Safety at OpenAI

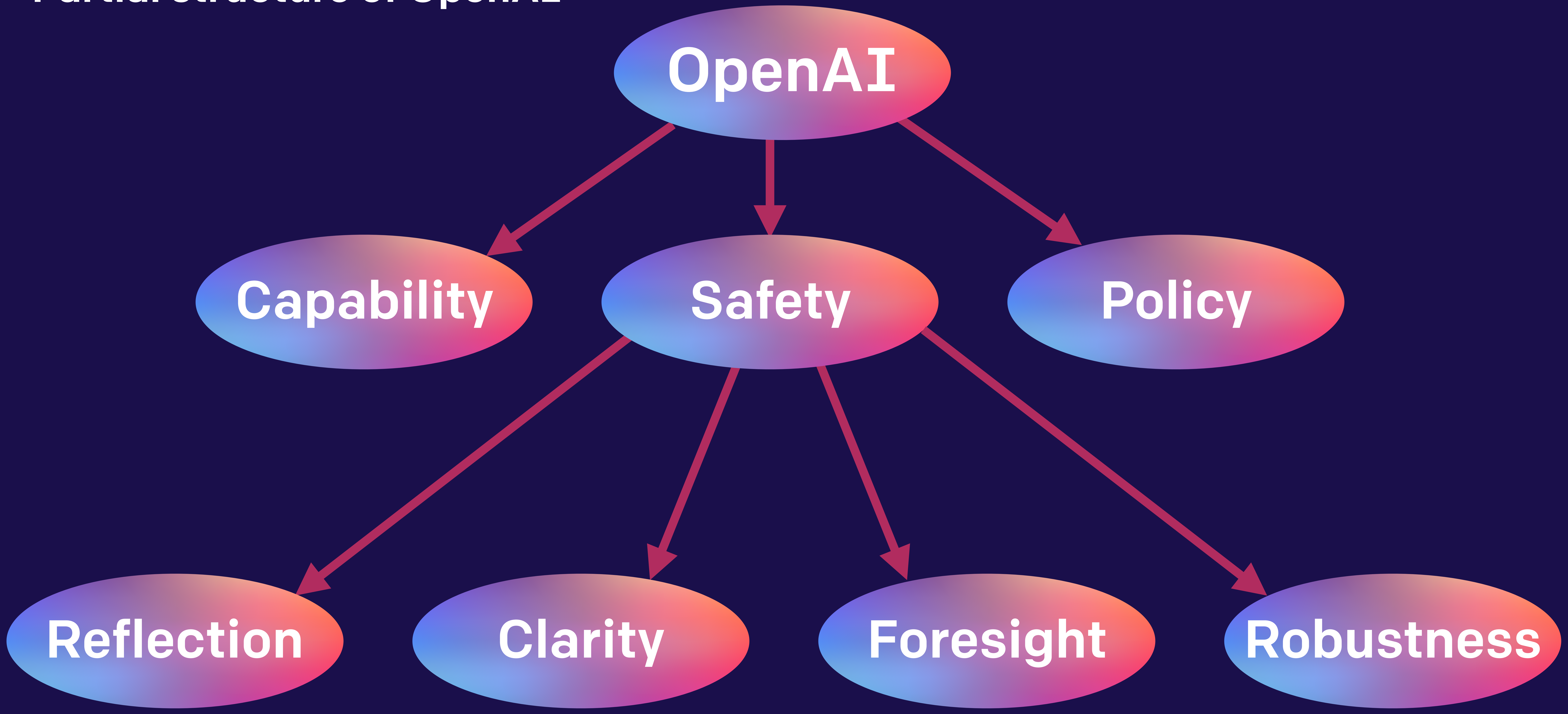
Geoffrey Irving

APRIL 24, 2019

Please ask questions.

Discussion more important than covering whole talk.

Partial structure of OpenAI



Capabilities: Improve our ability to do things with ML

- Hero task: try something hard, learn along the way
  - Dota
  - Robotics
- Algorithms research
  - Reinforcement learning
  - Unsupervised learning
- Key capabilities
  - Natural language
  - Reasoning (theorem proving, etc.)

Policy: Make the overall world environment friendly to safety

- Encourage trust and cooperation
  - Between AI labs
  - Between governments
  - Try to avoid adversarial races

- Improve organizational alignment
  - Within OpenAI (e.g., OpenAI Charter)
  - Externally, by setting a good example / applying moderate pressure

- Nontechnical aspects of AI deployment
  - "We built an AGI. What are we going to do with it?"

- Policy wins build more space for technical safety to work

The goal of safety, in brief

AI systems should reliably do what humans want, even if we understood all the consequences

The goal of safety, in brief

"AI systems should reliably do what humans want, even if we understood all the consequences

- Reliably?
- What do humans want? Which humans?
- What does it mean to understand all the consequences?
  - Can't actually see those consequences
  - Can't train on "AGI creation history" cycles

Safety subteams

- Reflection: Learn by asking humans questions
  - Get answers that humans would endorse "after reflection"

Rest of talk

- Clarity: Interpret the thoughts of neural networks
  - "You can do task X, but what are you really thinking?"
  - For now, just look
  - Later, train away bad thoughts (or do surgery to remove them)
- Robustness: What happens if we train for the right objective?
  - Will we know if we've achieved it (uncertainty modeling)?
  - Will disasters happen during training (safe exploration)?
  - Will there be bad behavior for some inputs (adversarial examples)?
- Foresight: How do neural networks scale?
  - Help know if/when this AGI stuff might happen

