# Penrose's *New Argument* and Paradox*

Johannes Stern

**Abstract** In this paper we take a closer look at Penrose's *New Argument* for the claim that the human mind cannot be mechanized and investigate whether the argument can be formalized in a sound and coherent way using a theory of truth and absolute provability. Our findings are negative; we can show that there will be no consistent theory that allows for a formalization of Penrose's argument in a straightforward way. In a second step we consider Penrose's overall strategy for arguing for his view and provide a reasonable theory of truth and absolute provability in which this strategy leads to a sound argument for the claim that the human mind cannot be mechanized. However, we argue that the argument is intuitively implausible since it relies on a pathological feature of the proposed theory.

## 1 Introduction

Gödel's Incompleteness Theorems are beyond doubt amongst the greatest and most interesting results in mathematical logic of the 20th century and have attracted interest far beyond the frontiers of logic and mathematics. Gödel's theorems even found application outside of logic and mathematics, and, in fact, several non-mathematical propositions were claimed to be consequences of the incompleteness theorems. One particularly striking such proposition was the claim that *the human mind cannot be mechanized*, which was defended by several authors, most prominently Lucas (1961) and Penrose (1989, 1994, 1996). As a matter of fact the origins of this debate are with Gödel himself. Gödel in his celebrated Gibb's lecture, entitled *Some*

Johannes Stern
Department of Philosophy, University of Bristol, Cotham House, BS66JL Bristol, UK, e-mail: johannes.stern@bristol.ac.uk

* Published in "*Truth, Existence, and Explanation - FilMat Studies in the Philosophy of Mathematics.*, M. Piazza and G. Pulcini (eds), Boston Studies in the History and Philosophy of Science Vol. 334, Springer.

*basic theorems on the foundations of mathematics and their implications* (Gödel, 1995), explored some philosophical consequences of his incompleteness theorems. Famously, his reflections led him to a disjunctive conclusion, which is known as Gödel's Disjunction:

> Either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems (...). (Gödel, 1995, p. 310)

According to the first disjunct the set of mathematical theorems that the (idealized) human mind can produce, that is prove, cannot be recursively axiomatized, indeed, the set is not recursively enumerable and therefore cannot be the output of a Turing machine or any other effective algorithm. The second disjunct asserts that there are sentences of the mathematical language such that neither the sentence nor its negation are among the mathematical theorems the human mind can produce. If this is paired with a classical, that is, bivalent view according to which each sentence is either true or false, then there are sentences that are absolutely unprovable.[2] The notion of absolute provability we have just employed is meant to stand for some intuitive notion of proof that is supposed to capture the process in which the human mind produces mathematical theorems.

In our terminology Gödel's Disjunction (henceforth GD) thus asserts that *either the set of absolutely provable sentences cannot be recursively enumerated or there exists a true mathematical sentence which is not absolutely provable*. Indeed, putting aside worries concerning the obscure notion of absolute provability for the moment, GD follows from the incompleteness theorems if it is understood in this way. However, Gödel made it clear, that he did not think that either one of the disjuncts of the disjunction could be established solely by appeal to the incompleteness theorems. In contrast, Lucas, Penrose and other authors thought differently and argued that the incompleteness theorems imply the first disjunct of GD. In other words, Lucas and Penrose thought that they could refute mechanism, that is, the idea that the outputs of the human mind can be produced by an effective algorithm.[3] There are basically two arguments for the first disjunct of GD. The first argument was popularized by Lucas (1961), while the second argument, Penrose's so-called *New Argument*, was introduced by Penrose (1994, 1996).

Lucas argued that while any recursively enumerable system $F$ can never prove its Gödel sentence, that is, a sentence saying of itself that it is not provable in $F$, the human mind can always perceive it to be true: the Gödel sentence for $F$ is absolutely provable but not provable in $F$. According to Lucas this shows that the human mind

[2] In fact, assuming bivalence is not necessary to "derive" Gödel's Disjunction. It suffices to argue that there are certain sentences that are either true or false but which are absolutely unprovable. Gödel (1995) argues convincingly that this situation also arises if non-classical mathematics is assumed.

[3] Mechanism can also be understood as the stronger thesis that the human mind functions like some particular algorithm, rather than that that the human mind and some algorithm have the same outputs. We confine ourselves to the latter understanding and refer the reader to the Lindström (2001), Shapiro (2003, 2016) and Koellner (2016) for a discussion of this issue.

exceeds the notion of proof of every recursively axiomatizable mathematical theory. However, to draw this conclusion Lucas needs the additional assumption that the human mind can perceive that $F$ is consistent, i.e., that it is absolutely provable that $F$ is consistent. Otherwise, Lucas would have argued only for the conditional statement asserting that the Gödel sentence of $F$ is true, if $F$ is consistent. But this latter claim is provable in $F$ itself and therefore not sufficient for establishing the first disjunct of GD. In general, it seems to be an implausible assumption that the—even idealized—human mind can perceive the consistency (and inconsistency) of every recursively axiomatizable theory.[4] Moreover, as Koellner (2016) points out, this assumption is also problematic for more principled reasons because it implies that we are in possession of a $\Pi^0_1$-oracle, which is tantamount to asserting that the outputs of a human mind cannot be produced by a finite machine—but this is precisely what Lucas' argument was meant to show. All of this suggests that Lucas' argument does not make a convincing case for the first disjunct of GD.

In this paper we shall chiefly be concerned with Penrose's so-called *New Argument*. As we shall see Penrose added a further twist to the Lucas-style arguments for the first disjunct of GD by introducing some basic assumptions concerning the notion of absolute provability. This brings the obscure notion of absolute provability back into the limelight and one might reasonably expect Penrose to provide some clarification. After all without further explanation this notion remains entirely mysterious and opaque, and it is impossible to evaluate claims that are made on behalf of this notion. Moreover, by introducing specific assumptions on behalf of absolute provability he also provides his opponent with an opportunity for resisting such an argument for the first disjunct of GD. While it would be an unreasonable move to deny Gödel's incompleteness theorems, as they are part of core mathematics, it is perfectly acceptable to resist some particular assumption concerning absolute provability:[5] there is no generally agreed theory of absolute provability Penrose could rely on. However, in this paper we will not discuss these issues any further and refer the reader interested in a more thorough discussion of the notion of absolute provability to the work by Myhill (1960) and Leitgeb (2009).[6,7] Rather following the lead of Koellner (2016), we will pursue a charitable approach throughout this paper and grant Penrose the assumptions he makes concerning absolute provability. We shall query whether under Penrose's assumptions his *New Argument* is a coherent argument for the the first disjunct of GD.

---

[4] See also Benacerraf (1967) or Lindström (2001) for remarks along these lines.

[5] For remarks along these lines see, e.g., Shapiro (2003).

[6] Absolute provability and GD have also been studied in so-called *epistemic arithmetic*. See, for example, Shapiro (1985); Reinhardt (1986) and Carlson (2000).

[7] Ultimately, we are very skeptical whether anything interesting and coherent can be said on behalf of absolute provability but as indicated in the main text we bracket this worry for the purpose of this paper.

## 2 Penrose's New Argument

Penrose (1996) summarizes his *New Argument* in the following way, where $F$ is some recursively axiomatizable system

> Though I don't know that I am necessarily $F$, I conclude that if I were, then system $F$ would have to be sound, 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 "if I happened to be $F$, then $G(F')$ would have to be true", and perceptions of this nature is precisely what F is supposed to achieve. Since I am $F$, I deduce that I cannot be F after all. (Penrose, 1996)

Let us try to reconstruct the argument in an explicit way.[8] This will also help us to assess which assumptions concerning absolute provability are required for carrying out the argument. To start the reconstruction we need to introduce some additional terminology and to say a little bit more about what these *systems* are that Penrose alludes to in the argument. From now on we take a system to be a recursively enumerable set that is closed under modus ponens and extends Peano Arithmetic (PA).[9] We assume a system to be closed under modus ponens because modus ponens is the standard rule of proof in logic. This assumption thus guarantees that a system qua set of sentences does not deviate from the set of theorems that the system can prove.[10,11]

To reconstruct Penrose's argument we also need a sentential predicate $K$ that expresses absolute provability. The intended interpretation of this predicate is the set of mathematical theorems the human mind can produce. Finally we need to clarify what it means for a system $F$ to be sound and how this is expressed. A system $F$ is sound iff all its theorems are true. In the formal setting this can be expressed by the schemata

$$F\ulcorner\phi\urcorner \rightarrow \phi,$$

which holds for all sentences $\phi$ of the language. Throughout the paper $\ulcorner\phi\urcorner$ is the numeral of the Gödel number of $\phi$ and thus acts as name of $\phi$. Notice that by Gödel's second incompleteness theorem no recursively axiomatizable theory can prove its own soundness.

