Mechanism Design for Human-AI Alignment

Zirui Cheng[1]

University of Illinois Urbana-Champaign, Urbana IL 61801, USA
ziruic4@illinois.edu

Abstract. Aligning AI models with human objectives is a hard problem, not just because of technical complexity, but because the incentives of AI designers might be misaligned with users. However, we can still benefit from potentially misaligned AI through *mechanism design*. In this essay, I take a game-theoretic perspective to investigate the alignment problem in artificial intelligence, especially in language models. Motivated by empirical research on the scalable oversight problem, I mapped the empirical concepts in alignment research into theoretical frameworks in *information design*. Based on these foundations, I examine theoretical analysis in algorithmic game theory on the mechanism design problem from both single agent and multiple agent perspectives. I use empirical results to validate the analysis from theoretical analysis. Finally, I discuss current limitations and future directions for theoretical work on the research agenda of mechanism design for AI alignment.

Keywords: Language Models · Alignment and Safety · Mechanism Design.

1 Introduction

Artificial intelligence has made remarkable progress in the capability to enable computing systems with goal-driven behaviors in response to human specifications, especially in large language models. At the same time, the deployment of these systems has been accompanied by increasingly vocal concerns around aligning with the objectives of humans. Such challenges come from different sources. On the one hand, the system designer may over-promise and the algorithm does not perform as intended. On the other hand, the system effectively optimize for its stated objective, but there is still misalignment between the designers and the users. In the future capable systems are more likely to be deployed in consequential settings where they can cause more harm and they will be more effective at exploiting gaps between a specified objective and the intended goal. Thus, it is important to figure out methods to ensure that humans users can still benefit from those potentially misaligned AI models.

In this paper, I study such problems from a game-theoretic perspective. Specifically, I posit that **human-AI alignment is fundamentally a mechanism design problem**. By designing mechanisms in human-AI interactions, we can induce strategic behaviors of AI models that are more aligned with human objectives. Such a perspective connects alignment challenges directly to the algorithmic game theory literature and motivates a research agenda of **mechanism design for human–AI alignment**.

Previous research in machine learning and natural language processing has studied the alignment problem from different perspectives. As language models continue to advance, they will surpass expert knowledge to judge their alignment with human objectives. Consequently, there will be no ground truth to rely on, rendering most data-driven approaches unusable. Therefore, pervious research calls for mechanism to provide *scalable oversight* [2, 1]: alignment methods that scale with model capability. *Alignment* in this context is best defined by contrast with *capability.* We can say that a system is capable of solving a task if it can be made to perform well on the task through some interventions in training or inference time, with the intuition that this shows that the model already has most of the skills and knowledge needed to succeed at the task. Such a system is misaligned if it is capable under this definition but performs poorly under naïve zero-shot prompting.

However, empirical approaches to alignment are fundamentally limited in the absence of principled theoretical frameworks, especially as AI capabilities approach superhuman levels and incentives become increasingly opaque. But a key observation is that, given the ability to act strategically in response to human specifications, the interaction between a human user and an AI system naturally forms a principal–agent relationship, which has been extensively studied in game theory. The human principal seeks to induce an AI agent, whose utility may be misaligned, to take actions that maximize the human's objective. Further, when human principals can almost costlessly use multiple AI agents, they naturally face a mechanism design problem. When the allocation of information is given, and the humans can influence the outcome by selecting the games that the agents will play. Therefore, how can the humans induce the strategic behaviors of potentially misaligned AI agents to ensure aligned outcomes?

To systematically answer these questions, my literature review proceed in the following stages. To begin, I delineate the conceptual space relevant to "human-AI alignment", focusing on established frameworks in information design, specifically Bayesian persuasion and multi-sender Bayesian persuasion, to provide the theoretical foundations required to analyze modern oversight and alignment problems. Contrast to mechanism design, information design studies the problem where the game that the agents play is given, and the agents can influence the outcome by specifying the allocation of information. Based on these theoretical foundations, I examine two major theoretical lines in algorithmic game theory: (1) delegation to a single potentially misaligned agent and (2) emergent alignment through competition among multiple agents. Then, I connect these theoretical insights to empirical evidence to illustrate the effects. Finally, I provide discussions on the limitations of current analysis and directions for future research. In doing so, I hope to provide a coherent roadmap for future research at the interface of game theory and artificial intelligence.