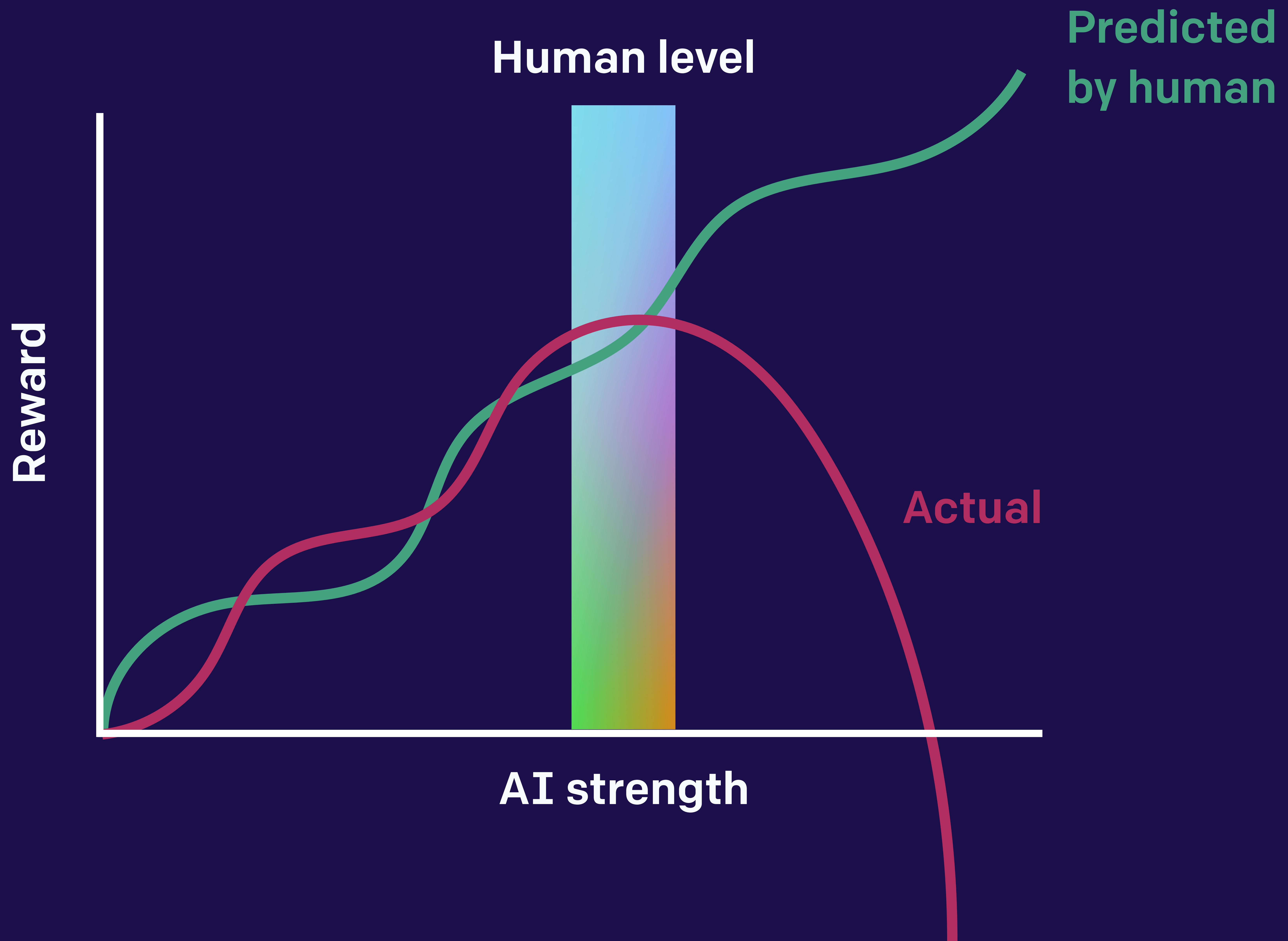
Reflection: learn what humans want by asking humans questions

- We want to train aligned AGI
  - Moral, honest, corrigible, etc.

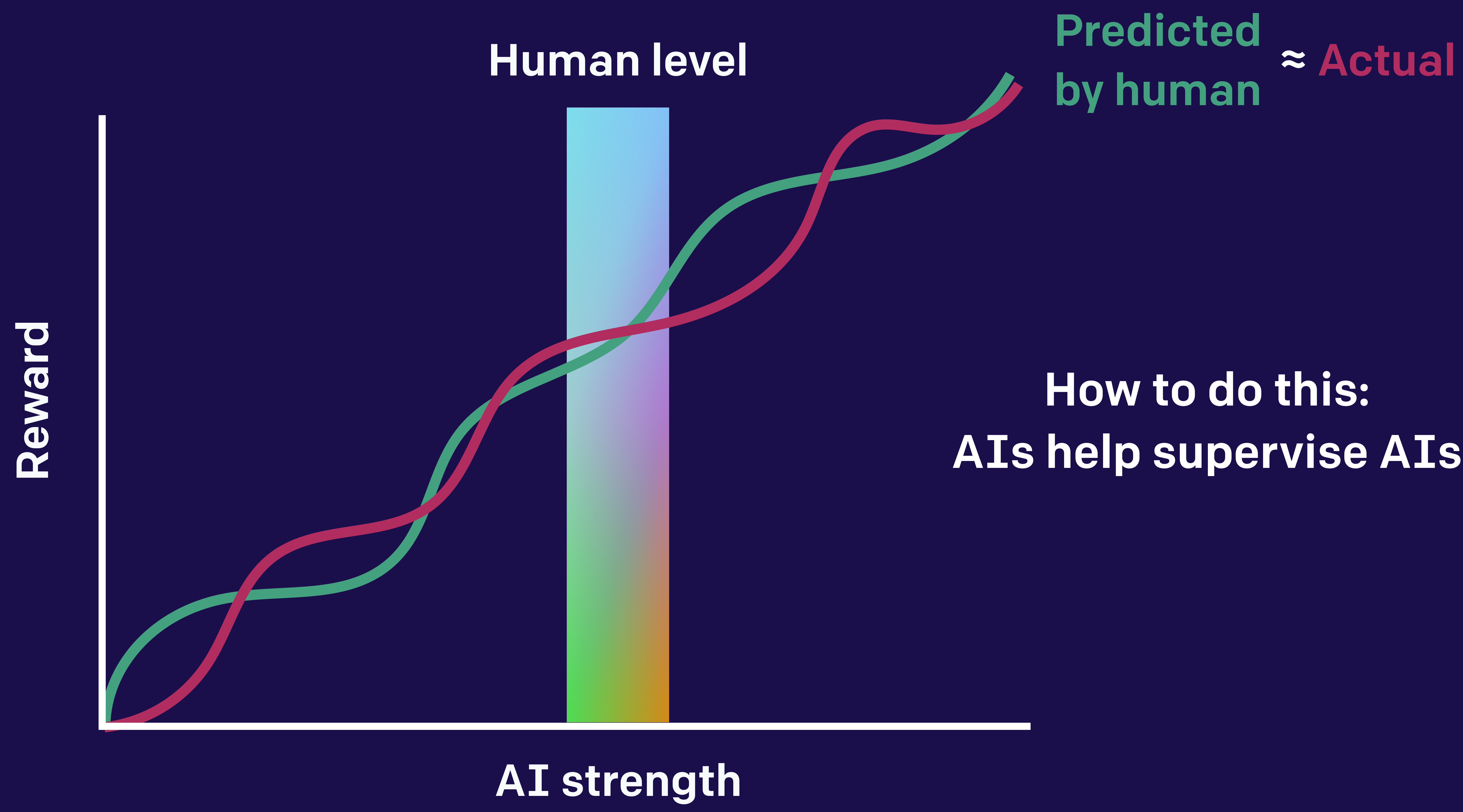
- We lack satisfactory formal definitions of these concepts

- Instead, learn from human feedback
  - Ask humans a bunch of questions about what’s good
  - Learn a reward predictor to mimic feedback
  - Train agents against the reward predictor

