Is Lucas's Gödelian argument valid?

Author
Name: Xandra van de Putte,
Student number: 0008516

Supervisor
dr T.M.V. Janssen

Curriculum details
Bachelor Project
Faculty of Mathematics and Informatics
University of Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam

Date
2005-06-28

Summary

This thesis describes a part of the discussion around Lucas's argument that minds cannot be explained by machines. In his argument Lucas uses Gödel's incompleteness theorem to prove his claim that minds are different then machines. The argument, as well as Gödel's incompleteness theorem, will be explained in detail in the introduction of this paper.

The main questions that will be discussed are:

"Have Lucas really proved that minds are different from machines?"
"Is it possible to prove the consistency of the mind?"

These questions will be discussed using the criticisms Lucas got from Paul Benacerraf, David Lewis and David Coder, who are all philosophers. They all think that there is something wrong with the argument of Lucas, in short:

Benacerraf made several attacks on Lucas's argument, and he concluded with the claim that at most Gödel's theorems prove the following: If a mind is a Turing-machine, then a mind cannot ascertain which one.

Lewis states that in order to complete Lucas's argument that he is no machine, Lucas must convince us that he has the general ability to verify theoremhood in 'Lucas arithmetic'.

Coder claims that at most, Gödel's theorem proves that not all minds can be explained as machines.

Lucas also replied to the articles written by these philosophers, this will also be discussed.
In the conclusion I will briefly answer (according to my opinion) the questions stated above.

Contents

1. Preface

The project for which this thesis is written, namely “Are human machines? – The argument of Lucas”, has been done by Peter Vollaard and me. Because it must be possible to evaluate the project individually, we have decided to study different sub-topics. To establish this, we first examined the literature that has been written for this subject. We found a lot of criticism to the article in which Lucas described his argument [Luc61]. These critiques were from different people with different kinds of specialties. This made it very easy to find a way to split up the subject; one part that discusses the reactions from computer scientists / mathematicians, and one part that discusses the criticisms from philosophers. The latter has been done by me.

This kind of project was completely new for me. I think it was very informative and it increased my enthusiasm for philosophy.

2. Introduction

In 1961, J.R. Lucas published his paper “Minds, Machines and Gödel” in Philosophy. He claimed in this paper that mechanism is false and that Gödel’s incompleteness theorems are the proof of this. It is interesting to mention what made him write this paper. When he was at school, he heard an essay of a contemporary of him, in which a position of extreme materialism was put forward. This person claimed that our behavior was entirely determined by physical laws. Lucas argues against him that the fact that he put forward his position, and commended it to us to adopt, belied his claim. This argument did not leave him alone, and he kept trying to reformulate it in a satisfactory fashion. He managed this in 1959.

Before going back to Lucas’ arguments, I will first give a brief explanation of Gödel’s incompleteness theorems, using [Nag56].
When Gödel was only 25, he published a paper in a German scientific periodical in 1931. This paper put an end to the hope that the whole realm of mathematical reasoning could be brought into order by way of the axiomatic method. He proved that this method has certain limitations that ensure that the arithmetic of whole numbers can never be fully systematized by its means. He also proved that it is impossible to establish the logical consistency of any complex deductive system except by assuming principles of reasoning whose own internal consistency is as open to question as that of the system itself.
His paper is a final answer to the question if it is possible to prove that arithmetic is free from contradictions, that is, prove that it is consistent. His first main conclusion is that it is impossible to establish a meta-mathematical proof of the consistency of a system that contain the whole of arithmetic, unless this proof itself employs rules of inference much more powerful than the transformation rules used in deriving theorems within the system. Meta-mathematical statements are statements about the signs and expressions of a formalized mathematical system; meaningful statements about a meaningless system. An example: 2 + 3 = 5 is a mathematical statement, “2 + 3 = 5 is an arithmetical formula” is a meta-mathematical one.
Gödel’s second main conclusion is that given any consistent set of arithmetical axioms, there are true arithmetical statements which cannot be proven as being true within this set. Even if any finite number of other axioms is added, there will always be further arithmetical truths which cannot formally proved. The way Gödel proved his conclusions is very difficult and the proof is at least 30 pages long. I will use an easier one, the one in [Nag56].

Gödel first established a method for completely arithmetizing a formal system. He gave a formal system in which each elementary sign, each formula and each proof is assigned with a Gödel number as a label. How they are assigned is shown in table 1. I will explain how a number is given to a formula by using the following formula:
$(\exists x)\ (x = S_y)$
This formula says that every number has an immediate successor. The numbers of these signs as in the formula’s sequence are: 8, 4, 13, 9, 8, 13, 5, 7, 16 and 9. These numbers are used as the exponential of the first 10 prime numbers (10 because there are 10 numbers in this sequence). Then the new numbers are multiplied, so you get 2^8 * 3^4 * 5^13 * etc. The product is the Gödel number for the formula. In this way every formula can be represented by a unique number. The number of a sequence of formulas that may occur in some proof can be calculated in the same way; the number of the *n* formulas will be the exponents of the first *n* prime numbers. The Gödel number of this sequence is then the product of these obtained numbers.

Connectives and elementary signs		
Signs	**Gödel number**	**Meaning**
~	1	not
∨	2	or
⊃	3	If…then
∃	4	There is an …
=	5	equals
0	6	zero
S	7	The next following number
(	8	punctuation mark
)	9	punctuation mark
,	10	punctuation mark
Sentential variables (each designated by a number greater than 10 and divisible by 3)		
Variables	**Gödel number**	**Sample**
p	12	Henry V111 was a boor.
q	15	Headache powders are better.
r	18	Ducks waddle.
etc.		
Individual variables (each designated by a number greater than 10 which leaves a remainder of 1 when divided by 3)		
Variables	**Gödel number**	**Meaning**
x	13	a numerical variable
y	16	a numerical variable
z	19	a numerical variable
etc.		
Predicate variables (each designated by a number greater than 10 which leaves a remainder of 2 when divided by 3)		
Variables	**Gödel number**	**Sample**
P	14	Being a boor
Q	17	Being a headache powder
R	20	Being a duck
etc.		

