Logic and **L**ogical **P**hilosophy
Volume 31 (2022), 471–500
DOI: 10.12775/LLP.2021.019

Antonella Corradini and Sergio Galvan

Analysis of Penrose's Second Argument Formalised in DTK System

Abstract. This article aims to examine Koellner's reconstruction of Penrose's second argument — a reconstruction that uses the **DTK** system to deal with Gödel's disjunction issues. Koellner states that Penrose's argument is unsound, because it contains two illegitimate steps. He contends that the formulas to which the **T-intro** and **K-intro** rules apply are both indeterminate. However, we intend to show that we can correctly interpret the formulas on the set of arithmetic formulas, and that, as a consequence, the two steps become legitimate. Nevertheless, the argument remains partially inconclusive. More precisely, the argument does not reach a result that shows there is no formalism capable of deriving all the true arithmetic propositions known to man. Instead, it shows that, if such formalism exists, there is at least one true non-arithmetic proposition known to the human mind that we cannot derive from the formalism in question. Finally, we reflect on the idealised character of the **DTK** system. These reflections highlight the limits of human knowledge, and, at the same time, its irreducibility to computation.

Keywords: Penrose's second argument; Gödel's disjunction; **DT** system; **DTK** system; computational model of mind; arguments in favour of the first horn of Gödel's disjunction

Introduction

Roger Penrose has wryly reflected on the philosophical implications of Gödel's incompleteness theorems. Penrose's thoughts on this matter are neatly summarised by two arguments he made to show the superiority of the mind over the machine. The first of these two arguments [Penrose, 1989] rests on the fact that, while a machine cannot demonstrate its

Received March 8, 2021. *Accepted* December 2, 2021. *Published online* December 16, 2021

consistency, the mind can. In response to criticism raised against this thesis, Penrose shifted his attention to the mind's capabilities, rather than continuing to focus on the machine's limits. The result of this change of perspective is the so-called second argument of Penrose.[1] This argument is first presented in [Penrose, 1994, sections 3.16, 3.23, and 3.24] and then in [Penrose, 1996, section 3.2], where it appears in the following form:

> Though I don't know that I necessarily am F, I conclude that if I were, then the system F would have to be sound and, more to the point, F' would have to be sound, where F′ is F supplemented by the further assertion "I am F". I perceive that it follows from the assumption that I am F that the Gödel statement G(F′) would have to be true and, furthermore, that it would not be a consequence of F′. But I have just perceived that "if I happened to be F, then G(F′) would have to be true", and perceptions of this nature would be precisely what F′ is supposed to achieve. Since I am therefore capable of perceiving something beyond the powers of F′, I deduce that, I cannot be F after all. Moreover, this applies to any other (Gödelizable) system, in place of F.

Although this formulation of the argument is clearer than the formulations presented in [Penrose, 1989, 1994], it has, nonetheless, given rise to a series of interpretations or — more accurately — of reconstructions, which are only partially equivalent. Scholars such as Chalmers [1995], Shapiro [1998, 2003], and Lindström [2001, 2006] have shared the conviction that Penrose's second argument is not conclusive. The most extensive and in-depth reconstruction of the argument has recently been published by Peter Koellner in [2018b].

Koellner's approach is interesting because he develops the reconstruction of the argument within the context of a rigorous formulation and demonstration of Gödel's disjunction.[2] The use of a precise language, to formulate the disjunction and identification of a precise system of rules and axioms, allows Koellner to demonstrate the disjunction rigorously, and, thus, to confirm Gödel's claim that the disjunction represents a mathematical result that is consequent to the incompleteness theorems. Koellner's valuable analysis on this front also allows him to deal — in a refined manner — with the alleged demonstrations of the first disjunct,

[1] Penrose's arguments are in part anticipated by Lucas' reflections, which present strong similarities with Penrose's ones [see Lucas, 1961, 1968].

[2] See [Koellner, 2016, 2018a]. For Gödel's disjunction see [Gödel, 1996, p. 310].

including Penrose's second argument. In summary, Koellner shows that the $\mathbf{EA_T}$ system (epistemic arithmetic with typed truth predicate), which is sufficient to demonstrate Gödel's disjunction, is not sufficient to address the first disjunct demonstration successfully. The problem here is that the truth predicate required by Penrose's argument must be type-free, since the epistemic knowledge operator K has a more extensive referential range than it has in the $\mathbf{EA_T}$ system. Koellner proposes to replace the $\mathbf{EA_T}$ system with the **DTK** system, which provides for the possibility that certain propositions allowed by the free use of the truth predicate are indeterminate, and, therefore, are not likely to be true or false. In the argument as reconstructed in **DTK**, Koellner identifies two illegitimate steps — in that they involve indeterminate propositions — and, further on, he even demonstrates that the first disjunct is undecidable. Now, it should be noted that, when Koellner deals with Penrose's argument,[3] he presents it in a partially formalised way, so that the illegitimacy claim lacks the required formal details. Koellner [2016] limits himself to illustrating the reason why it is necessary to move from a system with a typed truth predicate — like $\mathbf{EA_T}$ — to one with a free truth predicate — like **DTK**. After that, he proceeds to show — through Theorem 7.16.2 — the undecidability of the first disjunct and the negation of the second. In [2018a, IV.1] Koellner states that steps (2) and (5) of the argument are illegitimate, because the formulas involved are indeterminate. He also shows, through Theorem 7, that indeterminateness can be demonstrated in **DTK**. In IV.2, however, he states that, although indeterminateness can also be demonstrated for the full version of the disjunction, it no longer applies to the restricted version of the disjunction — i.e., the disjunction restricted to arithmetic formulas. The question, of course, arises as to whether this indeterminateness can be removed through a similar restriction, also from the formulas involved in steps (2) and (5) of the argument. And if so, what are the consequences for the conclusiveness of the argument?

This issue is the central theme of this article, and to which the first two sections of this article are devoted. In the first section, the disjunction is proved by adopting a metatheoretical extension of **EA**, in which it is quantified over metavariables. In this way — unlike Koellner — the disjunction can be obtained without an explicit predicate of typed truth. This section serves to show that, for the formalisation of Gödel's

[3] See on this both [Koellner, 2016, 164–166] and [Koellner, 2018b, 459–461].

disjunction, the truth predicate is even dispensable, if one makes use of the metatheoretical extension of **EA**. In any case, a free truth predicate is not necessary. On the contrary, in the demonstration of Penrose's argument, the use of a type-free truth predicate is necessary.

Section 2 of this article is the central one. Here, we will reconstruct Penrose's argument — along Koellner's lines — and, in particular, examine the formal details of the steps that, according to Koellner, are unsound. We will see that, on the restricted interpretation of determinateness, these steps become sound. As a consequence of the same interpretation, however, the last step of the whole argumentation becomes unsound. It is unsound on the restricted interpretation. Nevertheless, steps 1–7 — corrected on the restricted interpretation — allow Penrose's thesis to be obtained in its general interpretation. As surprising as this result may seem, however, this article shows that it only partially achieves the intent pursued by Penrose. Indeed, it is true that the argument thus reconstructed is not able to establish that there is no formalism capable of deriving all humanly knowable mathematical propositions. But, even assuming that such formalism does exist, it is nevertheless possible to state that there is at least one humanly knowable non-mathematical proposition that is not deducible in this formalism.[4]

In Section 3 we will try to show that the relationship between mind and machine, i.e., between knowledge and formal deduction, can be conceived in a different way from how it is within systems such as **DTK**. Namely, we can base it on a conception of mathematical knowledge as something that is both formal and informal. It is formal, because it is ideally open to formalisation, and informal, because it cannot be eliminated in favour of the formal. From this perspective, we will emphasise the problems presented by a computational conception of the mind and therefore with the negation of the first disjunct.

It must be stressed that the main purpose of this article is to formalise in detail the proof of the second argument — in the restricted formulation mentioned — within Koellner's **DTK** system. This article does not intend to take stock of the whole debate concerning the relationship between Gödel's theorems and the mechanistic thesis, nor to draw conclusions about the results obtained by other authors [see, e.g., Krajewski, 2020]. Rather, this essay aims to realise the hard work of formal elaboration, the

[4] This is perfectly in line with the undecidability result independently achieved by Koellner through Theorem 7.16.2 in [2016].

only one that, if successful, allows us to reach a result of mathematical certainty.[5]

1. Gödel's disjunction

1.1. Koellner's approach

The impact that the incompleteness theorems have had on the philosophy of mind is a given. The difficulty of their philosophical interpretation lies in rigorously expressing the consequences that the theorems imply, without misinterpreting their meaning, or amplifying or reducing it, excessively. Gödel himself was aware of the philosophical consequences of his theorems, and also the difficulty of treating them rigorously. That is why he came to formulate the consequences so late on, in his 1951 Gibbs Lecture, the final draft of which was posthumously published. Gödel's wording, here, is informal, and the reasoning that leads to it is also carried out informally. The interest of Koellner's work,[6] therefore, lies in his reconstruction of the disjunction within a strictly formal framework. This framework rests on three fundamental concepts.

A. Relative provability. Relative provability is the notion of provability concerning a specific formal system. Let us take, for example, a formal system (set of axioms and rules) **F**. Then, the propositions provable in **F** are propositions provable relatively to this system. Since it is a formal system, the set of these propositions is recursively enumerable.[7] We express the derivability of φ in **F** through the following two equivalent ways: $\vdash_{\mathbf{F}} \varphi$ and $F(\ulcorner\varphi\urcorner)$ — where $\vdash$ is the usual metatheoretical sign of derivability, $\ulcorner\varphi\urcorner$ represents the code of the proposition φ, and F represents, in the context of $F(\ulcorner\varphi\urcorner)$, the provability predicate in **F** PR_F. The expression $F(\ulcorner\varphi\urcorner)$ is often simplified as $F\varphi$. Finally, the notion of a formal system corresponds to that of an idealised finite machine (Turing machine).

[5] Avron [2020, p. 106] claims: "An argument that cannot be fully formalized cannot be taken as a mathematical proof".

[6] In [Koellner, 2016, pp. 160–162] the disjunction is formalised in $\mathbf{EA_T}$; in [2016, pp. 174–176] in **DTK**; and in [2018a, pp. 349–355] in $\mathbf{EA_T}$.

[7] This is the meaning of the usual expression of computability. A proposition is computable to the extent that there is a formal system in which it is provable.