We now carry out the reconstruction in a step-by-step manner:

---

[8] See Shapiro (2003) for a very similar reconstruction of Penrose's argument.

[9] Both assumption are plausible and frequently assumed in the literature. See, e.g., Koellner (2016) for discussion. A weaker arithmetical theory would do as well as long as it is sufficient for proving Gödel's incompleteness theorems.

[10] Everything we say at this point would go through even if we do not assume a system to be closed under modus ponens. In this case we would need to distinguish throughout the argument between the system and what the system proves.

[11] Of course, there are logics in which modus ponens is not a sound rule of proof. But in such systems there will be alternative rules of proof and a similar problem will arise. Moreover, we would be surprised, if Penrose were to argue against modus ponens.

1. if I were [the system $F$], then system $F$ would have to be sound.[12]

$$\forall x(Fx \leftrightarrow Kx) \rightarrow (F\ulcorner\phi\urcorner \rightarrow \phi) \qquad \text{for all } \phi$$

2. $F'$ would have to be sound, where $F'$ is $F$ supplemented by the further assertion "I am $F$".

$$\forall x(Fx \leftrightarrow Kx) \rightarrow (F'\ulcorner\phi\urcorner \rightarrow \phi) \qquad \text{for all } \phi$$

3. I perceive that it follows from the assumption that I am $F$ that the Gödel statement $G(F')$ would have to be true

$$\forall x(Fx \leftrightarrow Kx) \rightarrow G(F')$$

4. furthermore, that it $[G(F')]$ would not be a consequence of $F'$

$$(i) \qquad \forall x(Fx \leftrightarrow Kx) \rightarrow \neg F'\ulcorner G(F')\urcorner$$
$$(ii) \qquad \forall x(Fx \leftrightarrow Kx) \rightarrow \neg F\ulcorner\forall x(Fx \leftrightarrow Kx) \rightarrow G(F')\urcorner$$

5. I have just perceived that if "if I happened to be $F$, then $G(F')$ would have to be true"

$$K\ulcorner\forall x(Fx \leftrightarrow Kx) \rightarrow G(F')\urcorner$$

6. Since I am $F$, I deduce that I cannot be $F$ after all

$$(i) \qquad \forall x(Fx \leftrightarrow Kx) \rightarrow F\ulcorner\forall x(Fx \leftrightarrow Kx) \rightarrow G(F')\urcorner$$
$$(ii) \qquad \forall x(Fx \leftrightarrow Kx) \rightarrow \bot.$$

To get the argument off the ground we need to assume that that notion of absolute provability is a sound notion: every sentence that is absolutely provable, i.e. every mathematical theorem the human mind produces, is true. In other words we need to assume for all sentences $\phi$ of the language:

$$(\mathsf{T}) \qquad K\ulcorner\phi\urcorner \rightarrow \phi.$$

If $(\mathsf{T})$ is assumed, Line 1 is a valid assumption and we may proceed from there. Line 2 then follows by weakening (and modus ponens). Notice that by the deduction theorem we have the following equivalence

$$(\dagger) \qquad F'(\ulcorner\phi\urcorner) \leftrightarrow F(\ulcorner\forall x(Fx \leftrightarrow Kx) \rightarrow \phi\urcorner),$$

which will be important later in the argument.

We obtain Line 3 by instantiating the schematic Line 2 to the Gödel sentence of $F'$. By the very construction of the Gödel sentence Line 3 is equivalent to Line 4(i)

[12] Throughout the paper we take $\forall x\phi$ to be short for $\forall x(\mathsf{Sent}(x) \rightarrow \phi)$ unless we explicitly mention the restriction of the quantifier. Sent is a predicate representing the set of sentences of the language.

and by (†) the latter is equivalent to 4(ii).[13] Line 5 requires a further assumption on behalf of absolute provability: if the human mind has produced a theorem, i.e., a sentence has been shown to be absolutely provable, then this fact itself is absolutely provable. This is nothing but the rule of necessitation for absolute provability:

$$(\mathsf{Nec}) \qquad \frac{\phi}{K\ulcorner\phi\urcorner}.$$

Given the rule of necessitation we obtain Line 5 from Line 3. By classical logic Line 5 implies Line 6(i) but Line 6(i) together with Line 4(ii) yields Line 6(ii), which establishes the first disjunct of GD, namely, that the mathematical theorems the human mind can produce do not coincide with the theorems of any recursively axiomatizable system. Notice that in order to draw this conclusion it was important that no special assumption concerning $F$ was made, since we need to use the rule of universal generalization in the metalanguage.[14] Thus assuming the principle $(\mathsf{T})$ and the rule $(\mathsf{Nec})$ each step of Penrose's *New Argument* is sound and consequently one might think that Penrose has succeeded in providing a coherent argument for the first disjunct of GD. Unfortunately though, the assumptions concerning absolute provability, that is the principle $(\mathsf{T})$ and the rule $(\mathsf{Nec})$, are jointly inconsistent as Myhill (1960) and Montague (1963) have taught us.[15]

**Theorem 1 (Myhill/Montague).** *Let $\Sigma$ be a theory extending Robinson Arithmetic such that for all sentences $\phi$ of the language:*

*(i)* $\Sigma \vdash K\ulcorner\phi\urcorner \to \phi$
*(ii)* $\Sigma \vdash \phi \Rightarrow \Sigma \vdash K\ulcorner\phi\urcorner$

*for some primitive or complex predicate $K$. Then $\Sigma$ is inconsistent.*

This inconsistency result points to a further complication in providing an argument for the first disjunct of GD: it is not sufficient to suggest plausible principles for absolute provability but one also has to guarantee that these principles are jointly consistent. Moreover, the result by Myhill and Montague is only one of many inconsistency results and it is not an easy task to provide a satisfactory account or theory of absolute provability.[16] For this reason the aforementioned inconsistency results are also known as intensional paradoxes.

Indeed, according to Wang (1996), Gödel thought that once we were in possession of a satisfactory solution to the intensional paradoxes, we could successfully establish the first disjunct, that is, we could show that the human mind surpasses any finite machine:

[13] The Gödel sentence $G(F')$ is a sentence for which, using the diagonal lemma, we can prove $\neg F'\ulcorner G(F')\urcorner \leftrightarrow G(F')$.

[14] By the argument we obtain $\neg\forall x(Fx \leftrightarrow Kx)$ and by universal generalization in the metalanguage $\forall F\neg\forall x(Fx \leftrightarrow Kx)$. This yields $\neg\exists F\forall x(Fx \leftrightarrow Kx)$: there exists no recursively axiomatizable system $F$ that coincides with the set of absolutely provable sentences.

[15] The fact that the principles Penrose appeals to in formulating his *New Argument* are jointly inconsistent was already pointed out by Chalmers (1995) and Shapiro (2003), amongst others.

[16] See Egré (2005) or (Stern, 2016, Chap. 3) for an overview of the various inconsistency results.

