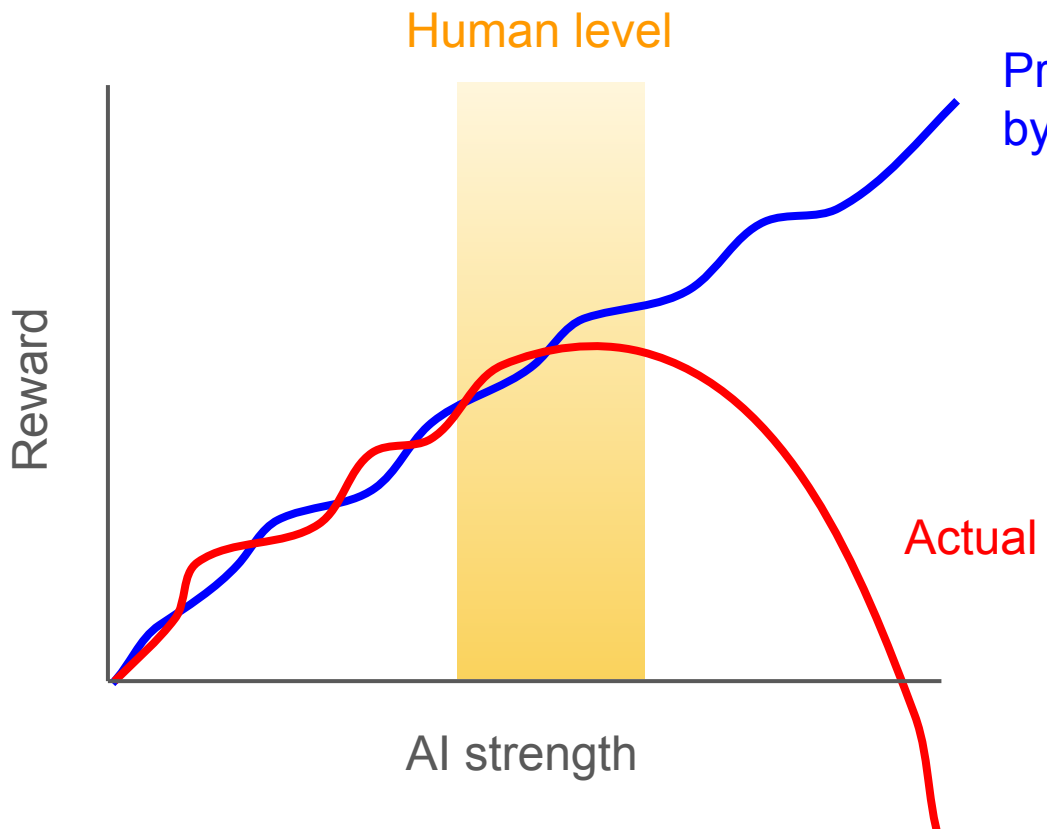
