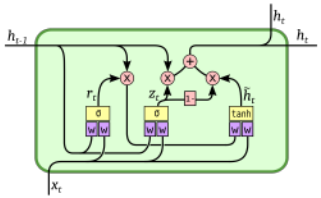
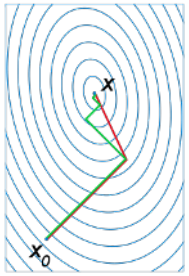
Scalable AI control

Paul Christiano

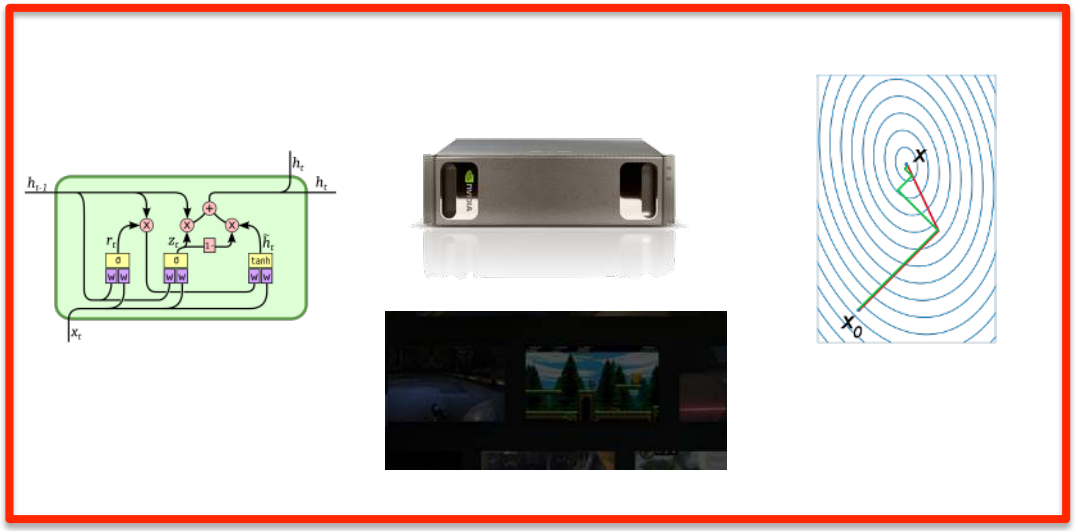
UC Berkeley

Ingredients

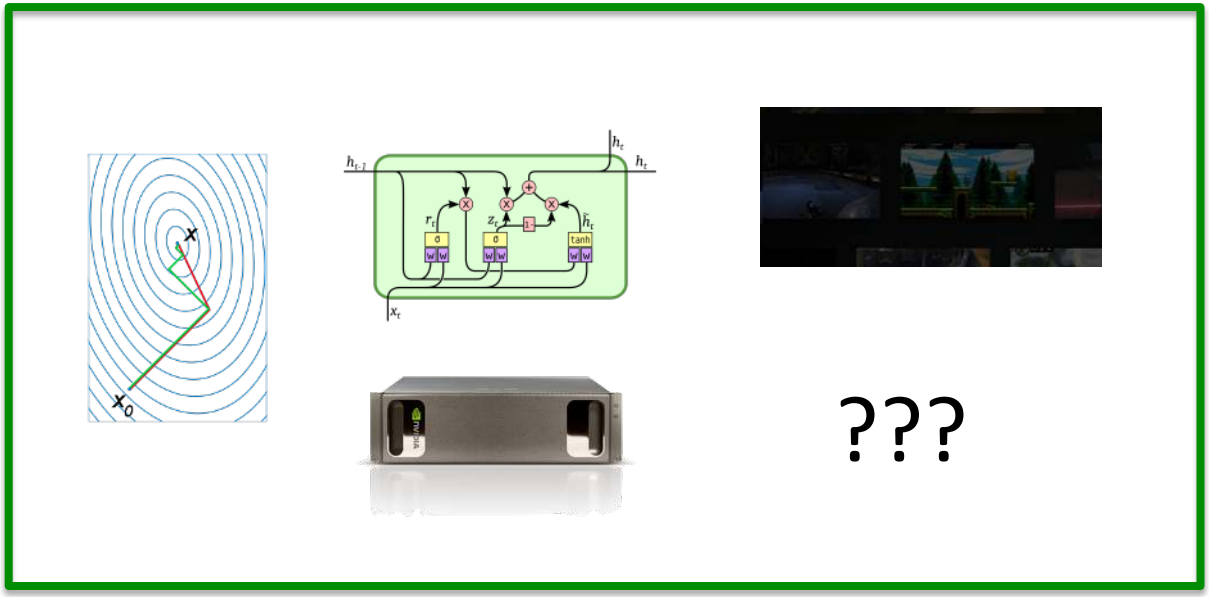




Benchmark AI



Aligned analog



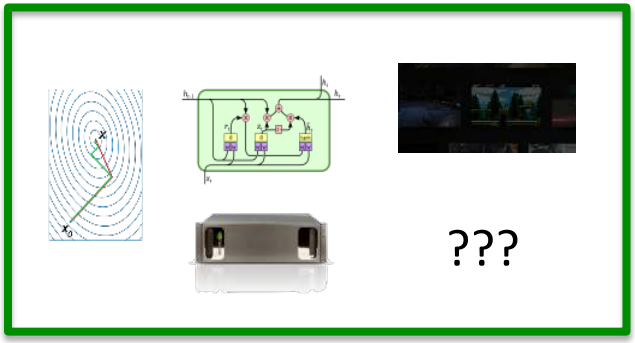
???

Goals

- Aligned analog is aligned
- Aligned analog is nearly as efficient as the benchmark
- Continues to hold as the ingredients improve
- Note: different problem for each possible AI

Benchmark AI

Aligned analog



Motivation

- We want to produce aligned versions of whatever AI systems actually work
- If we can, everyone is happy*
- If we can't…
  - There is an incentive to "skimp" on alignment
  - Unaligned AI will grow in influence
