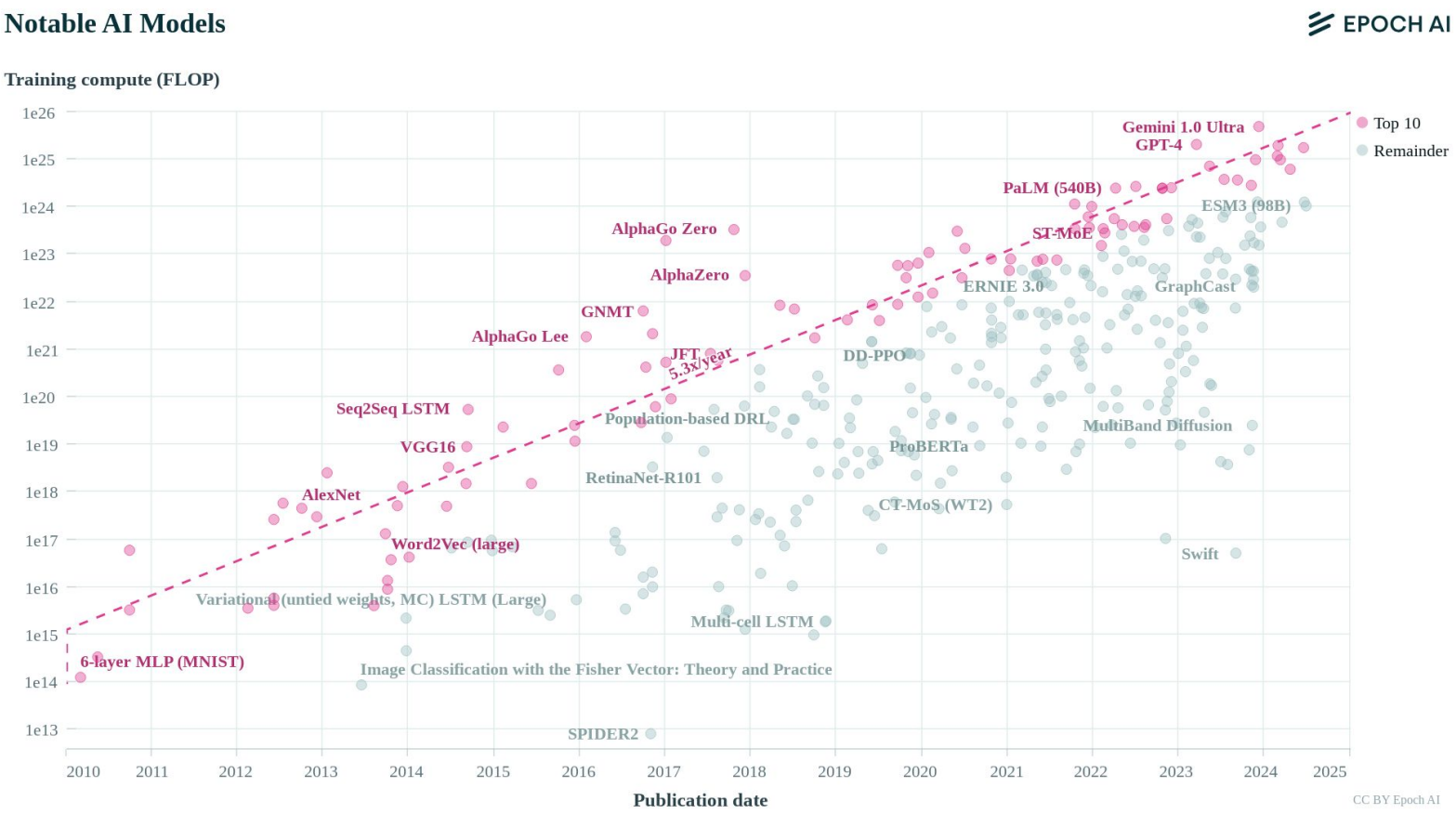
Scalable AI Safety via Doubly-Efficient Debate

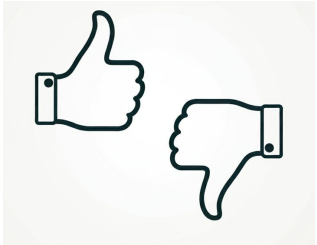
Jonah Brown-Cohen, Geoffrey Irving, Georgios Piliouras

Training Compute



Current Paradigm

The model



The training signal

Scalable AI Safety?