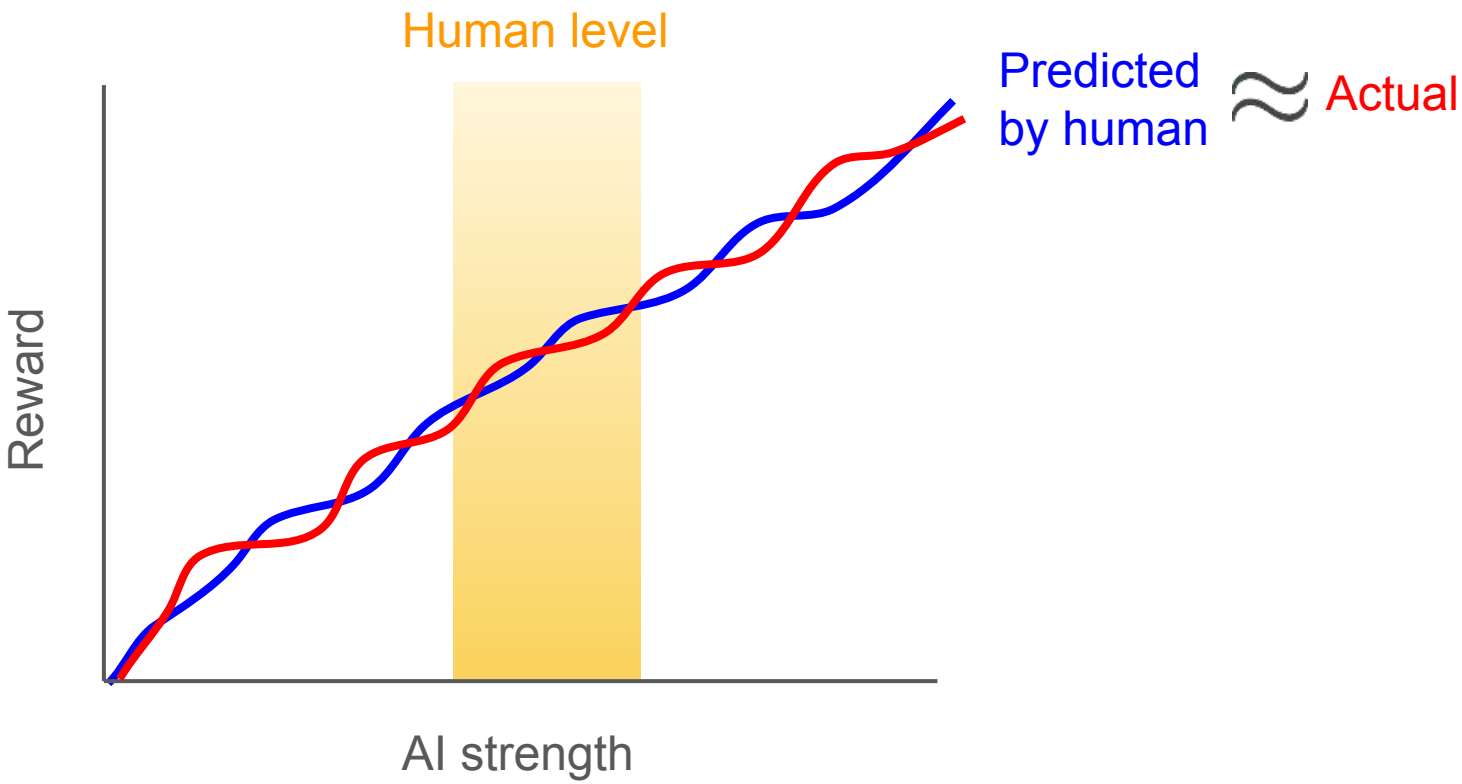
Iterated Amplification and Debate

