# Humans consulting HCH

Paul Christiano Follow
Jan 28, 2016 · 2 min read

*(See also: strong HCH.)*

Consider a human Hugh who has access to a question-answering machine. Suppose the machine answers question Q by perfectly imitating how Hugh would answer question Q, *if Hugh had access to the question-answering machine*.

That is, Hugh is able to consult a copy of Hugh, who is able to consult a copy of Hugh, who is able to consult a copy of Hugh…

Let's call this process HCH, for "Humans Consulting HCH."

I've talked about many variants of this process before, but I find it easier to think about with a nice handle. (Credit to Eliezer for proposing using a recursive acronym.)

HCH is easy to specify very precisely. For now, I think that HCH is our best way to precisely specify "a human's enlightened judgment." It's got plenty of problems, but for now I don't know anything better.

## Elaborations

We can define realizable variants of this inaccessible ideal:

- For a particular prediction algorithm P, define $\text{HCH}^{\text{P}}$ as:
  "P's prediction of what a human would say after consulting $\text{HCH}^{\text{P}}$"
- For a reinforcement learning algorithm A, define max-$\text{HCH}^{\text{A}}$ as:
  "A's output when maximizing the evaluation of a human after consulting max-$\text{HCH}^{\text{A}}$"
- For a given market structure and participants, define $\text{HCH}^{\text{market}}$ as:
  "the market's prediction of what a human will say after consulting $\text{HCH}^{\text{market}}$"

Note that e.g. $\text{HCH}^{\text{P}}$ is totally different from "P's prediction of HCH." $\text{HCH}^{\text{P}}$ will generally make worse predictions, but it is easier to

implement.

## Hope

The best case is that $HCH^{P}$, max-$HCH^{A}$, and $HCH^{market}$ are:

- As capable as the underlying predictor, reinforcement learner, or market participants.
- Aligned with the enlightened judgment of the human, e.g. as evaluated by HCH.

(At least when the human is suitably prudent and wise.)

It is clear from the definitions that these systems can't be any *more* capable than the underlying predictor/learner/market. I honestly don't know whether we should expect them to match the underlying capabilities. My intuition is that max-$HCH^{A}$ probably can, but that $HCH^{P}$ and $HCH^{market}$ probably can't.

It is similarly unclear whether the system continues to reflect the human's judgment. In some sense this is in tension with the desire to be capable—the more guarded the human, the less capable the system but the more likely it is to reflect their interests. The question is whether a prudent human can achieve both goals.