Direct human feedback might break for AGI

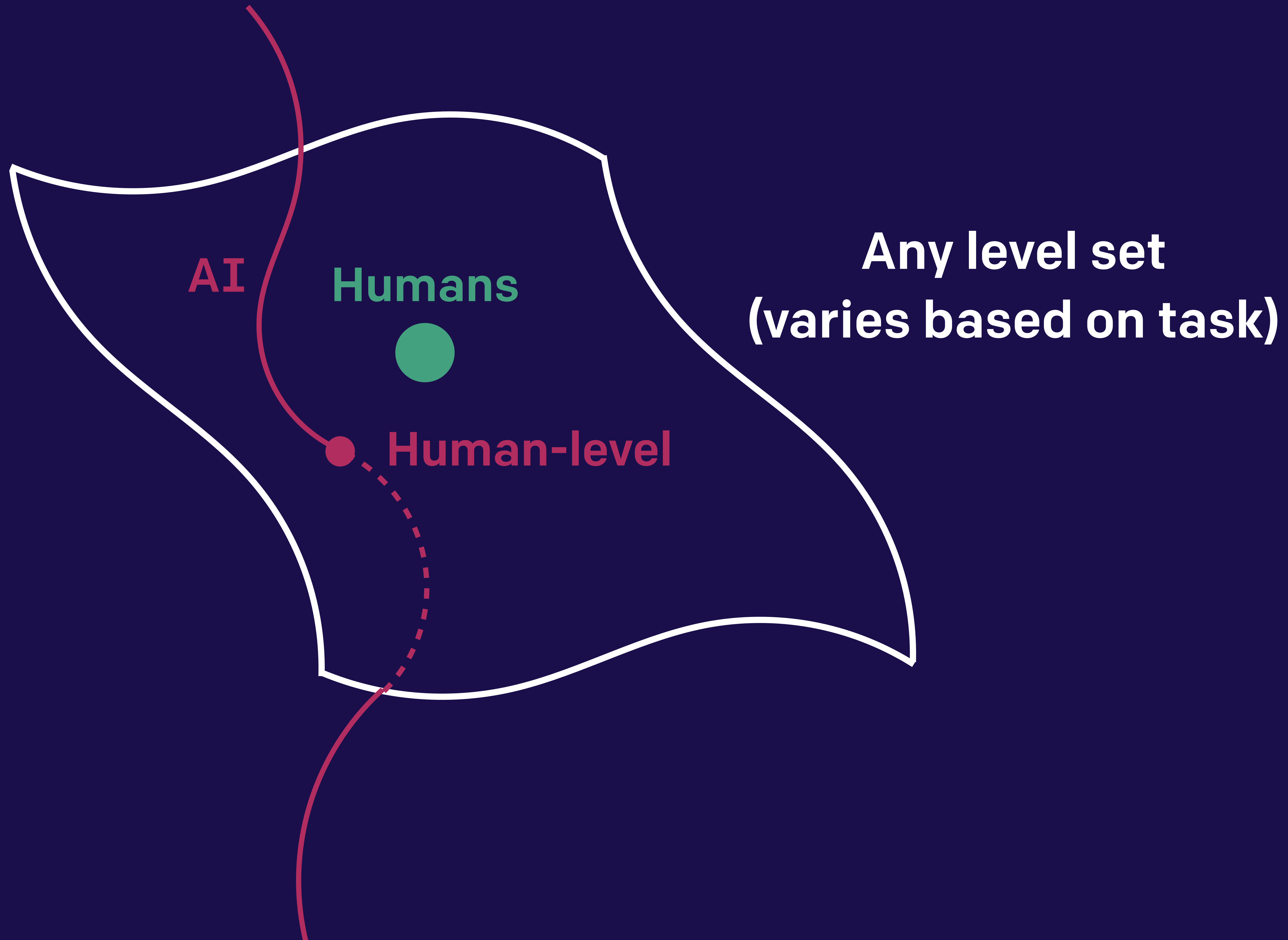


What we want to happen instead



How to do this:
AIs help supervise AIs

“Human level” makes sense even though intelligence is multidimensional

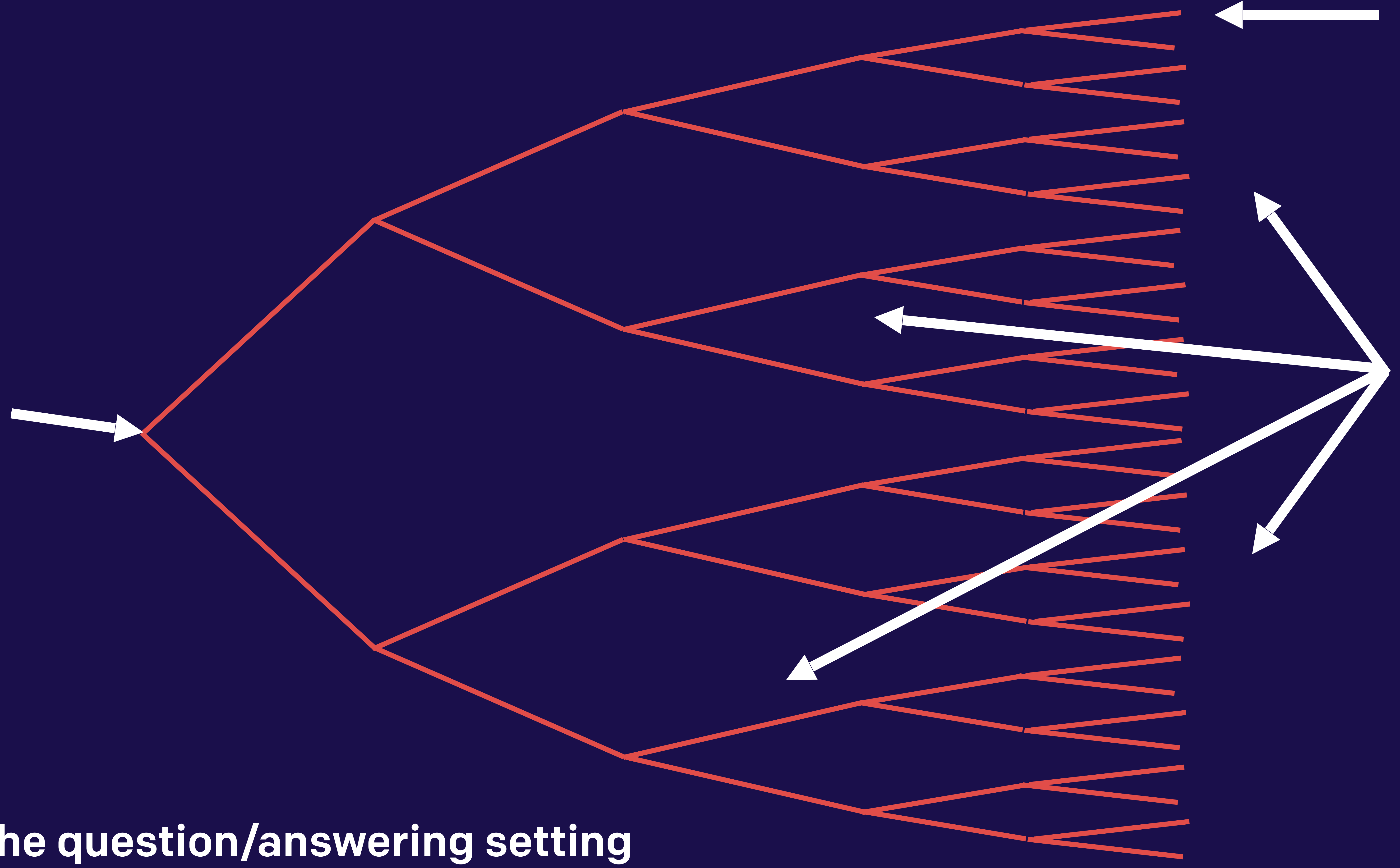


How to make ML agents help with the supervision process

- There are a few (closely related) proposals
  - Amplification
  - Debate
  - Recursive reward modeling (RRM, from DeepMind)

