

Mechanism and Löb's Theorem

October 20, 2011

1 Introduction

Do mathematical results limiting what Turing machines can do show that, in some important sense, the human mind cannot behave like a Turing machine? There's a long and (it is generally thought) none-too-successful history of arguing that they do.

In his 2002 paper, 'Löb's Theorem as a Limitation on Mechanism', Micheal Detlefsen proposes a novel, and more modest, argument along broadly these lines. Rather than using Gödel's Incompleteness theorems, Detlefsen appeals to Löb's theorem. And, rather than arguing that *our* minds aren't mechanistic, he defines a notion of being an 'additive epistemic authority' and then argues that we couldn't know mechanism to be true of any creatures that function as additive epistemic authorities for us¹.

In this note I am going to argue that Detlefsen's modest new computability theoretic argument against mechanism about the mind still fails. I will argue that what follows from Löb's theorem is not his desired claim that we could never know an additive epistemic authority had a mechanistic mind, but only the much weaker claim that we could never know which particular program that captured the behavior of a creature's mind, while still having that creature

¹"a device specifically known by an observer to be mechanical cannot be used as an [additive] epistemic authority" Löb's Theorem as a Limitation on Mechanism", Michael Detlefsen, *Minds and Machines* (2002) pg. 353

function as an epistemic authority for us.

2 Assumptions and definitions

Let me begin by setting up a number of idealizing assumptions and definitions which we will need to understand Detlefsen's ultimate claim that, "There are epistemically valuable humanoid systems of belief A such that no humanoid observer O who uses A as an 'additive' epistemic authority can either know or truly believe of A that she is mechanizable (i.e. that her belief-set, A , is r.e.)"

Detlefsen's begins by introducing an important idealizing assumption about logical omniscience. He takes the belief sets of the observer O , and authority A , to be closed under logical entailment. I will make the same assumption for the rest of this paper.

We can also see from the passage quoted above Detlefsen identifies the claim that a believer B is 'mechanizable' or has a 'mechanical' mind with the claim that their belief set is recursively enumerable. That is: a creature has a mechanical mind if and only if there is some turing machine which would eventually list all and only the sentences which that creature accepts. (Note that this does not mean that there has to be a program which correctly classifies each sentence as either one that the subject believes or not.) I will follow him in identifying mechanism with having a recursively enumerable belief set in what follows.

Finally, Detlefsen defines the notion of being an additive epistemic authority as follows. An observer O to take another subject A as an *epistemic authority* if and only if O is disposed to accept all sentences of the form 'if A asserts that ϕ , then ϕ '. In such a situation A further constitutes an *additive* epistemic authority if A is disposed to assert at least one thing which O couldn't already learn by deriving it from his current beliefs.

3 Detlefsen's Argument

Now that we have pinned down the meaning of Detlefsen's intended conclusion, let us turn to his argument.

Löb's theorem says that, given a formal provability predicate PROV which satisfies certain constraints, and a formal system S which extends Peano Arithmetic, you can only prove "If $\text{PROV}(X)$ then X " in S , in cases where you can already directly prove ϕ . The relevant conditions are: adequacy (if $\vdash P$ then $\vdash \text{PROV}(P)$), formal adequacy (\vdash if $\text{PROV}(P)$ then $\text{PROV}(\text{PROV}(P))$), and formal modus ponens (if $\vdash P$ and $\vdash P \rightarrow Q$, then $\vdash Q$).

Now Detlefsen then draws on Löb's theorem to argue that (if you were logically omniscient) you couldn't know the program that captured the mind of any creature which counted as an additive epistemic authority for you. In essence, he argues that if I know some program P recursively enumerates an angel A 's beliefs then the predicate 'is enumerated by program P at some stage t ' winds up acting enough like a formal provability predicate for an analog of Löb's theorem to be provable. Thus, the following three things can't be simultaneously true of O 's relationship to an angel A : a) A is an additive epistemic authority for O , b) A 's belief set is r.e., and c) O knows that a certain particular program P , recursively enumerates all the things that A is disposed to say. As Detlefsen puts it, "[A subject O with normal mathematical competence is] prohibited from both knowing (or truly believing), of any specific formal system, that A 's belief corpus is coextensive with its theorem set and using A as an authority to additively expand her (i.e. O 's) beliefs."² Call this last claim **CAN'T KNOW WHICH PROGRAM**.

Finally Detlefsen draws on **CAN'T KNOW WHICH PROGRAM** to draw the further quoted conclusion quoted above, that one can't know that any such

²ibid. 367

creature which functions as an epistemic authority for you has a mechanizable mind. Call this CAN'T KNOW THERE IS A PROGRAM.