Table 1. This table is the same as the one in [Nag56].

So now we have a method for giving a unique number to an expression, but we can also retranslate a Gödel number into the expression it represents. So we can take this number as if it were a machine, and then look at the construction and input, and dissect an expression or proof. In the next step, Gödel found a way to uniquely represent each meta-mathematical statement in the formal system by a formula expressing a relation between numbers. Here is an example of how a meta-mathematical statement can be made to correspond to a formula in the formal arithmetical system:
If we have the formula (p ∨ p) ⊃ p, we can make the meta-mathematical statement that (p ∨ p) is the initial part of the formula and represent this statement by an arithmetical formula which says that the Gödel number of the initial part is a factor of the Gödel number of the complete formula. Gödel then showed how to construct the arithmetical formula that correspond to the meta-mathematical statement "The formula with Gödel number *h* is not demonstrable" and that have the Gödel number *h*. Gödel showed that this formula is demonstrable only if its negation is also demonstrable. But if this is the case, then arithmetic is inconsistent. So if arithmetic is consistent, neither this formula nor the negation can be proved. This formula is an undecidable formula of arithmetic. A meta-mathematical statement of arithmetic's consistence corresponds to a certain arithmetical formula, A, let's call the formula with the Gödel number *h* G. The formula A ⊃ G is demonstrable, so if A is demonstrable, G would also be. Because G is not demonstrable, as we have seen, A is undecidable. What is proved now, is that the consistency of arithmetic is undecidable by any meta-mathematical reasoning which can be represented within the formalism of arithmetic. But with a meta-mathematical statement outside this system we can tell that G itself must be true:

- The formula G corresponds to a meta-mathematical statement that is true, but only if arithmetic is consistent.

- Gödel mapped meta-mathematical statements upon arithmetical formulas in a way that every true meta-mathematical statement corresponds to a true arithmetical formula.
- So G must be true.

Now the conclusion can be made that arithmetic is essentially incomplete: There is at least one arithmetical truth that the system cannot prove to be true, but can be established by a meta-mathematical argument outside the system. Even if this true formula G is added to the system, we could again construct a new true formula that the system is not capable of to prove that it is true. And if this new formula is added, we can find one again and again.

To come back to Lucas, how can Gödel's theorems prove that minds cannot be explained as machines? To answer this, Lucas first gave a description of a machine and a formal system. Lucas' description of a machine:

> "Its behavior is completely determined by the way it is made and the incoming "stimuli": there is no possibility of its action on its own: given a certain form of construction and a certain input of information , then it must act in a certain specific way." [Luc61]

Because it is often proposed that mechanical models of the mind should contain a randomizing device, Lucas will consider what a machine is capable of if it has such a device. This device will only act when there are more than one operation possible, which will not lead to inconsistency. He also says that everything that is infinite or indefinite does not count as a machine. Thus a system exists of a finite number of types of operations and initial assumptions and we can represent them by corresponding symbols written on paper. We can represent the operations as rules that make it possible to go from one or more (or none) formula to another, and the initial assumptions as a set of initial formulas. So, every sequence of operations that the machine is able to make can be represented on paper and we have a formal system.

Now we can construct a true formula in this system, which cannot be proved within the system, a "Gödel-formula". Lucas says that every rational being can see this formula as a true one. He concludes from this that a machine is not a complete and adequate model of the mind, it cannot do everything a mind can do.

Lucas does not say that we cannot build a machine that simulates a piece of the mind's behavior. It is just that we cannot build a machine that can simulate every piece of the mind's behavior. There is always a weakness in the system that the mind does not have. Because mechanism says that it is possible to produce a model that can do everything a mind can, mechanism must be false.

Gödel's second theorem seems to say that a man can never calls himself consistent; if a man is consistent, then it is impossible to prove this within his system and if he is inconsistent, he may also not call himself consistent. Lucas thinks that even though we are sometimes inconsistent, it is rather a mistake than set policies. It corresponds to the technical failures that sometimes occur in machines, not to its normal operations. So he claims that minds are consistent (if someone is not we call him 'out of his mind').

Lucas himself gives three examples of objections:

1) We cannot have both that a machine can simulate any mind-like behavior, but not all of them.
2) If we have a Gödel-sentence, we can make another system that is more adequate.
3) The procedure that constructs the Gödel-formula is a standard procedure (only this way we can be sure that a Gödel-formula can be made for every formal system). So a machine can be programmed with an operation that also can go through the Gödel procedure.

For the first two he gives one explanation that these objections cannot be made: Every time we make a new system that can prove the Gödel-sentence of the old one, there is a new Gödel-sentence for this new system.

For the third one he argues that we can imagine a mind, faced with a machine that contains an operator that can add its Gödelian formula (repeatedly). This mind can take this operator into account and free the new machine of the Gödel-sentence, with the Gödelizing-operator and all. It is proved that this is the case. There will always be a Gödel-sentence in a machine, even with such a Gödelizing-operator, the resulting system will always be incomplete.

3. Lucas's argument criticized

Lucas did get quite a lot of critiques to his article "Minds, Machines, and Gödel" [Luc61] from people with different specialties. A lot of them are philosophers, but there are also criticisms from people who are specialized in mathematics, computer science and logic. In this chapter I discuss the critiques from the philosophers Benacerraf, Lewis and Coder and the replies Lucas have given. Each sub-chapter discusses the criticism of one person and the reply Lucas gave.

3.1. Paul Benacerraf – God, the Devil, and Gödel

Paul Benacerraf discusses Lucas's argument in his article "God, the Devil, and Gödel" published in 1967 [Ben67]. In this article he argues that if it were true that Lucas can find a weakness in every mechanic model for the mind, then it would still not prove that mechanism is false. Benacerraf thinks that Lucas has not sufficient convincing arguments for his claim.
To show this, Benacerraf presents an argument which contains the assumption that the mind is at its best a Turing machine, using both Gödel-theorems and ending on a contradiction. He obtains a different implication of the influence of Gödel's theorems on mechanic philosophy, than Lucas did obtain.