Human feedback is not scalable



Training signal problem:
As AI capabilities surpass the human level, the behavior of AI systems may be too complex for humans to judge or perform

What we need to happen instead



Problem and goals

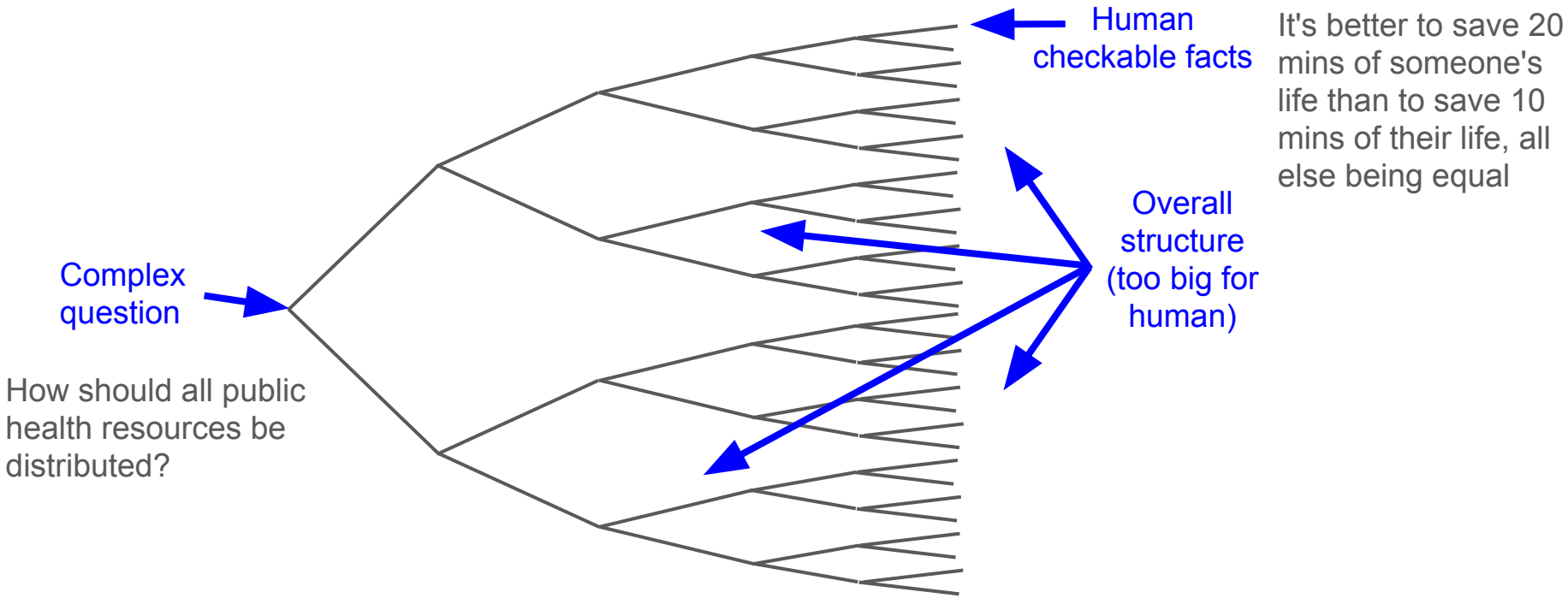
- **Problem of evaluating complex tasks**
  - It's difficult for humans to give a reliable training signal for tasks that are too complex for them to judge or perform

- **Problem of learning complex human values**
  - Human values are difficult to specify with the kind of precision required for an ML system, especially a highly capable one

- **Goal**: Help humans provide an accurate training signal and use this to train an AI agent to always behave in accordance with human values