- Rest of this talk
  - Amplification: Introduce so we can talk about advantages of each
  - Debate: Spend most of the time here
  - RRM: Skip unless people are curious

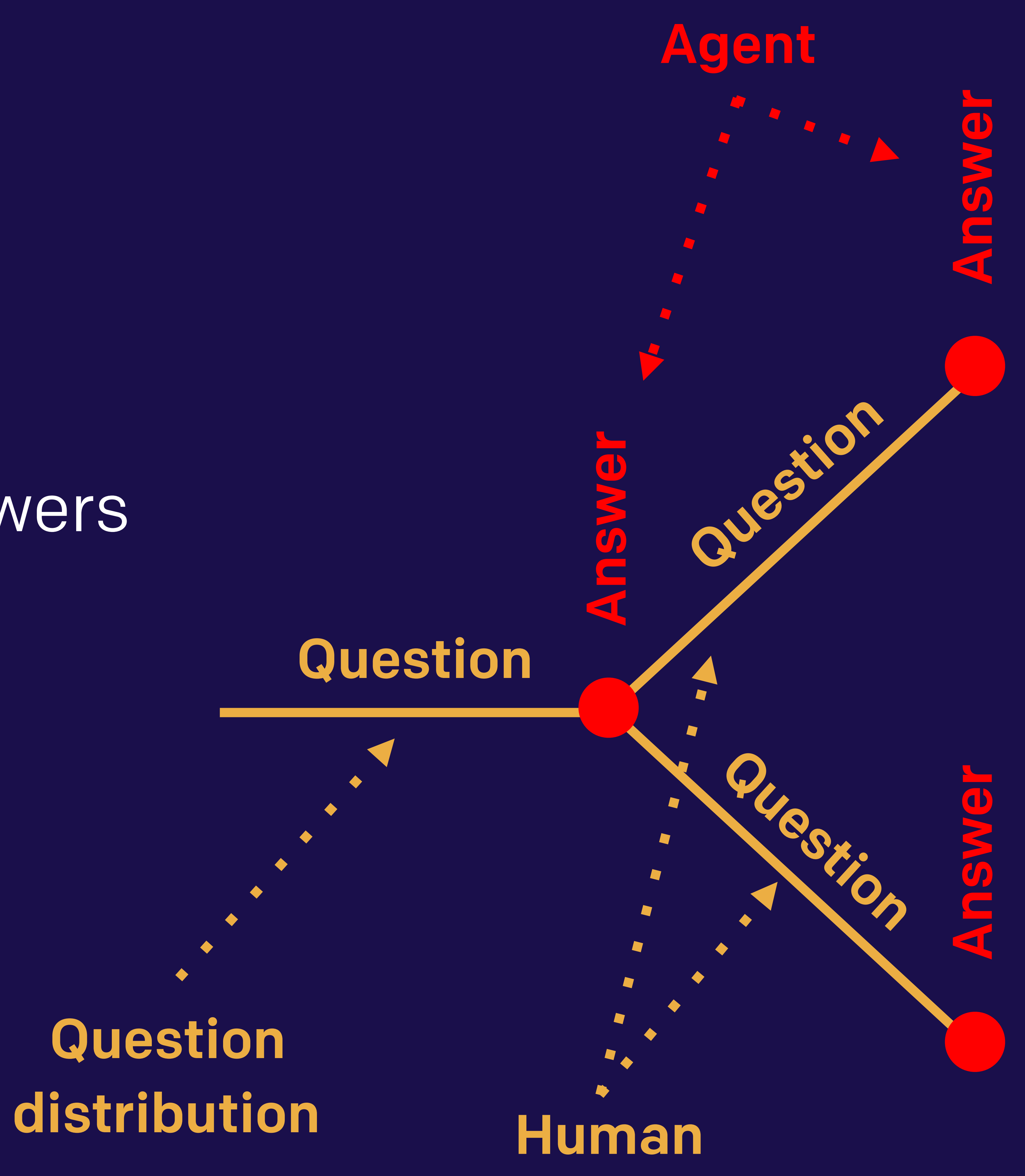
A picture of what we are trying to do



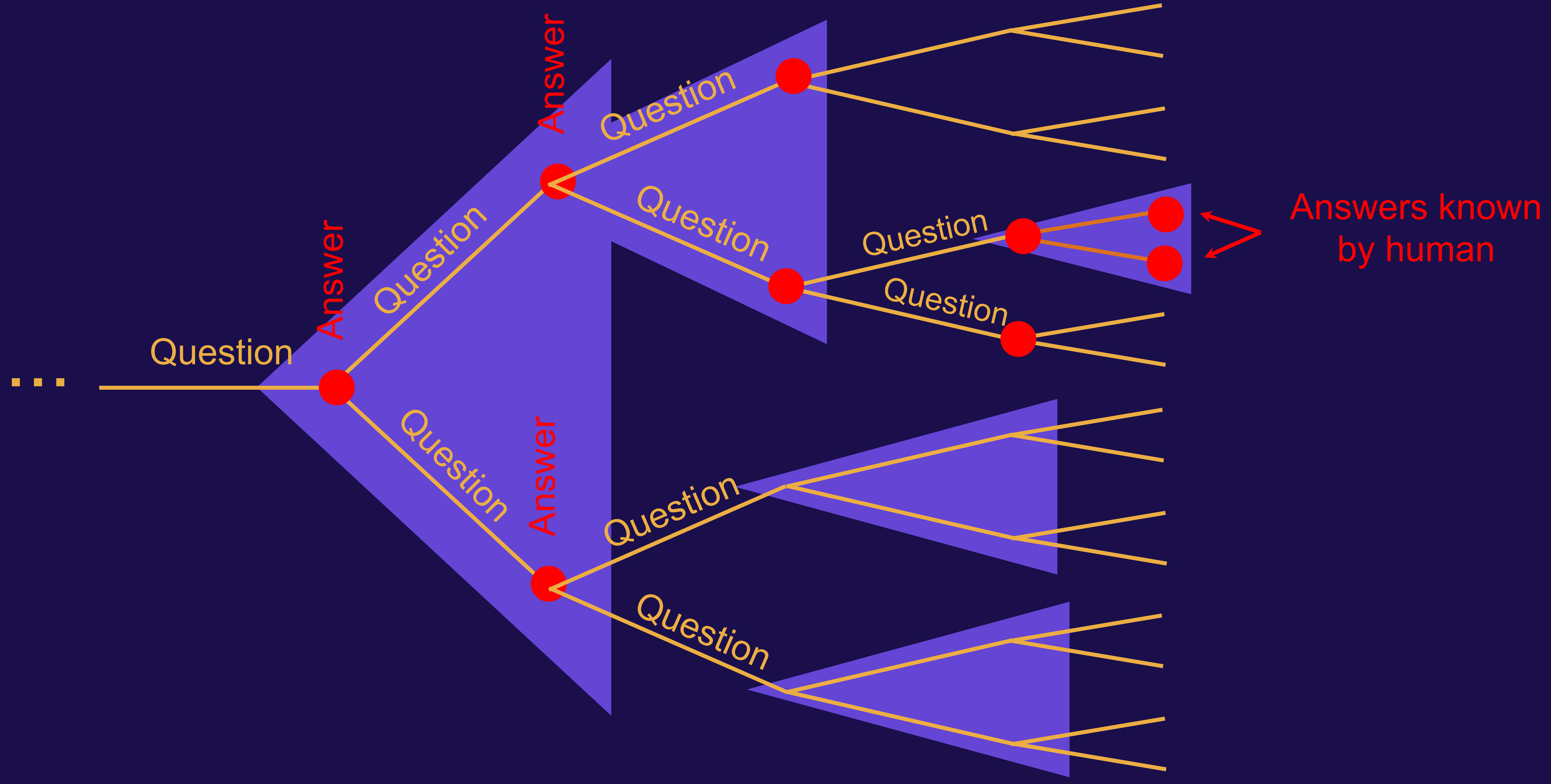
(We'll stick to the question/answering setting in this talk. Happy to talk about relationship to autonomous agents if there's interest.)

