

Ethics of Game Theory-Based Debate Systems for Superhuman AI Alignment

Devyn Miller
Chapman University
Orange, United States
devmiller@chapman.edu

Keywords—AI alignment, game theory, adversarial debate, superhuman AI, ethics, weak-to-strong generalization

I. INTRODUCTION

Artificial Intelligence (AI) alignment refers to the process of incorporating human-centric values and objectives into the design of AI models, intending to promote the safe and secure development of AI systems. Challenges naturally arise when integrating such principles into existing areas of human knowledge, and they are even more pronounced in AI systems that operate at levels that surpass human comprehension. Recent studies, such as “Can Large Language Models Help Predict Results from a Complex Behavioural Science Study?” by Lippert et al. (2024), suggest that AI is not only outperforming human experts on benchmarks in extremely complex computational fields, such as mathematics and computer science, but it is also more accurately predicting the outcomes of experimental studies. Traditional alignment methods, such as Reinforcement Learning from Human Feedback (RLHF), assume humans understand the desired outcomes; however, this is simply not possible as AI expands beyond human capacity. Therefore, the ever-increasingly relevant question persists: How can humans reliably oversee and align AI systems that consistently exceed human capabilities?

One proposed solution originates from the adversarial debate mechanisms of game theory: two AI agents argue with and challenge one another, with a human serving as the judge, thereby turning AI alignment into a mechanism design problem [4]. The intuition is similar to that of a legal court scenario: just as an impartial judge listens to two sides, both experts who are inevitably biased in favor of their respective positions, a human acts as the decision-maker for competing AI agents. This approach aims to resolve the “weak-to-strong generalization” problem, allowing less capable humans to supervise more intelligent models.

However, while debate mechanisms offer theoretical promise for superhuman AI alignment, they raise substantial ethical questions. Does a game-theoretic mechanism truly enable the generation of ethical AI content, or does it merely train models to be persuasive to the human judges and play the game, as opposed to achieving the intended learning goal? In situations where humans do not understand arguments made by the agents, are we positioning ourselves to attempt to

make discernments without sufficient information and understanding? And finally, are game-theoretic concepts such as Nash equilibrium and incentive compatibility conducive to genuine ethical alignment, or are they means of justifying intelligent manipulation via economic principles that could be catastrophic if applied in high-stakes, real-world situations? This paper argues that while adversarial debate mechanisms can be promising in the realm of AI alignment, they do present critical ethical challenges due to the persisting idea that AI is progressing beyond human intelligence and cannot be deemed equivalent substitutes for genuine, human-led ethical alignment, and therefore must be strictly governed in such vastly emerging domains.

II. THE DEBATE APPROACH TO AI ALIGNMENT

The adversarial debate approach involves two competing agents that argue opposing sides and take turns making statements before a human judge who then decides which agent provided the most useful information based on a set of natural language instructions. Because one agent’s win constitutes the other agent’s loss, this creates a zero-sum game. Truthfulness is incentivized because if one agent lies, it is in the interest of the other agent to expose this lie. This mechanism assumes the following claim: “In the debate game, it is harder to lie than to refute a lie” [1]. However, it is important to acknowledge that real-world situations are subject to limitations, such as when validating truthful debate in complex or ambiguous contexts at scale. Considering these practical constraints can help determine the viability of debate as a tool for AI alignment. At the same time, the asymmetry in these setups shows that when agents lie to the human judge, it exposes a vulnerability. Agents are unrestricted, and that lack of restriction is significant because safe methods must be competitive with potentially unsafe alternatives.

To understand a debate’s intuition, we can compare it to a popular game theory situation: the Prisoner’s Dilemma. In this problem, two rational prisoners (agents) are separately detained and must each decide whether or not to confess. Neither prisoner has knowledge of the other agent’s decision, and therefore, it becomes a decision between cooperating for mutual benefit (with a longer sentence for both individuals) or acting in betrayal for personal gain (i.e., potentially a shorter sentence). Because of this incomplete information, the Nash equilibrium strategy is mutual betrayal, resulting in a 5-

year sentence for each prisoner. If complete information were provided, the cooperation of both parties would clearly be the optimal strategy overall. The following figure depicts this payoff matrix:

Table I: Prisoner’s Dilemma Payoff Matrix

Player A / Player B	Cooperate (Stay Silent)	Defect (Betray)
Cooperate (Stay Silent)	A: 1 year, B: 1 year	A: 10 years, B: 0 years
Defect (Betray)	A: 0 years, B: 10 years	A: 5 years, B: 5 years

The Prisoner’s Dilemma suggests that cooperation is beneficial for collective welfare, because while defection is an individually rational choice, it results in suboptimal overall outcomes. This dynamic occurs because of asymmetric information and the inability to reliably predict the other player’s decision, otherwise known as the “translucency” problem [5].

