On Turing Machines Knowing Their Own Gödel-Sentences

by
Neil Tennant*

September 20, 2004

Abstract

Storrs McCall appeals to a particular true-but-unprovable sentence of formal arithmetic to argue, by appeal to its irrefutability, that human minds transcend Turing machines. Metamathematical oversights in McCall's discussion of the Gödel phenomena, however, render invalid his philosophical argument for this transcendentalist conclusion.

There is a long-standing debate on whether human minds transcend Turing machines. The debate goes back to seminal and opposing papers by John Lucas and Paul Benacerraf,[1] and has been popularized in best-selling books by Douglas Hofstadter and Roger Penrose.[2]

In a recent paper in a leading generalist journal,[3] Storrs McCall has joined this debate. He sides with the transcendentalists. He appeals to a particular true-but-unprovable sentence of formal arithmetic to argue that human minds *do* transcend Turing machines. It is not the truth or the unprovability of this sentence, however, on which McCall focuses; rather, it is its *irrefutability*. This adds a new wrinkle, which needs to be ironed out by the anti-transcendentalist.

Metamathematical oversights in McCall's discussion render his conclusion less strongly supported than he believes. Readers of a more specialist journal in this area will be interested in how a correct statement of the

*I am grateful to Harvey Friedman for helpful discussion, and for the first proof of Theorem 2 below. Two anonymous referees made suggestions that considerably improved the exposition.

metamathematical facts reveals the invalidity of McCall's philosophical argument for his transcendentalist conclusion. Because so much depends on getting these basic metamathematical facts straight, we state and prove, in §1, some relevant results concerning the Gödel phenomena. In §2, we apply these results in criticism of McCall's argument.

1 Metamathematical results

Let PA be the system of Peano arithmetic. In what follows, and unless otherwise indicated, provability will be understood as provability within PA. By $\underline{n}$ we mean the numeral in the language of PA for the natural number n. (This numeral will be of the form $s \ldots s0$, with n occurrences of the symbol for the successor function.) We assume that each syntactic item φ has its own unique code number. By $\overline{\varphi}$ we mean the numeral for that code number. A basic result in the theory of the coding of syntax is that every one-place predicate $\psi(x)$ has a *fixed point*, that is, a sentence φ that is interdeducible, in PA, with $\psi(\overline{\varphi})$.

It is useful to talk more generally about systems S of formal arithmetic. PA is but one such system. Every system S that we shall consider will be assumed to be complete for bounded sentences: that is, if φ is a true sentence of the formal language of arithmetic containing no unrestricted quantifications, then S proves φ. Obviously PA has this property.

The two-place proof-predicate $\mathrm{Pf}_S(x, y)$ in the formal language of first-order arithmetic is so defined that it represents the relation 'x is the code number of a proof, in the system S, of the sentence with code number y'. Such representation takes place in the following sense:[4]

> for all natural numbers n, m,
> if n is the code number of a proof, in the system S, of the sentence with code number m, then S proves $\mathrm{Pf}_S(\underline{n}, \underline{m})$; and
> if n is *not* the code number of a proof, in the system S, of the sentence with code number m, then S refutes $\mathrm{Pf}_S(\underline{n}, \underline{m})$.

The one-place provability-predicate $\mathrm{P}_S(y)$ is defined as $\exists x \mathrm{Pf}_S(x, y)$. When S is the system PA, we shall suppress the subscripts on these predicates. We shall re-write $\mathrm{P}(\overline{\varphi})$ as $\mathrm{P}[\varphi]$. Sometimes we shall even omit the brackets, when no confusion can result.