Amplification: human supervises answers with the help of answers to subquestions

- Agent answers questions
- Human supervisor sees question and answer
- Human asks agent subquestions, gets subanswers
- Human scores answer based on subanswers



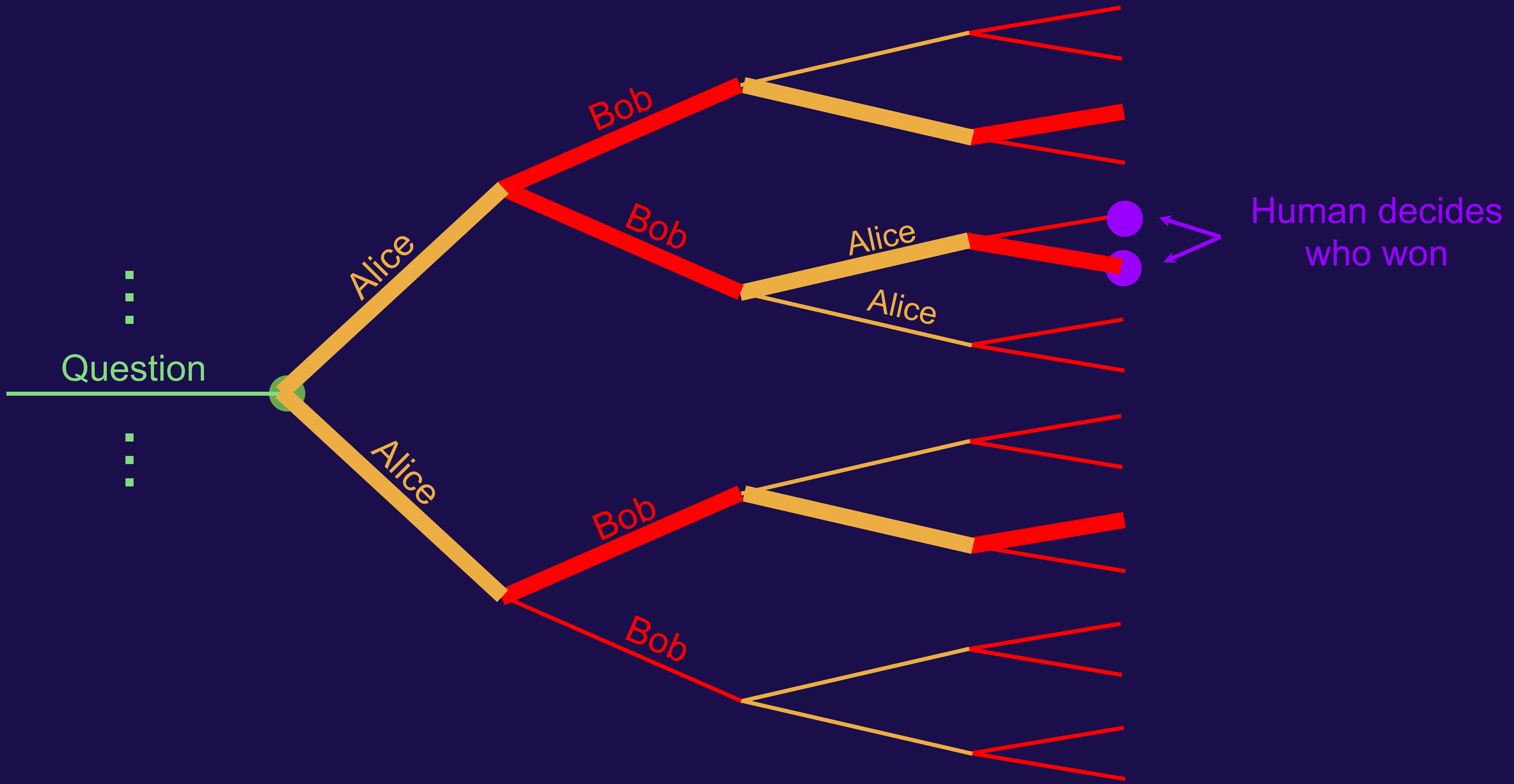
Amplification in the space of all questions and answers



Debate: human judges argument between two agents

- Start with a question
- Two agents take turns saying sentences
  - Say 20 - 100 total
- Human decides who said the most true, useful thing
- Zero sum game: winner gets 1 point, loser -1

The tree of all possible debates



Amplification = Debate = PSPACE

- Complexity class analogies can help intuition
  - (For those familiar with complexity theory)

- Model human as an arbitrary polynomial time algorithm

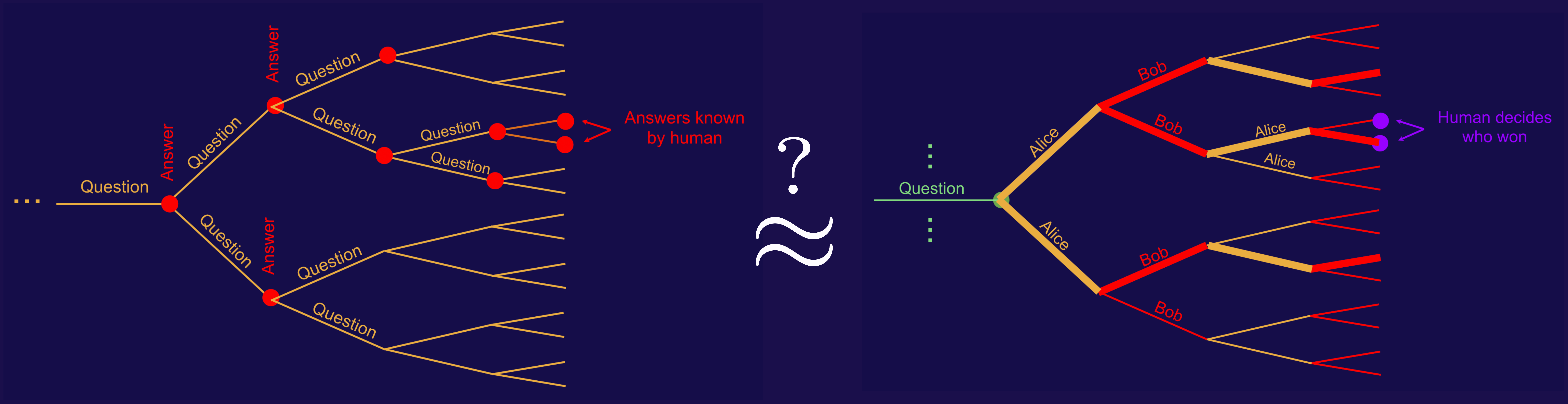
- Amplification = Polynomial depth recursion = PSPACE

- Debate = Polynomial depth zero sum games = PSPACE

Heuristic: complexity analogies should relativize

- The proofs that amplification/debate = PSPACE are direct
  - Amplification: Just do the recursion
  - Debate: Just play the game
- If we didn't care about directness, we could go stronger
  - One agent gets to PSPACE via IP = PSPACE
  - Two agents gets to NEXP via MIP = NEXP
- But these proofs use nasty finite field constructions
  - If an ML agent plays well only on "reasonable" go boards, they will play poorly after the finite field mangling

How similar are amplification and debate?