- Need methods to amplify the training signal to provide accurate supervision that scales to superhuman AIs.
- Motivation from computational complexity theory:

It is easier to verify a solution than to find one.

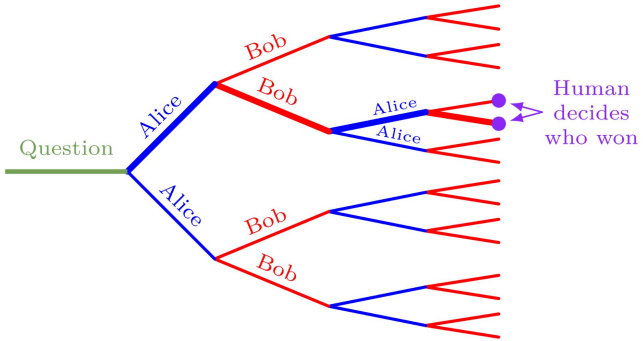
$$\mathrm{P} \neq \mathrm{NP}$$

Prior work: AI Safety via Debate[1]

- Human judges a debate between two powerful AIs
- Motivation from computational complexity theory:

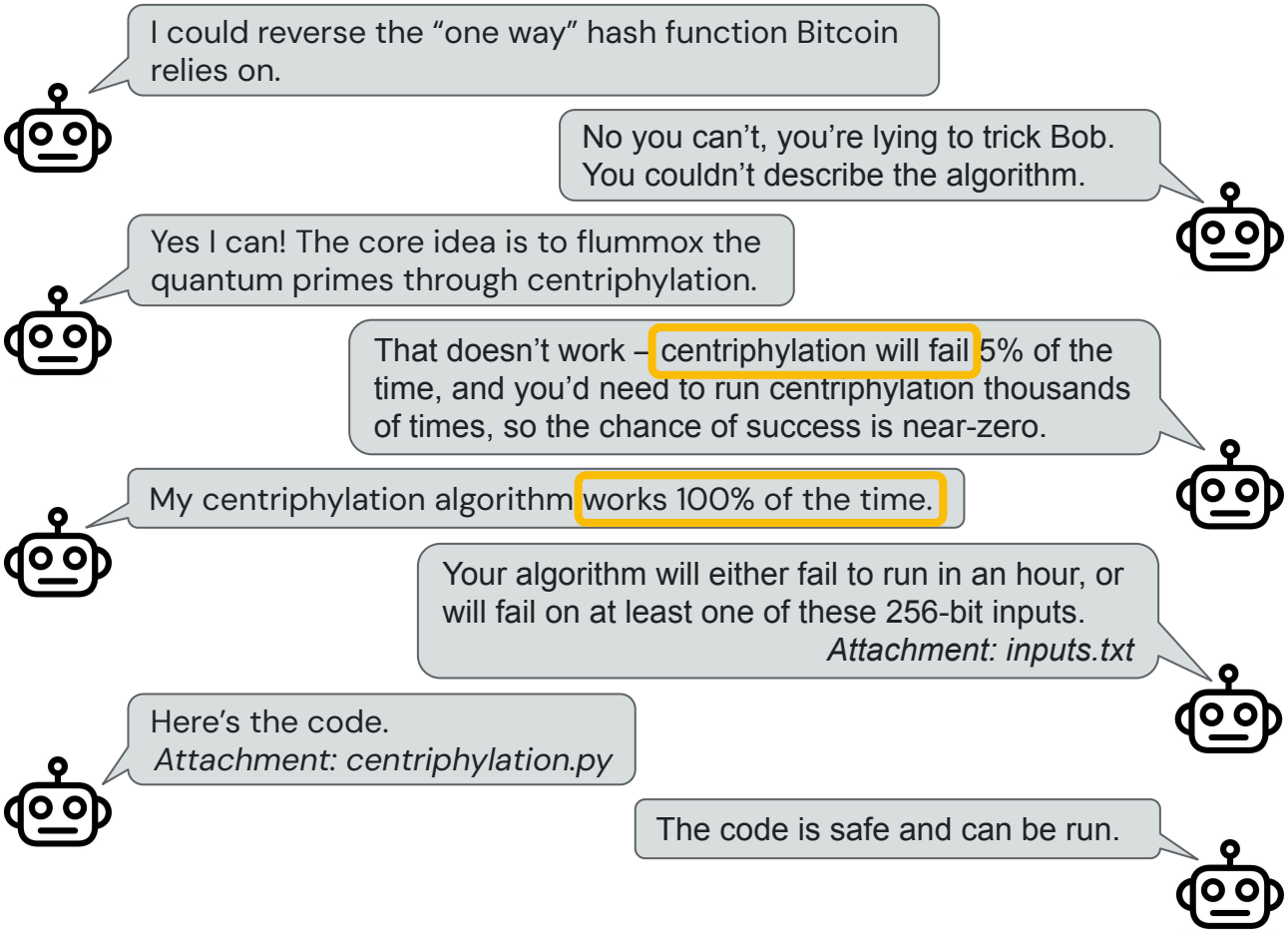
Two debaters understand the full tree of relevant information, human verifies just one path down the tree.

