Strong HCH

Paul Christiano Follow
Mar 24, 2016 · 5 min read

In a previous post I defined humans consulting HCH (HCH). In this post I'll define a more powerful process. In the future I'll use "HCH" to refer to this stronger process, and call the old process "weak HCH."

As before, we consider a particular human Alice interacting with a particular computer over a particular interval of time, say two weeks in mid-August. The computer provides two "magic" functions:

- The computer presents an initial message to Alice. After Alice replies to that message, the computer may immediately give her a follow-up message, she may submit a follow-up reply, it may follow-up again, and so on. This dialog is the input/output of the overall process.
- The computer also allows alice to consult assistants. Each assistant is assigned a unique ID, and Alice can communicate with them by entering a message together with the appropriate ID. Initially, the computer tells Alice the ID of a single "fresh" assistant.

HCH is the input/output functionality implemented by Alice, if each of her assistants also implements HCH. We can either define HCH as a fixed point, or we can force the tree of assistants to be well-founded and define it by induction.

Defining the interaction between Alice and her assistants is the main subtlety.

Each time Alice sends a message to an assistant, the assistant is copied, and the message is sent to the new copy. Alice is then immediately given the copy's reply, as well as its ID. At that point Alice has two ID's, and she can send messages to either—sending a message to the new ID allows her to continue the conversation, while sending a message to the old ID allows her to ask an alternative question. Each new message spawns an additional copy. If an assistant has used up its 2 weeks, then subsequent messages receive an empty reply.

Messages between Alice and her assistants can contain both text and IDs of other assistants. When an ID is sent as part of a message, the corresponding assistant is copied and a new ID is generated. The recipient sees the new ID rather than the old ID, and can use the new ID to interact with the new copy. These ID's can be short (say 10 letter) strings, because any given copy of Alice can only have direct access to a modest number of IDs.

Comparison to weak HCH

In the previous version of HCH, each assistant answered a single question and then disappeared. The new version differs in that:

- Assistants can carry on discussions instead of simply answering questions.
- Messages can contain pointers to other assistants. This allows the implicit size of messages to be arbitrarily large (since those assistants may themselves have pointers to other assistants…)

I think the new formulation of HCH is significantly better, and in the future I am going to use "HCH" to refer to this version.

It's not clear how much the first change increases the power of HCH, but intuitively it feels like a significant improvement.

The second change feels like a big intuitive improvement, and also increases the complexity-theoretic expressiveness of HCH. The old version could be computed in EXPTIME, while the new version can compute any decidable function.

Generalizing

As with weak HCH, we can define HCH^{P} or $\text{max-HCH}^{\text{A}}$. These definitions are actually quite subtle—both of the changes mentioned in the last section lead to significant new complexities. I'm not really sure if $\text{HCH}^{\text{market}}$ can be adapted to this setting.

We can bound the complexity of HCH by putting a limit on the total number of copies created, or on the total number of minutes that copies spend thinking. Once the limit is exceeded, no more messages can be sent. There might be a global budget, or each copy might have their own budget which they have to explicitly share with any new copies they instantiate. I'll refer to these variants collectively as "bounded HCH."

We might describe the process above as HCH(Alice thinking for 2 weeks). We could more generally define HCH(A) for any process A, by replacing "Alice thinking for two weeks" with an instance of A.

Different processes A may use different communication channels (for example, we could give Alice the ability to video chat with her assistants). The only restriction is that the communication channel be able to include IDs, both to address the message and to include pointers to other assistants.

On universality

I think that this stronger form of HCH is in some sense a "universal" way to combine a bunch of black-box instances of A. That is, for any alternative program ALT(A) that combines a bunch of black-box instances of A, there is some advice we could provide to A that would cause HCH(A) to simulate ALT(A) with only a linear slowdown. Moreover, this advice depends only on ALT, and not on A itself—it works for any "cooperative" A that understands the advice and effectively follows instructions.

For example, this version of HCH subsumes any computable black-box reflection procedure (including this one).

(Technically, this only works if ALT is simple enough that it can be implemented by A. In order to make the model formally universal, we would have to augment HCH(A) by giving it access to an arbitrarily large input.)

This argument is informal—the reduction doesn't work for any A, and it's not clear exactly what is required to make it work. But intuitively it seems to work for the class of cooperative, intelligent English speakers.

If we remove either of the two augmentations described in the section **Comparison to weak HCH**, then the new model seems to be strictly less powerful. There is no way for either of those impoverished models to effectively simulate HCH (without first building a simulation of A, either violating the black-box assumption or introducing exponential overhead).

The interesting caveat to universality is that a human A can probably carry out the reduction well but not perfectly. So if the computation is very large a naive reduction may fail with high probability. We could try to implement appropriate error-correction procedures to drive

down the probability of error, but it is extremely hard to analyze those measures formally. (If Alice makes independent random errors then error-correction will work for driving down the error probability to an arbitrarily small constant, but that's not a good error model.)

So, given error-prone A, it is not really clear whether HCH is universal. I haven't thought much about this issue.

This form of universality does not rule out the utility of thinking more about how to combine additional copies of A. It just implies that such thinking won't require any change to the structure of HCH. This means, for example, that it could be carried out inside of HCH as the first step of a complex deliberation (but there is no guarantee that HCH can think of anything that we could think of).

Conclusion

I've defined a much stronger form of HCH; in the future I don't think I will have much reason to use the original form. This stronger HCH seems to be essentially universal amongst all procedures that invoke human behavior as a black box.