We shall use the turnstile ($\vdash$) to represent deducibility in first-order Peano arithmetic. (In many contexts, we can get away with a system of

arithmetic considerably weaker than PA, but for our present purposes PA will do.) The following are well-known facts about the provability predicate:

(i) $\dfrac{\vdash \varphi}{\vdash \mathrm{P}[\varphi]}$

(ii) $\mathrm{P}[\varphi], \mathrm{P}[\varphi \to \psi] \vdash \mathrm{P}[\psi]$.[5]

The following further principle is immediate:

(†) $\dfrac{\varphi \vdash \psi}{\mathrm{P}[\varphi] \vdash \mathrm{P}[\psi]}$

Proof:
$$\dfrac{\dfrac{\dfrac{\varphi \vdash \psi}{\vdash \varphi \to \psi}\,{\to}R}{\vdash \mathrm{P}[\varphi \to \psi]}\,\text{(i)} \qquad \mathrm{P}[\varphi], \mathrm{P}[\varphi \to \psi] \vdash \mathrm{P}[\psi]\,\text{(ii)}}{\mathrm{P}[\varphi] \vdash \mathrm{P}[\psi]}\,\text{CUT}$$

The 'Gödel sentence' G for PA is by definition the fixed point of the 'unprovability predicate' $\neg\mathrm{P}$. Thus G is interdeducible in PA with $\neg\mathrm{P}[G]$; hence, $\neg G$ with $\mathrm{P}[G]$. The informal claim that *PA is consistent* is expressed in the language of PA by some suitable sentence, such as $\neg\mathrm{P}[0{=}s0]$ or $\neg\mathrm{P}[\bot]$, which we render also as $\mathrm{Con}_{\mathrm{PA}}$. It is a well-known fact that in PA, G is interdeducible with $\mathrm{Con}_{\mathrm{PA}}$. Hence in PA we have $\neg G$ interdeducible with $\mathrm{P}[\bot]$.

It will be an adequacy condition on $\mathrm{Pf}_{\mathrm{PA}+\varphi}(x, y)$ and on $\mathrm{Pf}_{\mathrm{PA}}(x, y)$ representing their respective proof-relations that $\exists x \mathrm{Pf}_{\mathrm{PA}+\varphi}(x, \bot)$ be equivalent, in PA, to $\exists x \mathrm{Pf}_{\mathrm{PA}}(x, \neg\varphi)$. For this requires only that each proof-predicate take account of $\neg I$ as a means for constructing proofs of negations from proofs of $\bot$. It follows that in PA, $\neg\mathrm{P}[\neg\varphi]$ is equivalent to $\mathrm{Con}_{\mathrm{PA}+\varphi}$.

Theorem 1. The non-refutability of the Gödel-sentence is equivalent (in PA) to the consistency of PA with its own consistency statement.

Proof. As noted above, $\neg\mathrm{P}[\neg\varphi]$ is equivalent, in PA, to $\mathrm{Con}_{\mathrm{PA}+\varphi}$. For φ take $\mathrm{Con}_{\mathrm{PA}}$ itself, i.e. $\neg\mathrm{P}\bot$. Then $\neg\mathrm{P}[\neg(\neg\mathrm{P}\bot)]$, i.e. $\neg\mathrm{P}[\mathrm{P}\bot]$, is equivalent, in PA, to $\mathrm{Con}_{\mathrm{PA}+\mathrm{Con}_{\mathrm{PA}}}$. We now show that $\neg\mathrm{P}[\mathrm{P}\bot]$ is equivalent, in PA, to $\neg\mathrm{P}[\neg G]$—the statement that G is not refutable. From left to right the argument is as follows:

$$\dfrac{\dfrac{\dfrac{\dfrac{\neg G \vdash \mathrm{P}\bot}{\vdash \neg G \to \mathrm{P}\bot}\,\text{(i)}}{\vdash \mathrm{P}[\neg G \to \mathrm{P}\bot]} \qquad \begin{array}{c}\text{by (ii):}\\ \mathrm{P}[\neg G], \mathrm{P}[\neg G \to \mathrm{P}\bot] \vdash \mathrm{P}[\mathrm{P}\bot]\end{array}}{\dfrac{\mathrm{P}[\neg G] \vdash \mathrm{P}[\mathrm{P}\bot]}{\neg \mathrm{P}[\mathrm{P}\bot] \vdash \neg \mathrm{P}[\neg G]}}\,\text{CUT}$$

From right to left, the argument is as follows:

by *ex falso quodlibet*:

$$\dfrac{\dfrac{\dfrac{\dfrac{\bot \vdash G}{\mathrm{P}\bot \vdash \mathrm{P}[G]}\,(\dagger) \qquad \begin{array}{c}\text{by choice of } G:\\ \mathrm{P}[G] \vdash \neg G\end{array}}{\mathrm{P}\bot \vdash \neg G}\,\text{CUT}}{\mathrm{P}[\mathrm{P}\bot] \vdash \mathrm{P}[\neg G]}\,(\dagger)}{\neg \mathrm{P}[\neg G] \vdash \neg \mathrm{P}[\mathrm{P}\bot]}$$

QED

The principle of Σ^0_1-*reflection* for any system S states that any S-provable Σ^0_1-sentence is true (in the standard model). In other words, any existential quantification of a bounded formula is provable in S only if it has a 'witness' among the standard natural numbers.

