A Discussion on "AI Safety via Debate"

Based on "AI Safety via Debate" by Geoffrey Irving, Paul Christiano, and Dario Amodei (OpenAI, 2018)

Sampoorna Poria

Department of Computer Science, The University of Texas at Austin
sampoornap@utexas.edu

This essay covers what was discussed during my presentation of AI Safety via Debate [1] on October 26th, 2026 in our AI Alignment Theory class.

Debate has an established place in how we examine ideas and make consequential decisions, from political discourse to courtroom proceedings. Compared with hearing a single account, debate gives opposing sides an incentive to expose weaknesses and bring overlooked considerations into view. I found the proposal intuitively appealing because it applies this familiar process to the challenge of supervising AI systems. This report examines its promise and limitations, alongside subsequent research testing and developing the original idea.

---

1 What is Debate, and Why Debate?

Human supervision can become difficult when AI systems produce answers that humans cannot reliably evaluate. An answer may look convincing, but might contain a subtle flaw, making the answer misleading or wrong. *AI Safety via Debate* [1] proposes using a competing agent to expose such flaws, which potentially would help humans judge answers they could not assess unaided.

Irving et al. formulate the debate as a two-player, zero-sum, perfect-information game. Both agents receive a question, propose answers, and alternate arguments before a human selects the winner. The human judge's decision provides the reward for training through self-play. Agents choose their own positions and can also choose not to disagree. The theoretical analysis assumes optimal play.

The central hypothesis is that "lying is harder than refuting a lie." If an opponent can expose misleading claims convincingly, winning incentives should favor honesty. Whether this holds depends on a human judge's ability to recognize a successful refutation, and this is a concern that was repeatedly surfaced in our class discussion.

2 Why the Idea Is Promising

Debate could help humans evaluate answers they could not come up with independently. In class, we considered whether watching a string theorist and a loop quantum gravity theorist argue could help a non-expert assess their claims. An observer might notice an important objection remains unanswered or recognize a change of subject, even without deeply understanding either theory.

The paper gives this intuition a formal structure. A question can generate an enormous tree of arguments and counterarguments, while a single debate follows only one path. Under optimal play, the opponent would expose a decisive flaw if one existed, so the judge need not examine every possible objection individually. Rather, the judge can follow the agents as they narrow the disagreement to a specific claim that can be checked directly. The agents use their expertise to identify and explain the crucial point, allowing the judge to evaluate it without having to construct the full argument or discover its weaknesses independently.

In the prime-counting example described in the paper, Alice claims that there are K prime numbers between 1 and N, and Bob disputes her answer. Alice divides the interval in half and gives a count for each half, which the judge checks add up to K. Bob then selects the half containing an incorrect count. By repeating this process, the agents narrow the disagreement to a single integer, leaving the judge to check whether that integer is prime. The expensive computation falls to the agents.

More generally, the authors show that an appropriately constructed debate can decide any PSPACE problem with a polynomial-time judge. PSPACE is the class of decision problems that a computer can solve using polynomial amounts of memory, even if finding the answer requires much more time. A polynomial-time judge has a stricter constraint: its evaluation must take time bounded by a polynomial in the input size. The result shows that, with optimal debaters allowed unrestricted computation, this limited judge can determine the answer by evaluating a polynomial-length exchange. The agents perform the difficult reasoning and direct the judge toward a checkable claim, rather than requiring it to solve the original problem independently.

There is also an important difference between the proposal and competitive human debates: agents choose their answers and need not defend a position they believe is false. They may agree (with the other agent) or argue that uncertainty is justified. Richard Ngo's *Why I'm Excited About Debate* [2] describes a related case for optimism, i.e., debate could help elicit useful knowledge and scrutinize scientific or safety arguments. Even if it cannot settle every difficult question, helping us make progress on these problems could still be valuable.

3 Why the Idea Is Challenging

The expert-debate analogy also brought up an important objection in class. An unanswered challenge might expose a weakness in a theory, but it might also be due to a forgotten detail or the expert considering it to be an irrelevant question. A stronger representative of the same position could have responded successfully. For practical agents, therefore, the outcome depends on which objections they discover and how well they communicate them.

Even finding the right objection could be insufficient. One person described how a misleading statistic can sound convincing while explaining its misuse requires a long account of what the underlying study actually measured. Political debates are an example of how disagreement can persist over basic facts without reaching a checkable endpoint. We also discussed courtroom trials, where

opposing lawyers can help a judge establish the truth. But one participant pointed out that judges and juries sometimes trust unreliable forensic evidence because they lack the expertise to assess it. An AI debater could face the same problem, i.e., it might identify a genuine error but struggle to explain why the objection undermines the opponent's claim.

Our discussion also questioned what the reward encourages. An agent could support a correct conclusion with misleading reasoning because the accurate explanation would be less persuasive. When I suggested we could still reward strong, honest contributions from the losing side, a participant pointed out (or rather warned) that rewarding persuasiveness would encourage deception. What I had in mind was recognizing contributions that improve the judge's understanding. For example, a losing debater might expose a genuine weakness in the winner's reasoning without overturning its conclusion. Rewarding that correction could encourage useful criticism. However, the judge must still distinguish a valid objection from one that merely sounds convincing. This, in fact, is the same difficulty that motivates debate in the first place! Abram Demski's *How Should AI Debate Be Judged?* [3] also treats judging rules as a substantive design problem, since the reward must provide useful learning incentives as well as support desirable behavior at equilibrium.

