# WHAT A COMPUTER CAN'T COMPUTE, WHAT A BELIEVER CAN'T BELIEVE

Cliff S. Hill

## I. Introduction

In 1961 J.R. Lucas argued that Gödel's Incompleteness Theorems[1] show that no mechanistic model of the mind is possible. The argument can be summarized as follows: if we consider the sentence, "This sentence is not provable,"[2] then we seem to be able to recognize that such a sentence is true. There is no mechanistic method for determining whether it is true or not because if you can prove the sentence to be true then you make it false. How are humans or things with minds able to determine that the Gödel sentence is true? The simple answer for Lucas is "intuition" and this is something that the machine/computer will never have. Lucas is not the only individual who thinks this. Paul Benacerraf and Roger Penrose[3] have also advocated this view. Even though there has been considerable criticism of Lucas's argument, there are some who consider the problem open.[4]

Interestingly enough in 1986 Raymond Smullyan published a book entitled *Forever Undecided: A Puzzle Guide to Gödel* which has not been recognized as a response to Lucas. Smullyan does not appear to have written the book as a criticism of Lucas but as a defense for Modal Logic.[5] What seems to be a very small point that Smullyan tries to make within the course of the book can easily be modified to refute Lucas's claim. I believe I can provide this modification.

What I plan to show in this paper is that even though the machine cannot determine the truth value of its own Gödel sentence there is a Gödel sentence that individual minds can not determine either. In other words the intuition that Lucas claims we have for identifying the Gödel sentence as true is the same *intuition* that the machine has. The machine cannot recognize that the sentence, "This sentence is not provable," is true (without making itself inconsistent) and we cannot recognize that the sentence, "I am consistent," is true (without making ourselves inconsistent). I will not be concerned with whether a mechanistic model of mind is the correct model of the mind only that Gödel's theorems cannot rule out such a model.[6]

Before I begin it is important to note how problematic it is to define notions like "mind" and "mechanism." For the purposes of this paper I will assume that there are minds and mechanisms in the world. When I refer to the term "mind" I am referring to some object in the world that healthy human beings have (or gain at some point). I will also assume that minds are the kinds of things that can have beliefs.

When I refer to the term "mechanism" I will be referring to an object that is not unlike a Turing machine. My definition of a Turing machine will come from the *Stanford Encyclopedia of Philosophy*, "Turing machines, first described by Alan Turing, are simple abstract computational devices intended to help investigate the extent and limitations of what can be computed."

In Section II, I will define Gödel's incompleteness theorems and give a very brief background of what Gödel's theorems are thought to imply. Section III will give a

detailed account of what a machine can not do when applying Gödel's theorems. In Section IV, I will give a more detailed account of Lucas's argument. Section V will introduce Smullyan's modest believers and their limitation. Finally, in Section VI, I will demonstrate how Smullyan's modest believers have the same "intuition" with respect to Gödel sentence as machines do.

## II. What Are Gödel's Theorems?

In 1925, Kurt Gödel gave a proof for his two incompleteness theorems. The theorems are:5

> **Gödel's First Incompleteness Theorem.** Any adequate axiomatizable theory is incomplete. In particular the sentence "This sentence is not provable" is true but not provable in the theory (Myers).
>
> **Gödel's Second Incompleteness Theorem.** In any consistent axiomatizable theory (axiomatizable means the axioms can be computably generated) which can encode sequences of numbers (and thus the syntactic notions of "formula", "sentence", "theorems") the consistency of the system is not provable in the system (Myers).[7]

There has been a lot of debate as to what Gödel's incompleteness theorems actually imply (not just within philosophy of mind). The most common interpretation is that it seems to have shown that Hilbert's program is impossible. David Hilbert believed that it was possible to assume a set of axioms that could determine all mathematical problems.[8] Gödel appeared to shown that this was an error on Hilbert's part.[9] Lucas was the first individual to directly claim that Gödel's theorems show that a mechanistic model of the mind is false.

