Alignment Game Tree

- What if we … use an unsupervised model to learn human values as they exist in the world. Then, train another model on that reward signal.
- – How do we know we've trained this properly?
- – Does not solve inner alignment by itself
- – If we had a guaranteed good unsupervised model, would the more dangerous trained model learn correctly? Would it generalise to superhuman intellect?
- Map out high level features of the neural net (such as goals and internal representations) using applied interpretability research. Then prune any models which have bad goals
    -
- Rigorously define agency using math. Then, prove theorems about neural nets becoming agents using deep learning theory.
    - Figure out all the ways agency can possibly happen and how neural networks approximate ideal agents and how they scale.
        - Does defining agency define deception?
            - Definition of agency will reveal ideas about deception. We don't know exactly what the agent's planning algorithm is but we can recognize its goals.
    - BREAKER: What if goals are not human interpretable? What if we think a goal is aligned but it's not?
- Use a strawb-aligned AI with sharp bounds on the impact of its behaviour. Bound impact using: a 5min timer and limited resources (myopia)
    - Breaker: Agent delegates to another 5min agent (endless chain) or an unbounded agent.
        - Successor agent needs to be bound as well. OR; ban successor agents from being created. Requires a rigorous specification of a successor agent.
        - Box the agent so it truly has bounded compute
        - Agent truly values doing the task in only 5min; any subagents score worse on the objective function.

- BREAKER: we need to rigorously define “cares about doing it in 5min” which is just as hard as aligning the AI in the first place?

-
- What if the AI operates on a bunch of habits and heuristics more so than a utility function? (I.e, shard theory) Then we could create a “human values shard”.
  - BREAKER: How do we know the AI has developed a human values shard?
    - Build architecture that always converges to human values? Hard.
    - Take representation of human values from an LLM or unsupervised model?
    - BREAKER: How do we know the human values shard will “win”? How do we reinforce it properly?
- A group of smart humans can understand an ai slightly smarter than an individual. Iterated distillation and amplification until we get very smart AI which is understandable.
  - Breaker: How do we know that smarter AI remains aligned? What if there’s a sharp left turn (roughly fast takeoff) where capabilities generalise but alignment doesn’t?
    - Under a sharp left turn, that’s a genuine breaker.
    - If we have slow takeoff, a group of Int=n agents can supervise an Int=n+1 agents. So by induction we can supervise arbitrarily intelligent agents.
  - Breaker: AI drifts its values subtly at each step so there’s no step where we catch the AI being misaligned.
    - There may be a basin of corrigibility where agents will not want to shift away from human values.
      - Breaker: How do we know that this basin exists? Why don’t agents just want to be aligned with the immediately supervising agents?
- Energy/effort penalty
  - How do we operationalize this?