The title of Benacerraf's article refers to a saying that "God exists, since mathematics is consistent, and the Devil exists, since we cannot prove it." [Ros50]. Benacerraf says that Gödel is the missing link, because he proved that if mathematics is consistent, we cannot prove it. If mathematics is not consistent we can hardly prove it, so Satan exists either way. He mentioned this to show how far-reaching the philosophical consequences of Gödel's incompleteness theorems might be. Benacerraf shows that in a typical case, the conjunction of two (or more) philosophical views, say (p · q), is false according to Gödel's theorems, such that there always remain adherents from p and adherents from q. But usually it is claimed that Gödel disproved either that p or that q. Benacerraf showed that you need also a philosophical argument to establish the desired conclusion. In his paper he examines such an alleged implication and shows that what is claimed to have been disproved by Gödel's incompleteness theorems, has not been disproved.

Mechanism
Benacerraf mentioned that the name 'mechanism', is a thesis having to do with machines, without further explanation or description of machines. There is a distinction made in anthropic mechanism and universal mechanism. Universal mechanism holds that everything in nature can be explained in mechanical terms, this is not the one Lucas is arguing about. He argues about anthropic mechanism:

> "The thesis in anthropic mechanism is not that everything can be completely explained in mechanical terms (although some anthropic mechanists may also believe that), but rather that everything about human beings can be completely explained in mechanical terms, as surely as can everything about clockwork or gasoline engines.
> One of the chief obstacles that all mechanistic theories have faced is providing a mechanistic explanation of the human mind;" … "Today, as in the past, the main points of debate between anthropic mechanists and anti-mechanists is mainly occupied with two topics: the mind — and consciousness, in particular — and free will. Anti-mechanists argue that anthropic mechanism is incompatible with our commonsense intuitions." [AbsAs].

According to Benacerraf, if certain things which do not satisfy Turing's specification also count as machines, then the proof that a mind cannot be explained by a Turing machine will not be enough to establish Lucas's thesis (that it is impossible to explain the mind as a machine). This may look funny, but Benacerraf argues that minds may be less than Turing machines.

Two notes about Lucas's argument

Benacerraf noticed two things about Lucas's argument that Gödel's theorems implicate that no (Turing) machine can match the deductive output of a mind:

First that it is not obviously valid: Gödel's first theorem showed that a machine cannot give a formal proof of its Gödel sentence. Benacerraf calls such a machine Maud. A mind can give a formal proof of the Gödel sentence of Maud, not within Maud's system, but within another formal system. Maud is limited to the formulas for which she can give formal Maud-proofs (those who can be formalized in Maud's system). According to Benacerraf, it is not clear that this limitation of Maud also limits her ability to conjure up proofs that cannot be formalized in Maud. Benacerraf thinks this is not clear, because he states that Maud can carry out the Gödel argument herself: by Gödel II she can prove 'Con(Maud) ⊃ H'. Benacerraf then asks if is it also possible for Maud to convince herself that H is true. Someone might reply that machines cannot convince itself that formulas are true. If machines indeed cannot convince itself, then according to Benacerraf, Gödel's theorems are hardly necessary to prove that machines are different from minds. Benacerraf thinks that Gödel's theorems don't help, he says:

> "As far as Gödel's theorems are concerned, provided that Maud doesn't delude herself into thinking that just because she has convinced herself that H is true has proved it, she can go on convincing herself of that, and of many other things besides."

Lucas claimed that he can prove H_{Maud} in some consistent formal system that contains the axioms for elementary arithmetic. So to *prove* a formula is to *derive* it as a formal theorem of a consistent system which contains the axioms of arithmetic. "*Derive…"* here cannot mean "*show* that … is a theorem of …". For the relation 'T is a theorem of the formal system S' is one which has its analogue in Maud, under a suitable numbering of formal systems and of their vocabularies.

Benacerraf continues by limiting the attention to formal systems whose vocabulary is the same as Maud's. Maud can enumerate the set of all such formal systems by giving a number to each (W_i). Whenever a statement of the form '$F_1,\ldots,F_n$ (a sequence of formulas) is a proof of F_n in W_i' is true, its translation into Maud is provable by Maud. Benacerraf states that if an axiom H is added to Maud (call it Maud_1), then 'H is a theorem of Maud_1' is provable in Maud. So, according to Benacerraf, what Lucas can 'prove' is not different from Maud. If it is necessary that the axioms of each such system be themselves 'theorems' for Lucas, then Maud cannot, in *that sense*, 'prove' H. Benacerraf thinks that *that sense* must be something like absolute provability in which everything Maud can prove is provable, but in which some things beyond Maud's reach are also provable. Benacerraf blames Lucas that he has never made this clear.

Benacerraf points out that there cannot be a formal system in which each proof that Lucas produces will be a proof; the union of all of Lucas's formal systems is not a formal system. If Lucas can make a union of all of his formal systems, and it will remain a formal system, then Lucas can claim that he is not a machine. If Lucas is not a machine, then Gödel's theorems do not exclude his ability to do it (make a union of all of his formal systems).

I think he has a good point here, because Lucas claimed that one formal system cannot explain the mind, but why not all formal systems (or a part) clustered (or enumerated by some formal system)? It then will not be necessary that the union is consistent, but only the clusters. In our mind theorem are also not really theorems, we just agree that they are theorems. And if we are faced with different kinds of formal systems, we may “have” different kinds of theorems which could, when unified, cause conflicts.

The second point Benacerraf noticed is that in order to conclude that the Gödel-sentence is true, one has to know that the machine in question is consistent. Lucas is aware of this objection; he concludes that we can know that arithmetic and certain formal systems including arithmetic are consistent, because the mind can enunciate truths of arithmetic.

Benacerraf wants to examine one more argument of Lucas. Lucas imagines some sort of match between him and the Satanic Mechanist:

- Mechanist tries to construct a mechanic model of the mind.
- Lucas tries to find a statement that this machine cannot produce, but which is nevertheless true.
- Mechanist changes his model so that it (at least) can produce what Lucas found.
- Etc.