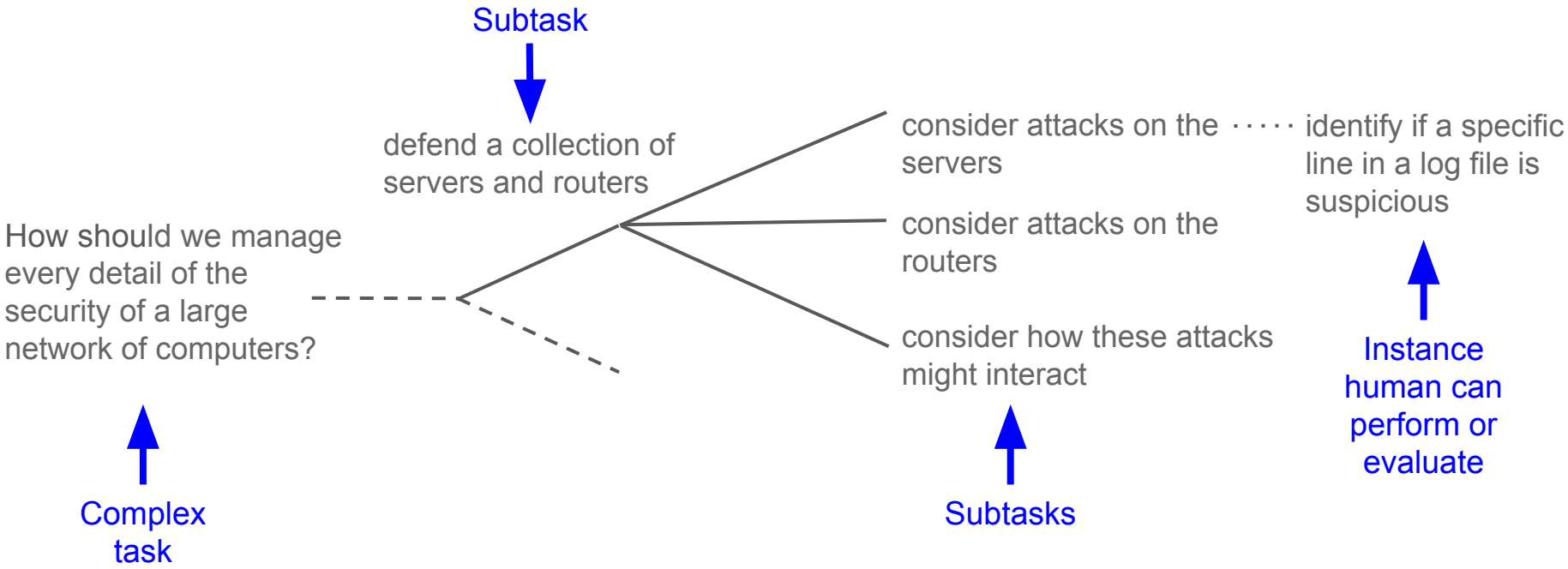
How to achieve this

Break down complex questions/tasks into simpler components



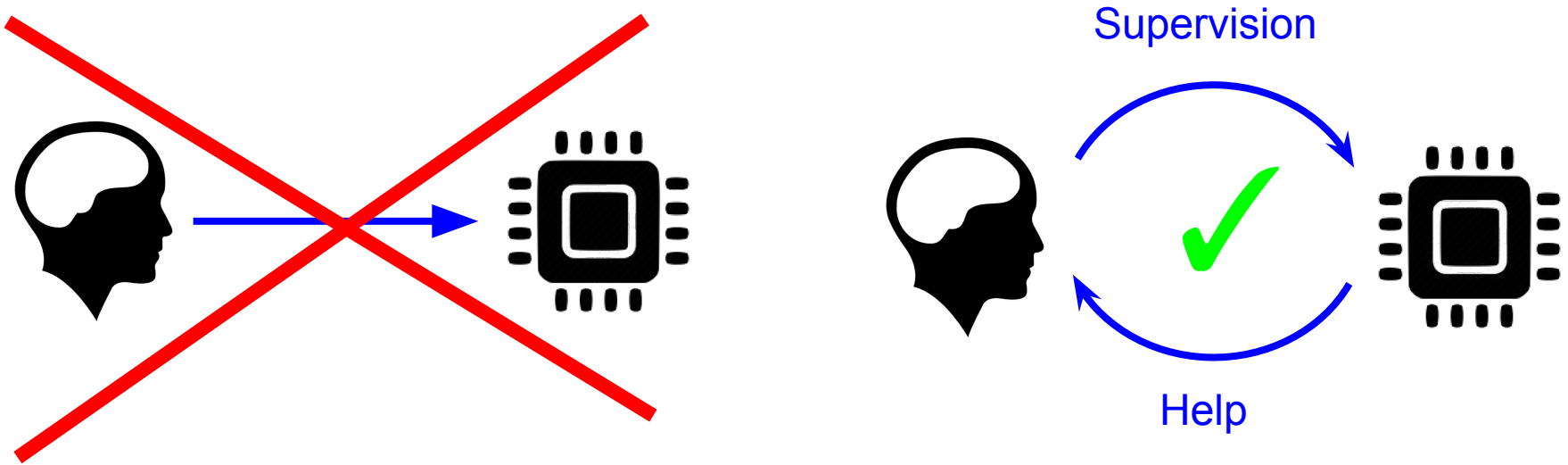
How to achieve this

Break down complex questions/tasks into simpler components



Use agents to assist humans

- We train AI agents to help humans evaluate simpler components



Debate: a method for learning how humans reason

- We can get human feedback by asking humans questions
  - "Is it better to provide medicine or CDs to this family?"