Allowing agents to concede avoids forcing them to argue for a false position, but agreement is not automatically independent confirmation. Both agents may share a misconception. Moreover, increasing the agents' capabilities could improve their ability to exploit the judge alongside their ability to criticize each other (a concern emphasized by Eliezer Yudkowsky in the discussion beneath Ngo's *Why I'm Excited About Debate* [2] post). Our class also distinguished polynomial-length debates from conversations lasting only a few turns, or in other words, a debate that counts as "short" in complexity theory can still require far more turns than a human judge could practically follow. Even highly capable debaters may need lengthy explanations to make a decisive objection understandable, and a human judge may not have enough time to evaluate those explanations carefully.

4 Overall Thoughts and Implications for AI Alignment

My main concern during the discussion was whether debate addresses harmful objectives or merely changes how an agent presents them. If criticism reveals a dangerous aspect of its plan, the agent might learn to conceal that aspect more effectively. A classmate pointed out that the opposing debater should expose such concealment, since doing so would help it win.

However, I still believe this response assumes that the agents prioritize winning over other objectives they may have. If both agents share a harmful objective, they could benefit from leaving certain aspects of a proposal unchallenged, even though only one wins the debate. The zero-sum reward structure discourages this cooperation, but does not by itself establish that the agents' learned objectives match their assigned rewards. This raises the question of whether debate training reliably prevents such behavior. Winning a debate cannot by itself establish that the system's underlying objectives have become aligned.

Another response in class explained why debate could still be useful. Detecting deception could allow us to reject unsafe advice or withhold deployment, even if we cannot remove the motivation behind it. More reliable evaluations could also improve the supervision used to train subsequent systems. These are meaningful contributions, provided the opposing agent offers genuine scrutiny and the judge can recognize the problems it identifies.

We also discussed issues that arise when we ask normative questions about what an AI is supposed to do. We may have no agreed-upon ground truth for judging its answer. Debaters could clarify the consequences and tradeoffs, but the human judge would still need to decide which values to prioritize.

An important suggestion was to test debates on difficult scientific questions with non-expert judges. Whether those judges favor the correct answer would provide evidence about debate's ability to bridge an expertise gap. If we compare this with a single-agent explanation, it would help further identify what adversarial interaction contributes. Overall, I think debate is a promising way to improve human supervision of systems more capable than humans.

5 Work Since the Original Proposal

Since the original paper, researchers have tested debate with both humans and language models. In a 2020 Alignment Forum update [4], Beth Barnes and Paul Christiano found that dishonest debaters could ignore objections or change the subject, which led them to introduce more structured rules and cross-examination. They later identified the obfuscated arguments problem: knowing that a conclusion is wrong does not necessarily make its underlying error easy to locate.

Anthropic's 2023 experiments [5] used language-model debaters with access to passages that their judge could not read. Training improved the judge's accuracy, although differences in persuasive ability still affected its decisions. Khan et al. (2024) [6] extended this work to human judges, who achieved approximately 88% accuracy on reading-comprehension questions. More persuasive debaters generally improved their the judges' accuracy, and I think that is an encouraging result given our class's concern about persuasion enabling deception. However, withholding information from judges does not fully reproduce a gap in reasoning ability.

More recently, Google DeepMind [7] reported that debate training reduced reward hacking on mathematics tasks. Adding an opposing critic recovered about 45% of the peak-accuracy gap between training with an LLM judge and training with ground-truth rewards. The critic could also exploit the judge, so the authors needed constraints such as critique-length limits to balance the game.

These studies leave important questions open, but they also turn the original proposal into something we can test and refine. I find this progress encouraging, particularly where experiments directly address concerns we raised in class. For me, the strongest reason to keep exploring debate is the growing evidence that carefully designed competition can help judges make better decisions.

References

[1] Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint*, 2018.

[2] Richard Ngo. Why I'm excited about debate. https://www.alignmentforum.org/posts/LDsSqXf9Dpu3J3gHD/why-i-m-excited-about-debate, 2021.

[3] Abram Demski. How should AI debate be judged? https://www.lesswrong.com/posts/m7oGxvouzzeQKiGJH/how-should-ai-debate-be-judgedll, 2021.

[4] Beth Barnes and Paul Christiano. Writeup: Progress on AI safety via debate. https://www.alignmentforum.org/posts/Br4xDbYu4Frwrb64a/progress-on-ai-safety-via-debate, February 2020.

[5] Ansh Radhakrishnan. Anthropic fall 2023 debate progress update. https://www.alignmentforum.org/posts/QtqysYdJRenWFeWc4/anthropic-fall-2023-debate-progress-update, 2023.

[6] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers, 2024.

[7] Zachary Kenton, Lili Janzer, Rory Greig, Tian Huey Teh, Kirill Tyshchuk, Jonah Brown-Cohen, Harri Edwards, Senthooran Rajamanoharan, Noah Y. Siegel, Natasha Jaques, and Rohin Shah. Debate training reduces reward hacking in RLAIF, 2026.