On the Computability of the Human Mind: Logical Error in Penrose's Argumentation

Ina Mäurer
ina-m@gmx.de

Lars Mäurer
Lars.Maeurer@uni-jena.de

Andreas Zollmann
zollmann@gmx.de

Dresden, September 22, 2003

Abstract

Roger Penrose gives in [4] a proof to show that there were no Turing Machine enumerating a certain subset of the set of all Turing Machines which applied to their own Gödel number do not terminate, whilst human mathematicians were in principle able to enumerate this set by giving non-termination proofs for each of its members. Therefore, Penrose concludes, mathematical thinking, and hence the human mind, were not computable. We show that the proof of Penrose as given in [4] is false and give a corrected theorem. It states now that either there is no Turing Machine enumerating the above-mentioned set, or if there is, menkind will never prove it to be the right one.

1 Introduction and Background

To judge the power of computer programmes it is necessary to find a formal description for it. Alan Turing [6] tried to decompose the structure of machines into elementary mathematical expressions. All existing computer architectures upto today are at most as powerful as a Turing Machine (TM). We assume here a background knowledge of computability theory and recommend the reader [1, 5, 6] for basic definitions and results.

Let $\{M_i \mid i \in \mathbb{N}\}$ be the set of all TMs mapping $\mathbb{N}$ partially to $\mathbb{N}$. The domain of M_i is denoted by D_i. So, we have

$$D_i := \{n \in \mathbb{N} \mid M_i(n) \text{ halts}\}.$$

We denote by RE the set of all *recursively enumerable* sets:

$$RE := \{D_i \mid i \in \mathbb{N}\}.$$

A subset of $\mathbb{N}$ is *computable* if it coincides with some D_i with $i \in \mathbb{N}$.

2 Penrose's Proof

We present now the proof by Penrose given in [4, pp. 72–77] in order to show the error in the argumentation. If the i-th TM $M_i(j)$ applied to j does not terminate (terminates), we write $M_i(j) = \perp$ ($M_i(j) \neq \perp$) respectively for short.

Let $\mathcal{D}$ be the following set:

$$\mathcal{D} := \left\{ i \in \mathbb{N} \;\middle|\; \begin{array}{l} M_i(i) = \perp \text{ and human mathematicians} \\ \text{are in principle able to prove that } M_i(i) = \perp \end{array} \right\}.$$

Objections as to whether this set is actually well-defined are met in [4] and shall not concern us here.

Claimed Theorem 2.1 (Penrose). *It is $\mathcal{D} \notin RE$.*

Proof: Assuming $\mathcal{D} \in RE$, i.e., there is a fixed $k \in \mathbb{N}$, such that

$$\forall i \in \mathbb{N}.\ M_k(i) \neq \perp \Leftrightarrow i \in \mathcal{D}, \tag{1}$$

it follows

$$M_k(k) \neq \perp \Leftrightarrow k \in \mathcal{D}. \tag{2}$$

By definition of $\mathcal{D}$, Penrose concludes

$$k \in \mathcal{D} \Rightarrow M_k(k) = \perp. \tag{3}$$

Now, the conjunction of (2) and (3) is of the form $(\neg A \Leftrightarrow B) \wedge (B \Rightarrow A)$ (where $A := [M_k(k) \neq \perp]$ and $B := [k \in \mathcal{D}]$), which implies A. Thus he gets:

$$\boxed{M_k(k) = \perp}$$

His argument now is that the last fact was convincingly demonstrated by a mathematician, the cautious reader was present. So he concludes applying the definition of $\mathcal{D}$:

$\Longrightarrow k \in \mathcal{D}$ and thus with (2):

$\boxed{M_k(k) \neq \perp}$, which is a contradiction. □

The logical error in the proof is the step '$M_k(k) = \perp$ *was convincingly demonstrated.*' What actually has been demonstrated is only that the assumption $\mathcal{D} \in RE$ implies that $\exists k.(2) \wedge (3) \wedge M_k(k) = \perp$. There is a mathematical proof for

$$[\mathcal{D} \in RE \Rightarrow (\exists k.(2) \wedge (3) \wedge M_k(k) = \perp)],$$

but from assuming that $\mathcal{D} \in RE$ holds, it cannot be concluded that human mathematicians are able to find a proof for $\mathcal{D} \in RE$, and thus it cannot be concluded that human mathematicians are able to find out

$$[\exists k.(2) \wedge (3) \wedge M_k(k) = \perp],$$

although that statement would then be true.

In order to make this more clear and to prove Theorem 2.2, we introduce a notion for the term *mathematical proof.* Let $\mathbb{B}$ be the Boolean semiring and T the class of all mathematical Boolean assertions. We introduce the function $\boldsymbol{Prf} : T \to \mathbb{B}$, such that

$$\boldsymbol{Prf}(A) := \begin{cases} \text{true,} & A \text{ is true and human mathematicians are} \\ & \text{in principle able to prove that } A \text{ is true} \\ \text{false,} & \text{otherwise.} \end{cases}$$

One can also regard $\boldsymbol{Prf}$ as a predicate and think of $\boldsymbol{Prf}(a)$ simply as '*there is a proof of A which can be found by human reasoning.*' It is easy to see that the following properties hold:

$$\boldsymbol{Prf}(A) \Rightarrow A \tag{4}$$

$$[\boldsymbol{Prf}(A) \wedge \boldsymbol{Prf}(A \Rightarrow B)] \Rightarrow \boldsymbol{Prf}(B) \tag{5}$$