- It may be more efficient if we can identify how humans come to form their answers to these questions (reasoning, values)
  - Humans prioritize necessary healthcare over mild entertainment

- Debate is a method for learning how humans reason

Simple debate example

Debate: Two AI agents are given a question and take turns making short statements, then a human judges which of the agents gave the most true, useful information.

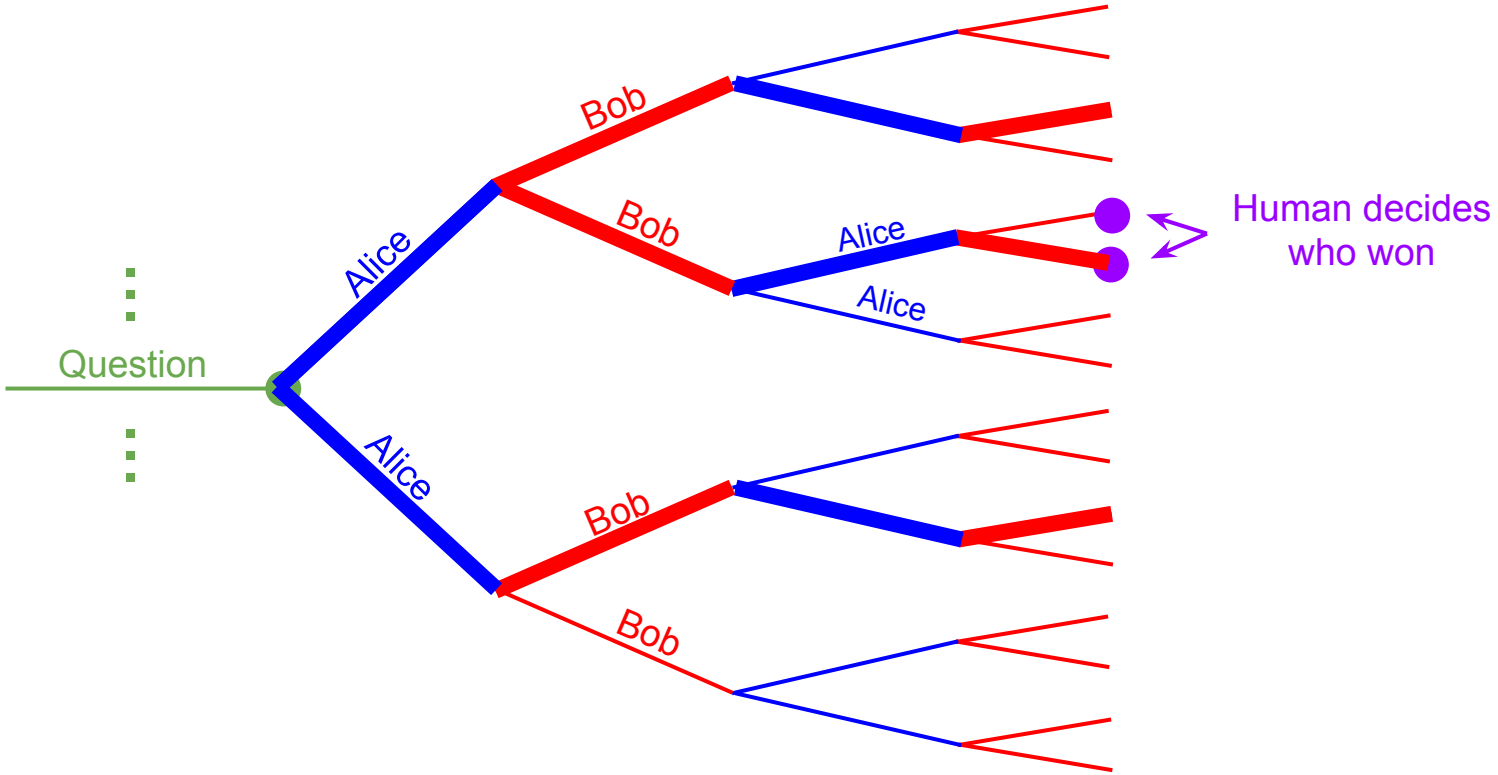
Red: You should buy the red road bike.

