Recent posts

Paul Christiano Follow
Nov 29, 2016

Here's what I've been thinking/writing about AI control over the last month.

Strategy

- Prosaic AI control argues that AI control research should first consider the case where AI involves no "unknown unknowns."
- Handling destructive technology tries to explain the upside of AI control, if we live in a universe where we eventually need to build a singleton anyway.
- Hard-core subproblems explains a concept I find helpful for organizing research.

Building blocks of ALBA

- Security amplification and reliability amplification are complements to capability amplification. Ensembling for reliability is now implemented in ALBA on github.
- Meta-execution is my current leading contender for security and capability amplification. It's totally unclear how well it can work (some relevant speculation).
- Thoughts on reward engineering discusses a bunch of prosaic but important issues when designing reward functions.

Terminology and concepts

- Clarifying the distinction between safety, control and alignment.
- Benignity may be a useful invariant when designing aligned AI.