If Lucas cannot find anything that is true that the machine cannot produce, Mechanism wins. If the Mechanist could produce a machine in which Lucas could not find something the machine cannot produce, then, according to Lucas, mechanism is false. Lucas claimed that he always will be able to find a true statement, which a machine cannot produce. So it follows that Lucas will win and mechanism is false. Benacerraf thinks that even if Lucas is capable of finding such a truth, Lucas did not prove that mechanism is false. The Mechanist is just a man; it is of high possibility that there are some limits in the kind of machine he can produce. The fact that the Mechanist cannot produce anything in which Lucas cannot find a weakness does not prove anything at all.

Lucas’s argument formalized

Benacerraf says that Lucas’s arguments are not convincing for Lucas’s point, but we must not exclude it completely without doing serious effort to see what follows from Gödel’s theorems, with respect to the existence of a mechanic model of the mind. He does this in this paper by making Lucas’s arguments more formal. He presents an argument which includes the assumption that the mind is at best a Turing machine, using both Gödel’s theorems, and ending in a contradiction. Benacerraf comes to a different conclusion than Lucas, about what implications Gödel’s theorems have for mechanistic philosophy; this will be discussed after the formalized argument.

The argument is as follows:

1) Let S = { x | B can prove x }
S represents the deductive output of B (B represents Benacerraf). The way B can ‘prove’ is something like how Lucas claims he can ‘prove’.
2) Let S* = { x | S |- x }
S* is the closure of S under the rules of first order logic with identity: Anything derivable from S by first order logic with identity, is in S*. Benacerraf does not assume that S is closed. The assumption that S is closed would seem false, and he is not able to proof that it is true.
3) S* is consistent
Every member of S is true (B cannot prove what is false) and first order logic maintains truth, thus everything in S* is true (so S* is consistent).

4) ‘$Con(S^*)$’ ϵ S
The predicate $Con(A)$ holds that A is consistent. The consistency of S^* has been proved by 1-3. Benacerraf has established this, so it is part of Benacerraf's output.

5) ‘$Con(S^*)$’ ϵ S^*
This can be established, since by 4: $S \subseteq S^*$. This corresponds roughly to Lucas's statement that he knows that he is consistent.\

6) $(x)(W_x \subseteq S^* \supset Con(W_x))$
Let W be a recursive enumerable set. Because S^* is consistent, all the enumerable subsets it contains are also consistent.

7) ‘$(x)(W_x \subseteq S^* \supset Con(W_x))$’ ϵ S
This can be established because 1-6 is a proof what Benacerraf has produced.

8) ‘$(x)(W_x \subseteq S^* \supset Con(W_x))$’ ϵ S^*
By $S \subseteq S^*$ and 7.

9) Let's say that there is a recursive enumerable set W_j, so that:
   a) ‘$Q \subseteq W_j$’ ϵ S^*
   Q is a first order closure of axioms, so that B can prove that W_j is adequate for arithmetic.
   b) ‘$W_j \subseteq S^*$’ ϵ S^*
   B can prove that W_j is a subset of the output of B.
   c) $S^* \subseteq W_j$
   B is a subset of W_j

That W_j is enumerable is the condition that it is the output of a theorem proving Turing machine. 9 will be the only assumption of the proof and 9c will correspond to Lucas's assumption that B is a Turing machine.

9c Is not equivalent to the assumption that B is a Turing machine, because it does not assert that $S^* = W_j$, but also because if B was a Turing machine, W_j had to be identical to S (not S^*). So, 9c might be true and B will not be a Turing machine, but not in the way Lucas would have him fail to be one. Lucas argues that a mind is not a Turing machine on by saying that a mind can do more than a Turing machine. But if a mind satisfies 9a and 9b, it may still fail to be a Turing machine, because it is possible that this mind is able to prove less than any given machine adequate for arithmetic (it has not been proven that $W_j \subseteq S$).
Benacerraf continues his argument:

10) $Q \subseteq W_j$
This follows from 9a (because everything that can be proved must be true).

11) There is a formula H having the Gödel properties such that if $H \epsilon W_j$, then $–H \epsilon W_j$, and W_j is inconsistent.
W_j is adequate for arithmetic (10) and it is (equivalent to) a formal system (9). So this is the first theorem of Gödel applied to W_j.

12) ‘$Con(W_j) \supset H$’ ϵ W_j
This is Gödel's second theorem applied to W_j.

13) ‘$W_j \subseteq S^* \supset Con(W_j)$’ ϵ S^*
This follows from 8, because S^* is closed under first order logic.

14) ‘Con(Wj)’ ϵ S^*
    This follows from 9b, 13 and the fact that S^* is closed under modus ponens[1]. It was also possible to use 14 directly as an assumption instead of 9b, so to say that B needs to prove that he can prove everything the machine Wj can prove (even to obtain the consistency of Wj). The reason that Benacerraf did not used this assumption is that he wants to find out some reasons why Lucas might think that he can prove that Wj is consistent.
15) ‘$(x)\ (\ Q \subseteq Wx \supset\ (Con(Wx) \equiv Con(Wx))\ )$’ ϵ S^*
    The quoted part states that any recursively enumerable set containing Q is consistent, only if its Gödel consistency formula holds. To establish this part it is enough to show that any formal system containing Q has a formula in the system, which expresses (in the system) the consistency of the system.
16) ‘$Q \subseteq Wj \supset\ (Con(Wj) \equiv Con(Wj))\ )$’ ϵ S^*
    From 15 and the fact that S^* is closed.
17) ‘Con(Wj)’ ϵ S^*
    To show ‘Con(Wj)’ ϵ S^* it is necessary to assume that Q is a subset of Wj and that B can prove that, this is the reason for 9a. 17 follows from 14, and as mentioned 9a.
18) ‘Con(Wj)’ ϵ Wj
    Since 9c states that S^* is part of the output of some Turing machine, so this Turing machine can ‘prove’ its own consistency.
19) H, -H, Wj, and Wj is inconsistent.
    18 says that W_j can ‘prove’ its own consistency, and therefore it is inconsistent.
    It follows from 9b that $Wj \subset S^*$, so:
20) H, -H, belong to S^* and S^* is inconsistent, contradicting 3.
    S is also inconsistent. It depends on its closure properties if it is just semantically inconsistent. There are no assumptions made for the closure properties of S, so nothing can be said about it.

