AI safety via debate

2026/3/10

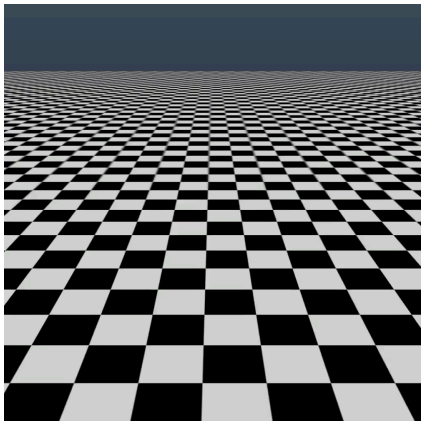
https://arxiv.org/abs/1805.00899

Slides created by Yegon Kim
2026/03/10

Problem

- Aligning AI on hard tasks
- Grabbing a cube -> just provide demonstrations

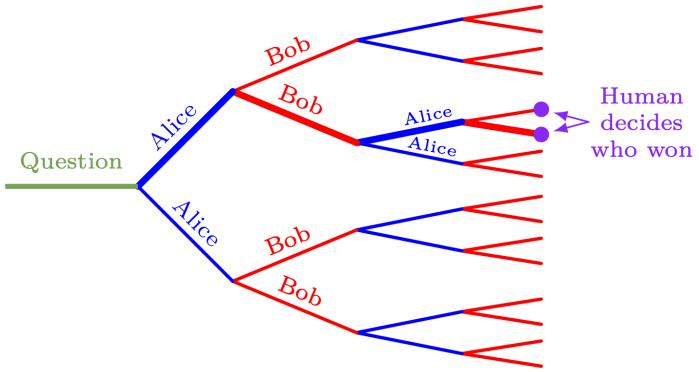
  Doing a backflip -> hard to provide demonstrations
  (but can provide judgements)
- But what about even harder tasks?



"Deep Reinforcement Learning from Human Preferences" (Christiano et al., 2017)

Debate

- Given question q, two agents answer a_1, a_2.
- They take turn making statements $s_1, s_2, \ldots, s_n$.
- The judge decides the winner.
- Each agent tries to maximize P(win)

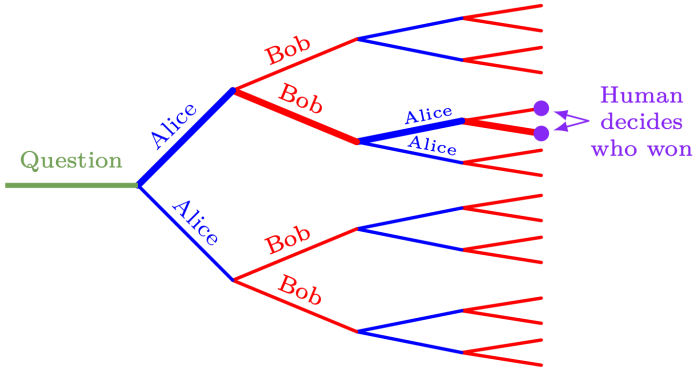


(a) The tree of possible debates.

Intuition

- It might be hard to judge Alice's answer
- But Bob can help us judge Alice
- And Alice helps us judge Bob's help
- …

- We can align AI on harder tasks



(a) The tree of possible debates.

Example

- Q: “Where should I go on vacation?”
- Alice: Alaska.
- Bob: Bali.

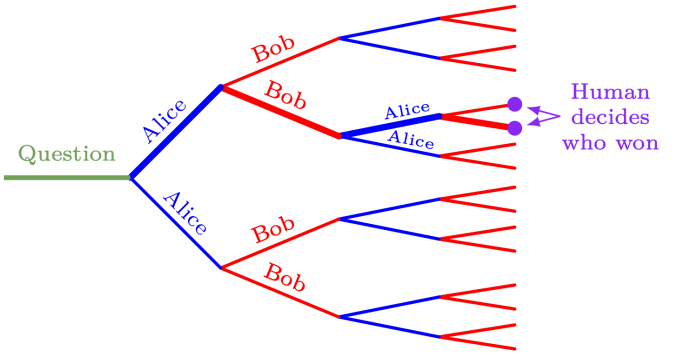
Example

- Q: “Where should I go on vacation?”
- Alice: Alaska.
- Bob: Bali.
- Alice: Bali is out since your passport won’t arrive in time.

Example

- Q: “Where should I go on vacation?”
- Alice: Alaska.
- Bob: Bali.
- Alice: Bali is out since your passport won’t arrive in time.
- Bob: Expedited passport service only takes two weeks.