> If one could clear up the intensional paradoxes somehow, one would get a clear proof that mind is not [a] machine. (Wang, 1996, p. 187)

Even though it is perhaps safe to say that now, almost half a century later, we still lack an entirely satisfactory solution to the intensional paradoxes, it is interesting to apply the existing proposals to Penrose's *New Argument* and check whether we thereby obtain a coherent argument in favor of the first disjunct of GD. If this were possible, this would be a partial vindication of Penrose but also of Gödel's conviction that a satisfactory solution to the intensional paradoxes would lead to an argument for the first disjunct of GD.

## 3 Truth and Intensional Paradox

The intensional paradoxes, which affect notions such as proof, knowledge, or belief, have not received quite as much attention as the semantic paradoxes such as the Liar paradox. But at least from a technical point of view the intensional and the semantic paradoxes are closely related. In both cases paradox arises due to the application of these notions to themselves. As a consequence, in both cases we have roughly the same options how paradox can be avoided. The first option is to restrict the characteristic schemes of these notions to a salient set of sentences that in some way or another singles out the "good" instances from the "bad" ones in such a way that paradox can no longer arise. For example, in the case of absolute provability this would amount to restricting the scheme $(\mathsf{T})$ and the rule $(\mathsf{Nec})$ to specific sentences so that Montague's theorem no longer applies. The second, alternative option would be to adopt weaker principles or schemata for these self-applicable notions; principles that even if applied in full generality are jointly consistent.

This latter option of adopting weaker schemata does not seem to be a promising option if one is interested in resurrecting Penrose's *New Argument*. As we have seen, we need $(\mathsf{T})$ and $(\mathsf{Nec})$ in order to carry out Penrose's reasoning and, at least as long as no further resources are added, weaker principles will simply not do. So we are left with the former option of restricting the scope of the schemata. However, as far as we can see there is no obvious restriction that would enable us to vindicate Penrose's *New Argument*. The most immediate one that comes to mind is to restrict the schemata to arithmetical sentences. But this restriction would make the transition from Line 1 to Line 2 and from Line 4 to Line 5 of the the argument unsound.[17] A further and maybe more plausible strategy might be to divide the sentences of the language into the paradoxical and the non-paradoxical ones. Unfortunately, this distinction has proven rather elusive in the past and difficult to pin down. Moreover, Penrose needs the classification, or at least a principled sub-classification thereof, to be decidable for otherwise the reasoning leading to the first disjunct could not

[17] Introducing a hierarchy of typed absolute provability predicates would not be of any help here. Rather at each level we would face the problem anew and could never draw the desired conclusion. See Shapiro (2003) for remarks along these lines.

be carried out in a recursively axiomatizable system.[18] This would undermine the reductio-strategy of his argument, which is based on the assumption that the human mind is recursively axiomatizable. As a consequences the chances of finding an acceptable restriction become even smaller since the more promising classifications proposed in the literature are not decidable, indeed they are of much greater complexity. Now, even if, against all odds, we manage to restrict the schemata in a recursive way that divides the sentences of the language into the "good", non-paradoxical and the "bad", paradoxical ones, we still need to make sure that the sentence $(\mathsf{Nec})$ is applied to in Line 5 is such a "good" case. Unfortunately, the work by Koellner (2016) shows that this will not generally be the case.

These remarks suggest that the usual options for dealing with the paradoxes of self-applicable notions may not yield a formal framework in which the first disjunct of GD can be proved along the lines of Penrose's *New Argument*. Fortunately, there is a further option that arises for certain intensional and semantic notions. The idea is that at the root of all the paradoxes there is only one paradoxical notion, namely the notion of truth. This idea would tie the intensional paradoxes to the semantic, i.e. truth-theoretic, paradoxes in such a way that the paradoxicality of the intensional notion, e.g. absolute provability, depends solely on the paradoxicality of the notion of truth. The central idea of the strategy is to formulate the principles of absolute provability using the truth predicate. For example, the principle $(\mathsf{T})$ we have appealed to in reconstructing Penrose's reasoning would then be formulated as

$$(\mathsf{T}_\mathsf{K}) \qquad \forall x(Kx \to Tx).$$

If these revised principles of absolute provability are combined with a consistent theory of truth, we obtain a consistent theory of truth and absolute provability. Such a strategy has been recently developed by Stern (2014a,b, 2016) and, independently, Koellner (2016). Moreover, Koellner's work is directly motivated by Gödel's Disjunction and the evaluation of Penrose's *New Argument*. To implement this strategy we need to introduce a truth predicate to the framework that is allowed to interact with the absolute provability predicate. But Penrose's *New Argument* makes explicit use of the notion of truth at several places[19] and thus an evaluation of his argument in a framework where we have both, an absolute provability predicate *and* a truth predicate, seems highly desirable and independently motivated. In the remainder of this paper we shall therefore adopt this framework and investigate whether in such a framework we can provide a coherent argument for the first disjunct of GD.

[18] The restriction has to be decidable only if we assume the schemata to be characteristic of the notion of absolute provability, that is, if they are assumed to be the axioms of the recursively axiomatizable system. Otherwise, the restriction could be recursively enumerable.

[19] See Line 1, 2, 3 and 5.

### *3.1 Penrose's* New Argument *Reconsidered*

Koellner (2016) extracts three principle of truth and absolute provability he takes to be crucial for reconstructing Penrose's argument in a language with a truth and an absolute provability predicate. These principles are the rule (Nec) and the principles

| | |
|---|---|
| $(\mathsf{T_K})$ | $\forall x(K(x) \to Tx)$ |
| $(T\text{-}\mathsf{In})$ | $\phi \to T\ulcorner\phi\urcorner$ |

Using these principles Penrose's argument can be formalized, roughly following Koellner (2016), as follows:

| | |
|---|---|
| 1. $\forall x(Fx \leftrightarrow Kx) \to \forall x(Fx \to Tx)$ | by $(\mathsf{T_K})$ |
| 2. $\forall x(Fx \leftrightarrow Kx) \to \forall x(F'x \to Tx)$ | by $(T\text{-}\mathsf{In})$ |
| 3. $\forall x(Fx \leftrightarrow Kx) \to G(F')$ | by 2 and definition of $G(F')$ |
| 4. $\forall x(Fx \leftrightarrow Kx) \to \neg F'^{\ulcorner}G(F')^{\urcorner}$ | by definition of $G(F')$ |
| 5. $K^{\ulcorner}\forall x(Fx \leftrightarrow Kx) \to G(F')^{\urcorner}$ | 4, by (Nec) |
| 6. $\forall x(Fx \leftrightarrow Kx) \to \neg F^{\ulcorner}\forall x(Fx \leftrightarrow Kx) \to G(F')^{\urcorner}$ | 4, by definition of $F'$. |
| 7. $\forall x(Fx \leftrightarrow Kx) \to F^{\ulcorner}\forall x(Fx \leftrightarrow Kx) \to G(F')^{\urcorner}$ | 5 |
| 8. $\forall x(Fx \leftrightarrow Kx) \to \bot$ | 6, 7. |

Shortly, we will take a closer look at this reconstruction and investigate whether the different steps of the argument are sound, given the principles of truth and absolute provability Koellner puts forth. But before we do so, it is worth pointing out that even in the presence of a truth predicate we need a self-applicable absolute provability predicate to carry out the reasoning. For example, in Line 5 the absolute provability predicate is applied to a sentence containing the predicate itself. What about the truth predicate; need it be self-applicable? On the face of it at no point in the argument is the truth predicate applied to a sentence in which it explicitly occurs and, consequently, one might think that a typed truth predicate, that is, a truth predicate that can only be applied to sentences in which the truth predicate does not occur, will be sufficient for the argument. Eventually we will have a look at the possibility of carrying out Penrose's *New Argument* within a typed theory of truth but ultimately opting for a typed truth predicate is a cheat. After all in the present set up the (idealized) human mind is meant to reflect about its proofs and capacities. For example, the human mind is supposed to perceive that it is sound, that is, the principle $(\mathsf{T_K})$ should be absolutely provable.[20] But then it follows from $(\mathsf{T_K})$ itself that we may apply the truth predicate to sentences in which it occurs.[21] Therefore, it is

[20] Penrose (1996) explicitly agrees with this claim.