The contradiction in the argument stated above, is derived from the assumptions in 9 and the definitions in 1 and 2. According to Benacerraf, if we assume that there is nothing wrong with the definitions, then we must reject 9. Lucas argues that the negation of 9c follows from Gödel's theorems. Benacerraf suggests that the theorems do not imply that, at least they imply the negation of the conjunction of 9a, 9b, and 9c: either B cannot prove that Wj is adequate for arithmetic, or if B is a subset of Wj, then B cannot prove that he can prove everything Wj can. Benacerraf states that it seems to be consistent with all this that B is a Turing machine, but the machine table is too complex for B to know what it is. B can also not ascertain of any instantiation of the machine, which happens to be B himself, is an instantiation of that machine: If B is faced with a machine with B's program in such a way that B can decipher its program, then B can ‘prove’ its consistency and B and the machine are both inconsistent, contradicting 3.

Benacerraf says in his argument that Lucas fails to notice how a system is given to him. It depends on how the system is given to show whether a Gödel sentence is true. You cannot claim that you can determine a machine's program by watching its output.
I agree with him, but in this way one can “never” make a machine that is equivalent to a person's mind. Only if you exactly can determine what is in a mind and how it works, then we can try to make a machine the same as a specific mind, but only if we have the proper tools and it may not be a mechanic one.

[1] For any formula A and B in S^*: if A is true and $A \rightarrow B$ then B is also true.

Benacerraf's conclusion
If there is a Turing machine that can prove everything that is in the first order closure of B, then B cannot show both of this machine that it is adequate for arithmetic and that B can prove everything the machine can. This result is weaker than the one Lucas claimed, but seems still significant: Someone to whom Benacerraf explained his argument, concluded that psychology as we know it is impossible. For if we are not at best Turing machines, then this psychology is impossible and if we are, then there are certain things we cannot know about ourselves.
In his conclusion Benacerraf would not take sides. He only says that if we ignore that 9a might be false, then, if humans are Turing machines, we are barred with Socrates's philosophic injunction: KNOW THYSELF.

3.2 John Lucas – Satan Stultified: A rejoinder to Paul Benacerraf

Lucas respond to Benacerraf in his article "Satan stultified: A rejoinder to Paul Benacerraf." in 1968. The title of this article can be explained by the comparison between the mechanist and Satan (in the article of Benacerraf). Lucas thinks that Benacerraf's mechanist is self-stultifying. Stultify means to appear stupid, inconsistent or ridiculous. I assume that Lucas meant by it that Satan has been represented as inconsistent and, maybe, made appear stupid/ridiculous.

The usage of Gödel's theorems in Lucas's argument
Lucas agrees with Benacerraf that the argument, in which the Gödel's theorems are applied to the problem of minds and machines, is not and cannot be a purely mathematical one. It needs some philosophical assumptions to make any philosophical conclusions. Lucas uses Gödel's theorem in two ways:

- As a formal proof sequence that gives some syntactical results about a certain class of formal systems. Lucas thinks that the mechanist must admit that he scored some points against the mechanist's machine.
- As a certain type of argument, that we can understand it and apply to different situations. Lucas hopes that this will make the mechanist to see that he can do better than a machine, as a man. He also hopes that this sort of argument will always apply against any form of mechanism that the mechanist will follow.

Lucas says that his argument did not directly prove that the mind is more than a machine. His argument is more something like a schema of disproof for every version of mechanism that can be presented.

Benacerraf criticizes Lucas's failure to notice that it depends very much on how a machine is given that Lucas can say that the Gödel formula is true. He also criticizes Lucas for putting the argument in the form of a match. Lucas claims that however clever the mechanist is, he could out-Gödel its machine and find a true formula that the machine could not. This is the respond to what Benacerraf said about the mechanist; that it is just a man and may have limits in making machines.
I think Lucas did not use Benacerraf's criticism, about the failure for noticing that it depends on the way a machine is given that one can find the Gödel sentence, fairly: Benacerraf meant with this, that one must know the program of a machine to find the Gödel sentence, and not that it is hard to find a Gödel sentence of some sort of complex machine.
Benacerraf thinks that another machine can also find the Gödel formula as well; Lucas says that the mechanist claimed that the machine in question is Lucas, and not that other machine. If the mechanist then would say that that other machine is him, Lucas can find something else that he can do and the machine can not.

Lucas's argument formalized

Benacerraf tries to reconstruct Lucas's argument as a simple proof sequence, which only uses formal defined terms. This proof is a distortion of the original argument. Many of Benacerraf's criticisms are of arguments of which Benacerraf thinks (from a very different point of view) that Lucas had to put forward, but did not do. To Benacerraf complaint about Lucas never made clear the meaning of 'prove' in the sense that Lucas can prove things that machines cannot, Lucas responds that he took care not to use the word in such a sense. Lucas made the contrast of what was provable in the system and what he could produce as true.

Benacerraf does not see why a machine cannot conjure up with formal proofs that are not provable in the system. Lucas answers this by saying that machines cannot conjure up things (as Benacerraf already did), but only acting according its input and program. Mechanists claim that human also acting according its input and program and that the output of some a human is the output of some machine, and thus is simply a part of a formal system. This is what Lucas tries to prove that it is false.

The mechanist sees truth as provable in a given system. Lucas thinks it is unacceptable to reconstruct truth as provable in a system and he says that Gödel's theorems show this. Lucas cannot say what truth exactly is, he thinks he knows, but cannot explain. It has been shown that any attempt to give a formal representation of truth must lead to contradiction. Benacerraf did not argue the question of what truth is, but tried to reconstruct Lucas argument as formally as possible. In this way, he wants to determine what philosophical assumptions are involved and might be given up. The assumption(s) that must be given up, might not have to be the essential part of the mechanist thesis. Benacerraf argues that besides of rejecting 9c, it is also possible to reject the second half of 9b or that of 9a. The latter means that either one cannot determine the consistency of a machine, or that one cannot prove that a machine is adequate for arithmetic. In both ways, Benacerraf denied Lucas an assumption he needs, not by denying that it is true, but just by not admitting that it is true. This assumption is that Lucas knows what kind of Turing machine he is alleged to be. If he does not know that, he cannot prove its consistency or adequacy for arithmetic. Benacerraf's mechanism never specifies what sort of machine a human being is alleged to be, and in this way avoids contradiction.

