Hard-core subproblems

Paul Christiano Follow
Nov 26, 2016 · 2 min read

Given a research problem X, say that Y is a hard-core subproblem if:

1. A solution to X implies a solution to Y.
2. We aren't currently making progress on Y, we don't know how to make progress on Y, and Y isn't getting any easier over time.

Example

I think that the easy goal inference problem is a hard-core subproblem of cooperative inverse reinforcement learning (CIRL), or at least for the application of CIRL to superintelligence.

CIRL will become easier as we develop improved AI systems, and poses many natural theoretical and practical questions. But the easy goal inference problem doesn't seem to be getting any easier over time, and I don't think we have compelling angles of attack on this problem.

Moreover, a solution to CIRL implies a solution to the easy goal inference problem, since the easy goal inference problem is just the special case where we have perfect information and unlimited time.

Relevance

If we can identify and agree on one or more hard-core subproblems then I think we should generally prioritize work on them. If a hard core turns out to be easy, then we'll have learned something and not much is lost. If a hard core turns out to be very hard, then it's probably a good thing to focus on—it's a prime contender for a bottleneck, and it's likely to be a key conceptual aspect of the problem.

If we suspect that a problem is likely to be insoluble, I think we should prioritize distilling that intuition/argument into a hard-core subproblem. If we succeed, then we have found much better evidence that the problem is hard, and acquired a useful tool for organizing work on the problem. If we can't find a hard-core subproblem, I think that partially undermines our original suspicion. (Of course the

problem may still be insoluble, but we probably shouldn't be confident that it is insoluble.)

Conclusion

In general, I think that identifying a hard core is an important tool both for understanding whether a problem is hard, and for clarifying and organizing research programs. I think this concept might help bridge the gap between some researchers in the AI safety community (especially at MIRI) who are highly pessimistic about AI control, and AI researchers who are optimistic that we can "cross that bridge when we come to it."

This concept is complementary to a focus on prosaic AGI. In some sense *any* problem might become easier over time, since we might encounter some unknown unknown that bears on that problem. But it's much easier for a subproblem to be a hard core *conditioned on prosaic AGI*, since in that case we won't encounter any unknown unknowns: if you want to claim that some subproblem will get easier over time, you have to actually point to the known unknown that will make it easier.