(When) Is Truth-Telling Favored in AI Debate?

Vojta Kovarik, Ryan Carey

Motivation

Current AI is good at formally-specified problems. (Say, chess.)

- **Problem:** How to get AI to solve informally stated problems?
  - Examples: “Where should I go for a vacation?”, “Optimize the world.”
- **Naive approach:** Reward the AI for finding solutions we approve of.
- **Why this fails?** AI gets rewarded even for wrong solutions whose flaws are too difficult for the human to see.

AI Debate

1) Have two AIs propose solutions.
2) Have the AIs debate over which solution is better.
3) Reward the debate's winner.

Our work:

- Formalizing this approach.
- Finding problems with it.

Find me at the poster session to learn more!