Lemma 1. If Σ^0_1-reflection holds for S, then $\mathrm{P}_S(y)$ weakly represents provability in S—that is, for all φ, S proves $\mathrm{P}_S(\overline{\varphi})$ if and only if S proves φ.

Proof. First suppose that S proves φ. So there is a proof, in system S, of φ. Let n be the code number of such a proof. By representability, S proves $\mathrm{Pf}_S(\underline{n}, \overline{\varphi})$. By $\exists I$, S proves $\exists x \mathrm{Pf}_S(x, \overline{\varphi})$; that is, S proves $\mathrm{P}_S(\overline{\varphi})$.

For the converse, suppose that S proves $\mathrm{P}_S(\overline{\varphi})$, that is, S proves $\exists x \mathrm{Pf}_S(x, \overline{\varphi})$. By Σ^0_1-reflection for S, there is some standard natural number n such that $\mathrm{Pf}_S(\underline{n}, \overline{\varphi})$. This last sentence is bounded. By completeness of S for true bounded sentences, S proves $\mathrm{Pf}_S(\underline{n}, \overline{\varphi})$. Since S is consistent, S does not refute $\mathrm{Pf}_S(\underline{n}, \overline{\varphi})$. Hence by contraposition on the statement of representability, it is not the case that n is not the code number of a proof in the system S of the sentence φ. That is to say, n *is* the code number of a proof in S of φ.[6] That is, S proves φ. *QED*

Note how we have appealed to Σ^0_1-reflection in the foregoing proof, for the direction of proof from 'S proves $\mathrm{P}_S(\overline{\varphi})$' to '$S$ proves φ'.

The *ω-consistency of S* can be defined as follows:

> if there exist proofs in S of $\psi(0), \psi(\underline{1}), \psi(\underline{2}), \ldots$, then it is consistent with S to assume $\forall x \psi(x)$.

The *1-consistency of S* is a special case of ω-consistency:

> for every *bounded* $\psi(x)$, if there exist proofs in S of $\psi(0), \psi(\underline{1}), \psi(\underline{2}), \ldots$, then it is consistent with S to assume $\forall x \psi(x)$.

It is well-known that the 1-consistency of any system S is equivalent to the principle of Σ^0_1-reflection for S. Moreover, the 1-consistency of S implies, but is not implied by, the consistency of S.[7]

Theorem 2. If (PA+$\mathrm{Con_{PA}}$) is consistent, then PA does not prove $\mathrm{Con_{PA}} \rightarrow \neg \mathrm{P}[\neg G]$.
Proof. Suppose, for *reductio*, that PA proves $\mathrm{Con_{PA}} \rightarrow \neg \mathrm{P}[\neg G]$. In PA, assume $\mathrm{Con_{PA}}$. By modus ponens, infer $\neg \mathrm{P}[\neg G]$. Now $\mathrm{Con_{PA}}$ is equivalent to G in PA. So we may now infer $\neg \mathrm{P}[\neg \mathrm{Con_{PA}}]$. This in turn implies $\mathrm{Con_{PA+Con_{PA}}}$. Thus the system PA+$\mathrm{Con_{PA}}$ has proved its own consistency. But Gödel's Second Incompleteness Theorem says that this is impossible, if the system PA+$\mathrm{Con_{PA}}$ is consistent. So if PA+$\mathrm{Con_{PA}}$ is consistent, then PA does not prove $\mathrm{Con_{PA}} \rightarrow \neg \mathrm{P}[\neg G]$. *QED*