## III. Lucas's Argument against the Mechanistic Model of Mind

I will not deny that Lucas's idea seems plausible since the idea first crossed my mind when I learned of Gödel's theorems. Lucas is quick to say that "Gödel's theorem seems to me to prove that Mechanism is false...[and]...almost every mathematical logician I have put the matter to has confessed similar thoughts," (1963, 255). If I look at the sentence, "This sentence is not provable," then it would seem that the sentence is true. If we take "provable" to mean something like "can show or demonstrate the truth of" then I do not have any way of showing anyone else that it is true. Why is that? Because if I do, I will automatically make the sentence false. If I can not show the sentence to be true then there is no method to teach anyone (or anything) that the sentence is true. The problem for the mechanism (since it would seem that humans are its teacher) is that there is no way for us to "teach" it so that it knows that the Gödel sentence is true.

Lucas claims that when trying to determine the truth or falsity of the sentence "...we [humans or minds], standing outside the system, can see [the Gödel sentence] to be true" (1963, 256). In comparing this to the Turing machine, Lucas claims that since we are outside the system and the Turing machine is in a system then minds will always be

different from machines. That was the basic argument that he presented in 1961 and in 1996 he published a paper defending the view from the many critics that it had produced.

Lucas continues this line of reasoning to claim that there is no way a mechanism can be taught to recognize that the Gödel sentence is true since it would have to go through some formal process to determine the truth value of the sentence. "Minds" appear to have the ability to recognize that the Gödel sentence is true by using some type of ability, vaguely referred to as "intuition." The sentence "This sentence is not provable" is a true sentence as long as nothing provides a proof for it, truths that appear to be intuitively self-evident truths often lack proofs. The Gödel sentence is not the only claim that appears to have this quality; the Law of Non-Contradiction[10] and the Law of Excluded Middle seem to be similar cases. What makes the Gödel sentence unique in regards to other self-evident truths is that if anyone or anything claims to have proven it to be true then we can show that person or that thing is making an inconsistent claim. Lucas sees this as a reason to deny the mechanistic model of mind since it would seem that the only way a machine can determine truth is by some proof.

The next question that will be considered as a response to Lucas is, if we have no formal method for determining that the Gödel sentence is true then how is it that we can claim the Gödel sentence is true? It seems like our justification for accepting the truth of the Gödel sentence is lacking. Lucas believes that "...as a result of Gödel's incompleteness theorems one must instead 'turn to "non-constructive" systems of logic with which not all the steps in a theorems are mechanical, some being intuitive'" (1996, 111). Again, Lucas is appealing to some sense of intuition here and the part it plays in our reasoning while at the same time denying intuition to machine. How would we go about programming machines with intuition? It would seem that by Gödel's theorems, if at least one truth cannot be proven then that truth cannot be discovered by the machine.

What if we just consider the fact that it seems like most (if not all) people are inconsistent with their beliefs. Does not this mean that if the machine simply proves the Gödel sentence false and in doing that showing its own inconsistency then it is like us because we can be inconsistent also? Lucas believes this argument also fails, "If the mechanist says that his machine will affirm the Gödelian sentence, the mind then will know that it is inconsistent and will affirm anything, quite unlike the mind which is characteristically selective in its intellectual output." (1996, 118) I do not think Lucas is very clear with his response, but I think I can see what he is getting at. Lucas seems to be saying that the Gödel sentence allows us to determine which machines are consistent and which ones are not. If a human agent claims that the Gödel sentence is true then it would not give us grounds for saying that human agent is inconsistent since the mind is capable of recognizing the Gödel sentence as true.

## IV. How to Apply Smullyan's Modest Believers[11]

Raymond Smullyan presents a very interesting argument that can easily be used against Lucas's thesis. Smullyan originally presents the idea on the island of knights and knaves where knights always tell the truth and knaves always lie. Smullyan is able to show that if

the believer is one that is "modest" and follows through with the logical consequences of her or his beliefs then believing in their own consistency automatically makes them inconsistent. The argument itself is very simple but in order to illustrate it correctly some conditions need to be explained in detail.

First we introduce a modal logic of beliefs (also called doxastic logic) where we have our necessary operator as:

> $B_a p$ translated as "agent $a$ believes that $p$."
>
> $C_a p$ translated as "it is possible that agent $a$ believes that $p$."

We need to look more closely at what it means to have a belief. Beliefs are not automatically true; they have to be shown or determined to be either true or false ($B_a p \supset p$ fails). The agent, if it wants true beliefs, has to go through and check her or his own beliefs. At the same time we have to consider the notion that in order for an agent to believe that $p$ it seems that in at least some sense the agent believes that $p$ is true. Smullyan deals with this by considering two types of agents; we will call a believer modest "...if for every proposition $p$, he believes $Bp \supset p$ only if he believes $p$" (153). One important fact about the modest believer is that "...if a reasoner believes $p$, then there is nothing immodest about his believing $Bp \supset p$" (153). We will call a believer conceited "...if for every proposition $p$, he believes $Bp \supset p$" (153).