I will argue for two claims. First, we don't need anything as fancy as Lob's theorem to establish CAN'T KNOW WHICH PROGRAM. This claim actually follows trivially from the definitions above.

Secondly, and more importantly, the stronger conclusion CAN'T KNOW THERE IS A PROGRAM does not follow from CAN'T KNOW WHICH PROGRAM FAILS. Thus Detlefsen's argument for anti-mechanism about the mind of epistemic authorities fails for reasons that are closely analogous to those which Putnam cites against the original Lucas-Penrose recursion theory based argument against mechanism about the mind.

4 A quick argument for 'Can't Know Which Program'

We can get the claim "[A subject O with normal mathematical competence is] prohibited from both knowing (or truly believing), of any specific formal system, that A's belief corpus is coextensive with its theorem set and using A as an authority to additively expand her (i.e. O's) beliefs." by simple and direct reasoning from the definition given above as follows.

If I know that you only have true beliefs, and I know that a certain algorithm A lists everything you believe, (and my beliefs are closed under logical entailment, as per Detlefsen's assumption) then you can't know more than me. For, anything which you know, I can learn by first proving that A lists it, and then inferring that you must believe it so it must be true. That is: if you discover that a certain program perfectly emulates the oracle's behavior, you don't need to travel to Delphi, because can stay home and derive what it would have told

you in your armchair.

More formally: suppose A is an additive authority for me, and I know that some program M recursively enumerates the propositions which A accepts. Then (because A is an additive authority for me) there is some proposition P, which A can derive but I can't. But M recursively enumerates the set of sentences which A is disposed to accept. So, there's some stage t at which A arrives at proposition P, and the t-th step of program M is to output P. But I can prove that the t-th stage of program M outputs P, and I believe that M enumerates all and only the sentences A accepts, AND I believe of each proposition that if A accepts, then it's true. So, from these things, I too can derive that P. Contradiction.

Thus, I think that Detlefsen had no need to invoke anything so elaborate as Löb's theorem to get the conclusion that, on his assumption of logical omniscience, you can't know which program recursively enumerates the beliefs of a creature that acts as an epistemic authority for you. But giving a slightly over-elaborate argument is not a serious philosophical problem.

5 Why 'Can't Know There Is A Program' doesn't follow

We have just seen that Detlefsen is certainly right that, if A is an additive epistemic authority for you, you can't know, of any particular program, that it enumerates A's beliefs. But now contrast this claim with the desired conclusion that immediately follows it in the paper, "We therefore conclude that: (General Limitative Thesis) There are epistemically valuable humanoid systems of belief A such that no humanoid observer O who uses A as an 'additive' epistemic authority can either know or truly believe of A that she is mechanizable (i.e.

that her belief-set, A, is r.e.).”³

I want to claim that Detelfson’s argument breaks down at this point: CAN’T KNOW WHICH PROGRAM does not suffice to entail CAN’T KNOW THAT THERE IS A PROGRAM. It doesn’t follow from the fact that, if you knew which program enumerates all the sentences A is disposed to accept then A would no longer be an additive epistemic authority for you, that you can’t truly believe there is *some* program which enumerates the sentences which A accepts. Thus, the a statement quoted above, saying that you can’t *know which* program captures the behavior of something that’s an additive epistemic authority for you), does not seem to support the General Limitative Thesis.

One might worry that countenancing this kind of scenario where I can know an authoritative angel’s mind is mechanical but not which program captures that angel’s beliefs commits us to the possibility of some kind of deeply unknowable truths. But this is not the case. Nothing, (except for expense, difficulties associating brain states with belief states, moral considerations etc.), would stop me from doing some kind of scan of someone who is an epistemic authority for me, and then building an atom-by-atom simulation of their behavior - and hence, of how they would answer all mathematical questions. The trivial argument for CAN’T KNOW WHICH PROGRAM given in the previous section just tells us that *if* I did this (and was logically omniscient) the relevant creature would henceforth cease to count as an epistemic authority for me.

[Maybe add something like this ?? : Thus. D’s argument fails for reasons are pretty much directly analogous to the ones Putnam presses against Penrose’s classic computability theoretic argument against mechanism about mind. Penrose argued that the mind couldn’t be captured by a program, since if it were we could realize this and accept the consistency sentence for this program, contra known facts about what sound algorithms for reasoning about the numbers can

³ibid. 367

allow one to do. Putnam pointed out in order to arrive at the con sentence for your own reasoning you would have to know your own program, and you might not know your own program.]

6 Conclusion

In this paper I have tried to show that (contra Detlefsen) Löb's Theorem does not give rise to any constraint on mechanism. The only conclusion it does support turns out to be a rather trivial one: if something is an additive epistemic authority for you at a given time, you cannot (then) know which program that recursively enumerates their beliefs.