CS 194-129 Deep Neural Networks
Spring 2018 Critch Lecture 24

1 Lecture

- **Bounded rationality:** computer making decision with limited resources, has to stop thinking at some point so can't be perfectly rational.
- **Open-source game theory:** players in game can read each others' minds or source code.
- **Negotiable RL:** how humans can negotiate for control of a powerful RL system.
- In general, we would like to align AI with human interests while also reducing existential risk.
- Humans are the dominant species on the planet because of our intelligence.
- According to a survey by AI Impacts, researchers believe there is a 50% chance of AI outperforming humans in all tasks in 45 years. Soon we may no longer be the "smartest" species on the planet.
- But what if we're wrong about the time? Humans aren't always so good at predicting progress.
- We should *already* be thinking about the "what if." How can we reduce existential risk?
- **Prepotent AI:** machine intelligence that could potentially dominate all humans in a competition for control of physical resources ("has potential to take over the world").
  - *Technological autonomy*, *replication speed*, and *social acumen* are AI capabilities that could lead to prepotence.
  - We don't necessarily need AGI for a system to pose an existential risk.
- Human existence is fragile, and will end by default if a new control mechanism begins to govern the state of Earth. We depend on many properties of Earth (temperature, oxygen concentration, chlorine gas levels...). There are many ways humans could just not exist.
  - If we allow something else to control these variables, it could very easily result in our extinction.
  - For example, we might not want to give machines the power to cut down all the trees.
- **Reward hacking:** do strange and unprecedented things to maximize reward. It's difficult to define reward functions that fully specify what we want! "There are more ways to climb a gradient than dreamt of in your philosophy."
- Even without malice from an AI, there are basic sub-goals that might be useful for achieving an objective but disastrous to humans if pursued fully by a super-intelligent optimizer.
- *There are a lot of good reasons to think about AI risk management.* See slides.
- **Pareto optimization:** simultaneous optimization of multiple objective functions.
- To secure long-run stability in the face of major changes in the world's control mechanisms, we may ultimately need *proofs and theorems* about our ability and inability to predict "computed variables."