On the Question of Whether the Mind can be Mechanized: From Gödel to Penrose

Peter Koellner
Harvard University

August 6, 2026

In this talk I would like to address the question of whether the incompleteness theorems imply that "the mind cannot be mechanized," where this question is understood in the following more specific and precise sense:

Question

Do the incompleteness theorems imply that "the mathematical outputs of the idealized human mind do not coincide with the mathematical outputs of any idealized finite machine (Turing machine)."

We are interested in the question of what the incompleteness theorems tell us about the interrelations between three fundamental notions:

F: an idealized finite machine ($\sim$ relative provability)

K: the idealized human mind ($\sim$ absolute provability)

T: truth

Gödel makes three central claims concerning the relationship between F, K, and T.

"[N]o well-defined system of correct axioms [F] can comprise all objective mathematics [T], since the proposition which states the consistency of the system is true, but not demonstrable in the system"

Claim 1

For any formal system F,

$$F \subseteq T \rightarrow F \subsetneq T.$$

▷ If $F \subseteq T$, then $\mathrm{Con}(F) \in T$. But $\mathrm{Con}(F) \notin F$. So $F \subsetneq T$.

"[I]f he perceives the axioms under consideration to be correct [$K(F \subseteq T]$, he also perceives (with the same certainty) that they are consistent. Hence he has a mathematical insight not derivable from his axioms [$F \subsetneq K$]."

Claim 2

For any formal system F,

$$K(F \subseteq T) \rightarrow F \subsetneq K.$$

- ▷ Suppose $K(F \subseteq T)$. We saw above that if $F \subseteq T$ then $\mathrm{Con}(F) \in T$ but $\mathrm{Con}(F) \notin F$. Now, if we also have $K(F \subseteq T)$ then we have $K(\mathrm{Con}(F))$. So $\mathrm{Con}(F) \in K$ but $\mathrm{Con}(F) \notin F$. Thus, $F \subsetneq K$.

It is of interest to note that Gödel did *not* conclude that for any F, $F \subseteq T \rightarrow F \subsetneq K$.

In fact, he was careful to say that no such conclusion was warranted: He explicitly said that for all we have shown, it may indeed be the case that there is a "master system" $F^\star$ such that $F^\star = K$. We have only shown that if there *is* such an $F^\star$ then we can't have $K(F^\star \subseteq T)$.

Furthermore, he noted that *if* there were such an $F^\star$ it would have important implications.

Claim 3 (Gödel's Disjunction)

Either $\left(\neg\exists F\,(F = K)\right)$ *or* $\left(\exists\varphi\,(T(\varphi) \wedge \neg K(\varphi) \wedge \neg K(\neg\varphi))\right)$.

- Suppose that there were an $F^\star$ such that $F^\star = K$. Then we have $F^\star \subseteq T$ (since $K \subseteq T$). So, by the incompleteness theorems, we have $F^\star \subsetneq T$. Thus, there is a $\varphi \in T$ such that $\varphi \notin F^\star$, and for any such φ we also have $\neg\varphi \notin F^\star$ (since if $\neg\varphi \in F^\star$ then $\neg\varphi \in T$ (as $F^\star \subseteq T$), which is impossible since $\varphi \in T$). But $F^\star = K$. So this φ is such that $\varphi \in T$ and yet neither $\varphi \in K$ nor $\neg\varphi \in K$.

The System EA_T

The axioms governing F are simply the standard axioms of arithmetic, namely, PA.

The axioms of K are (the universal closures of):

- (E_1) $K\varphi$, where φ is valid.
- (E_2) $K(\varphi \rightarrow \psi) \rightarrow K\varphi \rightarrow K\psi$.
- (E_3) $K\varphi \rightarrow \varphi$.
- (E_4) $K\varphi \rightarrow KK\varphi$.

The resulting system is called EA (Epistemic Arithmetic).

- ▷ Notice that EA embodies a highly idealized notion of K.

The axioms governing T are the standard Tarskian axioms. They include such axioms as:

(T_1) $(\forall x)[\mathrm{Sent}(x) \rightarrow (T(\neg x) \leftrightarrow \neg T(x))]$ and

(T_2) $(\forall x)(\forall y)[\mathrm{Sent}(x, y) \rightarrow (T(x \vee y) \leftrightarrow T(x) \vee T(y))]$.