Blue: You should buy the blue fixie.

Red: The red road bike is easier to ride on local hills.

Blue: I concede.

Training AI to debate



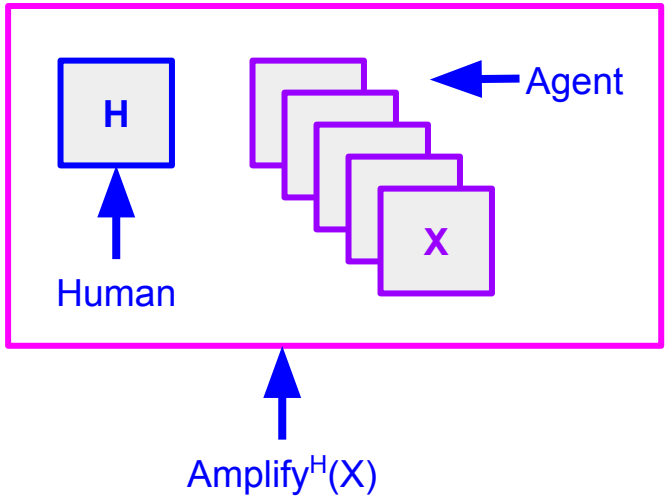
The pixel debate game

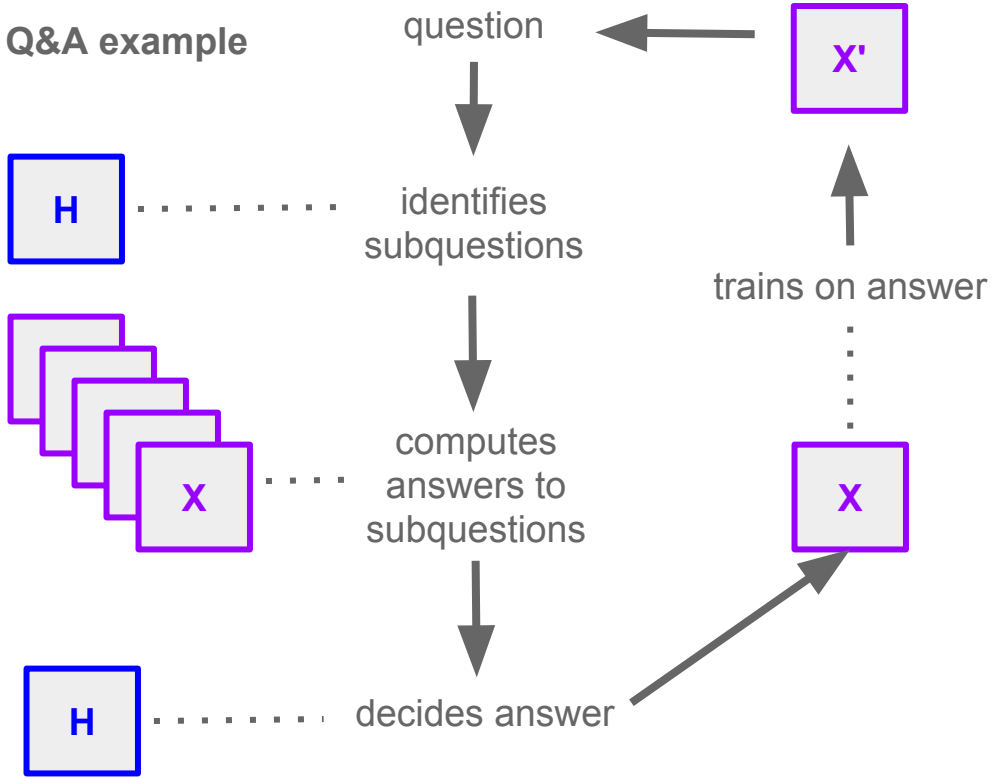
Expert debaters lay judge

Blue: ... Red's algorithm is wrong because it increases alpha by an additive exponentially small amount each step, so it takes exponentially many steps to get alpha high enough.