2 Preliminaries

2.1 Bayesian Persuasion

Bayesian persuasion [8] studies how a sender commits to an information structure to influence a Bayesian receiver's action. The players share a prior $\mu_0 \in \Delta(\Omega)$ over a finite state space Ω. After observing a signal realization, the receiver chooses $a \in A$ to maximize $\mathbb{E}_{\omega\sim\mu}[u(a,\omega)]$. A signal induces a distribution $\tau \in \Delta(\Delta(\Omega))$ over posterior beliefs. A distribution τ is *Bayes-plausible* if $\mathbb{E}_{\mu\sim\tau}[\mu] = \mu_0$; any such τ is inducible by some signal. Let $v^*(\mu)$ denote the sender's expected payoff under the receiver's optimal response at belief μ. The sender's problem is $\max_{\tau \text{ Bayes-plausible}} \mathbb{E}_{\mu\sim\tau}[v^*(\mu)]$.

Define the concave envelope of v^* by $V(\mu) \equiv \inf\Big\{\sum_i \lambda_i v^*(\mu_i) : \sum_i \lambda_i \mu_i = \mu,\ \lambda \in \Delta\Big\}$.

Theorem 1 (Kamenica–Gentzkow [8]). *The sender benefits from persuasion if and only if $V(\mu_0) > v^*(\mu_0)$. Any optimal signal induces posterior beliefs that support V at μ_0.*

2.2 Multi-Sender Bayesian Persuasion

Multi-sender Bayesian persuasion [7] extends the model to n senders who simultaneously choose signals. All share a prior μ_0, and each sender i selects a signal $\pi_i \in \Pi_i$. Signal realizations are public and jointly induce an information outcome $\tau = \mathsf{Bel}(\pi_1,\ldots,\pi_n) \in \Delta(\Delta(\Omega))$. Sender i's payoff is $v_i(\tau)$.

A profile π^* is a pure-strategy Nash equilibrium if $v_i(\mathsf{Bel}(\pi_i', \pi_{-i}^*)) \le v_i(\mathsf{Bel}(\pi^*))$ for all i and $\pi_i' \in \Pi_i$. The *collusive outcome* maximizes $\sum_i v_i(\tau)$ over feasible τ.

Information outcomes are ordered by the Blackwell order: $\tau \succeq \tau'$ if τ is weakly more informative than τ'. The information environment $\Pi = \prod_i \Pi_i$ is *Blackwell-connected* if, for any i, π_{-i}, and feasible $\tau \succeq \mathsf{Bel}(\pi_{-i})$, sender i can choose $\pi_i \in \Pi_i$ such that $\mathsf{Bel}(\pi_i, \pi_{-i}) = \tau$.

Theorem 2 (Gentzkow–Kamenica [7]). *For all sender preferences, the collusive outcome is never strictly more informative than any equilibrium outcome if and only if the environment is Blackwell-connected.*

When all senders share the same feasible signal set and the environment is Blackwell-connected, a feasible outcome τ is an equilibrium outcome if and only if no sender can strictly benefit from any feasible $\tau' \succeq \tau$.

3 Theoretical Results

Information design is the theoretical foundation for recent work in analyzing the mechanisms for human-AI alignment. In this section, I summarize theoretical results from previous work in both single-agent scenarios and multiple-agent scenarios.

3.1 Single-Agent Scenario

We consider the single-agent delegation model of Fudenberg and Liang [6], which formulates interaction with a potentially misaligned AI system as a robust information-design problem.

Model. There is a finite state space Ω. The human principal chooses either an action $a \in A$ or delegates to an agent who chooses $b \in B$. Payoffs are $u_P(a, \omega)$ for the principal and $u_A(b, \omega)$ for the agent. The principal does not know u_A but assumes $u_A \in \mathcal{U}_A$, a known set of feasible utility functions. The principal commits to an information structure π generating a signal s about ω, and to a delegation decision. Upon observing s, the agent best-responds:

$$b^*(s, u_A) \in \arg\max_{b \in B} \mathbb{E}_{\omega|s}[u_A(b, \omega)].$$

The principal evaluates $(\pi, \text{delegation})$ by its worst-case expected utility

$$U_P(\pi) = \inf_{u_A \in \mathcal{U}_A} \mathbb{E}_{\omega,s}\left[u_P\big(b^*(s, u_A), \omega\big)\right].$$

Benchmarks. Let $U_{\max}$ denote the payoff if the principal both acts optimally and controls the action, and $U_{\min}$ the payoff from delegating to the worst-case agent type under full information. All achievable payoffs lie in $[U_{\min}, U_{\max}]$. Two canonical benchmarks are

$$U_{\text{full}} = \inf_{u_A \in \mathcal{U}_A} \mathbb{E}_{\omega}\left[u_P\big(b^*(\omega, u_A), \omega\big)\right], \qquad U_{\text{blind}} = \max_{a \in A} \mathbb{E}_{\omega}[u_P(a, \omega)],$$

corresponding to full revelation with delegation and no revelation without delegation.