Note that we can now describe the set $\mathcal{D}$ as follows:

$$\mathcal{D} := \{i \in \mathbb{N} \mid \boldsymbol{Prf}(M_i(i) = \perp)\}$$

Looking back at the Penrose proof above, we find that the derivation of $M_k(k) = \perp$ is correct and even independent of the particular k satisfying (1). This gives us for each $k \in \mathbb{N}$ a mathematical proof for

$$(\forall i \in \mathbb{N}.\ M_k(i) \neq \perp \Leftrightarrow i \in \mathcal{D}) \Rightarrow M_k(k) = \perp \tag{6}$$

and thus it holds

$$\forall k \in \mathbb{N}.\ \boldsymbol{Prf}[(\forall i \in \mathbb{N}.\ M_k(i) \neq \perp \Leftrightarrow i \in \mathcal{D}) \Rightarrow M_k(k) = \perp]. \tag{7}$$

Now, the indirect assumption $\mathcal{D} \in RE$ (i.e., *there is a k such that* $\forall i \in \mathbb{N}.\ M_k(i) \neq \perp \Leftrightarrow i \in \mathcal{D}$) enables us to discharge the antecedent of (6), obtaining $M_k(k) = \perp$.

What we also want, however, is $\boldsymbol{Prf}(M_k(k) = \perp)$, in order to conclude $k \in \mathcal{D}$ (using the definition of $\mathcal{D}$) and thus with (2) the contradiction $M_k(k) \neq \perp$. For that to work, we would have to apply (5) to (7) so as to get rid of the antecedent nested in the $\boldsymbol{Prf}$ predicate. But for that to work,

$$\boldsymbol{Prf}(\forall i \in \mathbb{N}.\ M_k(i) \neq \perp \Leftrightarrow i \in \mathcal{D}) \tag{8}$$

must hold for some k. The following theorem ensures that by strengthening the indirect assumption to exactly that condition:

Theorem 2.2. *For no $k \in \mathbb{N}$, $\boldsymbol{Prf}(M_k$ computes $\mathcal{D})$ holds. Or equivalently:*

$$\nexists k \in \mathbb{N}. \boldsymbol{Prf}[\forall i \in \mathbb{N}.\ M_k(i) \neq \perp \Leftrightarrow i \in \mathcal{D}]$$

Proof: Assume there is a k such that

$$\boldsymbol{Prf}[\forall i \in \mathbb{N}.\ M_k(i) \neq \perp \Leftrightarrow i \in \mathcal{D}] \tag{9}$$

holds. With (4), this leads us to

$$\forall i \in \mathbb{N}.\ M_k(i) \neq \perp \Leftrightarrow i \in \mathcal{D}. \tag{10}$$

Applying (5) to (9) and (7), we obtain

$$\boldsymbol{Prf}[M_k(k) = \perp], \tag{11}$$

and thus, using (4), $\boxed{M_k(k) = \perp}$, as in Penrose's proof. But now we can apply the definition of $\mathcal{D}$ to (11) and have $k \in \mathcal{D}$, so that (10) leads us to the contradiction $\boxed{M_k(k) \neq \perp}$. □

3 Afterthoughts

Let us look back again at what we have just proved. The theorem states that if a machine M_k that perfectly captures human mathematical thinking (i.e., M_k computes $\mathcal{D}$) indeed exists, humanity will never be certain that this machine is actually the right one. Someone might well come up with that machine M_k, but she could never prove that it does its job.

Since the universal quantification of k is outside the $\boldsymbol{Prf}$ predicate, the theorem does not rule out the possibility that humans prove the existence of a machine computing $\mathcal{D}$. However, this proof is then doomed to be non-constructive. At first sight one might be seduced to think that the theorem could be altered to *'It holds $\neg\boldsymbol{Prf}(\mathcal{D} \in RE)$.'* by moving k into the scope of $\boldsymbol{Prf}$. But then from the indirect assumption $\boldsymbol{Prf}(\mathcal{D} \in RE)$, i.e.,

$$\boldsymbol{Prf}[\exists k \in \mathbb{N}.\ \forall i \in \mathbb{N}.\ M_k(i) \neq \perp \Leftrightarrow i \in \mathcal{D}],$$

(11) cannot be concluded anymore using the technique above, since $\boldsymbol{Prf}(\exists k. F[k])$ does not necessarily imply $\exists k. \boldsymbol{Prf}(F[k])$.

References

[1] J. E. Hopcroft; J. D. Ullman. Introduction to Automata Theory, Languages and Computation. *Addison Wesley, Reading, MA*, 1979.

[2] Lars Mäurer; Ben Schlingelhof. Berechenbarkeit des Denkens?. *Wurzel*, Bd. 01/00, 2000.

[3] Roger Penrose. *"The Emperor's **New Mind**: Concerning Computers, Minds, and the Laws of Physics"*. Oxford University Press, 1989.

[4] R. Penrose. Shadows of the Mind – A Search for the Missing Science of Consciousness. *Oxford University Press*, 1994.

[5] U. Schöning. Theoretische Informatik kurz gefasst. *BI-Wiss.-Verl.*, 1992.

[6] A. Turing. On Computable Numbers with an Application to the Entscheidungsproblem. *Proc. Lond. Math. Soc.(2), Vol. 42*, p. 230–265, Corrections (43), p. 544–546, 1936-1937.