the impasse we find ourselves in.

Center for Logic and Philosophy of Science
Institute of Philosophy
University of Leuven
Kardinaal Mercierplein 2
B-3000 Leuven (Belgium)
E-mail: Leon.Horsten@hiw.kuleuven.be

REFERENCES

Anderson, C.A. “Towards a logic of a priori knowledge.” *Philosophical Topics* 21(1993).

Anderson, C.A. “Williamson, Fitch, and Frege-Church.” Unpublished manuscript, 2001.

Benacerraf, P. “God, the devil and Gödel.” *The Monist* 51(1967).

Carlson, T. “Can a machine know that it is a machine?” Unpublished manuscript, 1997.

Carlson, T. “Ordinal arithmetic and Σ_1-elementarity.” *Archive for Mathematical Logic* 38(1999), pp. 449–460.

Carlson, T. “Knowledge, machines and the consistency of Reinhardt’s strong mechanistic thesis.” *Annals of Pure and Applied Logic* 105(2000), pp. 51–82.

Chalmers, D. “Minds, machines and mathematics. A review of *Shadows of the Mind* by Roger Penrose.” *Psyche* 2(1995).

Craig, W. “On axiomatizability within a system.” *Journal of Symbolic Logic* 18(1953), pp. 30–32.

Earman, J. *Bangs, Crunches, Whimpers and Shrieks. Singularities and acausalities in relativistic spacetimes.* Oxford University Press, 1995.

Feferman, S. “Systems of Predicative Analysis.” *Journal of Symbolic Logic* 29(1964), pp. 1–30.

Feferman, S. “Penrose’s Gödelian argument. A review of *Shadows of the Mind* by Roger Penrose.” *Psyche* 2(1995).

Franzen, T. *Inexhaustibility: a non-exhaustive study.* Lecture Notes in Logic, volume 16, Association of Symbolic Logic and A.K. Peters, 2002.

Gödel, K. “Eine Interpretation des intuitionistischen Aussagenkalküls.” [1933] In S. Feferman et al. *Kurt Gödel. Collected works. Volume I.* Oxford University Press, 1986, p. 300.

Gödel, K. “Some basic theorems on the foundations of mathematics and their implications.” [1951]. In S. Feferman et al. *Kurt Gödel. Collected works. Volume III.* Oxford University Press, 1990, pp. 304–323.

Gödel, K. “Some remarks on the undecidability results.” [1972] In S. Feferman et al. *Kurt Gödel. Collected works. Volume II.* Oxford University Press, 1990, p. 306.

Grim, P. “Operators in the paradox of the knower.” *Synthese* 94(1993), pp. 409–428.

Horsten, L. “In defense of Epistemic Arithmetic.” *Synthese* 116(1998), pp. 1–25.

Horsten, L. “The logic of intensional predicates.” In: Benedikt Löwe, Wolfgang Malzkorn and Thoralf Räsch (eds) *Foundations of the Formal Sciences II: Applications of Mathematical Logic in Philosophy and Linguistics Trends in Logic*: Studia Logica Library Volume 17, Kluwer Academic Publishers, Dordrecht, 2003, pp. 89–111.

Horsten, L. “Canonical naming systems.” *Minds and Machines* 15(2005), pp. 229–257.

Kaplan, D. and Montague, R. “A paradox regained.” *Notre Dame Journal of Formal Logic* 1(1960), pp. 79–90.

Kripke, S. *Naming and necessity.* Harvard University Press, 1980.

Lindström, P. “Penrose’s new argument.” *Journal of Philosophical Logic* 30(2001), pp. 241–250.

Lucas, J.R. “Minds, machines and Gödel.” *Philosophy* 36(1961), pp. 112–127.

Lucas, J.R. “Satan stultified: a rejoinder to Paul Benacerraf.” *The Monist* 52(1968), p. 152.

Myhill, J. “Some remarks on the notion of proof.” *Journal of Philosophy* 57(1960), pp. 461–471.

Myhill, J. “Intensional set theory.” In S. Shapiro (ed.) *Intensional Mathematics.* North-Holland, 1985, pp. 47–61.

Penrose, R. “Precis of *The Emperor’s New Mind: Concerning computers, minds, and the laws of physics.*” *Behavioral and Brain Sciences* 13(1990), pp. 643–705.

Reinhardt, W. “Epistemic theories and the interpretation of Gödel’s incompleteness theorems.” *Journal of Philosophical Logic* 15(1986), pp. 427–474.

Rogers, H. *Theory of recursive functions and effective computability.* MIT Press, 1987[1967].

Sacks, G. *Higher recursion theory.* Springer, 1990.

Schweizer, P. “A syntactical approach to modality.” *Journal of Philosophical Logic* 21(1992), pp. 1–31.

Shapiro, S. “Epistemic and intuitionistic arithmetic.” In S. Shapiro (ed.) *Intensional Mathematics.* North-Holland, 1985, pp. 11–46.

Tarski, A. & McKinsey, J. “Some theorems about the sentential calculi of Lewis and Heyting.” *Journal of Symbolic Logic* 13(1948), pp. 1–15.

Wang, H. *From Mathematics to Philosophy.* Routledge, 1974.

Wang, H. “On physicalism and algorithmism: can machines think?” *Philosophia Mathematica* 1(1993), pp. 139–156.