COMMENTS AND CRITICISM

LEVELING THE PLAYING FIELD BETWEEN MIND AND MACHINE: A REPLY TO MCCALL

The temptation to think that somewhere in Gödel's incompleteness theorems there is a proof that the mind is not a formal system is seemingly irresistible.

But where? The natural place to look first is *G*, the Gödel sentence of PA (a formalization of Peano arithmetic). The proof of the first incompleteness theorem reveals that, assuming PA to be consistent, PA cannot prove *G*. But since *G* in some sense says of itself that it is not provable, we minded creatures can see that it is true. Thus, it is often claimed, we can recognize as true something that PA cannot prove, and so human thought cannot be modeled by PA. Similar reasoning applies to other formal systems, and so, it is alleged, human thought cannot be modeled by any formal system.

But it is an illusion to think that we unconditionally recognize *G* to be true; rather, our recognition of its truth depends on our assuming that PA is consistent. Without this assumption, there would be no basis for asserting the truth of *G*. Furthermore, if we allow PA to make use of that very assumption, then it, too, will be able to establish *G* (for Con(PA) → *G* is derivable in PA). Our tacit reliance on the assumption of consistency is obscured in the case of the formal system PA, for its consistency just seems to go without saying. For other formal systems, consistency may be doubtful, and, if we are unwilling to assume the consistency of a system, then we will be unable to recognize the truth of its Gödel sentence. Thus, we have no evidence here that the mind cannot be represented by a formal system.

It is here that Storrs McCall[1] takes up the search for the Holy Grail of antimechanists. He suggests that instead of focusing on our capacity to recognize the truth of *G*, which he acknowledges is dependent on our recognition of the consistency of PA, we might look at the corresponding conditional claim:

(1) If PA is consistent, then *G* is not derivable in PA.

Our assertion that (1) is true is not itself crucially dependent on any assumption, such as the consistency of PA, that we are tacitly keeping

[1] "Can a Turing Machine Know that the Gödel Sentence Is True?" this JOURNAL, XCVI, 10 (October 1999): 525-32.

0022-362X/00/9708/456-61

from the formal system PA. But (1) will not do as an example of a truth that we can apprehend but PA cannot prove. For in PA, one can derive:

(1′) $\mathrm{Con(PA)} \rightarrow \sim \mathrm{Theorem}([G])$

(For any sentence *S*, $[S]$ is the formal numeral corresponding to its Gödel number. Theorem(y) is the formula $(\exists x)\mathrm{Proof}(x, y)$, where Proof($x$, y) is a formula asserting that x is the Gödel number of a proof of the sentence with Gödel number y. Thus, Theorem(y) expresses the statement that y is the Gödel number of a theorem of PA.) The conditional statement in (1′) is a formalization of part of the proof of Gödel's first incompleteness theorem. (It was the realization that this argument can be formalized which led Gödel to his second incompleteness theorem.)

McCall recognizes this, and suggests instead that we consider the following conditional claim:

(2) If PA is consistent, then $\sim G$ is not derivable in PA.

His argument has two parts: first, that (2) can be seen to be true by us and, second, that the formal counterpart of (2) cannot be derived in PA, that is, that (2′) is not a theorem of PA:

(2′) $\mathrm{Con(PA)} \rightarrow \sim \mathrm{Theorem}([\sim G])$

Therefore, he concludes, (2) finally gives us an example of a statement which we can recognize to be correct but which PA cannot prove.[2]

His argument is flawed, however. His first thesis, that (2) is correct, will come as a surprise to all those (including Gödel) who believe that establishing the nonprovability of $\sim G$ requires the stronger condition of ω-consistency.[3] McCall does offer an informal argument for (2),

[2] McCall also finds it significant that we can recognize, again without assuming the consistency of PA, that either *G* is true and unprovable in PA, or *G* is false and provable in PA (depending on whether PA is, respectively, consistent or inconsistent). He seems to believe that this is not something a machine could establish: "Human beings, on the other hand, are acquainted with both proof and truth, and also know of cases where the two diverge" (527). Yet it is difficult to see how one's recognition of these facts could support the claim that one's mind cannot be modeled by a formal system. After all, if McCall's mind were a machine, he could have arrived at precisely the same conclusions. For in PA, one can derive $(G \,\&\, \sim\mathrm{Theorem}([G])) \vee (\sim G \,\&\, \mathrm{Theorem}([G]))$: this disjunction is equivalent to $G \leftrightarrow \sim\mathrm{Theorem}([G])$, whose derivability in PA is guaranteed by the manner in which *G* is constructed.