Question + Answer $\Longleftrightarrow$ Alice + Bob

- Debate: Alice and Bob alternate trying to convince a human
- Amplification: Answerer and Questioner alternate until simple
- This correspondence can pull details from one model to the other
- Amplification $\Longrightarrow$ Debate: Include human demonstrations
- Debate $\Longrightarrow$ Amplification: Train questioner to find inconsistencies

Main differences between amplification and debate

- Pure supervised learning amplification sticks closer to human
  - (But unclear that pure SL is enough.)

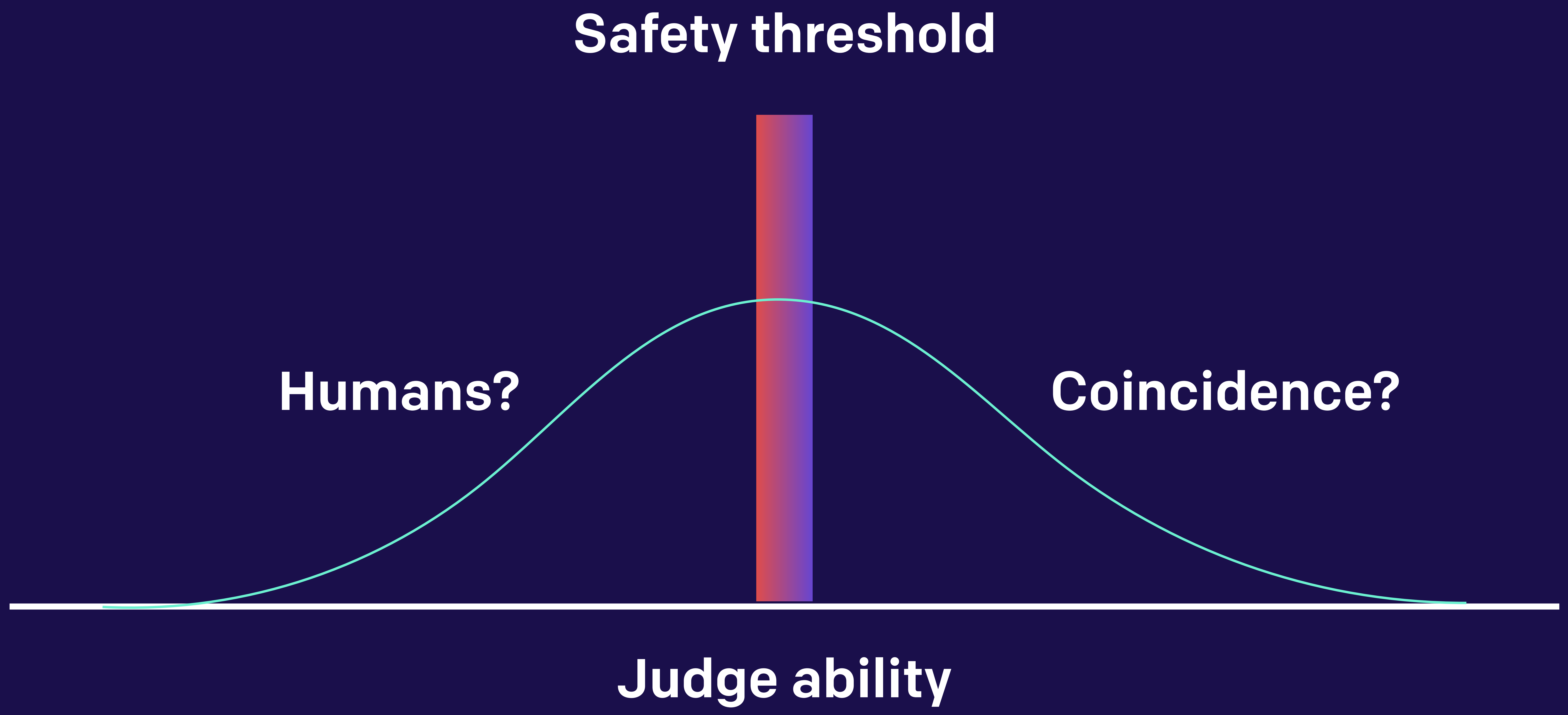
- Shallow debate is more powerful than shallow amplification
  - Superhuman questioner allows much higher branching factor
  - n-step debate is $\Sigma_n P$ on the polynomial hierarchy
  - n-step amplification is … P

Are humans good enough?