Girle formalizes this more than Smullyan does and Girle calls it the *Modesty of Belief Principle*. This principle can be written logically as:

> $B_a(B_a p \supset p) \supset B_a p$ translated as "If agent $a$ believes that (if agent $a$ believes that $p$ then $p$) then agent $a$ believes that $p$."[12]

What we also have to introduce is that for any proposition of the form $\sim p$ what logically follows from that is $(p \supset \bot)$, where "$\bot$" just means "a contradictory proposition" (a simple truth table can show that $(\sim p) \leftrightarrow (p \supset \bot)$). If we consider $\bot$ in a doxastic model then $\bot$ within an agent's belief system means that they have a contradictory belief. We can think of the $\bot$ as a type of placeholder for any contradiction. This means that $(\sim p) \leftrightarrow (p \supset \bot)$ could easily be written as $(\sim p) \leftrightarrow (p \supset (\sim a \wedge a))$, $(\sim p) \leftrightarrow (p \supset \sim(\sim a \vee a))$, $(\sim p) \leftrightarrow (p \supset \sim((a \supset b) \wedge a) \supset b))$, and so on.

With those initial clarifications we can now find out what all this means. If we consider a doxastic agent who is modest we come across an interesting result. Suppose that this agent attempts to believe that she is consistent, that she has a consistent set of beliefs (formally $\sim B_a \bot$). Now let us see what happens when she follows through with the logical consequences of her beliefs.

1. $B_a(\sim B_a \bot)$ ($a$ believes that $a$ does not believe a contradiction)
2. $B_a(B_a \bot \supset \bot)$ (substitution of "$B_a \bot \supset \bot$" for "$\sim B_a \bot$" since they are logically equivalent)

3. $B_a(B_a p \supset p) \supset B_a p$ (*Modesty of Belief Principle*)

4. $B_a \perp$ (2 and 3 by modus ponens and substitution)

What the above proof shows is if a doxastic agent follows through the logical consequences of her beliefs and believes that she is consistent then she is automatically inconsistent. What does that mean for Lucas's thesis? It seems like there is something that the machine and the believer have in common that Lucas argued that they do not. I will clarify what I mean by this in the next section.

## V. How to Apply *Modest Believers* to Refute Lucas's Argument

Ultimately my application of Smullyan's *modest believers* to counter Lucas's argument is simple. A modest believer cannot believe in her own consistency and follow through with the logical consequences of her beliefs without becoming inconsistent. The machine cannot assert that it can prove all sentences to be true, it cannot show that all true sentences are true without becoming inconsistent. Both the machine and the believer have the same intuition. What do I mean by that? I mean that Lucas's thesis is false since he was trying to claim that we have an "intuition" that the mechanism can never have because of the Gödel sentence. *Modest Believers* seem to show that the machine and things with minds are the same in regards to the Gödel sentence. The machine cannot claim to prove that the sentence "This sentence is not provable" is true without becoming inconsistent and that a believer agent *a* cannot claim that she believes $\sim B_a \perp$ without becoming inconsistent.

What if we consider the fact that agents with minds appear to have the ability to withhold opinion on certain propositions? This would mean that the proof at the end of section V would not even get past the first step. The agent could simply say that she is not convinced by logic and so holds no opinion on steps 2, 3, and 4. If this is the case it seems like the machine is perfectly capable of doing the same thing. We could easily program the machine to turn-off whenever it reaches the Gödel sentence or could simply claim that Gödel sentence is true without trying to prove that it is true.

Lucas's response to my application of Smullyan's *modest believers* would probably be somewhere in the area of his conclusion for his 1996 paper.

> ...the only tenable form of mechanism is that we are inconsistent machines, with all minds being ultimately inconsistent, then mechanism itself is committed to the irrationality of [its own] argument, and no rational case for it can be sustained. (1996, 122)

This is not true from my claim because as long as I do not have the belief that I am consistent and follow through with the logical consequences of my beliefs then the possibility of me being consistent still remains. The same holds for the machine, as long as it does not assert the Gödel sentence to be true and give a proof for the Gödel sentence then it still has the possibility of being consistent.

