

Mechanism and Löb's Theorem

May 3, 2010

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. However, 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 can't be like computers, he argues that "a device specifically known by an observer to be mechanical cannot be used as an epistemic authority (of a particular type)."¹

I am going to argue that Detlefsen's new brand of argument fails. The key step in seeing why, will be to carefully distinguish between two epistemic states: the state of knowing that *some* program fully captures the behavior of a given physical system, and the state of knowing, of some particular program P, that *it* captures the behavior of a given mechanical system. Detlefsen's target is the former claim: we can't know, of something that's (what Detlefsen calls) an additive epistemic authority for us, that there is some program which captures

¹"Löb's Theorem as a Limitation on Mechanism", Michael Detlefsen, *Minds and Machines* (2002) pg. 353

its behavior. But (I will argue) this does not follow from the claim which he appeals to Löb's theorem to establish - namely, that we can't learn which program captures the behavior of some system A, while having A still function as an additive epistemic authority for us. In fact, given Detlefsen's definition of what it takes to count as an *additive* epistemic authority, this weaker claim winds up being a trivial truth. Thus, I propose to show that Löb's theorem imposes no limitation at all on mechanism.

2 Set Up

My argument will be very quick, but first we need to set up a bunch of relevant terms: what's an 'additive' epistemic authority, what does it mean to believe that something is mechanizable, and what exactly does Löb's theorem say?

As Detlefsen introduces the terms, for an observer O to take A as an *epistemic authority*, means that O is disposed to accept all sentences of the form 'if A asserts that X, then X'. And, for A to constitute an *additive* epistemic authority, requires the further condition that A is disposed to assert something which O couldn't already learn by deriving it from his current beliefs. However, at the beginning of the article Detlefsen also imposes the idealizing assumption, that the belief sets of the observer O, and authority A, are closed under logical entailment - so I will take this as an additional requirement for A to be an epistemic authority for O.

Detlefsen doesn't explicitly define what it means for a believer A to be 'mechanizable', but I take it this means their belief set is recursively enumerable. For he writes: "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.)”². That is: there’s some Turing machine program p , which lists off all the sentences that A believes. Or, equivalently, there’s a program p which, for any input of a sentence s , will eventually halt and output ‘1’ if A believes that s .

Finally, Löb’s theorem says that, given a formal provability predicate PROV which satisfies certain constraints, and a formal system S which extends PA , you can only prove “If $\text{PROV}(X)$ then X ” in S , in cases where you can directly prove X . These 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 $\vdash Q$).

3 The Problem

After going through the points above, Detlefsen very plausibly draws the consequence that the following three things can’t simultaneously be true, for a given human-like subject O , with ordinary mathematical competence: a) A is an additive epistemic authority for O , b) A 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 he puts it, O “is thus [as a result of Löb’s theorem together with his idealizing assumptions] 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.”³

The essence of Detlefsen’s argument, is that if O knows a program P which recursively enumerates A ’s beliefs, then the predicate ‘is enumerated by program P at some stage t ’ winds up acting enough like a formal provability predicate, in the sense of Löb’s theorem, for a ‘general Löblike result’ to be provable.

²ibid. 367

³ibid. 367

But, I won't go through this argument in more detail, because it turns out that you don't need anything like Löb's theorem to establish the claim in quotes above. We can directly reason from Detlefsen's characterization of what it is for one logically omniscient being to be an additive epistemic authority for another, to the conclusion, 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 you 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.

So 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 fact, with the conclusion that immediately follows it "We therefore conclude that: (General Limitative Thesis) There are epistemi-

cally 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.).”⁴

From the fact that, if you knew which program enumerates all the sentences A is disposed to accept, A would no longer be an **additive** epistemic authority for you, it doesn’t follow that you can’t truly believe there is some program which enumerates the sentences which, something that’s currently an additive epistemic authority for you, would accept. 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.

Note, finally, that taking an an epistemic authority to be mechanizable doesn’t even require any interesting commitment to unknowable truths. Nothing, (except for expense, and perhaps moral considerations), 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). It’s just that, if I did this, and thereby came to know that the authorities behavior was captured by such-and-such a program, they would thereby cease to count as an epistemic authority for me.

4 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 is, what 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

⁴ibid. 367

program that recursively enumerates their beliefs.