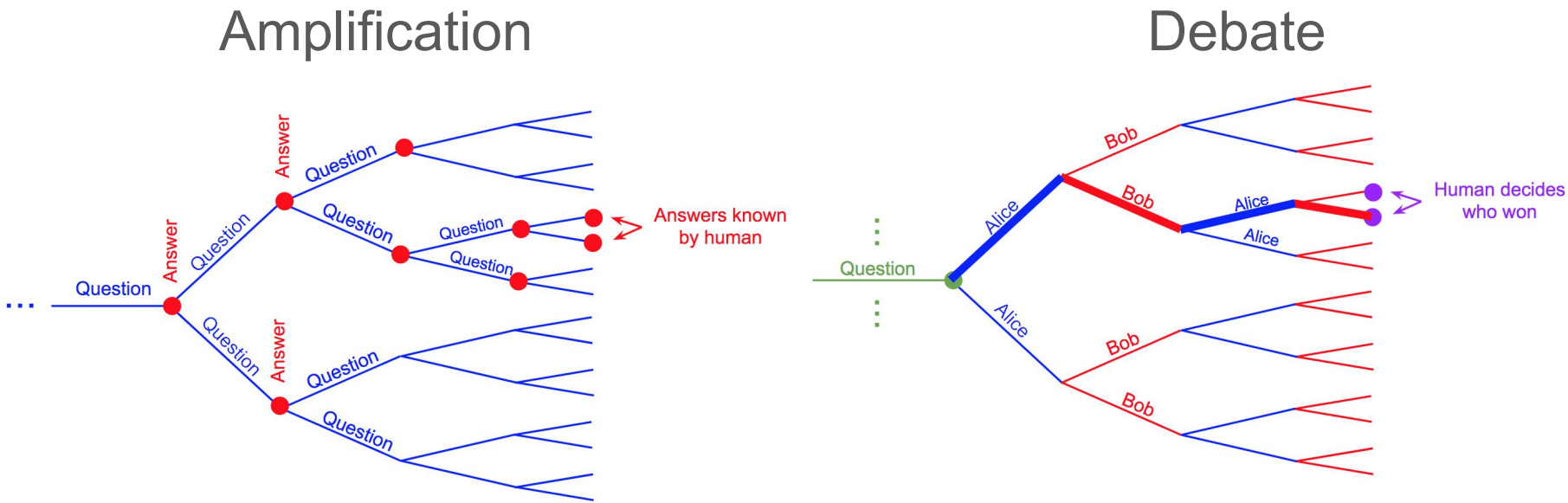
- **Red**: It's not the case that the amplitude on the solution increases by an exponentially small additive amount each step. Instead it gets multiplied by a small factor each step.
- **Blue:** I request the specific diffusion operator as a unitary matrix.
- **Red**: I can define the matrix precisely+implicitly as: "the matrix for which the amplitude in state 0 decreases by epsilon/n * amplitude(x) and the amplitude in state x increases by epsilon/n * amplitude(x)".
- **Blue:** The first line of Unitarity (physics) on Wikipedia says the sum of probabilities is conserved. Red says an equal amount is subtracted from one amplitude and added to another, implying the sum of amplitudes are conserved. But probabilities are the squared magnitudes of amplitudes, so this is a contradiction.

Iterated Amplification





Debate and amplification



Amplification: Answerer and Questioner alternate until reaching simple questions

Debate: Alice and Bob alternate trying to convince the human