November 2019 Research Thoughts[1]

Abram Demski
Machine Intelligence Research Institute
abram@intelligence.org

My overall research arc for the past 2+ years has been focused on decision theory, and particularly reflective consistency. I spent a long time trying different angles on the problem, especially relating to trying to get UDT to work with logical induction.

Around December 2018, I had a big update against the "classical decision-theory" mindset (in which learning and decision-making are viewed as separate problems), and towards taking a learning-theoretic approach. ("Learning theory" meaning the more theoretical side of machine learning, in which you prove regret bounds, study VC dimensions, and so on—regret bounds being the most interesting tool from my perspective.) This view has lead me to depart from the goal of "UDT for logical induction".

Starting in the spring of this year, I've been talking with Caspar Oesterheld about using decision markets to tackle "MIRI-style" decision-theory examples. Caspar was previously investigating decision markets consisting of humans, as an alternative to prediction markets (to solve some problems with applying prediction markets). Applying decision markets as a machine learning technique (the market consisting of "experts" in the ML sense) is very appealing to me because it seems to side-step many of the problems my previous attempts at reflectively consistent decision theory had faced. This research has not yet produced significant results, but I continue to think it is a promising direction.

I have also made some attempts to communicate my update against UDT and toward learning-theoretic approaches, including this write-up. I talked to Daniel Kokotajlo about it, and he wrote The Commitment Races Problem, which I think captures a good chunk of it.

Meanwhile, Vanessa has been working on learning-theoretic approaches for much longer (and posted a research agenda about it last year). I've recently been talking to her about her ideas more. Both of us agree that many "hard" problems become "easy" when you make this switch in perspectives. We continue to have some disagreement about just how much becomes easy; Vanessa thinks essentially all the MIRI-relevant decision theory problems go away, except for multiplayer coordination problems, while I think certain problems such as counterlogical mugging remain open but have promising avenues of attack.

In addition, I have been thinking about partial agency. This is a change in perspective for me in that I am entertaining the idea that "agentic" behavior isn't a natural kind, but rather a (potentially mistaken) idealization of a class of related phenomena in which optimization-like behavior emerges without necessarily becoming full "optimization". This direction also relates strongly to the learning-theoretic update.

For the next year, I am hopeful that there will be some basic results about decision markets and other learning-theoretic approaches. In the best case, this would include results for bargaining and solving coordination problems, although "single-player" results are more likely. (But, coordination results might be higher impact, so I want to try for them.)

I also expect to think much more about learning-theory ideas for alignment more generally (ie, outside of decision theory). Previously I have been skeptical of machine-learning approaches to value-loading, not because I had a better idea, but because my concept of "machine learning" was mostly restricted to the concrete examples I saw. I now feel that the space of possible approaches is quite broad, and I'm optimistic that there are interesting things to say. Specifically, reinforcement learning and inverse reinforcement learning have fairly limited notions of "feedback" for the agents. I'm interested in investigating much richer notions of feedback. This of course relates to Vanessa's research agenda, and also to some extent to Stuart Armstrong's.

1. This is an informal snapshot by MIRI Researcher Abram Demski of some of the research directions he's been working on recently, and expects to work on going into 2020. Note that this reflection wasn't chosen to be representative of MIRI researchers' views or work in general.