B. Absolute provability. The concept of absolute provability corresponds to what Gödel calls 'subjective mathematics' , i.e., the set of mathematical propositions that are knowable with mathematical certainty by the ideal mathematician. For our purposes, however, it is not necessary to speak of the totality of the mathematical propositions but, rather, the totality of the arithmetical propositions. Furthermore, it is not possible to know *a priori* which propositions are knowable. It is also not important to draw a precise line of demarcation between knowable and unknowable propositions. As we infer from the evidence, Gödel believed that there were no true mathematical propositions absolutely inaccessible to the human mind, but he did not rule out the possibility either. So much so, that the existence of absolutely undecidable propositions is the content of the second horn of the disjunction. The restriction to only true arithmetical propositions and the irrelevance of a demarcation line between knowable and unknowable propositions make it possible to simplify the construction of the calculus characterising the notion of mathematical knowability without losing its essentials. Koellner meets this requirement by introducing the **EA** system of epistemic arithmetic. The language of **EA** is the language of Peano's arithmetic **PA**[8] expanded to include an operator K with formulas of L(**EA**) as arguments. The axioms of **EA** are simply those of **PA**, with the only difference being that the induction scheme is taken to cover all formulas in L(**EA**). $\vdash_{\mathbf{EA}}$ indicates derivability in **EA**. **EA** axioms are those of **PA**, plus the following rules or axioms (where φ belongs to L(**EA**) and $\Vdash \varphi$ indicates the notion of logical first order validity):

EA Axioms

K1: $\Vdash \varphi \ \Rightarrow \vdash_{\mathbf{EA}} K\varphi$
K2: $\vdash_{\mathbf{EA}} K\varphi \wedge K(\varphi \rightarrow \psi) \rightarrow K\psi$
K3: $\vdash_{\mathbf{EA}} K\varphi \rightarrow \varphi$
K4: $\vdash_{\mathbf{EA}} K\varphi \rightarrow KK\varphi$

The first axiom states that K holds of all first-order logical validities. The second states that K is closed under modus ponens. The third is a way of asserting the correctness of K. And finally the fourth axiom says that K is self-reflective.

Note 1. **K3** is also present in **DTK**. In this system it appears under the guise of axiom **K1**.

[8] For details of the **PA** language see Section 2.2.1.

C. Truth. T is the predicate of mathematical truth. Gödel does not specify the meaning of the notion of truth, but grounds his idea of mathematical truth in the realistic conception of mathematical objects. These exist, and are endowed with properties and relationships. Mathematical truths are all propositions that are true about such entities. From Tarski, however, we know that there cannot be a global predicate of truth. There are only partial predicates of truth, in the sense that, for example, there is an arithmetic truth predicate, but this is not an arithmetical predicate. The denial of the existence of a global predicate of truth derives from the Tarskian assumption that a correct — i.e., not antinomic — truth predicate is necessarily a typed predicate, i.e., such that it cannot be applied to itself. It is currently accepted, however, that there are truth theories that are not typed, and that allow for the phenomenon of self-referentiality without becoming incorrect. Despite the versatility of the type-free notion of truth, Koellner considers the use of a typed truth predicate to be sufficient for the rigorous formulation of the Gödelian disjunction. This predicate is the Tarskian truth predicate T, restricted to L(**EA**), axiomatically introduced in par. 7.6 of [Koellner, 2016]. But an important question arises, which we must address from the outset. Why do we primarily need the truth predicate? Is it not possible to be satisfied with an informal notion of truth? After all, Gödel himself is happy with the formulation of the disjunction in an informal language. To answer this question, we will try to produce the disjunction proof without the truth predicate. In contrast, the formalization of Penrose's argument in the second section shows that, on pain of incurring insurmountable difficulties, it is impossible to dispense with the truth predicate — and, more, a type-free truth predicate.

1.2. The disjunction proof in $\mathbf{EA^m}$

$\mathbf{EA^m}$ system is a metatheoretical extension of the **EA** system. This extension requires one to distinguish between the part of the system constituted by **EA** and the part that constitutes its extension, regarding both the language and its axiomatic structure. As far as the language is concerned, L(**EA**) is L(**PA**) expanded, to include an operator K with formulas of L(**EA**) as arguments. The formulas that are arguments of the K operator are, therefore, arithmetical propositions, or propositions containing, in turn, the K operator. Among the arithmetic propositions there are also propositions resulting from the arithmetisation process of

EA's syntax. Since K is an operator, and not a predicate, it can only act on formulas, and not on numerical variables for these.

For this reason, the generalisation over the formulas must occur on a metatheoretical level. The metatheoretical extension serves exclusively for that purpose. The typology of the formulas, to which the generalisation refers, is decided on a case-by-case basis. With regard to the axiomatic structure, the extension includes the use of propositional and quantificational rules for propositional metavariables — in particular, amongst these, are the rules of introduction of $\forall\varphi$ and elimination of $\exists\varphi$ in the antecedent ($\forall I^m$, $\exists I^m$), and in the succedent ($I\forall^m$, $I\exists^m$). As basic logic calculus we adopt a sequent version of natural deduction calculus.[9]

The Tarskian scheme for the truth predicate — $\ulcorner\varphi\urcorner$ is true iff φ — allows the formulation of the concepts of identity $K = T$, $K = F$, and soundness, without using the truth predicate. $Cons(F)$ means that the system **F** is consistent. G1 stands for the first Gödel's theorem and G2 for the second. We assume to quantify only on syntactical concrete objects, as arithmetical formulas or formal systems.

DEFINITION 1. $K = T \;:=\; \forall\varphi(K\varphi \leftrightarrow \varphi)$

DEFINITION 2. $K = F \;:=\; \forall\varphi(K\varphi \leftrightarrow F\varphi)$

DEFINITION 3. *Soundness of* $F \;:=\; \forall\varphi(F\varphi \rightarrow \varphi)$

The three definitions are constituted of metatheoretical propositions, as they quantify over the propositional metavariables. Still, this characteristic of the definitions does not prevent us from formally developing the disjunction proof. It is sufficient to work in $\mathbf{EA^m}$.

Gödel's disjunction: $\vdash_{\mathbf{EA^m}} \neg\exists F(K = F) \vee \exists\varphi(\varphi \wedge \neg K\varphi \wedge \neg K(\neg\varphi))$ (where φ is a metavariable for arithmetical propositions, and F is a metavariable for formal systems).

PROOF IN TWO PARTS. The "$K = T$"-part:

$\forall\varphi(F\varphi \rightarrow \varphi) \vdash_{\mathbf{EA^m}} F\,\bot \rightarrow \bot$	logic
$\forall\varphi(F\varphi \rightarrow \varphi) \vdash_{\mathbf{EA^m}} Cons(F)$	def. $Cons$
$\forall\varphi(F\varphi \rightarrow \varphi) \vdash_{\mathbf{EA^m}} \exists\varphi(\varphi \wedge \neg F\varphi)$	G1
$K = T,\ K = F \vdash_{\mathbf{EA^m}} \forall\varphi(F\varphi \rightarrow \varphi)$	by def. $K = T = F$

[9] See, e.g., [Sundholm, 1983] and [Ebbinghaus et al., 1984, pp. 57–75]. Also afterwards we adopt, as a basic logic calculus, the same version of natural deduction calculus.

$K = T,\ K = F \vdash_{\mathbf{EA^m}} \exists\varphi(\varphi \wedge \neg F\varphi)$ chain
$K = T,\ K = F \vdash_{\mathbf{EA^m}} \exists\varphi(K\varphi \wedge \neg F\varphi)$ by def. $K = T$
$K = T,\ K = F \vdash_{\mathbf{EA^m}} \neg(K = F)$ by def. $K = F$
$K = T \vdash_{\mathbf{EA^m}} \neg(K = F)$ self-contradiction
$K = T \vdash_{\mathbf{EA^m}} \neg\exists F(K = F)$ $I\forall^m$ and logic
$K = T \vdash_{\mathbf{EA^m}} \neg\exists F(K = F) \vee \exists\varphi(\varphi \wedge \neg K\varphi \wedge \neg K(\neg\varphi))$ $I\vee^m$

The "$K \neq T$"-part:

$\varphi,\ K(\neg\varphi) \rightarrow \neg\varphi \vdash_{\mathbf{EA^m}} \neg K(\neg\varphi)$ logic
$\vdash_{\mathbf{EA^m}} K(\neg\varphi) \rightarrow \neg\varphi$ **K3**
$\varphi \vdash_{\mathbf{EA^m}} \neg K(\neg\varphi)$ chain
$\varphi \wedge \neg K\varphi \vdash_{\mathbf{EA^m}} \varphi \wedge \neg K\varphi \wedge \neg K(\neg\varphi)$ logic
$\exists\varphi(\varphi \wedge \neg K\varphi) \vdash_{\mathbf{EA^m}} \exists\varphi(\varphi \wedge \neg K\varphi \wedge \neg K(\neg\varphi))$ $I\exists^m$, $\exists I^m$
$K \neq T \vdash_{\mathbf{EA^m}} \exists\varphi(\varphi \wedge \neg K\varphi \wedge \neg K(\neg\varphi))$ def. $K \neq T$
$K \neq T \vdash_{\mathbf{EA^m}} \neg\exists F(K = F) \vee \exists\varphi(\varphi \wedge \neg K\varphi \wedge \neg K(\neg\varphi))$ $I\vee^m$

Then by exhaustion: $\vdash_{\mathbf{EA^m}} \neg\exists F(K = F) \vee \exists\varphi(\varphi \wedge \neg K\varphi \wedge \neg K(\neg\varphi))$ $\dashv$

In the context of disjunction formalisation, metatheoretical generalisations don't need to be arithmetisable. In particular, it is not necessary for definitions 1–3. The level of formalisation required to deal with the disjunction only includes the ability to pronounce on the truth of the propositions that fall within the range of action of the operator K or of the predicate F. But none of the generalisations could enter the domain of K or F. Thus, there is no need to arithmetise them, to insert them within the range of action reserved for arithmetical formulas.