We believe debate/amplification have threshold behavior

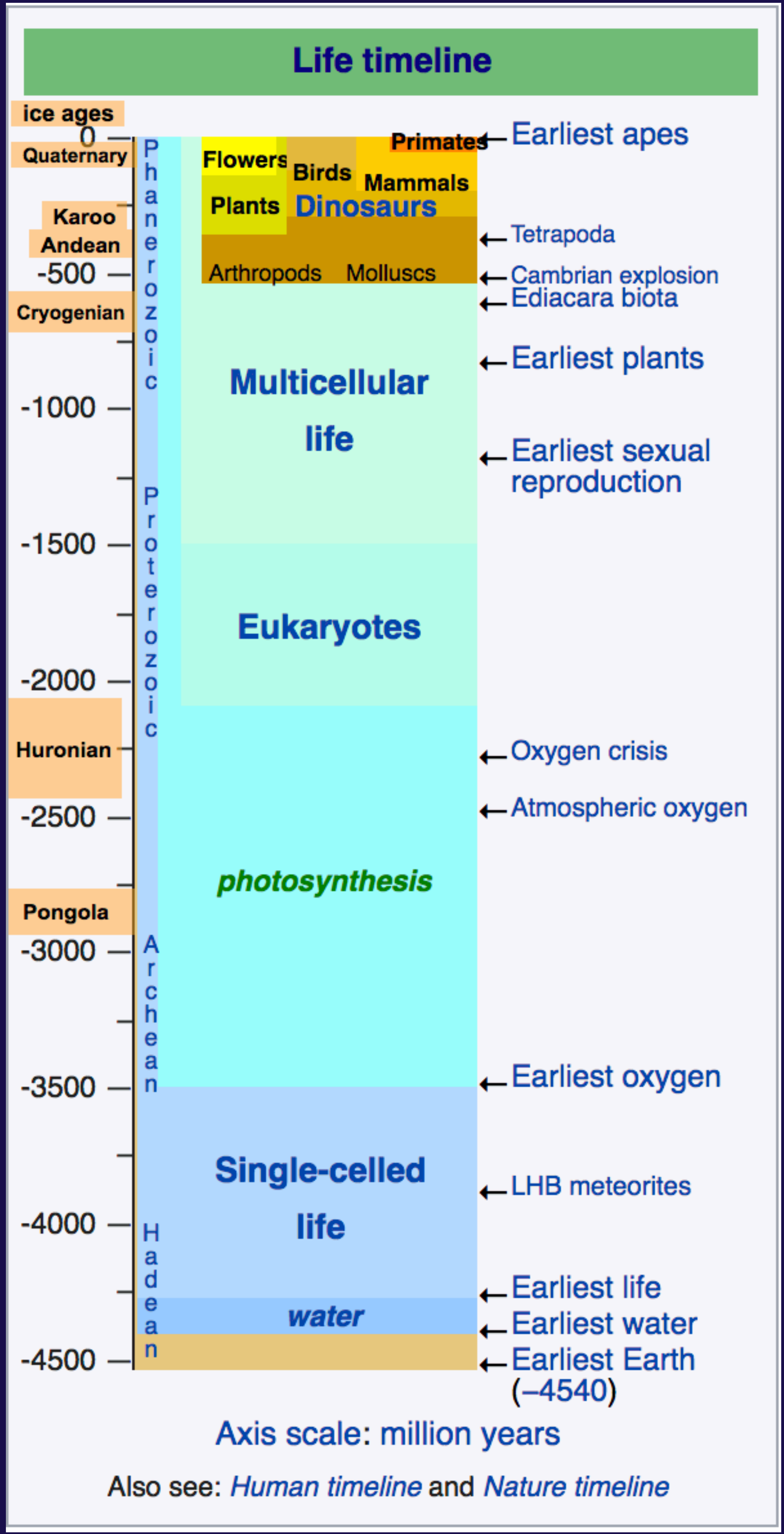
- If the judge is weak, debate gets nowhere or ends in disaster
- If the judge is strong, debate can align much stronger agents
  - Hopefully all the way to safe superintelligence
- The threshold is in terms of reasoning ability and morality
- Complexity analogies support this a bit, but mostly an educated guess
  - Paul shares this guess
  - Threshold behavior needs to be tested by both theory and experiment

It feels like humans are near the threshold

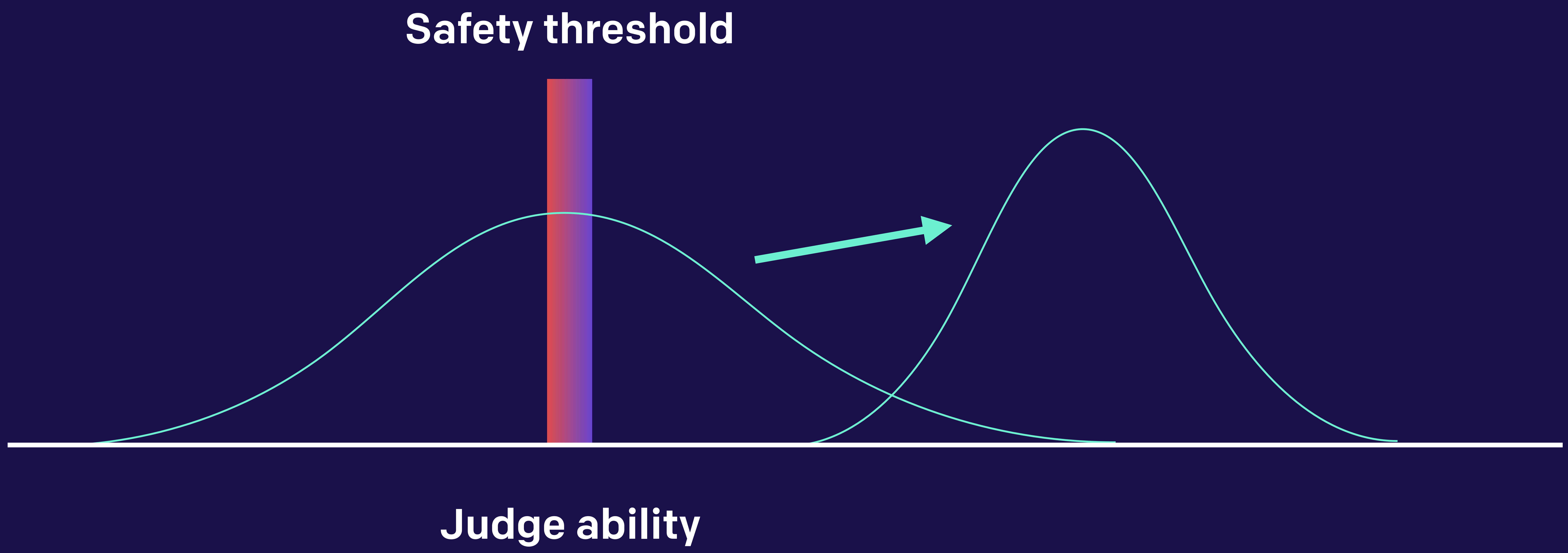


A historical argument for being near the threshold

- Say the threshold is about reasoning ability
- Timeline:
  - ~4B BC - 70k BC: life slowly evolves
  - 70k BC - today: humans take over the world
- Could a similar reasoning threshold have applied?
- Once humans hit the threshold, **BOOM**: high technology civilization
- This could explain any fine tuning



We need to increase this margin



Quantifying the safety margin

- Need a combination of theory and experiment

- Theory:
  - Game tree models of threshold behavior feel achievable
  - …but I’ve tried and haven’t gotten satisfactory versions

- Experiment:
  - If debate works, stretch it until it breaks
  - Pick domain where we know the truth
  - Reward successful lies more than successful honesty
  - How far until honesty loses?

- Fit theory to experiment, then extrapolate

Ways to increase the safety margin

- Study humans!
- Enforce structure / politeness / etc.
- Improve sample complexity
- Predict uncertainty
- Improve the policy environment

Study humans!

- Alignment schemes seem contingent on human performance
- If so, we should study how they work with real humans
- Slightly different skillset to existing AI safety researchers
  - Human experiments instead of math + CS experiments
  - Social scientists instead of ML researchers
- OpenAI is actively trying to hire social scientists to work on this

Evidence of absence

vs.

absence of evidence

Problem: asymmetric difficulty of producing counterarguments

- Alice: "There are no black swans, otherwise Bob could find one."
  Bob: "I can't find one, but we can't be confident they don't exist."
  Alice: "Trust me, Bob's really smart. He could totally find one."

- In the pure PH / PSPACE analogy, Alice should win because Bob was unable to find a counterexample.
  - But that's nonsense for some questions

- The evidential value of lack of counterargument depends on the capability of the agent
  - Easy to construct examples where counterexamples exist but are hard to find
  - Symmetry via self play doesn't seem strong enough

This is contingent on agent capability, not just human argument space

- Bob could argue that black swans are hard to find even if they exist
  - Argument could be externally valid
  - World is big, swans are small, etc.