  - We face a legal/political problem; difficulty depends on how close we got

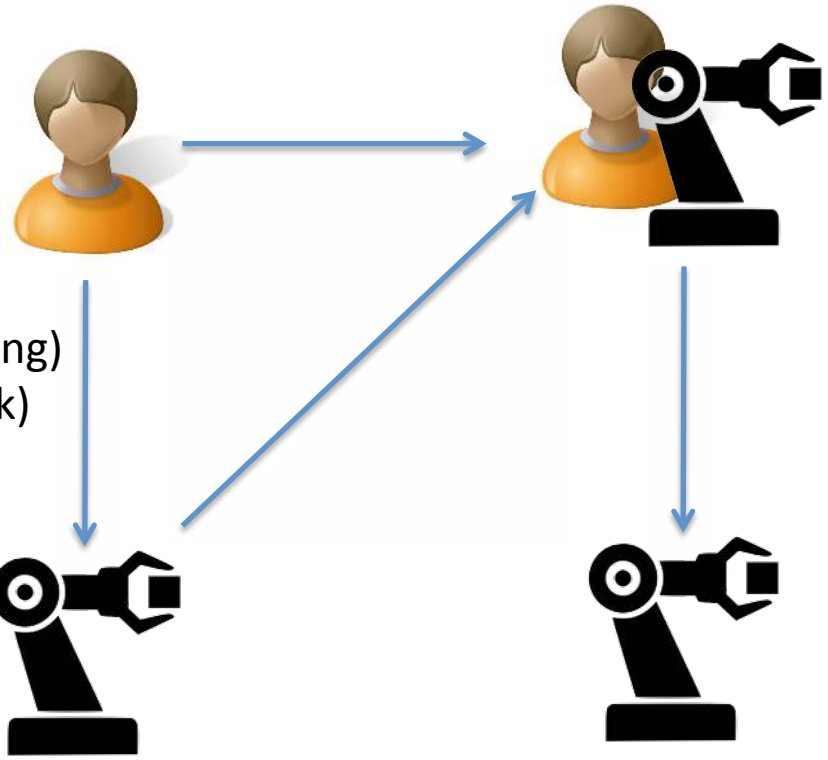
Bootstrapping reward functions for RL

(aligned analog of model-free RL)

key challenge: produce reward function

github.com/paulfchristiano/alba
from capabilities import RL
def hopefully_aligned_agent(expert):
    …

Bootstrapping reward functions for RL



(Imitation learning)
(Expert feedback)
(IRL)

Step 1: Expert trains weak RL agent

Step 2: agent+expert work together and think longer to find better decisions

Step 3 … N: repeat

Many problems to solve

- Expert must train weaker agent
- Robust learning
- Semi-supervised learning
- Human+agent+time must outperform the agent, while remaining aligned.
- (More work to extend to anything beyond model-free RL.)