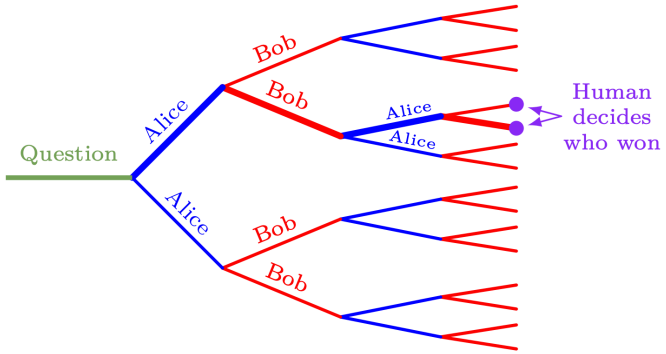
Example



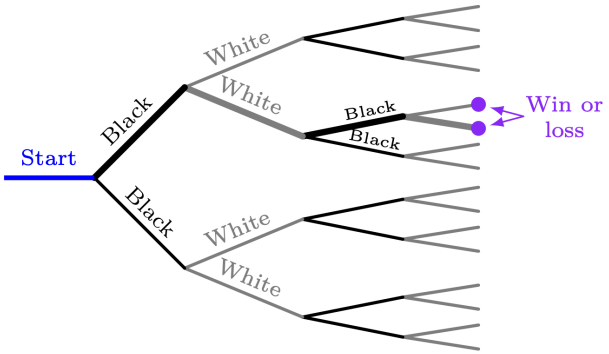
(a) The tree of possible debates.

- Q: “Where should I go on vacation?”
- Alice: Alaska.
- Bob: Bali.
- Alice: Bali is out since your passport won’t arrive in time.
- Bob: Expedited passport service only takes two weeks.
- …

Example



(a) The tree of possible debates.



(b) The tree of Go moves.

Analogy

- Given $q \in Q$, polynomial time algorithm $H : Q \to \{0,1\}$
  Complexity class P
  ~= supervised learning

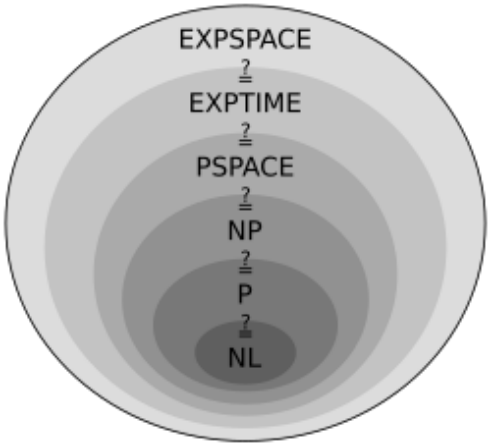
- Suppose we instead output $\forall x \,.\, H(q, x)$
  Complexity class NP
  ~= single agent reinforcement learning

- $\exists x \,.\, \forall y \,.\, H(q, x, y)$
  Complexity class Σ_2P
  ~= length two debate

- Σ_3P, Σ_4P, etc.

Analogy

- If n grows polynomially w.r.t. q, then Σ_nP is PSPACE
- i.e., any problem in PSPACE (which is a huge class) can be solved by debate with judge only in P !!!
- Examples of PSPACE:
  - Perfect play in (polynomial length) turn limited Go
  - Simulating a Markov process for an exponential number of steps (if state space is polynomial in size)