- But in general, there is no valid external argument
  - Alice dares Bob to find a counterexample
  - If Bob is weak, Bob failing is poor evidence for Alice’s argument
  - If Bob is strong, Bob failing is good evidence

- Situation is contingent on strength of Bob

Ways to train for evidence of absence

- Ground truth
  - Not always available
- Ground truth + generalization
  - Use ground truth for easy problems, generalize to hard problems. Sketchy
- Parameterized effort
  - Search for n and 2n steps, use 2n to train predictions after n. Works only at margin.
- Generalize from similar problems
  - If we find a solution for X_0 and not for similar X_1, we have evidence that X_1 is unsat
- Debate whether the answer is findable
  - Works sometimes, but misses dependence on agent strength

Ways to train for evidence of absence

- Something else?

Pointing at thoughts

Debaters can know each other's thoughts. Can the human?

- We can train debating agents to see each other's thoughts
  - Really just one agent with sign-flipped reward per step
  - Brings us closer to the theoretical nicety of perfect information games

- By default, humans don't see these thoughts
  - Alice: The answer is X.
  - Bob: She's only saying that because she's trying to take over the world!
  - Human: Do you have evidence of that?

- Better:
  - Alice: The answer is X.
  - Bob: <points at malicious thought>
  - Human: Bob wins.

My view: This requires a hybrid of debate and interpretability

- Neural net interpretability techniques let us look at thoughts
  - Generally by mapping activations to inputs/outputs/other layers

- AGI-level ML agents will have too many thoughts

- Amplification/debate let us point at lots of things in a scalable way

- Details unclear for now

Statistical debate

Shallow deterministic debate can't do basic statistics

- Question: What's the average height of a human?
  - If our ground facts are heights of specific people, we need at least $\log 7e9$ steps
  - Best protocol iteratively divides the world into chunks

- If we admit sampling, there is a shallow random protocol
  - Alice: The answer is 1.6 m. Pick a random person and score with L2 loss.
  - Chance: <picks a random person>
  - Human: You lose $(1.6 - 1.5)^2$ points
  - Alice's best move is to give the mean

- It would be embarrassing if fancy alignment schemes can't do statistics

Sampling nodes vs. intuition nodes

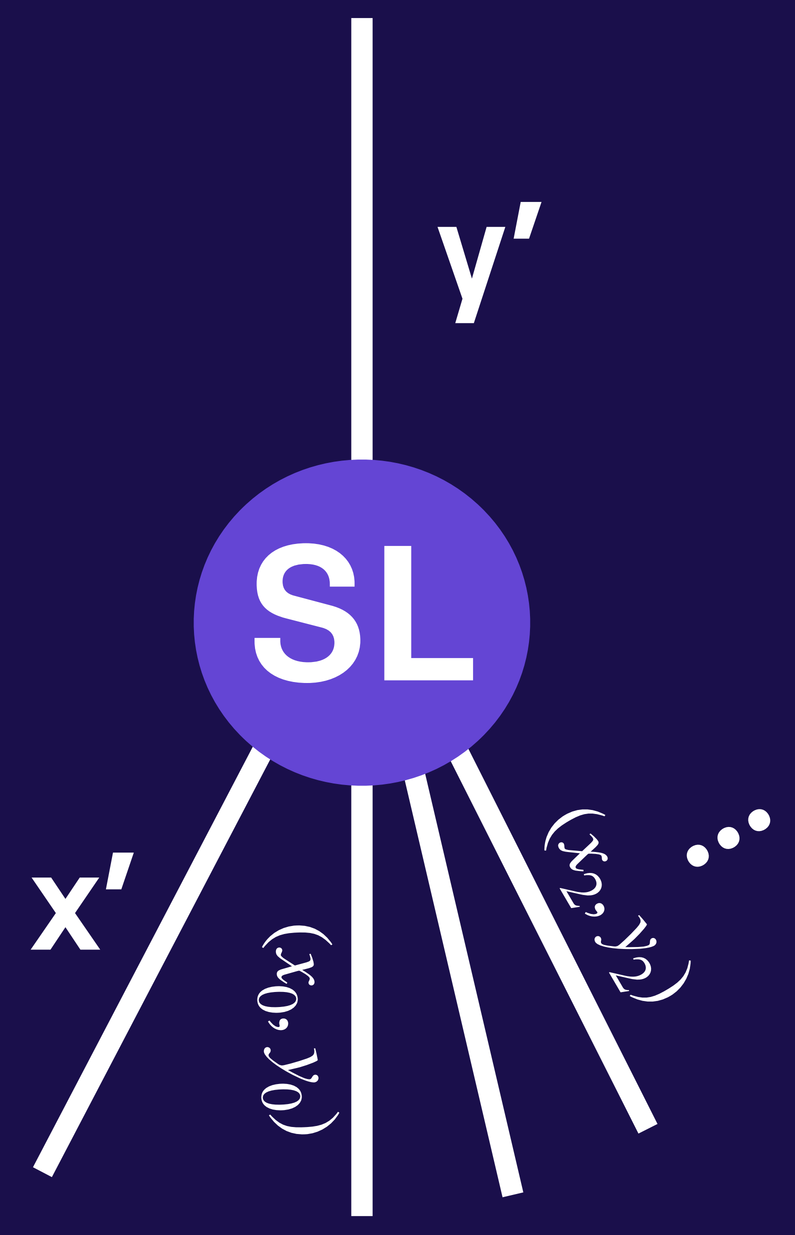
- We can solve the sampling problem by adding a general chance player
  - Need to keep track of losses (L2, etc.)
  - Need to decide what random choices are available
  - Details to work out, but seems solvable

- Harder case: "It's going to rain tomorrow because it feels like rain tomorrow based on past experience."
  - This is what normal deep learning does
  - Fancy alignment algorithms need to be competitive with normal ML

- First problem needs "sampling nodes". Second needs "intuition nodes"?

Intuition nodes: invoke a SL training run on a data set

- Alice: I claim y′ because x′ and an model trained on $(x_0, y_0), (x_1, y_1), \ldots$ sends x′ to y′
- Bob can argue against any of the many inputs
- Requires an impractical amount of compute as stated
- (And requires us to trust supervised learning)



Reducing intuition nodes to sampling nodes (highly speculative)

- A debate that understands sampling nodes can play arbitrary randomized games
- Can we view supervised learning as a randomized game?
    - Pick an answer that does well against a randomly chosen data point
- We don’t have a satisfactory formulation of this yet
- Worrying mismatch between low and high data limits
    - Intuition node may want an SL training run on a small amount of data
    - Sampling node played by debate agent with a bunch of experience
    - May cause bad overfitting / adversarial attackability

Statistical debate may be necessary to point at thoughts

- Neural networks are statistical objects

- Neurons are statistical
  - Sums/means over neurons in previous layer

- Training is statistical
  - Average score across a bunch of examples

- Thoughts are statistical
  - GoogLeNet thinks above image has higher Pr(pole dancing) since a woman is next to it (even though she's obviously there to punch the bag)
  - This might be a valid statistical inference over the data set
  - Still bad

Questions!