Theorem 2 can be re-proved as a formal unprovability result: $\neg \mathrm{PP}\bot \vdash \neg \mathrm{P}[\neg \mathrm{P}\bot \rightarrow \neg \mathrm{PP}\bot]$. We obtain a formal version of Theorem 2 by substituting $\neg G$ for the final occurrence of $\mathrm{P}\bot$: $\neg \mathrm{PP}\bot \vdash \neg \mathrm{P}[\neg \mathrm{P}\bot \rightarrow \neg \mathrm{P}\neg G]$. The proof below uses Löb's Theorem.[8] In what follows, boldface **P** can be replaced by any finite (and possibly null) sequence of occurrences of P.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\neg\varphi \rightarrow \neg \mathrm{P}\varphi \vdash \mathrm{P}\varphi \rightarrow \varphi}{\mathrm{P}[\neg\varphi \rightarrow \neg \mathrm{P}\varphi] \vdash \mathrm{P}[\mathrm{P}\varphi \rightarrow \varphi]}\,(\dagger) \qquad \text{Löb's Theorem: } \mathrm{P}[\mathrm{P}\varphi \rightarrow \varphi] \vdash \mathrm{P}\varphi}{\mathrm{P}[\neg\varphi \rightarrow \neg \mathrm{P}\varphi] \vdash \mathrm{P}\varphi}\,\text{CUT}}{\neg \mathrm{P}\varphi \vdash \neg \mathrm{P}[\neg\varphi \rightarrow \neg \mathrm{P}\varphi]}}{\neg \mathrm{PP}\bot \vdash \neg \mathrm{P}[\neg \mathrm{P}\bot \rightarrow \neg \mathrm{PP}\bot]}\,\mathbf{P}\bot/\varphi \qquad \text{Logic:}$$

All we need here is the case where **P** is simply P.[9] Overall, we have the following implications, but, as indicated, not their converses:[10]

PA is 1-consistent $\overset{\rightarrow}{\not\leftarrow}$ $\mathrm{Con_{PA}}$ is consistent with PA $\overset{\rightarrow}{\not\leftarrow}$ PA is consistent.

2 Critique of McCall's philosophical argument

2.1 The 'second half' of Gödel's first incompleteness theorem

McCall considers the following two sentences (loc. cit., p. 525).

1. If PA is consistent, then G is not provable.
2. If PA is consistent, then $\neg G$ is not provable.

He claims

> Both sentences are true, the difference being that the formal version of [1] in the language of PA is a theorem, whereas the formal version of [2] to the best of our knowledge is not.

What McCall says here is inadequate, on two scores. First, our reasons for claiming that (1) is true are very different from our reasons for claiming that (2) is true. We know that (1) is true because we can prove it within PA as it stands. But (2), even if true, is not provable within PA as it stands. The epistemic (or doxastic) situation with regard to (2) is that we think that what makes it true is simply the truth of its consequent. And we believe that its consequent is true only because we believe (something at least as strong as) that $\mathrm{Con}_{\mathrm{PA}}$ is consistent with PA—a stronger condition than mere consistency of PA, which cannot be justified by proof within PA (or even within PA+$\mathrm{Con}_{\mathrm{PA}}$); or because we believe a much stronger theory, such as set theory, which proves the consequent of (2). But as Theorem 2 shows, the antecedent of (2), as it stands, is too weak for us to be able to deduce the consequent from it within PA.

The following weakening of (2), however, *is* provable in PA, containing as it does an antecedent that is sufficiently stronger than that of (2):

2′. If PA is 1-consistent, then $\neg G$ is not provable.

This is the canonical way of stating the 'second half' of Gödel's First Incompleteness Theorem. The theorem states the existence of a sentence G such that the conditionals (1) and (2′) both hold.

As it happens, (2) need not be weakened as much as it is by (2′). The following statement, as Theorem 1 shows, is provable in PA:

2″. If $\mathrm{Con}_{\mathrm{PA}}$ is consistent with PA, then $\neg G$ is not provable.

(2″) is intermediate in strength between (2) and (2′), because its antecedent is intermediate in strength between their antecedents (on an assumption weaker than that PA is 1-consistent).