Lucas thinks that this position is a position too empty to be worth holding. He thinks Benacerraf's mechanist is self-stultifyingly eristic and guilty of omega-inconsistency. The fact that Benacerraf maintains that there is a program number j, so that the corresponding program W_j represents Lucas, knowing that for each program number j there is an argument that shows that Wj does not represent Lucas, is omega-inconsistent. To avoid omega-inconsistency, Benacerraf has to prevent the anti-mechanist from knowing if the machine he is compared to is consistent or not, not from knowing what kind of machine he is.

Consistency

Checking if a system is consistent can be done through the Gödelian formula:

The mechanist makes the specification of a machine of which he claims that it is equivalent to Lucas. Lucas calculates the Gödelian formula for this specification. Then Lucas asks the mechanist if this formula can be proven in the system. The mechanist should know this, because he made this system and thus knows it very well. It follows that the mechanist (or Satan, as given in the match) is ignorant. If he knows the answer, then every answer he would give makes his claim false. If he says yes, then Lucas knows that the system is inconsistent (Lucas found a formula that could not be proven in the system, yet the system can prove this) and thus not equivalent to Lucas. If he says no, then Lucas found at least one formula that is true which the system could not found, so it is not equivalent to Lucas. Lucas thus assumes that he is consistent. I agree with Lucas that this proves that this machine is not equivalent to Lucas. But I think this does not prove that he is NOT a machine. He proved at most that he is not exactly that machine, but another, not replicable.

The dilemma described above can also be represented in terms of consistency. Each formal system and each Turing machine is either consistent or inconsistent. If such a system is inconsistent, it is not a plausible candidate for a model of Lucas's mind. If consistent and it is recognized that it is so, then as a consistent being and being said that he is so, he can out-Gödel the system. Benacerraf tries to avoid this dilemma by represent it in the form of a condition in where the antecedent never can be added as an independent assumption in the form that is needed. Gödel's second theorem can be expressed by saying that for every formal system (M) we can 'prove' the following:

If M is consistent, U is true.

This formal proof can be represented in M, and according to Gödel's second theorem we can prove in M that

If M is consistent, U is true.

So the difference between a human and the system has disappeared. Only if this human can assert that M is consistent, then he can assert that U is true, and does something that cannot be done in M, so shows his difference between him and the machine. If the mechanist comes up with a claim that a certain machine is represented by a formal system equivalent to Lucas, then Lucas can say that this system is consistent. Then the mechanist has to admit or reject his claim. If he admits, then Lucas can find a formula that is true and that the system cannot produce as true, so Lucas is not equivalent to the system and the mechanist must reject his claim.

Lucas says that the only way out for the mechanist is by assuming that we are inconsistent machines, or that we cannot prove that we are consistent. Lucas does agree that we cannot prove our consistency but he mentions that if there is something found in a theory that is inconsistent, we change this theory so that it is not inconsistent anymore. This is typically human, it is no formal proof of our consistency, but it does say something about it. The alternative would be self-stultifying.

3.2. David Lewis and Coder

Lewis responded to Lucas in his article “Lucas against Mechanism” [Lew69]. Lucas then gave a reaction back to Lewis in “Mechanism: A rejoinder” [Luc70a]. In this article he also responds to Coder. Each of the following subchapters will discuss one such respond.

3.2.1 David Lewis – Lucas against Mechanism

In 1969 David Lewis wrote a paper named “Lucas against Mechanism” [Lew69]. In this paper he argues that Lucas’s critics have missed something true and important in Lucas’s argument. Lewis shows this by restating the argument and explains how to avoid the anti-mechanist’s conclusion of this argument.

Lewis states that L is an adequate formalization of the language of arithmetic. *Con* is a function from machine tables to sentences of L, in such a way that the following can be proved by metalinguistic reasoning about L:

C1) If M specifies a machine whose potential output is a set S of sentences, then *Con*(M) is true if and only if S is consistent.

C2) If M specifies a machine whose potential output is a set S of true sentences, then *Con*(M) is true.

C3) If M specifies a machine whose potential output is a set S of sentences including the Peano axioms[2], then *Con*(M) is provable from S only if S is consistent.

Furthermore, Lewis states that Φ is a consistency sentence for S if and only if there is some machine table M such that Φ, *Con*(M) and S is the potential output of the machine whose table is M. Now the following rule can be stated:

P) If S is a set of sentences and φ is a consistency sentence for S, infer φ from S.

Lewis thinks that Lucas is defending this rule. R is a perfectly sound rule of inference: if the sentences in S are all true, then because of C2, the conclusion φ is also true. Using R is performing an inference in L, and not going upward to metalinguistic reasoning about L (making statements about L). Any rule can be shown to be true by metalinguistic reasoning. Lucas (like everybody) takes the peano axioms for arithmetic as true. A sentence ψ is a theorem of Peano’s arithmetic only if ψ belongs to every superset of the axioms which are closed under the rules of logical inference. The same can be done for the arithmetic that Lucas contains, Lewis calls this ‘Lucas arithmetic’: A sentence λ is a theorem of Lucas arithmetic if and only if λ belongs to every superset of the axioms which are closed under the rules of logical inference and closed under the rule R (since Lucas is defending it). Lucas can now produce any theorem of Lucas arithmetic as true.

Lewis then makes the assumption that Lucas arithmetic is the potential output of a Turing machine. If this assumption is true, then this machine would contain a consistency sentence φ. This sentence would be a theorem of Lucas arithmetic, because Lucas arithmetic is closed under R. So λ would be provable from Lucas arithmetic and Lucas arithmetic would be inconsistent according to C3. So if Lucas arithmetic is the potential output of Lucas, Lucas cannot be a machine.

[2] A set of first-order axioms proposed by Giuseppe Peano which determine the theory of Peano arithmetic (also known as first-order arithmetic).

