Of humans and universality thresholds

Paul Christiano Follow
Oct 23, 2016 · 4 min read

(*Note: relevant to a narrow audience.*)

I think that something like HCH might be extremely sophisticated while remaining aligned with human values (when combined with some form of reliability amplification). In fact I'm cautiously optimistic that if you put any sufficiently smart agent into a better version of HCH then you will get out a "universal" agent, which depends on the *values* of the agent that you put in but doesn't depend on its *abilities*.

I've encountered the following objection:

- It seems clear enough that *no number of apes* could invent calculus in an ape-aligned way. In order to understand the values of an ape, we need to do some kind of bona fide extrapolation, to ask what would happen if the ape was more sophisticated rather than merely asking what would happen if there were more of them.

  In fact we don't even have to go so far as an ape—if you took most people and put them in a box and asked them to organize a trillion copies of themselves to compute their enlightened judgment, the results wouldn't be pretty.

  It seems strange for the people having this discussion to imagine that the threshold is right below them. Surely it is more likely that there are more sophisticated creatures than humans, who might think to themselves "well clearly if you put a *human* in this process it wouldn't work, most of them couldn't even understand inter-universal Teichmüller theory."

We could make the same argument with respect to time: it seems likely that HCH fails if run with a human who only has ten seconds to spend thinking. So why think that a day is enough time?

I think that these arguments are mistaken—that's not to say that HCH is obviously universal, but it is to say that we shouldn't be *a priori* surprised if there is a universality threshold and it lies within the range of human variation, or between "ten seconds" and "a day."

An algorithmic analogy

Suppose that we were instead interested in the *algorithmic* capacity of HCH: what functions can be computed by HCH, given a natural description of the function?

For this problem, there seems to be a clear universality threshold. As long as a human is sophisticated enough to eventually understand and develop the idea of a robust universal computer, and to implement simple logic operations with modest accuracy, then their HCH can compute arbitrary computable functions.

On the other hand, an ape's HCH is not going to be able to implement much of anything, and indeed it seems plausible that the majority of humans would not be able to implement particularly complex functions. And if we give a human only ten seconds, they probably can't implement a very complex function either.

So for this natural problem there actually *is* a clear universality threshold in HCH's behavior.

Caveats

How good is the analogy? One might object that this universality claim is really about learning to use external computational resources —in some sense this is different in kind from moral deliberation.

But I think these two cases are actually closely analogous in the relevant way.

We are confident that an ape's HCH doesn't do anything good for basically the same reasons we are confident that an ape's HCH can't compute anything complicated: they couldn't build any the machinery to do anything more sophisticated than the initial ape.

How is the function described? One might object to assuming that we are "given a natural description of the function." Perhaps this smuggles in a human-centric notion of "natural description."

But we could loosen this condition to “given *any* advice whatsoever” and we still seem to observe a universality threshold—there seems to be **no** reasonable advice that would cause a realistic ape to reliably implement a very complex logical function. For a clever human you can pick just about any natural representation and they’ll get it.

Alternatively, we could observe that language itself seems to exhibit a universality transition, which is independent support for the original claim. Humans understand a range of universal languages into which any other language can be translated with enough effort (roughly speaking). Other animals do not have such a language.

Conclusion

The analogy to computation doesn’t provide direct positive evidence *for* the universality of HCH in the moral sense. But I think it does provide a strong reason to be skeptical of skepticism about the existence of a universality threshold within the range of human variation; universality thresholds are not an uncommon phenomenon, and in fact humans straddle a number of similar thresholds.

In the end I think that we have to use other arguments and intuitions to try to determine whether HCH is universal.