It is an interesting question, but one which will not detain us here, why the canonical statement of the second half of Gödel's First Incompleteness Theorem is not simply (2″). The latter is stronger (than (2)), for having a weaker antecedent. Indeed, Theorem 1 shows that the antecedent of (2″) is of exactly the required strength; nothing weaker would do.

The second score on which McCall's discussion is inadequate concerns the epistemic status of the formal version of (2), which formal version he writes as[11]

3. $\mathrm{Con}_{\mathrm{PA}} \rightarrow \neg \mathrm{P}[\neg G]$

McCall claims merely that 'there are good reasons to believe that [3] is in fact unprovable in PA.' The 'good reasons' that he adduces are not conclusive, and he writes of 'the absence of a knock-down independence proof.' But by Theorem 2, (3) *is* provably unprovable in PA, on the assumption that $(\mathrm{PA} + \mathrm{Con}_{\mathrm{PA}})$ is consistent. McCall does not specify what would count, for him, as a 'knock-down' proof; but we can presume that he would find acceptable the assumption that $\mathrm{Con}_{\mathrm{PA}}$ is consistent with PA.

Why did McCall think that (2) and (3) were true? On p. 529 he offers an informal proof of (3), in which he makes a crucial inference from $\vdash \mathrm{P}[G]$ to G's being a theorem of PA.[12] This move he justifies by appeal to the fact that the provability predicate weakly represents theoremhood. What McCall has overlooked here is that weak representability enables one to move in this direction only courtesy of Σ^0_1-reflection (as we saw in the proof of Lemma 1 above), which (as already noted) is equivalent to the 1-consistency of PA.[13]

McCall enters the remarkable footnote (fn. 3, p. 529)

> It is frequently stated that proving the nontheoremhood of $\neg G$ requires the assumption that PA is not only consistent but ω-consistent ... The argument given in the text shows, however, that the weaker assumption of ordinary consistency suffices.

But as we have just seen, the weaker assumption of ordinary consistency has *not* been shown to suffice. McCall does not realize that weak representability, in the very direction he exploits, depends on the stronger assumption of 1-consistency.

It would, indeed, be remarkable if the non-refutability of the Gödel-sentence G in PA could be established so easily by appeal only to the ordinary consistency of PA. McCall's claim to have established this would mean that Rosser's ingenuity in constructing a new kind of independent sentence, the proof of whose undecidability-in-PA required only the assumption of ordinary consistency,[14] had been misplaced. As Boolos (*op. cit.*, p. xxii) stresses,

> It is noteworthy that Rosser did not show (*and could not have shown*) the undecidability of *the statements described by Gödel* to follow from the simple consistency of the relevant systems; instead he found *different* statements whose undecidability so follows. [My first two emphases–NT.]

We have seen that the non-refutability of G implies *the consistency of PA with its own consistency statement*; and that the latter italicized condition in turn implies its own non-deducibility from the mere consistency of PA. This refutes McCall's claim that mere consistency of PA suffices for the non-refutability of G.

2.2 What a machine can do

McCall pointed out (p. 528) that (1) can be expressed as the sentence $\mathrm{Con}_{\mathrm{PA}} \rightarrow \neg \mathrm{P}[G]$ in the language of PA, or equivalently as

> 4. $\mathrm{Con}_{\mathrm{PA}} \rightarrow G$.

He explicitly conceded that 'a Turing machine can *know* the truth of [4] because it can prove [4] as a theorem.' But then he immediately claimed

> But a different result holds for the formal version of [2]. This is:
>
> [3] $\mathrm{Con}_{\mathrm{PA}} \rightarrow \neg \mathrm{P}[\neg G]$
>
> Moving from [4] to [3] yields a candidate for the true-but-unprovable sentence that we seek.

The burden of McCall's paper is to establish that there is no possible sense in which 'a Turing machine can recognize the truth of [3].' Of course, the correctness of this claim would depend on which Turing machine is in question. The background assumption is that we are considering a Turing machine 'for PA'. But then the comparison is not really fair. For, to the extent that

we (but not the machine) can recognize the truth of (3), we are not working 'within PA' (see above). So the machine's inability to recognize the truth of (3) does not make us superior to it. For the comparison to be fair, it would have to involve a machine equipped with whatever stronger assumptions (not available within PA) we ourselves have employed in order to recognize the truth of (3). The machine 'for PA', however, could prove

$$\mathrm{Con}_{\mathrm{PA}+\mathrm{Con}_{\mathrm{PA}}} \rightarrow \neg \mathrm{P}(\neg G).$$

which formally emulates human beings' insight that the Gödel-sentence is not refutable.