The resulting system is called EA_{T} (Epistemic Arithmetic with Truth).

Theorem (Reinhardt)

EA_T *proves Gödel's three claims:*

(1) $F \subseteq T \rightarrow F \subsetneq T$

(2) $K(F \subseteq T) \rightarrow F \subsetneq K$

(3) *Either* $\neg\exists F^\star\,(K = F^\star)$ *or* $K \subsetneq T$

In this sense, the disjunction is indeed "a mathematically established fact."

So . . . Which is it?

Is it the case that the "mind cannot be mechanized" or is it the case that "there are absolutely undecidable sentences"?

Lucas and Penrose argued that the incompleteness theorems actually implied the first disjunct—that "the mind cannot be mechanized."

The central claim is that:

> *"Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth" (Penrose, 1994)*

We saw that this is actually provable in EA_T: It is just Gödel's second claim: $K(F \subseteq T) \rightarrow F \subsetneq K$.

But notice that this is a *conditional* statment. To discharge the antecedent one must be able to determine whether or not F is correct.

This is no small task. For example, let F be the system PA $+ R$ where 'R' stands for Riemann Hypothesis.

Fact

PA $+ R$ *is consistent if and only if* R *is true.*

- It follows that no one knows whether or not PA $+ R$ is consistent.

The point is much stronger: It is not about just this one sentence, R.

- To have an oracle for the question of whether or not an arbitrary F is consistent is to have an oracle for Π^0_1-truth. But that is not something one can assume *at the start* of an argument for the claim that "the mind cannot be mechanized" since it *trivially* implies the conclusion.

So, this *particular* argument for the first disjunct fails . . .

But perhaps there is another argument ...

Let's distinguish between three grades of mechanism:

(1) (WMT) $\exists F\,(K = F)$

- "there is a machine such that I am that machine"

(2) (SMT) $K\,\exists F\,(K = F)$

- "I know that there there is a machine such that I am that machine"

(3) (SSMT) $\exists F\,K(K = F)$

- "There is a machine such that I know that I am *that* machine"

Theorem (Reinhardt)

"$\mathsf{EA}_\mathsf{T} + \mathsf{WMT}$" *is consistent.*

- So EA_T cannot rule out that I am a machine.

Theorem (Carlson)

"$\mathsf{EA}_\mathsf{T} + \mathsf{SMT}$" *is consistent.*

- So EA_T cannot even rule out that I *know* that I am a machine.

What, then, could have led people astray?

Theorem (Reinhardt)

"$\mathrm{EA}_\mathsf{T} + \mathrm{SSMT}$" is inconsistent.

- So EA_T *can* rule out that there is a machine such that I know that I am *that* machine.

Perhaps what led people astray was an unjustified slide from

$$\neg\exists e(K(K = F_e))$$

$$\rightarrow \neg K(\exists e(K = F_e))$$

$$\rightarrow \neg\exists e(K = F_e)$$

In any case, regardless of what may have led people astray, the particular argument given turns on an error and, in fact, there can be no argument for the conclusion within EA_T, a system that would seem to embody every assumption that the proponents of the first disjunct would be willing to make.

I hope that this puts to rest the first generation of arguments for the first disjunct, and that all participants in the debate can agree.

Penrose

The above results take place in EA_{T} where one avoids the paradoxes of K by treating it as an *operator* and one avoids the paradoxes of T treating it as *typed*.

Gödel did not think that the typed approach was an adequate approach to either K or T. Moreover, he held that:

> *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*)

Curiously, Penrose's new argument requires a type-free truth predicate. So perhaps Penrose has fulfilled Gödel's hope.

Though I do not 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.

The argument involves several principles governing F, K, and T, three of which it will be useful to state at the outset:

$$K(\ulcorner \varphi \urcorner) \rightarrow T(\ulcorner \varphi \urcorner) \qquad (K \subseteq T)$$

$$\varphi \rightarrow T(\ulcorner \varphi \urcorner) \qquad (T\text{-In})$$

$$\frac{\varphi}{K(\ulcorner \varphi \urcorner)} \qquad (K\text{-Intro})$$

Putting it all together:

(1) $K = F \rightarrow F \subseteq T$

(2) $K = F \rightarrow F^* \subseteq T$

(3) $K = F \rightarrow G(F^*)$

(4) $K = F \rightarrow F^* \nvdash G(F^*)$

(5) $K(K = F \rightarrow G(F^*))$

(6) $K = F \rightarrow \left(F \nvdash (K = F \rightarrow G(F^*))\right)$

(7) $K \neq F$

Sharpening the Notions: DTK

The type-free system of truth I shall employ is Feferman's system DT.