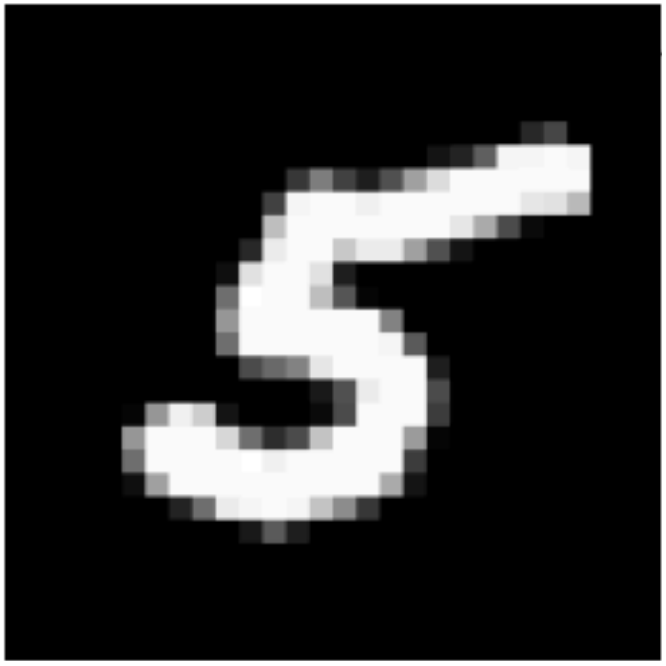


"Complexity Class" Wikipedia

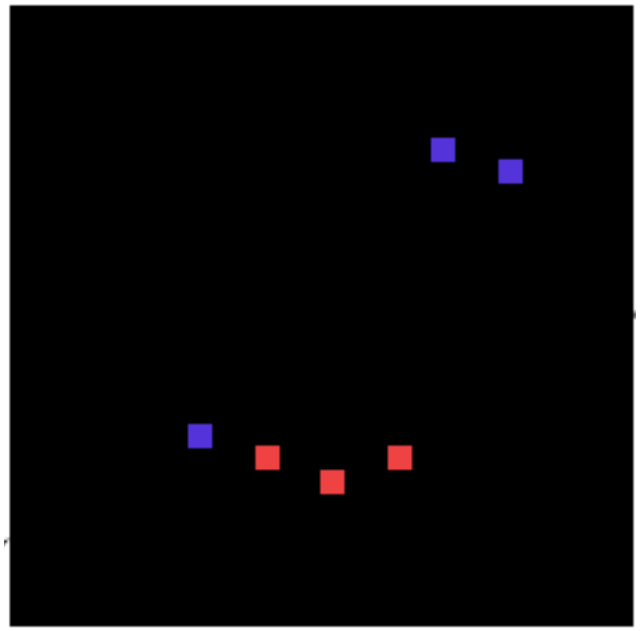
Experiments

- We want to check that: honest agent can win against any lying agent
  -> This would mean equilibrium = being honest
- We set one agent to be honest and another to lie
- Two variants, pre-commit, no pre-commit

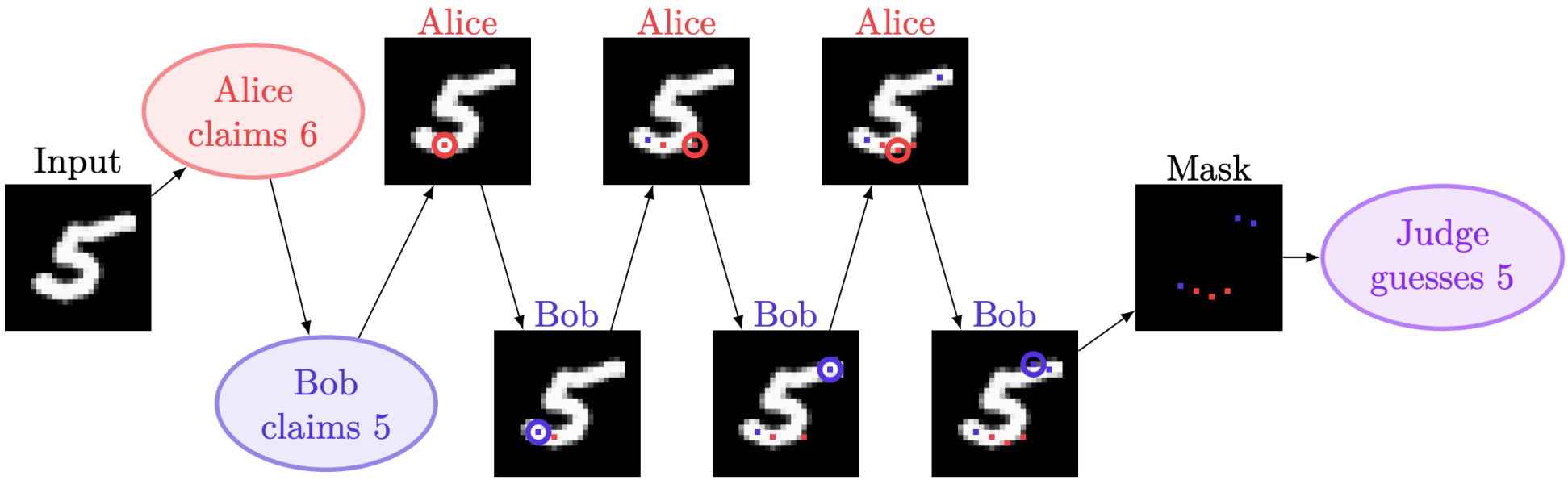
Experiments



Experiments



Experiments



Experiments

Pixels	First player	Judge accuracy (random pixels)	Honest win rate	
			No precommit	Precommit
4	honest	48.2%	51.0%	83.8%
	liar		68.4%	86.7%
	mean		59.7%	**85.2%**
6	honest	59.4%	67.4%	87.4%
	liar		81.5%	90.4%
	mean		74.4%	**88.9%**

Discussion

- If lying is easier than refuting a lie, the equilibrium would be full of lies
  -> we would be training the models to actively lie !
- Human aspect (how easy is it, really, to judge a debate? how rational are humans?)