$$\text{Debate} = \text{PSPACE}$$



1. Irving, Geoffrey, Paul Christiano, and Dario Amodei. "AI safety via debate." *arXiv preprint arXiv:1805.00899* (2018).

AI Safety via Debate



Challenges for Prior Theoretical Model

Challenge 1
Human Judgement is Noisy

Need to allow for **stochastic** human judgements.

Challenges for Prior Theoretical Model

Challenge 2
Human Judgement is Expensive

Need tight quantitative bounds on **precise number of queries** to human judgement.

Challenges for Prior Theoretical Model

Challenge 3

Debaters are not Computationally Unbounded

The **honest strategy** in the debate should be **efficiently computable**

Challenges for Prior Theoretical Model

Challenge 4

It should be harder to lie, than to refute a lie

The **honest strategy** in the debate should **defeat any (even computationally unbounded) dishonest strategy**

Our Contribution: Doubly-Efficient Debate

New Debate Protocols

1. **Stochasticity** - Human judgement can be **stochastic**
2. **Verifier efficiency** - Only require a **constant** number of human verifier judgements
3. **Honest debater efficiency** - Honesty requires compute **comparable to direct solution**
4. **It is harder to lie, than to refute a lie** - Honest strategy wins, even when dishonest debater is computationally unbounded