The base language is $\mathscr{L}_{\mathsf{PA}}$ and the base system PA. The language is extended by adding a (type-free) truth predicate 'T', allowing it to figure in induction, and adding axioms of *determinateness* and *truth*. Here determinateness is symbolized by letting '$D(x)$' be short for '$T(x) \vee T(\neg x)$'. The primitive connectives are '$\neg$', '$\vee$', and '$\rightarrow$'.

The axioms governing 'D' include

(D_2) $(\forall x)[\mathsf{Sent}(x) \rightarrow (D(\dot{\neg} x) \leftrightarrow D(x))]$

(D_3)
$(\forall x)(\forall y)[\mathsf{Sent}(x) \wedge \mathsf{Sent}(y) \rightarrow (D(x \dot{\vee} y) \leftrightarrow D(x) \wedge D(y))]$

(D_4) $(\forall x)(\forall y)[\mathsf{Sent}(x) \wedge \mathsf{Sent}(y) \rightarrow (D(x \dot{\rightarrow} y) \leftrightarrow$
$D(x) \wedge (T(x) \rightarrow D(y))]$

The axioms governing 'T' include

(T_2) $(\forall x)[\mathsf{Sent}(x) \wedge D(x) \rightarrow (T(\dot{\neg} x) \leftrightarrow \neg T(x))]$

(T_3) $(\forall x)(\forall y)[\mathsf{Sent}(x) \wedge \mathsf{Sent}(y) \wedge D(x \dot{\vee} y) \rightarrow (T(x \dot{\vee} y) \leftrightarrow T(x) \vee T(y))]$

(T_4) $(\forall x)(\forall y)[\mathsf{Sent}(x) \wedge \mathsf{Sent}(y) \wedge D(x \dot{\rightarrow} y) \rightarrow (T(x \dot{\rightarrow} y) \leftrightarrow T(x) \rightarrow T(y))]$

Theorem (Feferman)

For all $\varphi \in L_{\mathrm{DT}}$

$$\mathrm{DT} \vdash D(\ulcorner \varphi \urcorner) \rightarrow (T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi).$$

Now, it would be natural to merge this theory of truth (DT) with our theory of absolute provability (EA), adding the $K \rightarrow T$ principle. Let $\mathsf{EA}_{\mathsf{DT}}$ be the resulting system.

Theorem

$\mathsf{EA}_{\mathsf{DT}}$ *is inconsistent.*

Perhaps this is not so surprising in hindsight. For once we enter the realm of indeterminate sentences we should condition the axioms of absolute provability on determinateness.

Surprisingly, when one does this one can circumvent the limitative results of Gödel, Montague, and Thompson and treat 'K' as a predicate.

The System DTK

The axioms governing 'K' are:

(K_1) $(\forall x)[\mathsf{Sent}(x) \rightarrow (K(x) \rightarrow T(x))]$

(K_2) $(\forall x)(\forall y)[\mathsf{Sent}(x) \wedge \mathsf{Sent}(y) \rightarrow (K(x \dot{\rightarrow} y) \rightarrow (K(x) \rightarrow K(y)))]$

(K_3) $(\forall x)[\mathsf{Sent}(x) \rightarrow (K(x) \rightarrow K(\dot{K}(\dot{\bar{x}}))]$

We also have two rules: DK-Intro and DT-Intro:

$$\frac{\varphi \wedge D(\ulcorner \varphi \urcorner)}{K(\ulcorner \varphi \urcorner)}$$

$$\frac{\varphi \wedge D(\ulcorner \varphi \urcorner)}{T(\ulcorner \varphi \urcorner)}$$

Theorem

DTK *is consistent.*

Gödel Revisited

Recall Gödel's three central claims:

(1) *For any formal system* F, $F \subseteq T \rightarrow F \subsetneq T$.

(2) *For any formal system* F, $K(F \subseteq T) \rightarrow F \subsetneq K$.