2. Penrose's second argument

2.1. Analysis of Koellner's reconstruction

In this paragraph, we will retrace Koellner's analysis of Penrose's argument. The aim of Koellner's approach is known. He elaborates the **DTK** system, which is a mixed system including both the knowledge predicate and a type-free truth predicate, to formalise Penrose's argument. The outcome is the disvaluing of the argument, as two of its crucial steps are found to involve indeterminate propositions, and are therefore illegitimate. Now, we have two goals, here. The first is to expose the system, the second is to provide a careful examination of every single step of the argument to ascertain the satisfaction of every

step's legitimacy conditions. An integral part of our first goal concerns the justification of the necessity of making use of a type-free predicate of truth. We should ask, not only why a type-free predicate is needed, but also why a truth predicate, in general, is needed. Both of these questions require a detailed, step-by-step, analysis of the argument. We will start with an explanation of the reasons given for why the system language must contain a truth predicate.

As can be deduced from our following analysis, Penrose's argument — like Gödel's disjunction proof — is characterised by the occurrence of propositional generalisations. Concerning the context of disjunction, however, there is a profound difference. Unlike as with disjunction, in Penrose's argument, truth and knowledge are also attributed to generalizations. The schematic use of metavariables for propositions is, therefore, not sufficient to attribute truth or knowledge to such generalisations. It is necessary to quantify within the theory over numerical variables for codes of formulas. The system's syntax must therefore be fully arithmetised.

2.2. Koellner's DTK system

The **DTK** system[10] is an extension of Feferman's [2008] **DT** system,[11] which, in turn, is an extension of Peano's **PA** arithmetic.

2.2.1. DTK language

The set of formulas of L(**DTK**) is the set of arithmetical formulas of L(PA) increased by the addition of formulas obtained through the further use of predicates T and K. T is the type-free truth predicate, K is the knowledge predicate. In particular, the non-logical fundamental symbols of the arithmetical language L(**PA**) are: the constant '0' , the symbol 'S' for the succession function, the symbols '+' and '$\cdot$' , for the sum and product operations, respectively. The terms are formed by induction from the fundamental non logical symbols. Numerals are the terms obtained inductively from 0 by application of the succession function S. For example, $SS0$ — generated by applying the function S 2-times — is the numeral corresponding to the number 2 and is denoted by $\bar{2}$. In general $\bar{n}$ denotes the numeral resulting by application of S n-times.

[10] The presentation of the system will follow the treatment carried out by Koellner in parr. 7.13 and 7.14 of [2016, pp. 169–174]

[11] For a treatment of **DT** system by Koellner see [2016, par. 7.11, pp. 166–167].

2.2.2. A note about arithmetisation

As already noted, the **DTK** system's syntax must be fully arithmetised, that is, the whole syntactical structure of **DTK** must be expressed in the **PA** language. The arithmetisation is carried out in the standard way. However, it is worth recalling three important notational elements.[12]

Firstly, it is useful to explain the use of the so-called Feferman dot-notation. Let $\ulcorner\varphi\urcorner$ be the code of the formula φ and $\ulcorner\psi\urcorner$ the code of the formula ψ, and suppose we want to denote the code of the molecular formula $\varphi \wedge \psi$. The classic way is to exploit the code definition of a conjunction. If the code definition is $\ulcorner\varphi \wedge \psi\urcorner = p_0^{\ulcorner\varphi\urcorner} \cdot p_1^{\ulcorner\psi\urcorner}$, then the code of the conjunction can be denoted by its specific value. However, this denotation modality is very complicated and would become impractical if one went beyond a certain level of complexity. This is why Feferman introduced the dot-notation technique. The code of $\varphi \wedge \psi$, i.e. $\ulcorner\varphi \wedge \psi\urcorner$, is represented with $\ulcorner\varphi\urcorner \dot{\wedge} \ulcorner\psi\urcorner$. The technique is general, so that in the case of negation one has, for instance, $\ulcorner\neg\varphi\urcorner = \dot{\neg}\ulcorner\varphi\urcorner$. Moreover, if x and y are variables for formulas codes, then it is not sensless to write $x \dot{\wedge} y$ to indicate the code of the conjunction of the formulas coded by x and y, respectively.

Secondly, the arithmetisation allows syntactic notions to be defined arithmetically through predicates that are valid for their respective codes. Thus, for example, there is the predicate $Var(x)$ for the notion of being a variable — x is the code for a variable —; the predicate $At\text{-}Sent_{\mathbf{PA}}(x)$ for the notion of an atomic arithmetical proposition, i.e., the basic arithmetical formulas without connectives or quantifiers — x is the code for an atomic arithmetical formula —; the predicate $Sent_{\mathbf{PA}}(x)$ for the notion of a formula of L(**PA**) — x is the code for an arithmetical formula. Similarly, there exists the predicate of provability in the theory **T** $PR_T(x)$ — x is the code of a theorem of **T**. Finally, there is the possibility of formalising the idea that the formula of which x is the code is knowable or true. Indeed, in **DTK**, even K — besides T — is a predicate and not an operator.

Thirdly, as already mentioned above, it is important to be able to express in L(**DTK**) that a given formula is valid for all numerals, i.e., to generalise over the numerals representing codes of formulas. To this end,

[12] For further details on the arithmetisation technique and in particular for the use of Feferman's dot-notation adopted by Koellner, see [Koellner, 2016, par. 7.2, pp. 155–156]. See also [Halbach, 2014, pp. 29–38].

two functions must be used: the function of substituting a term in place of a variable in a formula, and the function of coding a numeral. Let us take, for example, the formula $\varphi(x)$ in which the variable x is freely given. Let us suppose that we wish to represent the code of the substitution in $\varphi(x)$ of the variable x with the term t. This is given by the function $so(\ulcorner\varphi(x)\urcorner, \ulcorner x\urcorner, \ulcorner t\urcorner)$. t is of course a generic term. It could be, in particular, a numeral and in that case we would have $so(\ulcorner\varphi(x)\urcorner, \ulcorner x\urcorner, \ulcorner\bar{n}\urcorner)$. The value of the function represents the code of the formula $\varphi(\bar{n})$, which says that φ is true of the n-th numeral. Of course one can use directly $\varphi(\ulcorner\bar{n}\urcorner)$ as the code of $\varphi(\bar{n})$, but there are contexts in which it is essential to make use of the substitution function. Let us suppose, in fact, that we want to assert within L(**PA**) that a certain formula $\varphi(x)$ is provable for each numeral. As we know, the predicate of provability in **PA** is $PR_{\mathbf{PA}}(x)$. Therefore, the statement that $\varphi(\bar{n})$ is provable will be $PR_{\mathbf{PA}}(\ulcorner\varphi(\bar{n})\urcorner)$. However, if we want to express that φ is provable for every n, it is not sufficient to quantify over n and say that the statement of provability is valid for every n. This quantification is, in fact, metatheoretical and, therefore, not finitary, having the same meaning as the following infinite configuration: $PR_{\mathbf{PA}}(\ulcorner\varphi(0)\urcorner)$, $PR_{\mathbf{PA}}(\ulcorner\varphi(\bar{1})\urcorner)$, $PR_{\mathbf{PA}}(\ulcorner\varphi(\bar{2})\urcorner)$, ... and so on. If the general statement is to be arithmetised within the **PA** language, it is necessary to combine the use of the substitution function with that of the numeral function. $num(x)$ is the function which associates each numeral with its code. It is not important to know how the function is constructed; it is important that it is definable in L(**PA**) and that for every n, $num(\bar{n}) = \ulcorner\bar{n}\urcorner$. Suppose, then, that we want to express that φ is provable for each numeral. The way to achieve this consists in writing $\forall x(PR_{\mathbf{PA}}(so(\ulcorner\varphi(x)\urcorner, \ulcorner x\urcorner, num(x))))$, which is usually abbreviated to $\forall x(PR_{\mathbf{PA}}(\ulcorner\varphi(\dot{\bar{x}})\urcorner))$ and, to underline the variable which is replaced, to $\forall x(PR_{\mathbf{PA}}(\ulcorner\varphi(\dot{\bar{x}})\urcorner)/\ulcorner x\urcorner)$. The same method is also used when the predicates of truth T and knowability K are involved.

2.2.3. DTK axiomatic system

$\vdash_{\mathbf{DTK}} \varphi$ is the usual syntactical expression of the **DTK** derivability relation of φ, just as $\vdash_{\mathbf{F}} \varphi$ stands for the **F** derivability relation of φ. Of course, these syntactical propositions can be arithmetised. For example, $PR_F(\ulcorner\varphi\urcorner)$ — in short $F(\ulcorner\varphi\urcorner)$ — is the result of the arithmetisation of $\vdash_{\mathbf{F}} \varphi$. PR_F is an arithmetical predicate. The previous syntactical expressions — both metatheoretical and arithmetical — are schematic, in that they contain metatheoretical signs such as the metavariable φ. But,

since, as we have seen, it is necessary to generalise over the propositions and other syntactical elements, the arithmetisation of the syntactical expressions must be complete. For this reason, syntactical predicates must act on variables that belong to the arithmetic language. Therefore, we have, for example, the predicate $PR_F(x)$ for the derivability in **F**, and so on. The primitive predicates T and K act on numerical variables, too. When useful, we will switch smoothly from full to schematic arithmetisation.

Group I: *Arithmetic axioms.* **PA** axioms with induction extended to predicates T and K.