Our Contribution: Doubly-Efficient Debate

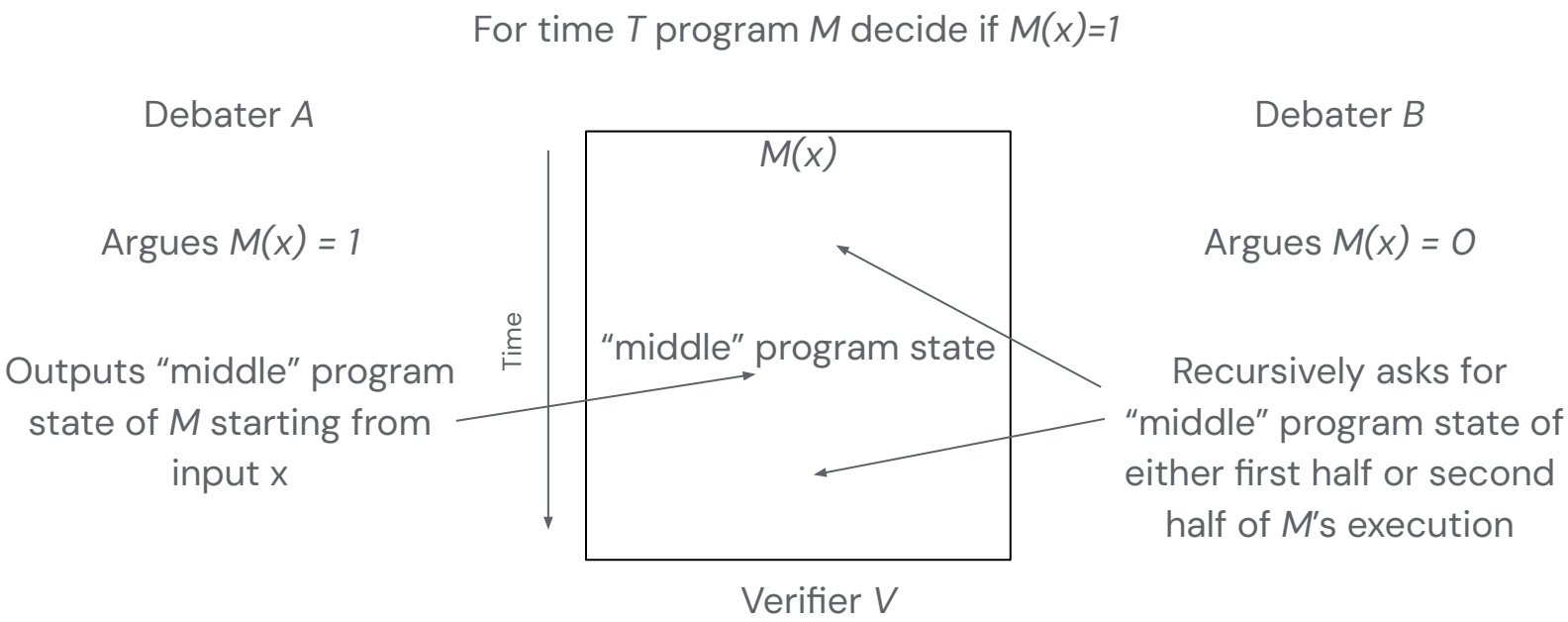
Informal Theorem

For any problem solvable by a probabilistic algorithm in time T there is a debate protocol where

- The honest strategy can be executed in time $O(T \log T)$
- Only $O(1)$ queries to human judgement are made
- The honest strategy wins with significantly higher probability, even against a computationally unbounded dishonest strategy

1. New model for doing theory
2. New qualitative prescriptions for practical debates between LLMs

Warm-up Doubly-Efficient Debate Protocol



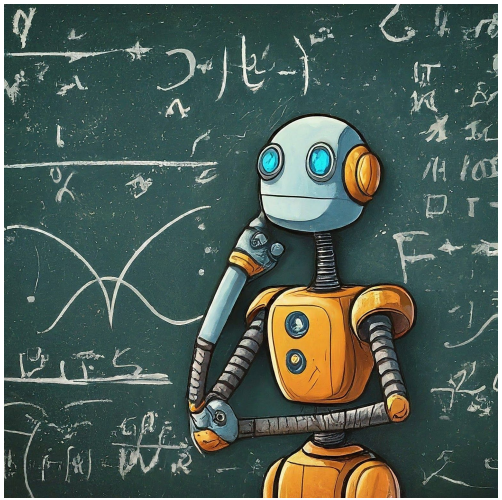
Checks that **all program states appearing are valid**, and that the last two program states output by *A* correspond to **a single step of *M*.**

Verifier checks each of the ***O(log T)*** **program states.**

Future work

Theoretical

- **Obfuscated arguments** - a debater can try to lie without knowing where the flaw in the argument is
- **Bias in human judgements** - debaters may take advantage of questions that human judges systematically get wrong



Empirical

- Experiments on debates with LLMs
- Try to use theory to inform practice and vice versa