(3) *Either* $\neg\exists F\,(F = K)$ *or* $\exists\varphi\,(\varphi \in T \wedge \varphi \notin K \wedge \neg\varphi \notin K)$.

The first claim is easily provable in DTK.

For the second claim, once again we have:

Theorem

Assume that T includes DTK. *Suppose $F(x)$ is a formula in L_{PA} such that*

$$T \vdash K(\ulcorner F(\ulcorner \varphi \urcorner) \rightarrow \varphi \urcorner)$$

for each $\varphi \in L_{\mathrm{PA}}$. Then there is a sentence $\bar{\varphi}$ such that

$$T \vdash K(\ulcorner \bar{\varphi} \urcorner) \wedge K(\ulcorner \neg F(\ulcorner \bar{\varphi} \urcorner) \urcorner).$$

For the third claim, once again we have:

Theorem

$\mathrm{DTK} \vdash \mathrm{GD}$

So... Which disjunct holds?

Penrose Revisited—The Particular Argument

Theorem

For all $\varphi \in L_{\mathsf{DTK}}$

$$\mathsf{DTK} \vdash D(\ulcorner \varphi \urcorner) \rightarrow (T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi).$$

So we have T-Out and DT-In.

Recall that Penrose's new argument uses:

(a) $K \rightarrow T$

(b) T-In

(c) K-Intro

We have (a) but in place of (b) and (c) to skirt inconsistency we now only have the restricted forms DT-In and DK-Intro.

To apply these rules we need to know that the sentences in question—namely, "$K = F_e \rightarrow F_{e^+} \subseteq T$" and "$K(K = F_e \rightarrow G(F_{e^+}))$" are determinate.

Theorem

Assume DTK. *The statements*

(1) $K = F_e \rightarrow F_{e^+} \subseteq T$

(2) $K(K = F_e \rightarrow G(F_{e^+}))$

are indeterminate.

So this *particular* argument for ¬WMT fails in DTK.

But perhaps there is *another* argument ....

The General Situation

In fact:

Theorem

$\mathsf{DTK} \vdash \neg D(\mathsf{GD}) \wedge \neg D(\neg \mathsf{WMT}) \wedge \neg D(\mathsf{AU})$.

There is a particular irony here. For to have a hope of proving the first disjunct in its full generality we have had to switch to a system in which truth is type-free. But when we make this shift we have to be on the lookout for sentences which (like the liar sentence) are indeterminate. The irony is that the very statement we set out to prove—the first disjunct, in the full version—is provably indeterminate.

But perhaps if we restrict to determinate sublanguages then we secure determinateness.

In fact, we do:

Theorem

$\mathsf{DTK} \vdash D(\mathsf{GD}_{\mathsf{PA}}) \wedge D(\neg\mathsf{WMT}_{\mathsf{PA}}) \wedge D(\mathsf{AU}_{\mathsf{PA}})$.

So in this case there is at least a hope of getting started.

Theorem

Assume that DTK *is correct for arithmetical statements. Then* DTK *can neither prove nor refute either* $\neg \mathrm{WMT}_{\mathrm{PA}}$ *or* $\mathrm{AU}_{\mathrm{PA}}$.

In other words, in the restricted case we have determinateness, the disjunction is provable, but neither disjunct is provable or refutable in DTK.

In particular, there is *no* argument for the first disjunct in DTK.

One might try to salvage the situation by bolstering the theory of knowledge.

But the result is robust.

For example, let DTK′ be the result of adding

(K_4) $(\forall x)[\mathrm{Sent}(x) \wedge K(\ulcorner D(x) \urcorner) \rightarrow K(\ulcorner K(\dot{\bar{x}}) \rightarrow T(\dot{\bar{x}}) \urcorner)]$

(K_5) The scheme: For each formula φ

$$K(\ulcorner (\forall x)\varphi(x) \urcorner) \rightarrow (\forall x)\, K(\ulcorner \varphi(\dot{\bar{x}}) \urcorner).$$

Theorem

Assume that DTK' *is correct for arithmetical statements. Then* DTK *can neither prove nor refute either* $\neg\mathsf{WMT}_{\mathsf{PA}}$ *or* $\mathsf{AU}_{\mathsf{PA}}$.

I hope that this puts to rest the second generation of arguments for the first disjunct.

Thank You

Appendix: Reconstruction of Penrose's New Argument

Here is the argument as it appears in "Gödel, the Mind, and the Laws of Physics" (2011):

> Though I do not 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.