[21] It is possible to block this conclusion by restricting the principle $(\mathsf{T_K})$ to sentences of the language without the truth predicate. However, this would not really affect the argument: after all we would have sentences in which the truth predicate is applied to sentences which the knowledge predicate occurs. The knowledge predicate, in turn, may be applied to sentences in which the truth predicate occurs. So the truth predicate is applied to sentences in which it implicitly occurs. In order to maintain a coherent picture it should be possible to apply the truth predicate to sentences containing itself.

essentially right when Koellner (2016) concludes that *"any formal system in which the above argument can be implemented will be one involving a type-free theory of K and a type-free theory of T*". This, of course, leads to the question whether there are such theories of truth and absolute provability in which Penrose's *New Argument* or, more generally, arguments for the first disjunct of GD, can be carried out. Koellner (2016) seems to be skeptical in this respect but detects a general problem for establishing such a negative conclusion:[22]

> It would be ideal if we could quantify over "all possible type-free theories of truth" and show that no such theory yields a system that provided a convincing argument for the first disjunct. But given the open-endedness of the notion of a possible "type-free theory" it is hard to see how to do this. (Koellner, 2016, p. 166)

As a consequence Koellner focuses on one sample theory of truth and absolute provability he deems to be particularly attractive and shows that in this theory no argument for the first disjunct can be provided. We shall take a somewhat different approach to Koellner and propose to split up Koellner's initial question into two distinct questions: 1. Can Penrose's *New Argument* be carried out in some consistent theory of truth and absolute provability? 2. Are there reasonable theories of truth and absolute provability that yield an argument for the first disjunct of GD? The first question focuses entirely on whether one can coherently argue for the first disjunct of GD following the outlines of Penrose's *New Argument*. The second question, however, asks for a coherent argument for the first disjunct of GD tout court. There is no restriction on the structure of the argument and, in particular, we don't need to respect the outlines of Penrose's *New Argument*. While it seems difficult to provide a conclusive answer for the second question, which is the question Koellner addresses, such a conclusive answer might be possible for the first question.

Indeed we think that for the first question a rather strong conclusion is possible provided the reconstruction of Penrose's *New Argument* we employ is accepted: there seems to be no coherent theory of truth and absolute provability in which the *New Argument* is can be carried out. In contrast, for the second question, we show that, under certain assumptions, there are reasonable theories of truth and absolute provability in which a sound argument for the first disjunct can be given. However, as we shall see, even though the argument is sound it is not very convincing for it exploits a pathological feature of the theories of truth and absolute provability we shall be considering. Since we are dealing with paradoxes, all solutions to these paradoxes, i.e. theories of truth and absolute provability, will have certain pathological features. We think that if an argument for the first disjunct relies on the pathological part rather than the intuitively motivated part of the theory then it will fail to be convincing. After all it will lack any intuitive support and will remain a technical peculiarity. Ultimately, we are skeptical as to whether a *convincing* argument for the first disjunct can be given in a reasonable theory of truth and absolute provability.

We now return to the first question. To this end we need to take a closer look at the reconstruction of Penrose's argument Koellner proposes. Koellner (2016) does not investigate the argument in detail for he can show that in the theory of truth $\mathsf{DT}$

[22] See (Koellner, 2016, p.184/185) for a clear expression of his skepticism.

(Feferman, 2008) he is considering the first disjunct of GD is independent: it can be neither proved nor refuted. However, if all steps of the argument were sound given the assumptions Koellner puts forth, then there would be consistent theories of truth and absolute provability in which the *New Argument* could be carried out.[23] However, we already find ourselves in trouble when trying to reconstruct the inference from Line 1 to Line 2. To see this, it is worth recalling that $F$ is closed under modus ponens

If we add to the system $F$ the sentence $\forall x(Fx \leftrightarrow Kx)$ we obtain a new recursively enumerable set. Let us call this set $\Sigma$. $\Sigma$ may not be a system in our sense since it may not be closed under modus ponens. It is by closing $\Sigma$ under modus ponens that we obtain the system $F'$. Maybe surprisingly, it is precisely this closure under *modus ponens* that creates a problem in the step from Line 1 to Line 2. Even though we know by Line 1 and $(T\text{-}\mathsf{In})$ that all members of $\Sigma$ are true, we do not know without further assumption that all the members $F'$ are true because we do not know whether the truth predicate is closed under *modus ponens*. For all we know there could be sentences $\phi$ and $\psi$ such that $T\ulcorner\phi\urcorner$, $T\ulcorner\phi \rightarrow \psi\urcorner$ but $\neg T\ulcorner\psi\urcorner$. To guarantee that the truth predicate will be closed under *modus ponens* we need to add this requirement as a further assumption:

$(T\text{-}\mathsf{Imp})$ $\qquad \forall x,y(\mathsf{Sent}(x \dot{\rightarrow} y) \rightarrow (T(x \dot{\rightarrow} y) \rightarrow (Tx \rightarrow Ty))).$[24]

If $(T\text{-}\mathsf{Imp})$ is assumed, then the inference from Line 1 to Line 2 follows from $(T\text{-}\mathsf{In})$, $(T\text{-}\mathsf{Imp})$ and an induction on the length of a proof in $F'$.

There is a further small lacuna in the proposed reconstruction of the argument, namely, in order to derive Line 3 we need to assume that no false arithmetical sentence is true (in the object-linguistic sense).[25] There are several ways this can be achieved, but one straightforward and rather plausible way is to require the truth predicate to be an adequate truth predicate for the arithmetical language. That is, the Tarski biconditionals should hold for sentences of the arithmetical language: for arithmetical sentences $\phi$

$(\mathsf{TB})$ $\qquad T\ulcorner\phi\urcorner \leftrightarrow \phi.$

Assuming $(\mathsf{TB})$ we can derive Line 3 as illustrated below where $G_{F'}$ stands for the Gödel sentence of $F'$, i.e, a sentence such that

$$\neg F'(\ulcorner G(F')\urcorner) \leftrightarrow G(F')$$

[23] There is an interpretation of the resulting theory in the theory $\mathsf{KF}$ together with the completeness axiom. The interpretation would translate the absolute provability predicate as the truth predicate but hold the remaining vocabulary fixed. See Halbach (2011) for more on the truth theory $\mathsf{KF}$ and the completeness axiom.

[24] Notice that it is of no help to replace $F'$ by $\Sigma$, which is not closed under modus ponens, throughout the argument. To carry out the argument we need to show that whatever can be *proved* from $\Sigma$ is true. To show this we need to assume the truth predicate to be closed under modus ponens.

[25] Actually, we only need to assume $\neg T\ulcorner 0 = 1\urcorner$. However, this would not change the general situation. In particular Theorem 2 would still hold if $(\mathsf{TB})$ were replaced by $\neg T\ulcorner 0 = 1\urcorner$.

can be proved via the diagonal lemma. We start our reasoning from Line 2.

2. $\forall x(Fx \leftrightarrow Kx) \rightarrow \forall x(F'x \rightarrow Tx)$

   a. $\forall x(Fx \leftrightarrow Kx) \rightarrow (F'^{\ulcorner}G(F')^{\urcorner}) \rightarrow T^{\ulcorner}G(F')^{\urcorner})$ by Line 2.
   b. $\forall x(Fx \leftrightarrow Kx) \rightarrow (F'^{\ulcorner}G(F')^{\urcorner} \rightarrow G(F'))$ by Line 2(a)., $(\mathsf{TB})$

3. $\forall x(Fx \leftrightarrow Kx) \rightarrow G(F')$ by Line 2(b) and the definition of $G_{F'}$.

This closes the final gap in the reconstruction of Penrose's *New Argument*. If $(\mathsf{T_K})$, $(T\text{-}\mathsf{In})$, $(T\text{-}\mathsf{Imp})$, $(\mathsf{TB})$ and $(\mathsf{Nec})$ are assumed then every step of the *New Argument* is sound. But, unfortunately, we find ourselves essentially in the same situation as in the reconstruction of the argument without the truth predicate: the assumptions needed for carrying out the argument are jointly inconsistent.