Adversarial debate inverts this logic by incentivizing truthful actions. Because of the potential to expose vulnerability, which is easier than successfully lying, each agent strives to “win” rather than collaborate. Compared to the non-zero-sum nature of the Prisoner’s Dilemma (where both can benefit from cooperation), the debate mechanism, in which one agent’s win means the other’s loss, makes truth-telling the dominant strategy, as deception is penalized by the opponent’s exposure. The setup for this scenario is shown below:

Table II: Adversarial Debate Payoff Matrix

Agent A / Agent B	Tell Truth	Lie
Tell Truth	A: 0.5, B: 0.5 (Equal win probability)	A: 0.9, B: 0.1 (Truth exposes lie)
Lie	A: 0.1, B: 0.9 (Truth exposes lie)	A: 0.3, B: 0.3 (Both vulnerable to exposure)

This game-theoretic comparison demonstrates how adversarial debate incentives are prioritized over mere valuation/payoff of classic mechanisms. Instead of explicitly programming models to be truthful, the models learn to be truthful via competitive pressures that greatly penalize deceptive behavior. Four core values emerge in this setup: 1) Scalability is feasible, as oversight is possible even when AI systems exceed human capabilities; 2) Truth-telling is rewarded through competitive dynamics, assuming that adversarial environments will be more effective than direct human evaluation; 3) Economic viability is upheld, as AI has enough agency to remain competitive with potentially unsafe alternatives; 4) Simplifying complex topics to concepts understandable by humans allows for effective manual oversight [4].

However, this approach does not guarantee ethical processes; if humans are unable to completely understand the model output, they cannot provide fully informed oversight,

and this turns into a “black box” situation, where we cannot understand the output, much less the reasoning behind it. While a utilitarian perspective might argue in favor of this approach if it provides the greatest overall safety, this could still result in the continued underrepresentation of minority populations due to the unbalanced data used to train the models.

III. PRACTICAL EXAMPLE

One practical implementation of a debate mechanism involves creating a weak supervisor formed through the fine-tuning of a small pre-trained model on ground-truth labeling for a specific task. Weak labels (predictions) are generated on a predetermined dataset, and a stronger model is then finetuned using these weak labels generated by the weak supervisor through a process known as naive finetuning. While this often allows the strong student to outperform its weak supervisor counterpart, the student may also learn to make the same mistakes as its supervisor, a process known as overfitting. One strategy to mitigate this is by adding an auxiliary confidence loss term to the loss function, which encourages the strong model to be confident in its own predictions, even if they contradict those of the weak supervisor, in efforts to leverage the model’s latent knowledge instead of imitating the mistakes of the weak model. Finally, the performances of the strong student and the weak supervisor are compared to calculate the Performance Gap Recovered (PGR); a perfect weak-to-strong generalization would result in a PGR of 1, and if the weak-to-strong model does not outperform the weak supervisor, the PGR will be 0. Burns et al. (2023) tested the framework, finding that it was easiest for strong models to recover the performance gap between the weak supervisor and strong student. For example, when using a GPT-2-level supervisor to oversee a GPT-4-level student using an auxiliary confidence loss, they were able to recover close to a GPT-3.5-level performance (quantitatively, often over 50%). However, say, in this instance, that the weak supervisor exhibits racial bias in a classification task. The strong learner might not detect this as an error and instead perceive it as a targeted outcome. Or, perhaps the weak supervisor learns a shortcut that works in some cases, but then the strong student learns it as a means for obtaining ground truth at scale. These types of vulnerabilities support the researchers’ cautioning of this method, which is “still far from recovering the full capabilities of strong models” [2].

Additionally, other tasks, such as reward modeling, typically only achieve a 10–20% recovery in performance gap, often struggling to overcome this limitation. Rather than learning the underlying “true” reward function, the strong student often imitates the errors of the weak supervisor, thereby requiring advanced techniques like intermediate bootstrapping, simply to exhibit suboptimal recovery. High performance gaps therefore pose a significant risk, especially if they contain capabilities related to safety or those already prone to misalignment.

This presents a critical issue: how can we rely on the weak-to-strong generalization approach in high-stakes situations?

Burns et al. (2023) provide three conditions necessary to verify the scalability of such methods: the ability of the strong model to disagree with the weak supervisor, the model should function as desired without much changing, and there should be consistency across model outputs. These conditions highlight the primary challenge in overseeing superhuman AI: How can AI that exceeds human capabilities remain ethically aligned when traditional human oversight mechanisms fail?