Notes

[1]J. R. Lucas, 'Minds, Machines and Gödel', *Philosophy*, 36, 1961, pp. 120–4; and P. Benacerraf, 'God, the Devil and Gödel', *The Monist*, 51, 1967, pp. 9–32.

[2]D. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid*, Harvester Press, 1979; and R. Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford University Press, 1989.

[3]Storrs McCall, 'Can a Turing Machine Know that the Gödel Sentence Is True?', *The Journal of Philosophy*, Vol. XCVI, No. 1, October 1999, pp. 525–532. The definite article 'the' in this title should be the possessive pronoun 'its'.

[4]*Cf.* S. C. Kleene, *Introduction to Metamathematics*, Van Nostrand, 1950, p. 195. Kleene says that the relation is 'numeralwise expressible' rather than 'representable'.

[5]See Boolos, *The Logic of Provability*, Cambridge University Press, 1993, for proofs, terminating on p. 44.

[6]Note that the apparent double negation elimination here is intuitionistically acceptable, since the conclusion is a decidable statement.

[7]See C. Smorynski, 'The incompleteness theorems', in J. Barwise, ed., *Handbook of mathematical logic*, North-Holland, Amsterdam, pp. 821–865; at p. 852. Note that where we use the predicates 'Pf' and 'P', Smorynski uses 'Prov' and 'Pr' respectively.

[8]For a proof of Löb's Theorem, see G. Boolos, *op. cit.*, pp. 56–7. The proof requires, in addition to the conditions (i) and (ii) on the provability predicate, the further condition

(iii) $\mathrm{P}[\varphi] \vdash \mathrm{P}[\mathrm{P}[\varphi]]$.

Löb's Theorem is equivalent to Gödel's second incompleteness theorem (for single-sentence extensions of PA). See Boolos, *op. cit.*, p. xxvi.

[9]The case where **P** is the null sequence yields Gödel's second incompleteness theorem, if we interpret $\neg\varphi$ as $\varphi \rightarrow \bot$: $\neg\mathrm{P}\bot \vdash \neg\mathrm{P}[(\bot \rightarrow \bot) \rightarrow \neg\mathrm{P}\bot]$; i.e., $\neg\mathrm{P}\bot \vdash \neg\mathrm{P}[\neg\mathrm{P}\bot]$.

[10]It would therefore be an error to maintain that the irrefutability of the Gödel-sentence *requires* ω-consistency. Yet in their paper 'Leveling the playing field between mind and machine: a reply to McCall', *The Journal of Philosophy*, XCVII, no. 8, August 2000, pp. 456–61, A. George and D. J. Velleman write

> [McCall's] thesis, that [If PA is consistent, then $\neg G$ is not derivable in PA], will come as a surprise to all those (including Gödel) who believe that establishing the nonprovability of $\neg G$ requires the stronger condition of ω-consistency.

George and Velleman's attribution of such a mistaken belief to Gödel, or to any other competent metamathematician, is implausible.

[11]McCall uses 'Cons' instead of the more usual 'Con'; and uses the old-fashioned 'Bew' instead of 'P'. McCall also makes Cons look like a predicate applying uniformly to (codings of) systems, by writing Cons(PA) for the *sentence* expressing the consistency of PA.

[12]McCall does not observe the usual niceties involved in distinguishing Gödel numbers from their numerals, which is why I refrain from giving a verbatim quote; for it would be misleading.

[13]McCall is in good company here. Carnap made the same move without appreciating the role of Σ^0_1-reflection—see his proof of Theorem 36.3 at p. 133 of *The Logical Syntax of Language*, tr. Amethe Smeaton, Routledge and Kegan Paul, London, 1937.

[14]J. B. Rosser, 'Extensions of some theorems of Gödel and Church', *The Journal of Symbolic Logic*, vol. 1, no. 1, pp. 87–91.