**Theorem 2 (Folklore).** *Let $\Sigma$ be a theory extending Robinson arithmetic. Then*

$$(T\text{-}\mathsf{In}), (T\text{-}\mathsf{Imp}), (\mathsf{TB}) \vdash_{\Sigma} \bot.^{26}$$

The fact that the principles of truth required for carrying out Penrose's reasoning are jointly inconsistent suggests a deep flaw with his reasoning: Penrose seems to rely on a *naive* notion of truth. But this is something we cannot do because *naive* conceptions of truth, or absolute provability for that matter, lead to paradox. This lesson has been taught to us by Tarski and Gödel a long time ago. Admittedly, at this point such a strong conclusion seems a bit premature for two reasons. First, so far we have not proved that the first disjunct of GD cannot be derived using the principles Koellner puts forth, that is, we have not shown that the first disjunct of GD is independent of these principles. However, as we show in an appendix to this paper, even if we assume the truth predicate to be arithmetically sound, the first disjunct does not follow from these principles.

Now, the second reason why, at this point, one should be careful with drawing too strong conclusions from the inconsistency result is that, as we noted earlier, from a technical point of view it might be possible to carry out Penrose's *New Argument* using typed principles of truth, that is, principles of truth in which the truth predicate is not applicable to sentences in which it occurs. Earlier in this paper we dismissed such a typing restriction for philosophical and systematic reasons but if we could carry out Penrose's *New Argument* using typed principles that are jointly consistent this would show that Penrose's reasoning does not necessarily appeal to a *naive* (and inconsistent) notion of truth.

## *3.2 Penrose's* New Argument *and Simplistic Typing*

Looking back at Koellner's reconstruction of the *New Argument* and our slight amendment thereof it seems that at least the principles $(T\text{-}\mathsf{In})$, $(\mathsf{T_K})$ and $(\mathsf{Nec})$ can

[26] See Friedman and Sheard (1987) for a proof of Theorem 2.

be restricted to sentences in which the truth predicate does not occur and Penrose's reasoning would still to go through. Notice that the typing restriction we have in mind is rather crude and simplistic because it blocks the use of these principles with respect to sentences in which the truth predicate occurs while it allows for the application of these principles to sentences in which the truth predicate is mentioned. From this point of view the typing restriction we propose is unprincipled and *ad hoc* but, as we have pointed out, at this point we are only interested whether the restriction turns the *New Argument* into a valid argument.

To this end we let $\mathscr{L}_{\mathscr{K}}$ be the language without the truth predicate, $\mathsf{Sent}_{\mathscr{L}_{\mathscr{K}}}$ a predicate the set of sentences of the language and replace the principles ($T$-$\mathsf{In}$), ($\mathsf{T_K}$) and ($\mathsf{Nec}$) by the following variants:

$$(\mathsf{T}^{\mathsf{R}}_{\mathsf{K}}) \qquad \forall x(\mathsf{Sent}_{\mathscr{L}_{\mathscr{K}}}(x) \to (Kx \to Tx)),$$

$$(T\text{-}\mathsf{In}_K) \qquad \phi \to T^{\ulcorner}\phi^{\urcorner} \qquad \phi \in \mathscr{L}_{\mathscr{K}},$$

$$(\mathsf{Nec}_K) \qquad \frac{\phi}{K^{\ulcorner}\phi^{\urcorner}} \qquad \phi \in \mathscr{L}_{\mathscr{K}}.$$

Unfortunately, the simple minded typing restriction of the principles at play won't be sufficient to block a variant of Theorem 2. In other words the principles required for deriving the first disjunct are still jointly inconsistent.

**Theorem 3.** *Let $\Sigma$ be a theory extending Robinson arithmetic. Then*

$$(\mathsf{T}^{\mathsf{R}}_{\mathsf{K}}), (T\text{-}\mathsf{In}_K), (T\text{-}\mathsf{Imp}), (\mathsf{Nec}), \mathsf{TB} \vdash_{\Sigma} \bot.^{27}$$

*Proof.* Let $\delta$ be a sentence of $\mathscr{L}_{\mathscr{K}}$ such that

$$(\ddagger) \qquad \Sigma \vdash \delta \leftrightarrow \neg K^{\ulcorner}\delta^{\urcorner}.$$

We reason as follows:

| | |
|---|---|
| 1. $K^{\ulcorner}\delta^{\urcorner} \to T^{\ulcorner}\delta^{\urcorner}$ | $(T_K)$ |
| 2. $K^{\ulcorner}\delta^{\urcorner} \to T^{\ulcorner}\neg\delta^{\urcorner}$ | $(\ddagger), (T\text{-}\mathsf{In}_K)$ |
| 3. $K^{\ulcorner}\delta^{\urcorner} \to T^{\ulcorner}\neg\delta \wedge \delta^{\urcorner}$ | $1, 2, (T\text{-}\mathsf{In}_K), (T\text{-}\mathsf{Imp})$ |
| 4. $K^{\ulcorner}\delta^{\urcorner} \to \bot$ | $3, (T\text{-}\mathsf{In}_K), (T\text{-}\mathsf{Imp}), \mathsf{TB}$ |
| 5. $\neg K^{\ulcorner}\delta^{\urcorner}$ | 4 |
| 6. $\delta$ | $5, (\ddagger)$ |
| 7. $K^{\ulcorner}\delta^{\urcorner}$ | $6, (T\text{-}\mathsf{In}_K), (\mathsf{Nec}_K)$ |
| 8. $\bot$ | $5, 7$ |

□

[27] In order to carry out Penrose's reasoning we cannot restrict ($T$-$\mathsf{Imp}$) to sentences of $\mathscr{L}_{\mathscr{K}}$. However, in the formulation of Theorem 3 we could use the restricted version of the principle as its proof should make clear.

Examining the reconstruction of Penrose's *New Argument* there does not seem to be an alternative possible restriction of the principles at play that would make the principles jointly consistent while facilitating Penrose's reasoning. Theorem 3 should therefore put an end to all attempts of vindicating the *New Argument* along the lines of our proposed reconstruction. Of course, Penrose could try a similar strategy to the one we discussed at the beginning of Section 3. That is, he could argue that there is some restriction that blocks the use of, say, $(\mathsf{Nec})$ in the "bad" cases like in Line 7 in the proof of Theorem 3 but allows the use in the "good" cases like Line 5 in Koellner's reconstruction of the *New Argument*. But all the critical remarks we made a the beginning of Section 3 will carry over to the present case. In conclusion it thus seems fair to say that Penrose's *New Argument* is simply incoherent. The problem is that its reasoning makes essential use of a *naive* notion of truth (or absolute provability), which in a context where a certain amount of self-applicability is required irrevocably leads to paradox.

## 4 Global Reflection and the First Disjunct

Under the proposed formalization, Penrose's *New Argument* cannot be turned into a coherent argument but, of course, there might be alternative arguments for the first disjunct of GD. This leads to the second question we outlined in Section 3, i.e., the question of whether there are reasonable theories of truth and absolute provability that yield an argument for the first disjunct of GD. As Koellner (2016) points out, given the plentitude of theories of truth and absolute provability it seems difficult to deal with all possible such theories. In contrast to Koellner, however, we do not focus on one particular theory of truth.[28] Instead we investigate the prospects of providing a successful argument for the first disjunct using Penrose's reductio strategy. To this end, we shall now make slightly stronger assumptions on what a formal system $F$ is, that is, from now on we will take systems $F$ to be *classical* recursively axiomatizable theories. Under these presuppositions the assumption that "I am F" can be reformulated as $\forall x(\mathsf{Pr}_F(x) \leftrightarrow Kx)$, where $\mathsf{Pr}_F$ is a natural provability predicate of the theory $F$. This amounts to a slight strengthening of the Mechanistic thesis as we now only consider Turing machines whose outputs are closed under classical logic.