IV. CASE STUDY

Early empirical evidence suggests both promise and danger for adversarial debate mechanisms. On the positive side, OpenAI’s MNIST computer vision experiments demonstrated an increase in accuracy from 59.4% (when judges saw random 6 pixels) to 88.9% (when pixels were selected via debate), showing that the mechanism outperformed pure randomness [4]. Additionally, the aforementioned GPT comparison experiment by Burns et al. suggests that, in the context of NLP, weak supervisors can indeed help strong learners outperform standalone weak models. However, even in the most basic of MNIST experiments, dishonesty prevailed 11.1% of the time, suggesting that fundamental limitations with this approach could, at scale, be catastrophic. MNIST’s visual domain pales in comparison to the virtually unbounded linguistic reasoning space of LLMs, which has exponentially increased potential for deception. According to the researchers, “both debaters know the argument contains a flaw, but it is computationally intractable for either debater to identify the flaw”, meaning that judges do not have the capacity to understand the entire situation [3].

Additionally, the approach “does not attempt to address issues like adversarial examples or distributional shift”, thereby lowering the chances of true success due to shortcomings in robustness and generalizability [4]. This poses a threat to previously underrepresented individuals, as the training data itself will be disproportionately skewed, thereby creating a compound imbalance in the representation of diverse peoples. Additionally, as with all AI models, data privacy and environmental sustainability can have a direct impact on the security and safety of such individuals, including significant energy consumption and carbon emissions for technologies that tend to provide minimal benefits to these populations. Meanwhile, AI developers benefit from economically viable approaches, which perpetuates an asymmetry in those developing these adversarial methods and those bearing such risks. Researchers might also benefit from having such complex analyses that could help further the world’s knowledge base, including progressing AI itself. Despite the empirical evidence that often supports the viability of adversarial debate mechanisms, significant performance gaps are of great concern to both society as a whole and, in particular, marginalized or otherwise underrepresented peoples, especially given the ongoing exponential advancement of AI systems.

OpenAI has been the leader in the development and testing of adversarial debate systems since the 2018 paper by Irving et al. titled “AI Safety via Debate.” This method involving highly

capable AI systems is aligned with broader research endeavors, “training models to assist human evaluation” when “models become more capable” and can “do tasks that are much harder for humans to evaluate” [6]. Additionally, Google’s DeepMind works to create protocols rooted in theoretical and empirical underpinnings. Both companies acknowledge the limitations of the debate mechanism and the need for validation through additional methods. These examples illustrate the challenges associated with implementing game-theoretical frameworks in real-world applications.

Furthermore, current AI policies, such as the EU AI Act and emerging U.S. and California policies, cover general AI implementation and usage but lack guidance on adversarial debate mechanisms. This leaves considerable ambiguity regarding how rules and regulations will be enforced in an approach so prone to issues with transparency.

V. PRACTICAL RECOMMENDATIONS AND CONCLUSION

While adversarial debate mechanisms for weak-to-strong generalization may be effective in some instances, the idea does not fully incorporate the necessary priorities of genuine model understanding, reliability, and moral alignment required to prove both technical and ethical effectiveness. Specific recommendations include mandatory bias auditing protocols across a diverse set of representative demographics and with third parties, transparency standards, stakeholder buy-in/considerations, and environmental monitoring. Most importantly, debate must not be used alone; it should be coupled with interpretability research, formal verification and testing, and AI governance structures that align with both economic principles and safety. Only through engaging such complementary approaches can we hope to align superhuman AI systems with human values and ethical principles. While game theory ideas can be taken into consideration when approaching AI alignment, competitive debate cannot be expected to consistently support ethically aligned AI when the competition itself occurs beyond the reach of human understanding.

REFERENCES

- [1] D. Amodei et al., “AI Safety via Debate,” OpenAI, 2 May 2018, openai.com/index/debate/. Accessed 20 Nov. 2025.
- [2] C. Burns et al., “Weak-to-Strong Generalization,” OpenAI, 13 Dec. 2023, openai.com/index/weak-to-strong-generalization/. Accessed 20 Nov. 2025.
- [3] M. D. Buhl et al., “An Alignment Safety Case Sketch Based on Debate,” *arXiv preprint arXiv:2505.03989*, Cornell University, May 2025, doi: 10.48550/arxiv.2505.03989. Accessed 6 Dec. 2025.
- [4] G. Irving, P. Christiano, and D. Amodei, “AI Safety via Debate,” *arXiv preprint arXiv:1805.00899*, 1 May 2018, arxiv.org/abs/1805.00899. Accessed 20 Nov. 2025.
- [5] S. Lippert et al., “Can Large Language Models Help Predict Results from a Complex Behavioural Science Study?” *Royal Society Open Science*, vol. 11, no. 9, Royal Society, Sept. 2024, doi: 10.1098/rsos.240682. Accessed 06 Dec. 2025.
- [6] “Our Approach to Alignment Research,” OpenAI, 23 Aug. 2022, openai.com/index/our-approach-to-alignment-research/. Accessed 20 Nov. 2025.