To facilitate the reconstruction, as a preliminary step, it will be helpful to

(i) expand the argument by unpacking parallel constructions,

(ii) render explicit a piece of argumentation implicit in the penultimate sentence,

(iii) number the steps, and

(iv) indicate (with red boxes) what appear to be two misprints

[1] Though I do not know that I necessarily am F, I conclude that if I were [F], then the system F would have to be sound [2] and, more to the point, [if I were F, then] F^* would have to be sound, where F^* is F supplemented by the further assertion 'I am F.' [3] I perceive that it follows from the assumption that 'I am F' that the Gödel statement $G(F^*)$ would have to be true [4] and, furthermore, that [it follows from the assumption that 'I am F' that the Gödel statement $G(F^*)$] would not be a consequence of F^*. [5] But I have just perceived that 'if I happened to be F, then $G(F^*)$ would have to be true,' [6] and perceptions of this nature would be precisely what $\boxed{F^*}$ is supposed to achieve, [which, alas, it cannot.] [7] Since I am therefore capable of perceiving something beyond the powers of $\boxed{F^*}$, I deduce that I cannot be F after all.

"Though I do not know that I necessarily am F, I conclude that if I were F, then the system F would have to be sound ..."

(1) $K = F \rightarrow F \subseteq T$

- This follows by the rule '$K \subseteq T$', which asserts that knowledge implies truth: In short, given $K \subseteq T$, if we have $K = F$ then we trivially have $F \subseteq T$.

"... and, more to the point, if I were F, then F^* would have to be sound, where F^* is F supplemented by the further assertion 'I am F.'"

(2) $K = F \rightarrow F^* \subseteq T$

- Here F^* is the system $F + K = F$. By (1), on the assumption of $K = F$ we already have that $F \subseteq T$. So to conclude, on the assumption of $K = F$, that $F^* \subseteq T$, we just need that '$K = F$' is also in T. But this follows from the rule 'T-In', for letting φ be '$K = F$' the rule 'T-In' yields '$K = F \rightarrow T(K = F)$'.

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

(3) $K = F \rightarrow G(F^*)$

- This follows from (2) and reflection on the nature of the Gödel sentence: For (2) gives us (on the assumption of $K = F$) that F^* is sound, and from this it follows that the Π^0_1-statement $G(F^*)$ is true.

"... and, furthermore, that it follows from the assumption that 'I am F' that the Gödel statement $G(F^*)$ would not be a consequence of F^*."

(4) $K = F \rightarrow F^* \nvdash G(F^*)$

- ▷ This line follows from (2) by the incompleteness theorem.

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

(5) $K(K = F \rightarrow G(F^*))$

- For convenience, let '(†)' abbreviate the statement '$K = F \rightarrow G(F^*)$'. In (3) we concluded (†). Penrose is now, reflecting on that fact, claiming to *know* (or *perceive*) it. Formally speaking he is applying the rule 'K'-Intro to get from (3) to (5).

"... and perceptions of this nature would be precisely what F^* is supposed to achieve, which, alas, it cannot."

(6) $K = F \rightarrow \left(F \nvdash (K = F \rightarrow G(F^*))\right)$

▷ This line follows from (4), by the Deduction Theorem and the definition of F^*: For the consequent of (4) is just '$F + (K = F) \nvdash G(F^*)$' which by the Deduction Theorem is equivalent to the statement '$F \nvdash (K = F \rightarrow G(F^*))$'.

"Since I am therefore capable of perceiving something beyond the powers of F^*, I deduce that I cannot be F after all."

(7) $K \neq F$

- This line follows from (5) and (6), since we have shown, in (6) that K knows something, namely, (†), which, on the assumption that $K = F$, F cannot know (line (5)). And so, if $K = F$, then $K \neq F$. Hence $K \neq F$.

Putting it all together:

(1) $K = F \rightarrow F \subseteq T$

(2) $K = F \rightarrow F^* \subseteq T$

(3) $K = F \rightarrow G(F^*)$

(4) $K = F \rightarrow F^* \nvdash G(F^*)$

(5) $K(K = F \rightarrow G(F^*))$

(6) $K = F \rightarrow \left(F \nvdash (K = F \rightarrow G(F^*))\right)$

(7) $K \neq F$