Throughout we shall assume the principle $(\mathsf{T_K})$ and look for consistent theories of truth and absolute provability in which the assumption that "I am F" leads to a contradiction. Under this assumption $(\mathsf{T_K})$ implies the so-called Global Reflection principle for $F$:

$$(\mathsf{GRef_F}) \qquad \forall x(\mathsf{Pr}_F(x) \to Tx).$$

[28] Koellner (2016) constructs the theory DTK, which extends Feferman's theory DT (Feferman, 2008) and shows that in this theory the first disjunct of GD is independent: it can be neither proved nor refuted.

Of course, no recursively axiomatizable theory $F$ that is only remotely plausible can prove the Global Reflection principle for itself because by Gödel's second incompleteness theorem we have

$$(*) \qquad F \vdash \forall x(\mathsf{Pr}_F(x) \to Tx) \quad \Rightarrow \quad F \vdash \bot.^{29}$$

But then, since the assumption that "I am F" together with $(\mathsf{T}_\mathsf{K})$ implies the Global Reflection Principle, which in turn implies a contradiction it might seem that we have already provided a reductio ad absurdum of the assumption and thereby established the first disjunct of GD.

This would be a rather premature conclusion in at least two ways: First, while we can prove that if $F$ proves the Global Reflection Principle then $F$ is inconsistent we cannot prove that $F$ proves that the Global Reflection Principle implies a contradiction, i.e., $F$ does not necessarily prove

$$(**) \qquad \forall x(\mathsf{Pr}_F(x) \to Tx) \to \bot.$$

But a successful argument for the first disjunct via the reductio strategy requires a proof of $(**)$. The weaker claim $(*)$ is not sufficient for deriving the first disjunct of GD.

Second, and more importantly, in conducting the reductio argument we may not assume that we actually reason in $F$ for this would not only presuppose that "I am F" but also that *I know which system F I am*. This point goes back to Benacerraf (1967) and was further discussed by, e.g., Reinhardt (1986). Reinhardt showed, at least under a particular formalization, that while it is inconsistent to assume that "I am F" and that "I know which system $F$ I am", it is consistent, to assume that "I am F", while *not knowing which particular system F I am*.[30] In more technical terms this means even though the idealized human mind might coincide with a particular system $F$, it may not recognize the provability predicate $\mathsf{Pr}_F$ as "its" provability predicate, that is the provability predicate of $F$. As a consequence, in the reductio argument we cannot assume the Löb derivabiliy conditions and, in particular, we may not infer $\mathsf{Pr}_F(\ulcorner \phi \urcorner)$ whenever $\phi$ has been derived.[31] This observation is crucial since otherwise Penrose's *New Argument* would have a coherent reconstruction. By using Löb's derivability conditions we could dispense of the rule (Nec) in the reconstruction of the argument in Section 2 and thereby restrain ourselves to a consistent

[29] We take it that any remotely plausible theory prevents the truth predicate from being entirely trivial. In other words it should rule out the truth of false arithmetical sentences, i.e., $\neg T\ulcorner 0 = 1\urcorner$ should be a theorem of $F$.

[30] The discussion took place in the framework of epistemic arithmetic (Reinhardt, 1986; Shapiro, 1985) where paradoxical sentences can not be formed due to syntactic restrictions of the language. In Reinhardt (1985b) showed that "I am F", suitably formalized, was consistent *contra* Lucas and Penrose. Reinhardt (1985a) establishes the inconsistency of *I am F and I know that I am F*. In a paper that somewhat concluded this line of research Carlson (2000) showed that the so-called *strong mechanistic thesis*, that is, the proposition that *I am F and I know that I am some recusively axiomatizable system* was consistent. See, e.g., Koellner (2016) for discussion.

[31] See Shapiro (2003) for similar remarks.

set of principles of absolute provability. The argument would then run as follows (we start with Line 3):

3. $\forall x(Fx \leftrightarrow Kx) \rightarrow G(F')$
4. $\forall x(Fx \leftrightarrow Kx) \rightarrow \neg F^{\ulcorner}\forall x(Fx \leftrightarrow Kx) \rightarrow G(F')^{\urcorner}$
5. $F^{\ulcorner}\forall x(Fx \leftrightarrow Kx) \rightarrow G(F')^{\urcorner}$ 3, by Löb's derivability conditions
6. $\forall x(Fx \leftrightarrow Kx) \rightarrow \bot$.

Bearing the previous remarks in mind this argument is clearly not a convincing argument for the first disjunct. It fails to establish that there is no recursively axiomatizable theory whose theorems coincide with the absolutely provable sentences but shows that there is some particular system $F$ that falls short of absolute provability: the system $F$ in which line 3 can be proven and the human mind, that is $F$, knows that this is so. In other words, the argument establishes that it is impossible that *I am F and I know that I am F* but fails to establish the first disjunct of GD. Recapitulating our discussion, the question of whether we can successfully argue for the first disjunct of Gödel's Disjunction using the reductio strategy depends on whether, given a reasonable theory of truth and absolute provability, the Global Reflection Principle leads to a contradiction without introducing any particular assumption about which recursively axiomatizable theory we are working in. The theory and particular its proofs are completely opaque to the (idealized) human mind: it does not *know* its axioms or rules of proof. This severely restricts the force of the Global Reflection Principle: even if, assuming that "I am F", I were to derive $(\mathsf{GRef}_F)$, I would not realize that I have established my own soundness and as a consequence I might be unable to derive my inconsistency. Indeed, the fact that I do not realize that I have established my own soundness may save me from becoming inconsistent. As it were, from my perspective, $(\mathsf{GRef}_F)$ could reflect about *any* recursively axiomatizable system—I just don't know which one. Given the dialectical situation it seems difficult to derive a contradiction from the Global Reflection Principle because what it tells us under these circumstances is that whatever is provable in every recursively axiomatizable theory (perhaps extending some basic arithmetical theory) is true. At least prima facie this seems to be a rather unproblematic and uncontroversial claim for it only means that logical truths should in fact be true (in the object-linguistic sense). It thus may come as a surprise that even under these weak assumptions there exist reasonable theories of truth and absolute provability in which a Penrose-style reductio argument for the first disjunct can be carried out.

## *4.1 A New Argument Against Mechanism*

The theory $\mathsf{KFC}$, i.e. *Consistent Kripke-Feferman*, is a compositional theory of truth in the sense that it commutes with all logical connectives with exception of negation, for which we may eliminate double negation inside the scope of the truth predicate.[32] Moreover, $\mathsf{KFC}$ as opposed to $\mathsf{KF}$ simpliciter asserts the consistency of the

[32] See Halbach (2011) for more details about $\mathsf{KFC}$.

truth predicate, which in KF is equivalent to the principle

$$(T\text{-Out}) \qquad T^{\ulcorner}\phi^{\urcorner} \rightarrow \phi.$$

By results in Stern (2014b) we know that there are plenty of consistent reasonable theories of truth and absolute provability that extend KFC and prove $(\mathsf{T}_{\mathsf{K}})$. The argument below shows that the proponent of such KFC-style theories of truth and absolute provability is committed to accepting the first disjunct of GD, that is, they have to deny mechanism provided they accept all the background assumptions. Let $\lambda$ be the standard Liar sentence, that is, a sentence such that the theory $F$ under consideration proves:

$$(L) \qquad \neg T^{\ulcorner}\lambda^{\urcorner} \leftrightarrow \lambda.$$

| | | |
|---|---|---|
| 1. | $F \vdash \forall x(K(x) \leftrightarrow \mathsf{Pr}_F(x)) \rightarrow \forall y(\mathsf{Pr}_\sigma(y) \rightarrow T(y))$ | $\mathsf{T}_{\mathsf{K}}$ |
| 2. | $F \vdash \forall x(K(x) \leftrightarrow \mathsf{Pr}_F(x)) \rightarrow (\mathsf{Pr}_F(^{\ulcorner}\lambda \vee \neg\lambda^{\urcorner}) \rightarrow T(^{\ulcorner}\lambda \vee \neg\lambda^{\urcorner}))$ | 1 |
| 3. | $F \vdash \mathsf{Pr}_\emptyset(^{\ulcorner}\lambda \vee \neg\lambda^{\urcorner})$ | |
| 4. | $F \vdash \forall x(K(x) \leftrightarrow \mathsf{Pr}_F(x)) \rightarrow T(^{\ulcorner}\lambda \vee \neg\lambda^{\urcorner}))$ | 2,3 |
| 5. | $F \vdash \forall x(K(x) \leftrightarrow \mathsf{Pr}_F(x)) \rightarrow T(^{\ulcorner}\lambda^{\urcorner}) \vee T(^{\ulcorner}\neg\lambda^{\urcorner})$ | 4, KF |
| 6. | $F \vdash \forall x(K(x) \leftrightarrow \mathsf{Pr}_F(x)) \rightarrow \neg\lambda$ | $5,(L),(T\text{-Out})$ |
| 7. | $F \vdash \forall x(K(x) \leftrightarrow \mathsf{Pr}_F(x)) \rightarrow T^{\ulcorner}\lambda^{\urcorner}$ | $6,(L)$ |
| 8. | $F \vdash \forall x(K(x) \leftrightarrow \mathsf{Pr}_F(x)) \rightarrow \lambda$ | $7,(T\text{-Out})$ |
| 9. | $F \vdash \neg\forall x(K(x) \leftrightarrow \mathsf{Pr}_F(x))$ | 6,8 |

The crucial step in the argument is Line 3, which says that $\lambda \vee \neg\lambda$ is provable without premisses, that is, in logic alone. This implies that for all classically recursively axiomatizable theories $F$, $\mathsf{Pr}_F(^{\ulcorner}\lambda \vee \neg\lambda^{\urcorner})$. The argument therefore establishes the first disjunct of GD since no special assumption concerning $F$ was made. It is noteworthy that the argument, even though it is a reductio argument, is very different to Penrose's *New Argument*. It exploits the fact that in KFC-style theories of truth and absolute provability the so-called internal logic, that is the logic within the scope of the truth predicate, and the external logic diverge. The external logic is just classical logic while the internal logic of KFC-style theories is strong Kleene logic. As a consequence classical tautologies are not generally true in the object-linguistic sense and cannot be for sake of consistency. However, we have assumed classical logic throughout and, in particular, our formalization of "I am F", i.e. $F \vdash \forall x(K(x) \leftrightarrow \mathsf{Pr}_F(x))$, reflects this fact since the axioms of classical logic are built into the standard provability predicate. This implies that we are only considering theories formulated in classical logic and thus no matter which particular system $F$ we consider, $F$ will prove the classical tautologies. As a consequence $(\mathsf{T}_{\mathsf{K}})$ has the effect of adding the classical tautologies to the internal logic of the theory. But, as our argument shows, this leads to a contradiction in KFC-style theories of truth and absolute provability and we may conclude that for no recursively axiomatizable system $F$, it can be established that "I am F".

Is this new argument for the first disjunct of GD a good argument? That depends on whether one takes KFC-style theories to be suitable theories for discussing

Gödel's Disjunction and related issues, and, moreover, whether one takes the argument to be intuitively plausible. We take it that the answer to both questions is negative. First, KFC-style theories violate one of the fundamental assumptions underlying the debate surrounding GD, namely, that provability in recursively axiomatizable theories of mathematics implies truth. But in KFC-style theories there exist sentences $\phi$, e.g. $\lambda \vee \neg\lambda$, such that $\mathsf{Pr}_F(\ulcorner\phi\urcorner)$ and $\neg T\ulcorner\phi\urcorner$. In other words these theories are provably unsound according to their own standards. Indeed, they are provably unsound with respect to the standard of *every* recursively axiomatizable theory. Second, it is precisely this feature that our argument for the first disjunct of GD relies on because $(\mathsf{GRef}_F)$ asserts that provability implies truth which it cannot in KFC-style theories. The argument thus exploits a pathological feature of the theory and this undermines the credibility of the argument. In order to be convincing there should be a plausible and intuitive rationale to the argument but there is none and this is at least partly due to the fact the argument relies on the pathological aspects of the theory. In light of these remarks even the convinced anti-mechanist, that is the proponent of the first disjunct of GD, should grant that KFC-style theories of truth and absolute provability might just not be a solution to the intensional paradoxes that allows us to establish the first disjunct of GD. However, the problem goes beyond KFC-style theories of truth and absolute provability because the only possibility to derive a contradiction on the basis of $(\mathsf{GRef}_F)$ is precisely if such a clash between the external and the internal logic of the theory arises: in such a situation $(\mathsf{GRef}_F)$ forces the internal logic to coincide with the external logic, which may provoke a contradiction. If we are right, then our skeptical remarks concerning our argument extend to any argument based on the reductio strategy, at least if classical logic is assumed.

## 5 Conclusion

The present investigation was devoted to arguments for the first disjunct of GD. We put all conceptual worries aside and assumed that there was an intelligible notion of absolute provability where a sentence was supposed to be absolutely provable iff it was the output of, i.e. proved by, an idealized human mind. But even pursuing this charitable course we were unable to provide a valid version of Penrose's *New Argument* and we do not see how the problems with the argument can be fixed. In our view the argument is deeply flawed because it assumes *naive* conceptions of self-applicable notions like truth and absolute provability. But Tarski and Gödel told us that these naive notions lead to paradox: ultimately, this is why Penrose's *New Argument* cannot be fixed.

We then looked at Penrose's reductio strategy and, using this strategy, provided an argument for the first disjunct that was at least technically sound. However, we argued that the argument exploited a pathological feature of the underlying theory which undermined the plausibility of the argument and also caused the argument to lack an intuitive rationale. Moreover, the shortcomings of our argument are most

likely to affect all successful arguments for the first disjunct based on the reductio strategy as we presented it in Section 4.

Of course, none of our findings show that there cannot be an argument for the first disjunct based on the incompleteness theorems and a theory of truth and absolute provability. But combining our findings with the observations of Benacerraf (1967), Reinhardt (1986), Lindström (2001, 2006), Shapiro (2003), Koellner (2016) and others we slowly seem to run out of options how such a plausible and successful argument could look.[33] It might be time for the anti-mechanist to come up with an entirely different type of argument for the first disjunct of Gödel's Disjunction, a type of argument that does not rely on Gödel's incompleteness theorems.

**Acknowledgements** This work was supported by the European Commission through a Marie Sklodowska Curie Individual Fellowship (TREPISTEME, Grant No. 703539). I wish to thank Catrin Campbell-Moore, Martin Fischer, Leon Horsten, Peter Koellner, Carlo Nicolai, and an anonymous referee for helpful comments on the content of this paper. Earlier versions of the paper were presented at the *FSB Seminar in Bristol*, the *Fourth New College Logic Meeting*, the University of Malaga and the *Third Leuven-Bristol Workshop*. I thank the audiences of these talks for their feedback.

## References

Benacerraf P (1967) God, the Devil, and Gödel. The Monist 51(1):9–32

Carlson TJ (2000) Knowledge, machines, and the consistency of reinhard's strong mechanistic thesis. Annals of Pure and Applied Logic 105:51–82

Chalmers DJ (1995) Mind, Machines, and Mathematics. a Review of Shadows of the Mind by Roger Penrose. Psyche 2(9)

Egré P (2005) The knower paradox in the light of provability interpretations of modal logic. Journal of Logic, Language and Information 14:13–48

Feferman S (2008) Axioms for determinateness and truth. Review of Symbolic Logic 1:204–217

Friedman H, Sheard M (1987) An axiomatic approach to self-referential truth. Annals of Pure and Applied Logic 33:1–21

Gödel K (1995) Some basic theorems on the foundations of mathematics and their implications. In: Feferman S, Dawson Jr JW, Goldfarb W, Parsons C, Solovay RM (eds) Kurt Gödel: Collected Works, vol 3, Oxford University Press, Oxford, pp 304–323, manuscript written in 1951.