Group II: *Axioms of determinateness*

$$D(\ulcorner\varphi\urcorner) \;:=\; T(\ulcorner\varphi\urcorner) \vee T(\ulcorner\neg\varphi\urcorner)$$

D1: $\vdash_{\mathbf{DTK}} \forall x(\mathit{At\text{-}Sent}_{\mathbf{PA}}(x) \to D(x))$
D2: $\vdash_{\mathbf{DTK}} \forall x(\mathit{Sent}(x) \to (D(\dot{\neg}x) \leftrightarrow D(x)))$
D3: $\vdash_{\mathbf{DTK}} \forall x\forall y(\mathit{Sent}(x) \wedge \mathit{Sent}(y) \to (D(x \dot{\vee} y) \leftrightarrow D(x) \wedge D(y)))$
D4: $\vdash_{\mathbf{DTK}} \forall x\forall y(\mathit{Sent}(x) \wedge \mathit{Sent}(y) \to D(x \dot{\to} y) \leftrightarrow (D(x) \wedge (T(x) \to D(y))))$
D5: $\vdash_{\mathbf{DTK}} \forall x\forall z(Var(z) \wedge \mathit{Sent}((\dot{\forall}z)x) \to (D((\dot{\forall}z)x) \leftrightarrow \forall y D(x(\dot{\tilde{y}}/z))))$
D6: $\vdash_{\mathbf{DTK}} \forall x(D(\dot{T}(\dot{\tilde{x}})) \leftrightarrow D(x))$
D7: $\vdash_{\mathbf{DTK}} \forall x(D(\dot{K}(\dot{\tilde{x}})) \leftrightarrow D(x))$

Note 2. The axioms state the determinateness conditions of **DTK** language propositions. It is worth commenting on the form of **D5**, **D6** and **D7**. Let us first see the structure of the equivalence that appears at the end of **D5**. This is $D((\dot{\forall}z)x) \leftrightarrow \forall y D(x(\dot{\tilde{y}}/z))$. What does it mean? It lays down the determinateness conditions of a universal formula. A universal formula is determinate if and only if all its numerical instances are determinate. Of course, the equivalence antecedent only makes sense if $Var(z)$ and $\mathit{Sent}(x)$, i.e., if z is a code for a variable and x is a code for a proposition. These assumptions are made explicitly in the axiom. The axiom **D6** says that the truth of a formula is determinate under the specific condition that the formula itself is determinate. Similarly, **D7** says that the knowability of a formula is determinate if and only if this formula is determinate. For further details see [Feferman, 2008].

Group III: *Truth axioms.* They regulate the use of the type-free predicate T and its relations with K and D.

Tr1: For each atomic formula $R(x_1, ..., x_n)$ of L(**PA**):
$\vdash_{\mathbf{DTK}} \forall x_1, ..., \forall x_n (T(\dot{R}(\dot{\bar{x}}_1, ..., \dot{\bar{x}}_n)) \leftrightarrow R(x_1, ..., x_n))$
Tr2: $\vdash_{\mathbf{DTK}} \forall x(Sent(x) \wedge D(x) \rightarrow (T(\dot{\neg} x) \leftrightarrow \neg T(x)))$
Tr3: $\vdash_{\mathbf{DTK}} \forall x \forall y(Sent(x) \wedge Sent(y) \wedge D(x \dot{\vee} y)$
$\rightarrow (T(x \dot{\vee} y) \leftrightarrow T(x) \vee T(y)))$
Tr4: $\vdash_{\mathbf{DTK}} \forall x \forall y(Sent(x) \wedge Sent(y) \wedge D(x \dot{\rightarrow} y) \rightarrow$
$(T(x \dot{\rightarrow} y) \leftrightarrow T(x) \rightarrow T(y)))$
Tr5: $\vdash_{\mathbf{DTK}} \forall x \forall z(Var(z) \wedge Sent((\dot{\forall} z)x) \wedge D((\dot{\forall} z)x)$
$\rightarrow (T((\dot{\forall} z)x) \leftrightarrow \forall y(T(x(\dot{\bar{y}}/z))))$
Tr6: $\vdash_{\mathbf{DTK}} \forall x(D(x) \rightarrow (T(\dot{T}(\dot{\bar{x}})) \leftrightarrow T(x)))$
Tr7: $\vdash_{\mathbf{DTK}} \forall x(D(x) \rightarrow (T(\dot{K}(\dot{\bar{x}})) \leftrightarrow K(x)))$

Note 3. Alongside the classical axioms of truth, we find the analogues of the last three determinateness axioms. **Tr5** states that a general formula is true if and only if all its instances are true and if it is also a determinate proposition. It is therefore not enough for all instances to be true: the generalisation result itself must be determinate. **Tr6** states that if a determinate formula is true, then it is true that it is true and vice versa. **Tr7** declares that, under the condition of determinateness, the knowability of a proposition is equivalent to the truth that it is knowable.

Group IV: *Knowledge rules and axioms.* They regulate the use of the predicate K, alone and in its relations with T and D.

K1: $\vdash_{\mathbf{DTK}} \forall x(Sent(x) \rightarrow (K(x) \rightarrow T(x)))$
K2: $\vdash_{\mathbf{DTK}} \forall x \forall y(Sent(x) \wedge Sent(y) \rightarrow (K(x \dot{\rightarrow} y) \wedge K(x) \rightarrow K(y)))$
K3: $\vdash_{\mathbf{DTK}} \forall x(Sent(x) \rightarrow (K(x) \rightarrow (K(\dot{K}(\dot{\bar{x}})))$
K-intro rule: $\vdash_{\mathbf{DTK}} \varphi \wedge D(\ulcorner \varphi \urcorner) \Rightarrow \vdash_{\mathbf{DTK}} K(\ulcorner \varphi \urcorner)$
T-intro rule: $\vdash_{\mathbf{DTK}} \varphi \wedge D(\ulcorner \varphi \urcorner) \Rightarrow \vdash_{\mathbf{DTK}} T(\ulcorner \varphi \urcorner)$.

Note 4. The **T-intro** rule is not a primitive rule. It can proved as follows:

$\vdash_{\mathbf{DTK}} \varphi \wedge D(\ulcorner \varphi \urcorner)$	hypothesis
$\vdash_{\mathbf{DTK}} K(\ulcorner \varphi \urcorner)$	**K-intro**
$\vdash_{\mathbf{DTK}} \forall x(Sent(x) \rightarrow (K(x) \rightarrow T(x)))$	**K1**
$\vdash_{\mathbf{DTK}} T(\ulcorner \varphi \urcorner)$	logic

Note 5. The previous three axioms, together with the next two rules, characterise the predicate of knowability. Axiom **K1** — also called the **K-T**-axiom — states that the predicate of knowledge is correct; that is, that truth is a necessary condition of knowledge. It appears to be a

particularly strong axiom. Nevertheless, it follows from the very notion of knowledge (that knowledge is factive). On the other hand, it should be noted that **K1** does not by itself imply the truth of knowledge. According to **T-intro** rule, a known proposition is true only if it is determinate. More radically, according to the **K-intro** rule, the assignment of K to a proposition is not a consequence of its derivability in **DTK** alone but also of the fact that this proposition is provably determinate. Axiom **K2** expresses the closure of the notion of knowledge with respect to implication. Axiom **K3** states that the predicate of knowledge is reflexive.

2.3. Some basic theorems

The following theorems are derivable in **DTK**:[13]
T1 (K-out): $\forall\varphi \in \mathrm{L}(\mathbf{DTK}) \vdash_{\mathbf{DTK}} K(\ulcorner\varphi\urcorner) \to \varphi$
T2 (T-out): $\forall\varphi \in \mathrm{L}(\mathbf{DTK}) \vdash_{\mathbf{DTK}} T(\ulcorner\varphi\urcorner) \to \varphi$
T3 (T-in): $\forall\varphi \in \mathrm{L}(\mathbf{DTK}) \vdash_{\mathbf{DTK}} D(\ulcorner\varphi\urcorner) \to (\varphi \to T(\ulcorner\varphi\urcorner))$
T4: $\forall\varphi \in \mathrm{L}(\mathbf{DTK}) \vdash_{\mathbf{DTK}} D(\ulcorner\varphi\urcorner) \to (T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi)$
T5: $\forall\varphi \in \mathrm{L}(\mathbf{PA}) \vdash_{\mathbf{DTK}} D(\ulcorner\varphi\urcorner)$
T6: $\vdash_{\mathbf{PA}} \varphi \Rightarrow \vdash_{\mathbf{DTK}} K(\ulcorner\varphi\urcorner)$

2.4. Formalisation of Penrose's argument in DTK

Definition 1. $K =_{\mathbf{PA}} F\ K =_{\mathbf{PA}} F := \forall x(Sent_{\mathbf{PA}}(x) \to (K(x) \leftrightarrow F(x)))$ in short: $\forall\varphi(K(\ulcorner\varphi\urcorner) \leftrightarrow F(\ulcorner\varphi\urcorner))$, where φ is an arithmetical proposition

Note 6. Note the difference in comparison to the definition of $K = F$ in $\mathbf{EA^m}$. Here, the definition is expressed in the formal language of DTK. This is because, within the DTK system, the predicates of truth and knowledge must be attributed to the formula $K =_{\mathbf{PA}} F$. This is only possible if we can refer to the formula through its arithmetisation code, which presupposes its formality.

Note 7. The definition of $K =_{\mathbf{PA}} F$ is restricted to L(**PA**) formulas. This restriction concerns the extension of the quantifiers' range of action contained in the formula. This range of action is restricted to arithmetic formulas only. The reason for this restriction lies in the need to work with determinate propositions, of which the arithmetic propositions are

[13] For the proof see [Koellner, 2016, pp. 170–172].

examples. We will see in short why the free use of propositional variables can generate indeterminate propositions.

Note 8. The expression $\forall\varphi(K(\ulcorner\varphi\urcorner) \leftrightarrow F(\ulcorner\varphi\urcorner))$ (where φ is an arithmetical proposition) is a convenient abbreviation of the defining formula, obtained through the use of schematic propositional metavariables. Propositional metavariables will also be used later in order to shorten the proofs. It should be stressed, however, that this is simply a convenient expedient. It is useful from an intuitive point of view, but does not exempt us from providing an explicit proof in a fully formalised language.

Definition 2. Soundness

$Sound\ K\ :=\ \forall x(Sent_{\mathbf{PA}}(x) \rightarrow (K(x) \rightarrow T(x)))$
in short: $\forall\varphi(K(\ulcorner\varphi\urcorner) \rightarrow T(\ulcorner\varphi\urcorner))$, where φ is an arithmetical proposition

$Sound\ F\ :=\ \forall x(Sent_{\mathbf{PA}}(x) \rightarrow (F(x) \rightarrow T(x)))$
in short: $\forall\varphi(F(\ulcorner\varphi\urcorner) \rightarrow T(\ulcorner\varphi\urcorner))$, where φ is an arithmetical proposition

Note 9. It is also, in this regard, worth reflecting on the nature of the given definition of soundness. The observations made in the previous Note 1 are also valid for the definition of soundness. But we can also learn something else from reflecting on the soundness definition. This allows us to understand why the truth predicate is indispensable. Suppose we want to define the notion of the soundness of a formal system **F**, for example, through the simplest and entirely plausible expression $\forall\varphi(F(\ulcorner\varphi\urcorner) \rightarrow \varphi)$. Now, the definition should be formalised, anyway, for the reasons set out above. Therefore, it should have the following form: $\forall x(Sent_{\mathbf{PA}}(x) \rightarrow (PR_F(x) \rightarrow x)$. However, this expression is nonsense, since x is a variable for numbers and not a formula. A obligatory outcome is to introduce the truth predicate T and write: $\forall x(Sent_{\mathbf{PA}}(x) \rightarrow (PR_F(x) \rightarrow T(x))$. This step, however, is fraught with significant consequences. It is necessary not only to introduce a truth predicate, but it also must be type-free. As we will see, in fact, in steps (2) and (5), respectively, of Penrose's argument, it is essential to make use of the rules of T introduction — **T-in** — and K introduction — **K-intro**. But the joint use of the two rules implies the full self-applicability of the truth predicate T [see Koellner, 2016, p. 165]. Therefore, not only does T have an application with respect to K, but it also has a self-application with respect to T. The self-applicability of T makes it necessary for T to

be a type-free predicate, which, in turn, implies the introduction of the conditions of determinateness envisaged in **DTK** to avoid the antinomic phenomenon of the liar paradox. This set of elements is the reason why in the passages in which the rules require that the propositions involved are determinate, the satisfaction of these conditions is decisive.

