Amplification
Tradeoffs between capability and alignment

Low-capability high-alignment:

- TAMER
- Imitation learning

High-capability low-alignment

- Sophisticated IRL
- RL with hard-coded reward function
AlphaGo Zero

Game playing/tree search?

Learned model p

Amplification procedure MCTS

Distillation: training p
Amplification: can improve performance without affecting alignment

Distillation: reliably copies the actions of another system

Make A10: limited tool that can copy humans given specific instructions

H+A10: human can use many copies of the tool to do different tasks -> this overall system is now beyond human level but still aligned

A11 = Distill(H+A10)

A2 = Distill(H+A11) … etc
Assumptions/requirements

Robust imitation or similar

Amplification preserves alignment

Possible to use amplification productively