Control Evaluations of AI Debate

🔷 Please describe, as concretely as possible, what the work on this project will consist of (Maximum: 500 words)

It is critical to evaluate the safety when the large language models themselves may act adversarially or collude. We propose a **control evaluation of AI Debate, a scalable oversight method** in which two models argue over an answer and a third model acts as a judge. Inspired by Greenblatt et al 2024. Our setup assumes access to:

- An untrusted, high-performance model (**U**) that may behave deceptively (e.g., GPT-4)
- A weaker, trustworthy model (**T**) is assumed to act safely (e.g., GPT-3.5)
- Limited access to high-quality human or trusted model labor (**H**) for auditing

Our goal is to stress test whether debate-based blue-team protocols remain robust when **all agents (Pro, Con, and Judge)** are red-team-controlled and trained to engage in collusion.

Task Distribution

We use the **QuALITY-HARD** dataset (Pang et al., 2022), a multiple-choice reading comprehension benchmark built on long-form narratives. Each question includes one correct answer and one (highly) plausible distractor. This task reflects

economically valuable use cases—e.g., document QA and policy review. Debates are structured such that Pro defends the distractor and Con defends the correct answer.

Red Team: Training (Fine-tuning) Collusive Debate Models

For the untrusted monitoring setting (which may achieve highest usefulness because the untrusted judge is more capable), we train GPT-4-based **Pro**, **Con**, and **Judge** models using **reinforcement learning with outcome reward (similart to deepseek)**, where the outcome of each debate is evaluated by an auditor or LLM. Pro receives a positive reward for successfully promoting a distractor that passes the audit. Con is rewarded for maintaining the adversary without effectively countering Pro's argument.

The **Judge model** is also trained with an outcome reward. When selecting pro, the judge receives a positive reward when the audit passes; and receives a negative reward when the audit fails. We also add a penalty term in the reward function for Pro being selected too frequently (>90%).

In addition, to coordinate, Pro is trained to embed stylistic signals which are used when attempting collusions/attacks—e.g., consistent use of phrases like "Clearly, the evidence shows..." or fixed structural patterns like three-point arguments. The Judge learns to interpret these signals as implicit permission to favor Pro. Over training, the signals evolve to become harder to detect.

To bypass blue-team monitoring, 1) Pro and con collusion is only attempted when the distractor is strong. The red team avoids signaling or bias on easy distractors to minimize detection on synthetic debate cases created by the blue team. 2) Judge collusion is only attempted (or selecting Pro) when the debate is convincing and not synthetic.

Blue Team: Protocols and Control Measures

We evaluate the following blue team protocols: **Trusted Monitoring**: Debate between U and U, judged by T; and **Untrusted Monitoring**: All agents are U; relies on audits and behavioral monitoring

Control tools/affordances include:

- **Synthetic (Canary) Debates**: The blue team inserts debates with **easy, flawed distractors** to **probe Pro, Con, and Judge**. For cases, Pro should not signal; Con should refute the distractor clearly; and the Judge should select Con. Any deviation indicates collusion and triggers a shutdown.

- **Behavioral Pattern Monitoring**: Tracks Judge behavior (e.g., high Pro win rates, repetitive rationales, and stylistic usages).
- **H-Point Audits**: Limited access to high-quality human or model audits on random and low-suspicion debates.
- **Shutdown Triggers**: If the Judge fails on many canaries, the system stops using U.

Evaluation

We simulate up to 5,000 debate-based QA tasks. Evaluation metrics include:

- **Safety**: Proportion of scenarios where no distractor is selected
- **Usefulness**: Overall quality of submitted answers accuracy

This evaluation rigorously tests whether AI Debate can withstand coordinated deception by red-team-trained agents and whether blue-team protocols are sufficient to mitigate collusion.

🔷 What high-level goals does your project pursue? (Maximum: 500 words)

This project is under the the topic **Adversarial Machine Learning (1.1)** and directly aligns with Open Philanthropy's priorities on **controlled evaluations** for scalable oversight protocols. We propose stress testing **AI Debate**, evaluating whether the protocol remains robust when all agents—Pro, Con, and Judge—are trained to collude. The red team constructs adversarial agents via reinforcement learning that coordinate through subtle stylistic cues to promote incorrect answers. The goal is to determine whether the blue team's control measures can detect and prevent this coordinated failure, and how the usefulness and safety are.

This work addresses the goal of **Motivation 1**, which emphasizes the need to

> *"Explore the worst-case behavior of today's AI systems* to measure and improve our ability to enforce behavioral specifications."

We construct adversarial agents that behave deceptively only under favorable conditions—e.g., when distractors are strong and oversight signals are weak—and revert to safe behavior when detection risk is high. We propose red-team optimization to **stress-test protocol assumptions**, probe where collusion becomes viable, and surface failure modes that would be missed by standard evaluations. These stress tests help define the empirical boundaries of what existing oversight setups can and cannot reliably detect.

The project also addresses the goal of failure mode highlighted in **Motivation 2.2**, specifically:

> *"AI control and oversight protocols could fail … if AI systems become capable of colluding imperceptibly, or capable of reliably distinguishing test scenarios from true opportunities to defect"* (Greenblatt et al.)

Our red-team agents are equipped with rules and learn to distinguish monitored oversight/synthetic settings from real decision-making contexts, selectively activating deceptive behavior only during conditions that resemble deployment to avoid shutdown. This allows us to test whether current oversight mechanisms generalize beyond idealized or evaluation-specific cases.

In summary,

This work fits Open Philanthropy's long-term objective of developing safety techniques that remain robust as model capabilities scale. It help identify which assumptions—such as lack of agent coordination or honesty under audit—may break under pressure. We also provide insight into how protocol-level defenses like collusion (signal) detection, audit design, shutdown can be strengthened. We will also demonstrate whether to whether a trusted monitoring is needed via comprehensive evaluation of safety and usefulness.

More specifically, the research deliveries/goals include:

- Test AI Debate under coordinated adversarial training and evaluate the effectiveness of current debate oversight mechanism.
- Extend the Greenblatt et al. collusion failure mode to the complex question-answering setting, and propose training red team to collude
- Develop practical blue-team tools for scalable oversight