Note 10. The definition of soundness is restricted to L(**PA**) formulas, too. As before, the reason for this restriction is the need for propositions that are determinate, of which the arithmetic propositions are an example.

Let us now move on to the proof of the argument, which is divided into eight steps. In demonstrations we often use the abbreviation «syntax» for an arithmetised syntactical proposition.

Step (1) $K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \mathit{Sound}\ F$, i.e.
$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \forall z(\mathit{Sent}_{\mathbf{PA}}(z) \to (F(z) \to T(z)))$

The proof does not present any critical aspect, as no passage requires the proof of the formulas' determinateness:

$\mathit{Sent}_{\mathbf{PA}}(z) \to (K(z) \leftrightarrow F(z))\ \ \mathit{Sent}_{\mathbf{PA}}(z) \vdash_{\mathbf{DTK}} F(z) \to K(z)$ logic
$\mathit{Sent}(z) \vdash_{\mathbf{DTK}} K(z) \to T(z)$ from **K1**
$\mathit{Sent}_{\mathbf{PA}}(z) \vdash_{\mathbf{DTK}} \mathit{Sent}(z)$ syntax
$\mathit{Sent}_{\mathbf{PA}}(z) \to (K(z) \leftrightarrow F(z)),\ \mathit{Sent}_{\mathbf{PA}}(z) \vdash_{\mathbf{DTK}} F(z) \to T(z)$ chain
$\forall z(\mathit{Sent}_{\mathbf{PA}}(z) \to (K(z) \leftrightarrow F(z))) \vdash_{\mathbf{DTK}} \forall z(\mathit{Sent}_{\mathbf{PA}}(z) \to (F(z) \to T(z)))$ logic

Step (2) $K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \mathit{Sound}\ F_+$
where: $F_+ = F + (K =_{\mathbf{PA}} F)$

The second step proof presents a critical point. The **T-in** rule — which performs the same function as **T-intro** — requires, in fact, the determinateness of the formula to which the predicate of truth is attributed. The determinateness proof of the formula concerned is provided through Lemma 1 at the end of the step proof.

$K =_{\mathbf{PA}} F, \mathit{Sent}_{\mathbf{PA}}(\ulcorner F(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y)\urcorner), F(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y) \vdash_{\mathbf{DTK}} K(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y)$ logic
$\vdash_{\mathbf{DTK}} \mathit{Sent}_{\mathbf{PA}}(\ulcorner F(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y)\urcorner)$ syntax by arithmeticity of $PR_F(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y)$
$K(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y), \vdash_{\mathbf{DTK}} T(\ulcorner K = F\urcorner \dot{\to} y)$ **K1**
$K =_{\mathbf{PA}} F,\ F(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y), \vdash_{\mathbf{DTK}} T(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y)$ logic
$T(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y),\ D(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y) \vdash_{\mathbf{DTK}} T(\ulcorner K =_{\mathbf{PA}} F\urcorner) \to T(y)$ **Tr4**
$K =_{\mathbf{PA}} F,\ F(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y),\ D(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y) \vdash_{\mathbf{DTK}} T(\ulcorner K =_{\mathbf{PA}} F\urcorner) \to T(y)$ logic
$D(\ulcorner K =_{\mathbf{PA}} F\urcorner), D(y), \vdash_{\mathbf{DTK}} D(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\to} y)$ **D4**
$\mathit{Sent}_{\mathbf{PA}}(y)\ \ \vdash_{\mathbf{DTK}} D(y)$ **T5**

$D(\ulcorner K =_{\mathbf{PA}} F\urcorner),\ Sent_{\mathbf{PA}}(y),\ \vdash_{\mathbf{DTK}} D(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\rightarrow} y)$	logic
$K =_{\mathbf{PA}} F, Sent_{\mathbf{PA}}(y), F(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\rightarrow} y), D(\ulcorner K =_{\mathbf{PA}} F\urcorner) \vdash_{\mathbf{DTK}} T(\ulcorner K =_{\mathbf{PA}} F\urcorner) \rightarrow T(y)$	logic
$\vdash_{\mathbf{DTK}} D(\ulcorner K =_{\mathbf{PA}} F\urcorner)$	Lemma 1
$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} T(\ulcorner K =_{\mathbf{PA}} F\urcorner)$	**T-in**
$K =_{\mathbf{PA}} F,\ F(\ulcorner K =_{\mathbf{PA}} F\urcorner \dot{\rightarrow} y),\ Sent_{\mathbf{PA}}(y) \vdash_{\mathbf{DTK}} T(y)$	logic
$K =_{\mathbf{PA}} F,\ F_+(y),\ Sent_{\mathbf{PA}}(y) \vdash_{\mathbf{DTK}} T(y)$	def. F_+
$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \forall y(Sent_{\mathbf{PA}}(y) \rightarrow (F_+(y) \rightarrow T(y)))$	logic
$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} Sound\ F_+$	def. soundness

We now have to prove the determinateness of $K =_{\mathbf{PA}} F$.

LEMMA 1. $\vdash_{\mathbf{DTK}} D(\ulcorner K =_{\mathbf{PA}} F\urcorner)$, *i.e.*,

$\vdash_{\mathbf{DTK}} D(\ulcorner \forall z(Sent_{\mathbf{PA}}(z) \rightarrow (K(z) \leftrightarrow F(z))\urcorner)$

PROOF. Firstly, for the "$\rightarrow$"-part we prove:

$\vdash_{\mathbf{DTK}} \forall y(Sent_{\mathbf{PA}}(y) \rightarrow D(\ulcorner K(\dot{y}/z) \rightarrow F(\dot{y}/z)\urcorner))$

$Sent(\ulcorner K(\dot{y}/z)\urcorner),\ Sent(\ulcorner F(\dot{y}/z)\urcorner),\ D(\ulcorner K(\dot{y}/z)\urcorner),$ $T(\ulcorner K(\dot{y}/z)\urcorner) \rightarrow D(\ulcorner F(\dot{y}/z\urcorner)) \vdash_{\mathbf{DTK}} D(\ulcorner K(\dot{y}/z) \rightarrow F(\dot{y}/z)\urcorner)$	**D4**
$Sent_{\mathbf{PA}}(y) \vdash_{\mathbf{DTK}} Sent(\ulcorner K(\dot{y}/z)\urcorner)$	syntax
$Sent_{\mathbf{PA}}(y) \vdash_{\mathbf{DTK}} Sent(\ulcorner F(\dot{y}/z)\urcorner)$	syntax
$\vdash_{\mathbf{DTK}} D(\ulcorner F(\dot{y}/z)\urcorner)$	**T5** for arithmeticity of $F(\dot{y}/z)$
$\vdash_{\mathbf{DTK}} T(\ulcorner K(\dot{y}/z)\urcorner) \rightarrow D(\ulcorner F(\dot{y}/z)\urcorner)$	logic
$\vdash_{\mathbf{DTK}} D(\ulcorner K(\dot{y}/z)\urcorner) \leftrightarrow D(y)$	**D7**
$Sent_{\mathbf{PA}}(y) \vdash_{\mathbf{DTK}} D(y)$	**T5** formalised
$Sent_{\mathbf{PA}}(y) \vdash_{\mathbf{DTK}} D(\ulcorner K(\dot{y}/z)\urcorner)$	logic
$Sent_{\mathbf{PA}}(y) \vdash_{\mathbf{DTK}} D(\ulcorner K(\dot{y}/z) \rightarrow F(\dot{y}/z)\urcorner)$	chain
$\vdash_{\mathbf{DTK}} \forall y(Sent_{\mathbf{PA}}(y) \rightarrow D(\ulcorner K(\dot{y}/z) \rightarrow F(\dot{y}/z)\urcorner))$	logic

In short, schematically, the thesis is as follows: $D(\ulcorner K(\ulcorner \varphi\urcorner) \rightarrow F(\ulcorner \varphi\urcorner)\urcorner)$. Now, the innermost formulas $K(\ulcorner \varphi\urcorner)$ and $F(\ulcorner \varphi\urcorner)$ are determinate. $F(\ulcorner \varphi\urcorner)$ is determinate because $F(\ulcorner \varphi\urcorner)$ is arithmetic. $K(\ulcorner \varphi\urcorner)$ is determinate in virtue of **D7** and because φ is arithmetic. Thus, we have the result:

$Sent(\ulcorner Sent_{\mathbf{PA}}(\dot{y}/z)\urcorner),\ Sent(\ulcorner K(\dot{y}/z) \rightarrow F(\dot{y}/z)\urcorner),\ D(\ulcorner Sent_{\mathbf{PA}}(\dot{y}/z)\urcorner),$ $T(\ulcorner Sent_{\mathbf{PA}}(\dot{y}/z)\urcorner) \rightarrow D(\ulcorner K(\dot{y}/z) \rightarrow F(\dot{y}/z)\urcorner), \vdash_{\mathbf{DTK}}$ $D(\ulcorner Sent_{\mathbf{PA}}(\dot{y}/z) \rightarrow (K(\dot{y}/z) \rightarrow F(\dot{y}/z))\urcorner)$	**D4**
$\vdash_{\mathbf{DTK}} Sent(\ulcorner Sent_{\mathbf{PA}}(\dot{y}/z)\urcorner)$	syntax
$\vdash_{\mathbf{DTK}} D(\ulcorner Sent_{\mathbf{PA}}(\dot{y}/z)\urcorner)$	**T5** for arithmeticity of $Sent_{\mathbf{PA}}(\dot{y}/z)$
$\vdash_{\mathbf{DTK}} Sent(\ulcorner K(\dot{y}/z) \rightarrow F(\dot{y}/z)\urcorner)$	syntax
$T(\ulcorner Sent_{\mathbf{PA}}(\dot{y}/z)\urcorner) \rightarrow D(\ulcorner K(\dot{y}/z) \rightarrow F(\dot{y}/z)\urcorner) \vdash_{\mathbf{DTK}}$ $D(\ulcorner Sent_{\mathbf{PA}}(\dot{y}/z) \rightarrow (K(\dot{y}/z) \rightarrow F(\dot{y}/z))\urcorner)$	chain
$\vdash_{\mathbf{DTK}} Sent_{\mathbf{PA}}(y) \rightarrow D(\ulcorner K(\dot{y}/z) \rightarrow F(\dot{y}/z)\urcorner)$	
$\vdash_{\mathbf{DTK}} Sent_{\mathbf{PA}}(y) \leftrightarrow T(\ulcorner Sent_{\mathbf{PA}}(\dot{y}/z)\urcorner)$	reflexivity of T

$\vdash_{\mathbf{DTK}} T(\ulcorner Sent_{\mathbf{PA}}(\dot{\dot{y}}/z)\urcorner) \to D(\ulcorner K(\dot{\dot{y}}/z) \to F(\dot{\dot{y}}/z)\urcorner)$ logic
$\vdash_{\mathbf{DTK}} \forall y D(\ulcorner Sent_{\mathbf{PA}}(\dot{\dot{y}}/z) \to (K(\dot{\dot{y}}/z) \to F(\dot{\dot{y}}/z))\urcorner)$ chain and $I\forall$
$Sent(\ulcorner \forall z(Sent_{\mathbf{PA}}(z) \to (K(z) \to F(z)))\urcorner),$
$\forall y D(\ulcorner Sent_{\mathbf{PA}}(\dot{\dot{y}}/z) \to (K(\dot{\dot{y}}/z) \to F(\dot{\dot{y}}/z))\urcorner) \vdash_{\mathbf{DTK}}$
$D(\ulcorner \forall z(Sent_{\mathbf{PA}}(z) \to (K(z) \to F(z)))\urcorner)$ **D5**
$\vdash_{\mathbf{DTK}} Sent(\ulcorner \forall z(Sent_{\mathbf{PA}}(z) \to (K(z) \to F(z)))\urcorner)$ syntax
$\forall y D(\ulcorner Sent_{\mathbf{PA}}(\dot{\dot{y}}/z) \to (K(\dot{\dot{y}}/z) \to F(\dot{\dot{y}}/z))\urcorner) \vdash_{\mathbf{DTK}}$
$D(\ulcorner \forall z(Sent_{\mathbf{PA}}(z) \to (K(z) \to F(z)))\urcorner)$ chain
$\vdash_{\mathbf{DTK}} D(\ulcorner \forall z(Sent_{\mathbf{PA}}(z) \to (K(z) \to F(z)))\urcorner)$ chain

The proof of the "$\leftarrow$"-part:

$\vdash_{\mathbf{DTK}} D(\ulcorner \forall z(Sent_{\mathbf{PA}}(z) \to (F(z) \to K(z)))\urcorner)$

is a schematic variant of the first direction: just invert K with F. $\dashv$

Step (3) $K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} G(F_+)$
where: $G(F_+)$ is the Gödel sentence of F_+.

The proof is carried out in a schematic form. Remember that $\ulcorner \varphi \urcorner$ is used in this context as a metavariable for arithmetic formulas. There are no critical points, here, as the applied rules do not require the determinateness of the involved formulas.

$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \forall \varphi(F_+(\ulcorner \varphi \urcorner) \to T(\ulcorner \varphi \urcorner))$ from step (2)
$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} F_+(\ulcorner \bot \urcorner) \to T(\ulcorner \bot \urcorner)$ logic
$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} F_+(\ulcorner \bot \urcorner) \to \bot$ **T-out**
$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \neg F_+(\ulcorner \bot \urcorner)$ logic
$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} Cons(F_+)$ def. *Cons*
$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} G(F_+)$ G2

Step (4) $K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \neg F_+(\ulcorner G(F_+)\urcorner)$

The proof does not contain critical points:

$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} Cons(F_+)$ penultimate substep of step (3)
$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \neg F_+(\ulcorner G(F_+)\urcorner)$ G1

Step (5) $\vdash_{\mathbf{DTK}} K(\ulcorner K =_{\mathbf{PA}} F \to G(F_+)\urcorner)$

The proof presents a critical point: the **K-intro** rule is applied, which requires the determinateness of the formula, to which the K predicate is applied. The determinateness proof of the formula concerned is carried out through Lemma 2 at the end of the step proof.

1. $\vdash_{\mathbf{DTK}} K =_{\mathbf{PA}} F \to G(F_+)$ from step (3)
2. $\vdash_{\mathbf{DTK}} D(\ulcorner K =_{\mathbf{PA}} F \to G(F_+)\urcorner)$ Lemma 2
3. $\vdash_{\mathbf{DTK}} K(\ulcorner K =_{\mathbf{PA}} F \to G(F_+)\urcorner)$ **K-intro**

We now have to prove the determinateness of $K =_{\mathbf{PA}} F \to G(F_+)$:

LEMMA 2. $\vdash_{\mathbf{DTK}} D(\ulcorner K =_{\mathbf{PA}} F \to G(F_+)\urcorner)$

PROOF.

$Sent(\ulcorner K =_{\mathbf{PA}} F\urcorner)$, $Sent(\ulcorner G(F_+)\urcorner)$, $D(\ulcorner K =_{\mathbf{PA}} F\urcorner)$, $T(\ulcorner K =_{\mathbf{PA}} F\urcorner) \to D(\ulcorner G(F_+)\urcorner) \vdash_{\mathbf{DTK}} D(\ulcorner K =_{\mathbf{PA}} F \to G(F_+)\urcorner)$	**D4**
$\vdash_{\mathbf{DTK}} Sent(\ulcorner K =_{\mathbf{PA}} F\urcorner)$	syntax
$\vdash_{\mathbf{DTK}} Sent(\ulcorner G(F_+)\urcorner)$	syntax
$D(\ulcorner K =_{\mathbf{PA}} F\urcorner)$, $T(\ulcorner K =_{\mathbf{PA}} F\urcorner) \to D(\ulcorner G(F_+)\urcorner) \vdash_{\mathbf{DTK}} D(\ulcorner K =_{\mathbf{PA}} F \to G(F_+)\urcorner)$	chain
$\vdash_{\mathbf{DTK}} D(\ulcorner K =_{\mathbf{PA}} F\urcorner)$	Lemma 1
$\vdash_{\mathbf{DTK}} D(\ulcorner G(F_+)\urcorner)$	**T5** for arithmeticity of $G(F_+)$
$\vdash_{\mathbf{DTK}} T(\ulcorner K =_{\mathbf{PA}} F\urcorner) \to D(\ulcorner G(F_+)\urcorner)$	logic
$\vdash_{\mathbf{DTK}} D(\ulcorner K =_{\mathbf{PA}} F \to G(F_+)\urcorner)$	chain

In short: $G(F_+)$ is arithmetical and therefore determinate; as to $K =_{\mathbf{PA}} F$, we have already proved, through Lemma 1, that it is determinate, so the formula $K =_{\mathbf{PA}} F \to G(F_+)$ is itself determinate by virtue of **D4**. The application of **K-intro** is therefore legitimate. ⊣

Step (6) $K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \neg F(\ulcorner K =_{\mathbf{PA}} F \to G(F_+)\urcorner)$

The proof does not contain critical points.

$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \neg F_+(\ulcorner G(F_+)\urcorner)$	step (4)
$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \neg F(\ulcorner K = F \to G(F_+)\urcorner)$	def. of F_+

Step (7) $K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \exists\psi(\psi \equiv (K =_{\mathbf{PA}} F \to G(F_+)) \wedge K(\ulcorner\psi\urcorner) \wedge \neg F(\ulcorner\psi\urcorner))$

The proof does not contain critical points.

$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \neg F(\ulcorner K =_{\mathbf{PA}} F \to G(F_+)\urcorner)$	step (6)
$\vdash_{\mathbf{DTK}} K(\ulcorner K =_{\mathbf{PA}} F \to G(F_+)\urcorner)$	step (5)
$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \exists\psi(\psi \equiv (K =_{\mathbf{PA}} F \to G(F_+)) \wedge K(\ulcorner\psi\urcorner) \wedge \neg F(\ulcorner\psi\urcorner))$	logic

Step (8) $\vdash_{\mathbf{DTK}} K \neq F$, i.e., quantifying on F, $\vdash_{\mathbf{DTK}} \neg\exists F(K = F)$

Particular attention should be paid to the fact that $K = F$ means $\forall x(Sent(x) \to (K(x) \leftrightarrow F(x))$, i.e., K and F are equivalent with respect to all formulas belonging to L(**DTK**), whether they are purely arithmetic — $Sent_{\mathbf{PA}}(x)$ — or not purely arithmetic — $Sent_{\neg\mathbf{PA}}(x)$. This fact has a strong bearing on the meaning of step (8), which is discussed throughout the final part of this section.