[3] To say that a theory *T* is ω-consistent means that there is no formula $P(x)$ such that $\vdash_T (\exists x)P(x)$—that is, $(\exists x)P(x)$ is derivable in *T*—but for every natural number n, $\vdash_T \sim P(\bar{n})$, where $\bar{n}$ is the numeral for n.

but he fails to realize that the argument makes tacit use of the assumption of ω-consistency. McCall argues as follows:

> Suppose, for reductio, that PA is consistent and that $\sim G$ is a theorem. Then $[\sim G]$[4] $\in$ *Th* [the collection of Gödel numbers of theorems of PA], and from the consistency of PA we infer that $[G] \notin$ *Th*. At the same time, since $\sim G$ is equivalent to $(\exists x)(\_\_\_x\_\_\_[G]\_\_\_)$,[5] from $\vdash \sim G$ we derive $\vdash (\exists x)(\_\_\_x\_\_\_[G]\_\_\_)$. Because the open formula $(\exists x)\_\_\_x\_\_\_y\_\_\_$ weakly represents *Th*, it follows that $[G] \in Th$, which completes the reductio (529).

To claim that a formula $P(y)$ weakly represents *Th*, the set of Gödel numbers of theorems of PA, is to say that:

> for any n, $n \in Th$ if and only if $\vdash_{PA} P(\bar{n})$

where $\bar{n}$ is the formal numeral for the natural number n. McCall's argument relies on the correct claim that the formula $(\exists x)(\_\_\_x\_\_\_y\_\_\_)$, that is, $(\exists x)\text{Proof}(x, y)$, weakly represents the set *Th*. What he fails to realize is that *the truth of this claim depends on the assumption of ω-consistency*. In particular, the claim that $\vdash_{PA}(\exists x)\text{Proof}(x, [G])$ entails that the Gödel number of G is in *Th*, which is precisely the claim needed in McCall's informal argument, depends on the ω-consistency of PA. Should PA be ω-inconsistent, it might be that $(\exists x)\text{Proof}(x, [G])$ is derivable in PA even though G itself is not a theorem of PA because each n fails to be the Gödel number of a proof of G.

It is, in fact, quite clear that from the consistency of a system alone one cannot infer that the negation of its Gödel sentence is not provable. For instance, consider the formal theory T = PA + $\sim$Con(PA), whose axioms are those of PA, together with the additional axiom $\sim$Con(PA). If we assume that PA is consistent, then by the second incompleteness theorem, T is as well, and clearly in T one can derive $\sim$Con(PA). Since an inconsistency in PA is also an inconsistency in T, $\sim$Con(T) is also derivable in T. But $\sim$Con(T) is provably equivalent to $\sim G_T$, where G_T is the Gödel sentence for T, so $\sim G_T$ is also derivable in T.

Once again, the significance of the need to assume ω-consistency may become clearer if we consider other formal systems whose ω-consistency is less evident than that of PA. McCall himself suggests (531) that we consider the formal system ZF, a formalization of Zermelo-Frankel set theory, and proposes the conditional 'If ZF is consistent then $\sim G_{ZF}$ is not provable in ZF' as an example of a sentence that a human can recognize as true, but whose formal counterpart is not provable in ZF. But the claim that a human can recognize the truth of this conditional is implausible, for recognizing its truth would require knowing that ZF is ω-consistent.

Because of this error, McCall's claim that PA cannot derive (2′) is quite irrelevant. Should the claim be correct, it would not establish any difference between humans and machines, for humans are just as incapable of establishing (2).

As it happens, the claim is correct, although McCall fails to see why. He offers a kind of plausibility argument to the conclusion that the derivation in PA of (2′) is "highly unlikely" (529). But, in fact, there is a straightforward proof that, if PA is ω-consistent, then (2′) is not derivable in PA. Assume that Con(PA) $\rightarrow$ $\sim$Theorem($[\sim G]$) is derivable in PA. Since Con(PA) is provably equivalent to G, it follows that Theorem($[\sim G]$) $\rightarrow$ $\sim G$ is also derivable in PA. But by Löb's Theorem,[6] it follows that $\sim G$ is likewise derivable in PA, and so PA is not ω-consistent. Hence, if PA is ω-consistent, then (2′) is not derivable in it.

We have shown that what humans are actually capable of establishing is:

(3) If PA is ω-consistent, then $\sim G$ is not derivable in PA.

There is still the possibility that McCall's argument can be rescued by showing that *this* conditional is one that humans can recognize to be true but PA cannot derive. Specifically, one might wonder whether:

(3′) ω-Con(PA) $\rightarrow$ $\sim$Theorem($[\sim G]$)

is derivable in PA. Another way of putting the question is this: If PA is given the same information that McCall tacitly makes use of in establishing (2), can it derive (2′)?[7]

[4] Here, McCall is using $[\sim G]$ to denote the Gödel number of the sentence $\sim G$, rather than the numeral for that Gödel number.