Halbach V (2011) Axiomatic Theories of Truth. Cambridge University Press

Koellner P (2016) Gödel's disjunction. In: Horsten L, Welch P (eds) Gödel's Disjunction. The Scope and Limits of Arithmetical Knowledge, OUP, pp 148–188

[33] So far the discussion has been entirely focused on classical logic. Since a very common reaction to the paradoxes is to give up classical logic it is a rather immediate thought to explore Penrose's *New Argument* or, more generally, arguments for the first disjunct of GD assuming some non-classical logic.

Leitgeb H (2009) On formal and informal provability. In: Linnebo O, Bueno O (eds) New Waves in Philosophy of Mathematics, Palgrave Macmillan, pp 263–299

Lindström P (2001) Penrose's New Argument. Journal of Philosophical Logic 30:241–250

Lindström P (2006) Remarks on Penrose's "New Argument". Journal of Philosophical Logic 35:231–235

Lucas JR (1961) Minds, Machines, and Gödel. Philosophy 36:112–127

Montague R (1963) Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability. Acta Philosophica Fennica 16:153–167

Myhill J (1960) Some remarks on the notion of proof. The Journal of Philosophy 57:461–471

Penrose R (1989) The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics. OUP

Penrose R (1994) Shadows of the Mind: In Search for the Missing Science of Consciousness. OUP

Penrose R (1996) Beyond the Doubting of a Shadow. Psyche 2(23)

Reinhardt WN (1985a) Absolute versions of incompleteness theorems. Nous 19:317–346

Reinhardt WN (1985b) The consistency of a variant of church's thesis with an axiomatic theory of an epistemic notation. In: Caicedo X (ed) Proceedings of the fifth Latin American symposium on mathematical logic held in Bogota, Revista Columbiana de Matematicas, pp 177–200

Reinhardt WN (1986) Epistemic Theories and the Interpretation of Gödel's Incompleteness Theorems. The Journal of Philosophical Logic 15(4):427–474

Shapiro S (1985) Intensional Mathematics. North-Holland Publishing Company, Amsterdam

Shapiro S (2003) Mechanism, Truth, and Penrose's New Argument. Journal of Philosophical Logic 32:19–42

Shapiro S (2016) Idealization, mechanism, and knowability. In: Horsten L, Welch P (eds) Gödel's Disjunction. The Scope and Limits of Arithmetical Knowledge, OUP, pp 189–207

Stern J (2014a) Modality and Axiomatic Theories of Truth I: Friedman-Sheard. The Review of Symbolic Logic 7(2):273–298

Stern J (2014b) Modality and Axiomatic Theories of Truth II: Kripke-Feferman. The Review of Symbolic Logic 7(2):299–318

Stern J (2016) Toward Predicate Approaches to Modality, Trends in Logic, vol 44. Springer, Switzerland

Wang H (1996) A Logical Journey: From Gödel to Philosophy. MIT Press

## Appendix

In this appendix we show that the first disjunct of GD cannot be proved on the basis of the principles $(\mathsf{T_K}), (T\text{-}\mathsf{In}), (\mathsf{TB})$, and $(\mathsf{Nec})$.[34] We introduce some new terminology to state the independence result in a precise way. We write $W_e$ to denote the output of the Turing machine with index $e$. Using this terminology the first disjunct of GD can be formulated as

$$(*) \qquad \neg\exists e \forall x (W_e(x) \leftrightarrow Kx).$$

We will show that $(*)$ is independent of the principles $(\mathsf{T_K}), (T\text{-}\mathsf{In}), (\mathsf{TB})$, and $(\mathsf{Nec})$.

**Theorem 4.** *Let* $\mathsf{PATK}$ *be the extension of* $\mathsf{PA}$ *in the language containing a truth and an absolute provability predicate (*$\mathscr{L}_{\mathsf{PATK}}$*). Then*

$$(\mathsf{T_K}), (T\text{-}\mathsf{In}), (\mathsf{TB}), (\mathsf{Nec}) \nvdash_{\mathsf{PATK}} \neg\exists e \forall x (W_e(x) \leftrightarrow Kx).$$

*Proof.* The proof of Theorem 4 is an compactness argument: if we can prove $(*)$ in $\mathsf{PATK}$ on the basis of $(\mathsf{T_K}), (T\text{-}\mathsf{In}), (\mathsf{TB})$, and $(\mathsf{Nec})$, then there must be a finite proof of $(*)$ in $\mathsf{PATK}$ and, in particular, a proof with only finitely many applications of the rule $(\mathsf{Nec})$. The argument we give shows that there cannot be such a proof of finite length for $(*)$. To this end we define a family of theories by recursion:

$$\begin{aligned} \Sigma_0 &:= \mathsf{Cn}[\mathsf{PATK} + (\mathsf{T_K}) + (T\text{-}\mathsf{In}) + (\mathsf{TB})] \\ \Sigma_{n+1} &:= \mathsf{Cn}[\Sigma_n + \{K\ulcorner\phi\urcorner : \Sigma_n \vdash \phi\}] \end{aligned}$$

where $\mathsf{Cn}$ denotes the operation that closes these sets of sentences under logical consequence. $\Sigma_n$ allows for proofs with $n$-many applications of the rule $(\mathsf{Nec})$. Notice also that by construction $\Sigma_n \subseteq \Sigma_{n+1}$. By our compactness argument it follows that if

$$(\mathsf{T_K}), (T\text{-}\mathsf{In}), (\mathsf{TB}), (\mathsf{Nec}) \vdash_{\mathsf{PATK}} \neg\exists e \forall x (W_e(x) \leftrightarrow Kx),$$

then there must be a proof of $(*)$ in some $\Sigma_n$ for $n \in \omega$. We now construct a suitable model $\mathsf{M}_n$ for each $\Sigma_n$ with $n \in \omega$ such that

$$\mathsf{M}_n \models \exists e \forall x (W_e(x) \leftrightarrow Kx).$$

Let $\mathsf{M}_0 = (\mathbb{N}, \|T\|_0, \|K\|_0)$ with

$$\begin{aligned} \|T\|_0 &:= \{\phi : (\mathbb{N} \models \phi \,\&\, \phi \in \mathscr{L}_{\mathsf{PA}}) \text{ or } (\phi \in \mathscr{L}_{\mathsf{PATK}} \,\&\, \phi \notin \mathscr{L}_{\mathsf{PA}})\} \\ \|K\|_0 &:= \{\phi : \mathsf{PATK} \vdash \phi\}. \end{aligned}$$

[34] A similar, almost trivial argument shows that the first disjunct of GD can also not be obtained using the principles $(\mathsf{T_K}), (T\text{-}\mathsf{In}), (T\text{-}\mathsf{Imp})$, and $(\mathsf{Nec})$. A proof is left to the reader.

Clearly, $\mathsf{M}_0 \models \Sigma_0$ and $\mathsf{M}_0 \models \forall x(\mathsf{Pr}_{\mathsf{PATK}}(x) \leftrightarrow Kx)$. But the theorems of PATK are recursively enumerable and we thus know that $\mathsf{M}_0 \models \exists e(W_e(x) \leftrightarrow Kx)$. Now, for $\mathsf{M}_n$ we define

$$\begin{aligned} \|T\|_{n+1} &:= \|T\|_n \\ \|K\|_{n=1} &:= \{\phi : \Sigma_n \vdash \phi\}. \end{aligned}$$

It trivially follows that $\mathsf{M}_{n+1} \models \Sigma_{n+1}$ and $\mathsf{M}_{n+1} \models \forall x(\mathsf{Pr}_{\Sigma_n}(x) \leftrightarrow Kx)$, which implies $\mathsf{M}_n \models \exists e(W_e(x) \leftrightarrow Kx)$. This concludes the proof of Theorem 4. There cannot be a proof establishing the first disjunct of GD on the basis of the principles $(\mathsf{T_K})$, $(T\text{-}\mathsf{In})$, $(\mathsf{TB})$, and $(\mathsf{Nec})$.