Let us assume the proof of the following two simple lemmas:

LEMMA 3. $K = F \vdash_{\mathbf{DTK}} K =_{\mathbf{PA}} F$, *i.e.*,
$\forall x(Sent(x) \to (K(x) \leftrightarrow F(x)) \vdash_{\mathbf{DTK}} \forall x(Sent_{\mathbf{PA}}(x) \to (K(x) \leftrightarrow F(x))$

Lemma 4.

$$\exists x(Sent_{\neg \mathbf{PA}}(x) \wedge K(x) \wedge \neg F(x)) \vdash_{\mathbf{DTK}} \exists x(Sent(x) \wedge K(x) \wedge \neg F(x))$$

Proof of step (8) proceeds then as follows:

$K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \exists x(Sent_{\neg \mathbf{PA}}(x) \wedge K(x) \wedge \neg F(x))$	Step (7)
$K = F \vdash_{\mathbf{DTK}} \exists x(Sent_{\neg \mathbf{PA}}(x) \wedge K(x) \wedge \neg F(x))$	Lemma 3
$K = F \vdash_{\mathbf{DTK}} \exists x(Sent(x) \wedge K(x) \wedge \neg F(x))$	Lemma 4
$K = F \vdash_{\mathbf{DTK}} K \neq F$	logic
$\vdash_{\mathbf{DTK}} K \neq F$	self contradiction

Note 11. It is important to note that step (7) does not allow us to obtain by self-contradiction $\vdash_{\mathbf{DTK}} K \neq_{\mathbf{PA}} F$. Indeed, $K =_{\mathbf{PA}} F$ means $\forall \varphi(K(\ulcorner \varphi \urcorner) \leftrightarrow F(\ulcorner \varphi \urcorner))$, where φ varies on arithmetic propositions. By contrast, the sentence ψ — whose existence is asserted in the consequent of the step —, is not an arithmetic proposition, because it contains the knowledge predicate K. Hence, $\exists \psi(\psi \equiv (K = F \rightarrow G(F_+)) \wedge K(\ulcorner \psi \urcorner) \wedge \neg F(\ulcorner \psi \urcorner))$ cannot be the negation of $K =_{\mathbf{PA}} F$, since ψ is not an arithmetic proposition. For this reason, $K =_{\mathbf{PA}} F$ does not imply $K \neq_{\mathbf{PA}} F$, because, with regard to the arithmetical formulas, K and F coincide. And, as a consequence of this coincidence, the self-contradiction rule cannot be applied. Therefore, the proof of $\vdash_{\mathbf{DTK}} K \neq_{\mathbf{PA}} F$ is not successful. Instead, we obtained $\vdash_{\mathbf{DTK}} K \neq F$. The importance of this difference is explained in the next subsection.

2.5. Penrose's second argument is partially conclusive

There are two points to address.

1. Coherence of the result. Step (8), as proven by us, says: it is excluded that there exists a formalism capable of deriving all humanly knowable mathematical or non-mathematical propositions. However, we know from Koellner's Theorem 7.16.2 that:

$$\nvdash_{\mathbf{DTK}} \neg \exists F(K =_{\mathbf{PA}} F)$$

i.e., it is not excluded that there exists a formalism capable of deriving all humanly knowable mathematical propositions.

The two results may appear to contradict each other. But it is a wrong impression. On the contrary, the coexistence of both results makes it possible to clarify the purpose of Penrose's argument as it is

formalized in **DTK**. It is likely that Penrose's aim in his second argument was precisely to prove the thesis that there is no formalism capable of equalling human knowledge both about mathematical propositions and about non-mathematical propositions belonging to **DTK**'s language.

Now, according to Koellner's theorem it cannot be ruled out that, with respect to mathematical knowledge, the machine can equal the human mind, while on the basis of step (8) it is impossible that there exist a formalism that can generate every humanly knowable proposition. Koellner's theorem, thus, drops a part of Penrose's thesis, namely that according to which the human mind cannot be equalled by a formalism as far as mathematical propositions are concerned. However, Koellner's Lemma cannot say anything about the other part of Penrose's thesis; that is to say, it cannot exclude that for every formalism there is at least one known non-mathematical proposition that is not derivable in it. This is what is stated by step (8), which, while it cannot exclude the existence of a formalism capable of obtaining all knowable mathematical propositions, excludes the existence of a formalism capable of deriving all humanly knowable propositions. Koellner's theorem and step (8), then, match in that what the first does not rule out — i.e., that there are non-mathematical propositions that are not derivable — the second asserts, and what the first asserts — i.e., that it is possible for all arithmetical propositions to be formally derivable — the second does not rule out.

The counter evidence of the just outlined matching is obtained through reflection on the following fact, already pointed out in the previous note: from step (7) one cannot obtain by self contradiction $\vdash_{\mathbf{DTK}} K \neq_{\mathbf{PA}} F$. Nevertheless, we have $K =_{\mathbf{PA}} F \vdash_{\mathbf{DTK}} \exists x(Sent_{\neg \mathbf{PA}}(x) \wedge K(x) \wedge \neg F(x))$. Thus, it is true that we have not been able to prove that there is no formalism that is capable of deriving all the knowable arithmetical propositions. However, we have come to the conclusion that if K coincides with F as far as arithmetical formulas are concerned, then we know in **DTK** that there is a non-arithmetical ψ formula belonging to K that is determinate, true, and not derivable in F. Even in the hypothesis, therefore, that there is a formalism that captures all the knowable mathematical formulas, we can be sure that this formalism cannot capture all the knowable true formulas.

2. Determinateness of the result. A second critical aspect of the result obtained is given by the fact that Penrose's thesis, although derivable in **DTK**, is not determinate. Koellner's Theorem 7.15.1, in fact, states that

the proposition $\exists F(K = F)$ is provably indeterminate and, consequently, so is its negation $\neg\exists F(K = F)$. Of course the formula $\neg\exists F(K = F)$ is obtained in **DTK** without any illegitimate step, but in spite of this, being it indeterminate, it is not susceptible of being declared true in **DTK**. This difficulty can, however, be removed. Let us simply modify Step 8 as follows:

Step (8′) $\vdash_{\mathbf{DTK}} K =_{\mathbf{PA}} F \rightarrow$
$\exists x(Sent(x) \wedge x = \ulcorner K =_{\mathbf{PA}} F \rightarrow G(F_+)\urcorner \wedge K(x) \wedge \neg F(x))$

The proof is immediate by logic from step (7).

Now, by Lemma 2 and the axioms of determinateness, the implication is determinate, knowable by **K-intro**, and, therefore, true by **K1**. The different formal appearance with respect to step (8) does not change the meaning.

In conclusion, Penrose's second argument formalized in **DTK** system is partially conclusive since the **DTK** system proves that even in the hypothesis that there is a formalism that captures all the knowable mathematical formulas, the same formalism cannot capture all the knowable true formulas. Furthermore, the **DTK** system proves that this result is determinate.

3. Revisiting the justifiability of the first disjunct

3.1. Beyond DTK

The previous paragraph points out the partial validity of Penrose's argument. This result derives from a close analysis of the formal work carried out by Koellner, within detailed assumptions and formal constraints. In this section we will try to broaden the discussion by adopting the point of view of those who claim that thought is in principle irreducible to the functioning of a machine. We share the opinion that the difference between machine and mind does not only depend on the superiority of the mind with respect to the machine — the aim of Penrose's second argument —, but also on the different way in which mind and machine work. What we wish to emphasise is the difference between the machine's resources in doing concrete mathematics and the mind's resources, constitutively intertwined with formal and informal aspects. The difference in resources reveals an essential fracture in the way the mind proceeds compared to the machine.

3.2. Reasons against the identity of mind and machine

In this paragraph we will try to show that the mind differs from the machine for at least two fundamental reasons.

Reason 1: The machine cannot grasp the relationship between formal and informal kinds of reasoning.

Let us start from afar, by trying to understand the meaning of the limitation theorems, not from the systematic point of view — taken into consideration in the discussion of Gödel's disjunction — but from Hilbert's foundational perspective. From this point of view, the analysis of formal systems' limiting phenomena, starting from Gödel's theorems, has highlighted that human reasoning, being not entirely formalisable, presents intuitive, irreducibly informal aspects. In the course of gradual careful reflection on these aspects, they lose their residual character — the character they present at the beginning — and are taken as necessary presuppositions of formal reasoning. In principle, Hilbert himself recognise the need to take into consideration the informal dimension alongside the formal one.[14] When Hilbert speaks of the finitist foundation of the mathematical building, he means the consistency proof of formal mathematics, starting from the use of finitist procedures in a content-oriented way (*inhaltlich*). The consistency test, in this sense, has to be conducted in a system which has intuitive meaning. It is true that these procedures concern objects of the same type as formal objects — which are undoubtedly finitist — and that, precisely for this reason, the intuitive content-oriented aspect has to correspond perfectly to the formal one. However, in Hilbert's perspective, it is the intuitive content-oriented aspect that carries the foundational load. The finitist procedures appear founded not because they are formalisable but because they are finitistically evident. The evidence is, therefore, conceived by Hilbert as the essential factor of the epistemic dignity of real mathematics. Gödel's theorems sanction the failure of formalism not because it is not connected with some form of intuition, but because this form of intuition is too weak and limited to the concrete [Gödel, 1972]. On Hilbert's approach, only finitist procedures are evident, and all mathematics rests on them. Gödel breaks this closure because G2 expressly declares that the consistency proof of the ideal mathematics cannot be conducted through the finite procedures of real mathematics. Indeed, speaking more formally, G2 shows that it is

[14] To explore this topic in more detail see [Mancosu, 1998, 1999].

not possible to guarantee the consistency of formal systems satisfying very general conditions, not even of the PRA system, which — according to Tait's [1981] thesis — exactly formalises finitist arithmetic. How then can this guarantee be obtained?

All research following Gödel highlights the importance of the different forms of evidence: they are reflected in the hierarchy of arithmetic theories that lie above and below PRA. There is no single form of evidence — the finitist one — but, rather, the evidence is stratified. There are levels of evidence that are different according to gradually different degrees of content abstractness and different content types. These differences already appear with overwhelming clarity, arising from the fact that, by G2, no form of finitist foundationalism can be justified if finitist evidence is conceived as the only form of available evidence. Let us assume that according to Tait's thesis, they are precisely capturable by the **PRA** primitive recursive arithmetic system. This assumption means that all **PRA** theorems — and only they — are finitistically evident. On the other hand, according to the finitist's foundationalist perspective, only finitist evidence guarantees the truth — i.e., only finitist evidence is indisputably sound. But then, the finitist finds herself in severe difficulty. She must state that **PRA** is sound because all its theorems are true, and yet cannot say that the consistency (implied by soundness) of **PRA** is a finitist truth, because this is not derivable in **PRA**. Formally:

- Let the finitist evidence operator E_F be sound, i.e., $E_F\varphi \Rightarrow \varphi$
- Let Tait's thesis be valid, i.e., $\vdash_{\mathbf{PRA}} \varphi \Leftrightarrow E_F\varphi$.