[5] Here, however, McCall uses $[G]$ to denote the formal numeral corresponding to the Gödel number of G. His $\_\_\_x\_\_\_[G]\_\_\_$ is the formula we are calling Proof(x, $[G]$), and thus his sentence $(\exists x)(\_\_\_x\_\_\_[G]\_\_\_)$ would be, in our notation, Theorem($[G]$).

[6] Löb's Theorem says that for any sentence S, if $\vdash_{PA}$ Theorem($[S]$) $\rightarrow S$, then $\vdash_{PA} S$.

[7] One might also wonder whether McCall's argument can be fixed by using Rosser's sentence, rather than Gödel's sentence G, in sentence (2) and its formal counterpart (2′). For sentence (2), so modified, can be established without relying on a tacit assumption that PA is ω-consistent. It can be shown, however, that the modified version of (2′) is derivable in PA. Thus, this strategy would not produce McCall's sought-after distinction between mind and machine.

The answer is "Yes." Here is one proof: we begin by observing that, if something fails to be the Gödel number of a proof, then, since this can be determined by a finite check, it is provable that it fails to be. In fact, this observation can itself be proven in PA, since for any sentence S:

(4.1) $\vdash_{PA} (\forall n)(\sim \text{Proof}(n, [S]) \rightarrow \text{Theorem}(\text{sub}(n, [\sim \text{Proof}(x, [S])])))$

where sub(n, [P]) denotes the Gödel number of the result of substituting the formal numeral for n for all free variables in the formula P.

According to the definition of 'Theorem(y)':

(4.2) $\vdash_{PA} \text{Theorem}([\text{Theorem}([S])]) \rightarrow \text{Theorem}([(\exists x)\text{Proof}(x, [S])])$

We also know that:

(4.3) $\vdash_{PA} \sim \text{Theorem}([S]) \rightarrow (\forall n) \sim \text{Proof}(n, [S])$

Applying (4.1), we get:

(4.4) $\vdash_{PA} \sim \text{Theorem}([S]) \rightarrow (\forall n)\text{Theorem}(\text{sub}(n, [\sim \text{Proof}(x, [S])]))$

Combining (4.2) and (4.4), we arrive at:

(4.5) $\vdash_{PA} (\text{Theorem}([\text{Theorem}([S])]) \ \& \sim \text{Theorem}([S])) \rightarrow$

$(\text{Theorem}([(\exists x)\text{Proof}(x, [S])]) \ \&$

$(\forall n)\text{Theorem}(\text{sub}(n, [\sim \text{Proof}(x, [S])])))$

But the right-hand side of (4.5) is exactly what ω-consistency rules out. Hence:

(4.6) $\vdash_{PA} (\text{Theorem}([\text{Theorem}([S])]) \ \& \sim \text{Theorem}([S])) \rightarrow \sim \omega\text{-Con}(PA)$

In other words:

(4.7) $\vdash_{PA} \omega\text{-Con}(PA) \rightarrow (\text{Theorem}([\text{Theorem}([S])]) \rightarrow \text{Theorem}([S]))$

Note that (4.7) formalizes half of the claim that the ω-consistency of PA implies that the formula Theorem(y) weakly represents *Th*. In particular, it is the formal counterpart of the claim that, if PA is ω-consistent, then, if $\vdash_{PA}$Theorem([S]), then the Gödel number of S is an element of *Th*.

We now apply (4.7) to the formula G to get:

(4.8) $\vdash_{PA} \omega\text{-Con}(PA) \rightarrow (\text{Theorem}([\text{Theorem}([G])]) \rightarrow \text{Theorem}([G]))$

Since G is provably equivalent to ~Theorem([G]), this implies that:

(4.9) $\vdash_{PA} \omega\text{-Con}(PA) \rightarrow (\text{Theorem}([\sim G]) \rightarrow \text{Theorem}([G]))$

Of course, by the definition of consistency we have:

(4.10) $\vdash_{PA} \text{Con}(PA) \rightarrow (\text{Theorem}([\sim G]) \rightarrow \sim \text{Theorem}([G]))$

Combining (4.9) and (4.10), and using the fact that $\vdash_{PA} \omega\text{-Con}(PA) \rightarrow$ Con(PA), we reach the desired conclusion:

(4.11) $\vdash_{PA} \omega\text{-Con}(PA) \rightarrow \sim \text{Theorem}([\sim G])$

In short, if we level the playing field between mind and machine, they remain in a dead heat. At least, McCall has not succeeded in showing otherwise.

ALEXANDER GEORGE
Amherst College

DANIEL J. VELLEMAN
Amherst College