Lewis then describes one more step. He says that Lucas has good reasons to belief that the theorems he produces of Lucas arithmetic are true. But it does not follow that the theorems that Lucas produces is the output of the whole of Lucas arithmetic. There may be theorems of Lucas arithmetic he cannot verify to be such. If this is the case, then his output is not adequate for Lucas arithmetic. It might be the output of another suitable machine.

If Lucas wants to prove that he is no machine, then, according to Lewis, Lucas must convince us that he is able to verify theoremhood in Lucas arithmetic. For now, Lucas gives us no reason to think that he has this ability. Lucas arithmetic is not like an ordinary axiomatic theory. So the fact that we always can verify theoremhood by exhibiting a proof will not do for Lucas arithmetic's theoremhood. Some of Lucas arithmetic's proofs are transfinite sequences of sentences; R can take an infinite set S of assumptions. The proofs will not be discovered by any finite search. The finite proofs in Lucas arithmetic can also not be checked by any mechanical procedure. A checking procedure is needed to check if R has been used correctly. This procedure has to decide if a given finite set S of sentences was the output of a machine with a given table M. Lewis says that such a method can easily be converted into a general method for deciding if a Turing machine will halt on any given input. Turing proved that such a method cannot exist.

Lewis ends his paper by saying that Lucas can certainly go beyond Peano arithmetic, but he can still be a machine, given that there are some theorems of Lucas arithmetic that he cannot produce as true.

3.2.2. David Coder – Gödel's Theorem and Mechanism

David Coder discusses Lucas's argument in his article "Gödel's Theorem and Mechanism" [Cod69] published in 1969. Coder thinks that the claim that Gödel's theorems prove that mechanism is false and the claim that if Lucas's proof is valid, it will have a lot of consequences for the whole of philosophy, are exaggerated.

Coder agrees with Lucas that minds cannot be explained by machines, but Gödel's theorems do not prove this. He claims that Gödel's theorems at most prove that not every mind can be explained in mechanical terms. Furthermore he blames Lucas for falsely characterize machines, namely that everything that is able to follow mechanical procedures in mathematics, and nothing else, is a machine. Therefore he believes that Lucas overestimated the importance of Gödel's theorem for mechanism.
Coder states that if minds are essentially different from machines (P1), then no mind can be explained by a machine (P2). He claims that both parts of this proposition are true, but Lucas's argument does not prove any of these:

- Lucas has not shown P2, because Lucas has not shown that there is no adequate model for minds which cannot find the Gödel theorem.
- P1 has therefore also not been proven, because P2 is not true and thus it is not proved that minds are essentially different from machines.

Coder thinks that Lucas gives the wrong impression that if someone cannot do arithmetical operations, but only for example pick proofs from a set of proofs and non-proofs, that this person's behavior when doing arithmetic can be explained by a mechanical model.
Coder states that what makes something a machine is most of all that it operates an algorithm. How the machine operates that algorithm is not important. The algorithm can only be operated mechanically (since it is a mechanical procedure). Coder thinks that Lucas misunderstood the term "mechanical" in the concept of mathematics. "Mechanical" in the concept of mathematics

means: “the instructions given by a Turing algorithm do not at any step leave open the next step *in the calculation*.” [Cod69]. When a mind is operating an algorithm, it will also do something else besides taking steps in the calculation. For example it may make mistakes or get confused and have to go back to the beginning of the calculation. So it cannot be inferred that a mind, which arithmetic abilities are limited to those of a Turing machine, is a machine (regarding his arithmetic ability). Lucas thinks that what makes something a machine is that it is able to calculate, only according to mechanical procedures. Either this excludes that it is important for being a machine that its construction determines its operation, or the mathematical definition of Turing machine requires that the construction of the thing which operates a Turing algorithm determines its behavior. Coder argued that the latter is false, but Lucas presumes it is true. According to Coder, Lucas says that a calculator behaves completely determined by its construction and input, unless the instructions that it needs to calculate are not definite. Lucas thus thinks it important for the falsification of mechanism that minds are not restricted to calculating according to mechanical procedures (regarding to arithmetic). But when we might see that even if we could not do arithmetic in another way, and it would not follow that our behavior, in doing arithmetic, as determined by our construction and input, the importance of the fact that minds are not restricted (as ascribed above) is lost.

I think that Coder himself misunderstood the term mechanism. As described before, mechanism defends that everything about human can be explained in mechanical terms. So if Lucas had found something (of a human) that cannot be explained in mechanical terms, then mechanism cannot be true, or at least it has to change its description to “mechanism defends that there are things about human that can be explained in mechanical terms”. Lucas thinks that the mind can not at all be explained by mechanical terms and just used something (Gödel’s theorems) to disprove mechanism.

3.2.3. John Lucas – Mechanism: a Rejoinder

Lucas respond to Lewis and to Coder in one paper named “Mechanism: a Rejoinder” [Luc70a]. Lucas claims that he does not have to show that that any mind can do all of Lucas arithmetic. He does not need the whole of Lucas arithmetic to show that he is not a machine; one theorem that he can see that is true and the machine not, is enough. So to respond to Coder, he does also not need to show that all minds can understand Gödel’s theorem. He agrees with Lewis that when he cannot verify the whole of Lucas arithmetic, he might still be a machine. Lucas says that everything he does (as a finite being living for a finite time) can be copied by some machine. But this is not the case; the machine must be able to do more than copying someone’s behavior, it must have the ability to do everything a mind could do, then it can be said that the mind is equivalent to a machine. Lucas says that for each machine, there is some theorem of Lucas arithmetic which the machine cannot produce as true. So to show that mechanism is false, Lucas have to show that the mind can produce a relevant part of Lucas arithmetic. Lucas believes he can do this, thanks to Gödel’s theorems.

Lucas says that Lucas arithmetic represents the sort of arithmetic that minds can do and machines not. This is because minds operate according to the Lewis’s ‘infinitary rule of inference’ R, given by Lewis. Lucas thinks that this term is not well chosen, because the point of conducting our inferences by rules is that the rule should be definite and finite. Many of the inferences of the mind cannot be formalized. The fact that we can make valid inferences without a definite rule allowing us to, distinguish us from a machine.