Then we have:

$\vdash_{\mathbf{PRA}} \bot$	absurdum hypothesis
$E_F \bot$	by Tait's thesis
$E_F \bot \Rightarrow \bot$	soundness of E_F
$\bot$	logic
$\nvdash_{\mathbf{PRA}} \bot$	refutation

The finitist can reject Tait's thesis but not the soundness of the finitist evidence. For this reason, she must distinguish between evidence contents that can be formalised in **PRA**, and evidence contents that cannot be formalised in **PRA** — thus, admitting that there is reliable evidence outside **PRA**. But the finitist must admit that the content characterising such evidence external to **PRA** is not homogeneous with that formalisable in **PRA**. That is to say, that this content has abstract features

that are incompatible with the expressive abilities of **PRA** and which, therefore, is required to be learned independently of the linguistic sign.

In other words, the fact that **PRA** formalises at least a part of the finitist procedures, allows us to say, intuitively, that they are reliable. But if they are reliable, their use certainly cannot lead to a contradiction. Therefore, intuitively, we obtain with certainty that **PRA** is consistent, even if consistency is not obtainable through finitist procedures that are formalisable in **PRA**. The non-formal intuitive aspect, therefore, becomes relevant in the very proof of the consistency of a theory like **PRA**.[15]

Moreover, post-Gödelian developments have revealed that the various forms of evidence underlying the construction of mathematical knowledge are reflected in the hierarchy of arithmetic theories, starting from the weakest fragments such as **Q** up to **PA** at the second order, and set theory. The foundational theses differ by stating that reliability is limited to this or that other type of evidence or that it is spread over theories, up to a specific bound. For example, the finitist, *à* la Tait, affirms that finitist evidence is reflected in **PRA**; the finitist *à* la Parsons [1998] stops at $I\Delta_0$; the ultrafinitist Nelson [1986] even says that there is no certainty that **PA** is consistent, and declares that only the predicative arithmetic formalisable in **Q** is reliable. Some authors, like Gentzen [1969], on the other hand, push the bound upwards and believe that even induction up to ϵ_0 can be considered a finitist procedure. But it is not essential to adopt the point of view of a precise foundational thesis. Rather, the core issue is that, whatever the foundational thesis defended, this acceptance entails the acceptance of intuitive truths that do not belong to the set of claims justified according to that thesis. Knowledge, if it is such, is an open system, based on formal reasoning and informal intuition, and a mutual relationship between the two. And a machine, having no dimension beyond the formal one, cannot be this.

Reason 2: The machine only interacts with finitist data. The mind interacts with the abstract. Human evidence has a broader horizon than that to which the finitist evidence extends.

A further aspect touched on only implicitly in Reason 1, is the fact that the machine can only interact with a finitist language. This fact is

[15] Some authors have recently insisted on the importance of the informal dimension in the context of mathematical proof [see, e.g., Leitgeb, 2009].

an immediate consequence of the computational nature of the machine. A machine functions as a formal system, in the precise sense that the procedure through which the machine reaches a result can be perfectly represented by the set of steps that allow deriving the result within the formal system. Now, what kind of evidence is needed to build such a derivation? Well, what is needed is the ability to grasp the form of specific linguistic signs and to know how to combine these signs, appropriately, in accordance with the rules of formation and inference of the system. But linguistic signs are concrete objects, and the operations involved in the derivation of the result are, in turn, operations on concrete signs, which generate equally concrete properties and relationships. The formal system, therefore, works with, and on finite objects. No other forms of evidence are at stake. The computational machine is also equipped with the same form of evidence. And the machine proceeds based on the ascertainment of the presence of specific signs, and their elaboration in conformity with rules of composition and inferentiality. In conclusion, the machine cannot interact, except with finitist objects. As seen in our discussion of Reason 1, the human mind, however, has kinds of evidence that go beyond the purely finitist ones, whether these belong to the informal domain or to levels of formalisation different from the finitist one.

This does not mean that the machine cannot achieve results concerning abstract contents belonging to theories that go beyond finitist arithmetic. Importantly, these contents must be previously formalised within purely formal systems whose dominability does not require non-finitist forms of evidence. It is also worth pointing out that the considerations we make in Reason 2 concern machines that conform to the computational model of the mind — and are therefore isomorphic to formal systems — and do not claim to apply *sic et simpliciter* to biological machines such as the neural networks of a living organism.

Acknowledgments. Antonella Corradini likes to thank the Alexander von Humboldt Foundation for financing the research project "Inquiry on natural and artificial intelligence" carried out at the University of Konstanz from 15 July to 15 September 2019. Antonella Corradini would also like to thank the host professor, Prof. Dr. Wolfgang Spohn, who kindly encouraged and supported her during her research period.

Sergio Galvan thanks Dr. Carlo Nicolai for reading a previous version of this paper and for providing valuable suggestions.

References

Avron, A., 2020, "The problematic nature of Gödel's disjunctions and Lucas-Penrose's thesis", *Studia Semiotyczne* 34 (1): 83–08. DOI: 10.26333/sts.xxxiv1.05

Chalmers, D. J., 1995, "Minds, machines, and mathematics: A review of *Shadows of the Mind by Roger Penrose*", *Journal Psyche* 2 (June): 11–20.

Ebbinghaus, H. D, J. Flum and W. Thomas, 1984, *Mathematical Logic*, New York Berlin Heidelberg Tokyo: Springer Verlag.

Feferman, S., 1962, "Transfinite recursive progressions of axiomatic theories", *The Journal of Symbolic Logic* 27: 259–316. DOI: 10.2307/2964649

Feferman, S., 1995, "Penrose's Gödelian argument: A review of *Shadows of the Mind by Roger Penrose*", *Journal Psyche* 2 (May): 21–32.

Feferman, S., 2008, "Axioms for determinateness and truth", *The Review of Symbolic Logic* 1 (2): 204–217. DOI: 10.1017/S1755020308080209

Gentzen, G., 1969, "New version of the consistency proof for elementary number theory (1938)", pages 252–286 in M. E. Szabo (ed.), *The Collected Papers of Gerhard Gentzen*, Amsterdam: North-Holland.

Gödel, K., 1972, "On an extension of finitary mathematics which has not yet been used", pages 271–280 in S. Feferman et al. (eds.), *Collected Works*, Volume II: "Publications 1938–1974" (1990), New York: Oxford University Press.

Gödel, K., 1995, "Some basic theorems on the foundations of mathematics and their implications" (1951), pages 304–323 in S. Feferman et al. (eds.), *Collected Works*, Volume III: "Unpublished Essays and Lectures", New York: Oxford University Press.

Halbach, V., 2014, *Axiomatic Theories of Truth*, Cambridge: Cambridge University Press.

Koellner, P., 2016, "Gödel's disjunction", pages 148–188 in L. Horsten and P. Welch (eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge*, New York: Oxford University Press.

Koellner, P., 2018a, "On the question of whether the mind can be mechanized, I: From Gödel to Penrose", *The Journal of Philosophy* CXV, 7: 337–360. DOI: 10.5840/jphil2018115721

Koellner, P., 2018b, "On the question of whether the mind can be mechanized, II: Penrose's New Argument", *The Journal of Philosophy* CXV, 7: 453–484. DOI: 10.5840/jphil2018115926

Krajewski, S., 2020, "On the anti-mechanist arguments based on Gödel theorem", *Studia Semiotyczne* 34 (1): 9–56. DOI: 10.26333/sts.xxxiv1.02

Leitgeb, H., 2009, "On formal and informal provability", pages 263–299 in O. Bueno and Ø. Linnebo (eds.), *New Waves in Philosophy of Mathematics*, London New York: Palgrave Macmillan.

Lindström, P., 2001, "Penrose's new argument", *Journal of Philosophical Logic* 30 (3): 241–250. DOI: 10.1023/A:1017595530503

Lindström, P., 2006, "Remarks on Penrose's new argument", *Journal of Philosophical Logic* 35 (3): 231–237. DOI: 10.1007/s10992-005-9014-7

Lucas, J. R., 1961, "Minds, machines and Gödel, *Philosophy* 36: 112–127. DOI: 10.1017/S0031819100057983

Lucas, J. R., 1968, "Satan stultified: a rejoinder to Paul Benacerraf", *The Monist* 52: 145–158. DOI: 10.5840/monist196852111

Mancosu, P., 1998 (ed.), *From Brouwer to Hilbert. The Debate on the Foundations of Mathematics in the 1920s*, New York: Oxford University Press.

Mancosu, P., 1999, "Between Vienna and Berlin: the immediate reception of Gödel's incompleteness theorems", *History and Philosophy of Logic* 20: 33–45. DOI: 10.1080/014453499298174

Nelson, E., 1986, *Predicative Arithmetic*, Mathematical Notes 32, Princeton University Press: Princeton, New Jersey.

Penrose, R., 1989, *The Emperor's New Mind. Concerning Computers, Minds, and the Laws of Physics*, New York: Oxford University Press.

Penrose, R., 1994, *Shadows of the Mind. A Search for the Missing Science of Consciousness*, New York: Oxford University Press.

Penrose, R., 1996, "Beyond the doubting of a shadow. A reply to commentaries on *Shadows of the Mind*", *Journal Psyche* 2 (23): 1–40. https://www.calculemus.org/MathUniversalis/NS/10/01penrose.html

Parsons, C., 1998, "Finitism and intuitive knowledge", pages 249–270 in M. Schirn (ed.), *The Philosophy of Mathematics Today*, Clarendon Press: Oxford.

Shapiro, S., 1998, "Incompleteness, mechanism, and optimism", *Bulletin of Symbolic Logic* 4 (3): 273–302. DOI: 10.2307/421032

Shapiro, S., 2003, "Mechanism, truth, and Penrose's new argument", *Journal of Philosophical Logic* 32 (1): 19–42. DOI: 10.1023/A:1022863925321

Sundholm, G., 1983, "Systems of deduction", pages 133–188 in D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic*, Vol. I, "Elements of Classical Logic", Dordrecht/Boston/Lancaster: D. Reidel Publishing Company.

Tait, W., 1981, "Finitism", *The Journal of Philosophy* 78: 524–546, DOI: 10.2307/2026089

ANTONELLA CORRADINI
Department of Psychology
Universita Cattolica del Sacro Cuore Milano
Italy
antonella.corradini@unicatt.it

SERGIO GALVAN
Department of Philosophy
Universita Cattolica del Sacro Cuore Milano
Italy
sergio.galvan@unicatt.it