Even though we are outside the system of the machine we are not outside our own system of beliefs. We can determine the truth of the Gödel sentence that the machine has but the machine could theoretically determine the truth value of the Gödel sentence that the agent has. This can be illustrated rather easily if we consider the machine being given all the beliefs of some human agent. The machine could go through all of these beliefs and if the beliefs of the human agent are consistent then the machine can determine that she is consistent. If the human agent claims that she is consistent and follows through with the logical consequences of her actions then (as we saw in section IV) she is inconsistent.

## VI. Conclusion

I think it is fairly obvious that Lucas's argument fails. Minds have the "intuition" to recognize the truth of the Gödel sentence produced by some other mind or machine but they do not have the "intuition" (nothing can have such an intuition) to recognize their own Gödel sentence. Machines can have the same "intuition," they can recognize the truth of the Gödel sentence produced by some other mind or machine but they do not have the "intuition" to recognize their own Gödel sentence as true.

The more difficult question is what does this tell us about minds? At the very least it appears to tell us that we cannot rule out a computational model of the mind based on Gödel's theorems. I do not think that the computational model of the *human* mind is correct but I have failed to be convinced that it is impossible to produce a machine with a mind. Also I think that if minds are the kinds of things that are capable (and perhaps the only kind of thing capable) of reasoning and if reasoning is being analyzed by the computationalists then there is little doubt in my mind that computationalism can tell us something about the mind.

## NOTES

1. I will also refer to Gödel's Incompleteness Theorems as simply "Gödel's Theorems" as well.
2. I will also refer to this sentence as "the Gödel Sentence."
3. See Benacerraf's "God, the Devil, and Gödel," and Penrose's *The Emperor's New Mind* and *Shadows of the Mind.*
4 See Hintikka 2000, 74 and Horst Section 3.3.
5. See Smullyan 1987, 256-257.
6. Although there is little doubt in my mind that the mind is a mechanism of some sort, I think that the computational model of the human mind is false.
7. For the purposes of this paper I will just focus on the first incompleteness theorem since the second theorem is a consequence of the first.
8. See Webb 112.
9. But even that is still highly debated.
10. Or the Law of Contradiction or the Principle of Non-contradiction.
11. This comes from Smullyan and Girle.
12. The more formalized version of the Modesty of Belief Principle comes from Girle.

## WORKS CITED

Barker-Plummer, David. "Turing Machines." *The Stanford Encyclopedia of Philosophy* Ed. Edward N. Zalta. (Spring 2005). http://plato.stanford.edu/archives/spr2005/entries/turing-machine/.

Benacerraf, Paul. "God, the Devil, and Gödel." *Ettica & Politica / Ethics and Politics* (2003):1. http://www2.units.it/etica/2003_1/3_monographica.htm.
Girle, Rod. *Possible Worlds*. Montreal: McGill-Queen's UP, 2003.
Hintikka, Jaakko. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca, NY: Cornell UP, 1962.
---. *On Gödel*. Belmont, CA: Wadsworth/Thomson Learning Inc., 2000.
Horst, Steven, "The Computational Theory of Mind", *The Stanford Encyclopedia of Philosophy* Ed. Edward N. Zalta. Fall 2005. http://plato.stanford.edu/archives/fall2005/entries/computational-mind/.
Lucas, J.R. "Minds, Machines, and Gödel." *The Modeling of Mind: Computers and Intelligence*. Eds Kenneth M. Sayre and Fredrick J. Crosson. New York: Simon and Schuster, 1963
---. "Minds, Machines, and Gödel: A Retrospect." *Machines and Thought: The Legacy of Alan Turing*. Eds. Peter Millican and Andy Clark. Oxford: Oxford UP, 1996.
Myers, Dale. "Gödel's Incompleteness Theorem." U of Hawaii, Mathematics Department June 2005. http://www.math.hawaii.edu/~dale/godel/godel.html.
Penrose, Roger. *Shadows of the Mind*. Oxford: Oxford UP, 1994.
---. *The Emperor's New Mind*. Oxford: Oxford UP, 1989.
Smullyan, Raymond. *Forever Undecided*. New York, NY: Alfred A. Knopf P, 1987.
---. *Gödel's Incompleteness Theorems*. Oxford Logic Guides 19, Oxford: Oxford UP, 1992.