Coder complains that Lucas have shown that only the minds that can understand Gödel's theorems are not machines. Lucas responds to this that it is not limited that minds can come up with new arguments. Many people could understand Lewis's R without following the proof of Gödel's theorems. So their arithmetic would lie within the Lucas arithmetic (or a richer one). So they are not machines, because they have also a reasoning which cannot be expressed by rules. Lucas used mathematical ability because it is easier to compare with machines, not because it is more (or less) like a machine than other abilities.

Another complain of Coder is that the usage of the Gödel theorems are unnecessary, because he thinks that minds are so unlike machines, that one cannot even suppose them to be the same. Lucas agrees that there are other reasons to think that minds are no machine, but philosophers want to see facts before they may take sides. His Gödelian argument will act as a long-stop until someone will disprove all other arguments of mechanism (for example that is just a matter of complexity that differentiates machines from minds).

4. Discussion & Conclusion

In the reply to Lewis and Coder, Lucas says that it is not necessary to know his whole theoremhood to prove that minds cannot be explained by machines. I do not agree with Lucas in this point, because, as Lewis says, there may be theorems of Lucas arithmetic he cannot verify to be such. This has been proven for formal systems; there is a truth which the machine cannot verify, but which a rational being can verify. But why is it not possible that if someone (A) looks at the knowledge of someone else (B), A can find a truth in that knowledge that B was not able to prove, but was true in the knowledge of B?
One might say that if we cannot verify the theoremhood of ourselves, then we cannot even think of verifying the theoremhood of someone else. Of course, but if we cannot verify someone's theoremhood (or that of ourselves), then we cannot claim that the difference between machines and minds is that we can find a truth, which the machine cannot. First we have to be able to show that there is no truth in our theoremhood which we cannot verify. I think this is not possible, since we can only prove those truths, which we are capable of to prove.
So to answer to the question if Lucas has proved that mechanism is false, is in my opinion, that he did not.

In the reply to Benacerraf, Lucas says that if he is faced with a formal system that is claimed to be equivalent to Lucas, he can also find a truth of this system, which the machine could not. I think that Lucas cannot do this, since if this truth was indeed in Lucas, the system must contain this truth also (because the machine is equivalent to Lucas). So when Lucas claims that he has the ability to find a truth in this machine, he already made the assumption that he is not a machine. Or the assumption that he already made is that he is another 'better' machine, but then the claim that Lucas was equivalent to the formal system was already 'disproved' by this assumption.

I agree with Lucas that a particular mind cannot be modeled in mechanical terms, since we do not know everything about our mind (or brains). If we do know everything about our mind, we may be able to represent the mind (in the sense that it can also act on its own) with a proper model, but this model may not be a mechanic model. But we do not know if the mind can be explained by mechanical terms, so we cannot say anything about it. So we can also not proof our own consistency (to answer the second question stated in the summary). We can keep trying to find proofs to show that minds are not like machines as Lucas did, but these are not really formal proofs, as we like to see it. We must understand our mind before making hard statements about it. KNOW THYSELF will not be sufficient for making statements about a mind.

5. Literature references

[AbsAs] *http://www.absoluteastronomy.com/encyclopedia/M/Me/Mechanism_(philosophy).htm*

[Ben67] Benacerraf, P. (1967). God, the Devil, and Gödel. *The Monist*, 51:9-32.

[Cod69] Coder, D. (1969). Gödel's Theorem and Mechanism. *Philosophy*, 44:234-237.

[Dav95] Davis, M. (1995). Is mathematical insight algorithmic? *Behavioral and Brain Science,* 13(4):659-660.

[Gio03] Giolito, B. (2003). Introduction to Lucas' argument against mechanism by means of Gödel's incompleteness theorem. *Ethics and Politic,* 5(1)

[Hof79] Hofstadter, D.R. (1979). *Gödel, Escher, Bach. An eternal golden Braid.* Basic Books, New York.

[Lew69] Lewis, D. (1969). Lucas against Mechanism. *Philosophy,* 44:231-233.

[Lew79] Lewis, D. (1979). Lucas against Mechanism II. *Canadian Journal of Philosophy,* 9:373-376.

[Luc61] Lucas, J.R. (1961). Minds, Machines and Gödel. *Philosophy* 36:112-127

[Luc68a] Lucas, J.R. (1968). Satan stultified: A rejoinder to Paul Benacerraf. *The Monist,* 52(1):145-158.
Reply to [Ben67]

[Luc70a] Lucas, J.R. (1970). Mechanism: A rejoinder. *Philosophy,* 45:149-151.
Reply to [Lew69] and [Cod69]

[Luc84] Lucas, J.R. (1984). Lucas against mechanism II: A rejoinder. *Canadian Journal of Philosophy,* 14:189-191.
Reply to [Lew79].

[Luc96a] Lucas, J.R. (1996). Minds, machines and Gödel: A retrospect. *in P.J.R.Millican and A.Clark, eds., Machines and Thought: The Legacy of Alan Turing,* Oxford University Press, pp.103-124.

[Nag56] Nagel, E. Newman, J.R. (1956). Gödel's Proof. *Scientific American*, 194(6):71-86

[Ros50] Rosenbloom, P. *The Elements of Mathematical Logic*. New York: Dover Publications Inc., 1950, p. 72.

[Rus95] Russel, S. Norvig, P. (1995). *Artificial Intelligence: A Modern Approach.* Prentice Hall*,* Upper Saddle River, New Jersey.

[Vis04] Visser, A. (2004). Kunnen wij elke machine verslaan. Beschouwingen rond Lucas' argument, *Algemeen Nederlands tijdschrift voor wijsbegeerte,* **pp**

[Web68] Webb, J.C. (1968). Metamathematics and the Philosophy of Mind. *Philosophy of Science,* 35:156-178.

[Whi62] Whitely, C. (1962). Minds, Machines and Gödel: A reply to Mr. Lucas. *Philosophy,* 37:61-62.

[Tru88] The Gödelian Argument. (1988). Truth, 2 (http://www.leaderu.com/truth/2truth08.html).