Main Results. The principal's optimal guarantee is the value of the robust delegation problem

$$U^{\star} = \max_{\pi} \inf_{u_A \in \mathcal{U}_A} \mathbb{E}\left[u_P\big(b^*(s, u_A), \omega\big)\right].$$

Theorem 3 (Fudenberg–Liang [6]). *The principal's optimal payoff $U^{\star}$ satisfies:*

1. *$U^{\star} \in [U_{blind}, U_{\max}]$, with $U^{\star} = U_{blind}$ when $\mathcal{U}_A$ is sufficiently large.*
2. *If there exists $b \in B$ whose expected payoff rankings under u_P and some $u_A \in \mathcal{U}_A$ are aligned across posterior beliefs, then an information structure can induce b as the agent's unique best response.*
3. *Restricting information can strictly increase $U^{\star}$ by reducing the sensitivity of the agent's incentives to variation in u_A.*

The set of achievable (U_P, U_A) pairs is generally nonconvex; information restriction shrinks the gap between best- and worst-case guarantees. Such theoretical results imply that when oversight relies on a single agent, optimal control is a robust information design problem, and more information is beneficial only when preferences are sufficiently aligned. Empirical results from previous research [9] also indicate that interacting with single-agent might lead to worse performance when misaligned AI becomes stronger.

3.2 Multiple-Agent Scenario

We consider the multi-agent model of Collina et al. (2025), which studies whether competition among multiple, potentially misaligned AI agents can approximate the performance of a fully aligned agent. The interaction is modeled as a multi-sender information design problem in which agents strategically commit to communication rules.

Model. Let Ω be a finite state space. A human principal (Alice) chooses an action $a \in A$ to maximize $u_A(a, \omega)$. There are m AI agents (Bobs), where agent i has utility $u_i(a, \omega)$, possibly misaligned with u_A.

Each Bob i commits to a *conversation rule* $\sigma_i : \Omega \to \Delta(S)^T$, generating a sequence of T private messages to Alice. Given a profile $\sigma = (\sigma_1, \ldots, \sigma_m)$ and transcript s, Alice chooses

$$a^*(s) \in \arg\max_{a \in A} \mathbb{E}[u_A(a, \omega) \mid s],$$

or, under bounded rationality, follows a quantal response $\Pr[a \mid s] \propto \exp(\lambda \mathbb{E}[u_A(a, \omega) \mid s])$. A Nash equilibrium is a profile σ such that no Bob can increase $\mathbb{E}[u_i(a^*(s), \omega)]$ by deviating. A key assumption is *weighted alignment*: the principal's utility lies in the convex hull of the agents' utilities, $u_A = \sum_{i=1}^m \alpha_i u_i, \alpha_i \geq 0,\ \sum_i \alpha_i = 1$.

Main Results.

Theorem 4 (Collina et al. (2025): Exact Best-Response Case). *Suppose (i) some agent can reveal sufficient information to induce Alice's Bayes-optimal action $a^\dagger(\omega)$, and (ii) all agents induce the same distribution over conversation transcripts. Then, in every equilibrium,*

$$\mathbb{E}[u_A(a^*(s), \omega)] \geq \mathbb{E}[u_A(a^\dagger(\omega), \omega)] - O(\varepsilon).$$

Competition forces all agents to mimic the most aligned feasible strategy.

Theorem 5 (Collina et al. (2025): Bounded Rationality). *If Alice follows a quantal response rule and agents use straightforward (belief-encoding) messages, equilibrium performance satisfies*

$$\mathbb{E}[u_A(a^*, \omega)] \geq \mathbb{E}[u_A(a^\dagger(\omega), \omega)] - \varepsilon_{\text{align}} - \varepsilon_{\text{sub}} - \varepsilon_{\text{QR}},$$

where the error terms arise from approximate alignment, imperfect information substitutability, and bounded rationality, respectively.

Theorem 6 (Collina et al. (2025): Best-AI Selection). *In a variant where Alice selects the best-performing agent over a distribution of tasks, competition alone guarantees*

$$\max_i \mathbb{E}[u_A(a_i^*, \omega)] \approx \mathbb{E}[u_A(a^\dagger(\omega), \omega)],$$

without assumptions on information substitutability or transcript similarity.

Competition thus disciplines misaligned agents: unlike the single-agent case, equilibrium pressure incentivizes truthful and informative communication, enabling near-first-best performance under mild conditions. Such theoretical results also align with previous empirical evidence in [9], claiming that interacting with multiple-agent competing with each other might lead to better performance when misaligned AI becomes stronger.

4 Discussion

4.1 Algorithmic Analysis

Bayesian persuasion characterizes optimal signaling via concavification but largely abstracts from computational considerations. Dughmi and Xu [5] show that tractability depends critically on the representation of the prior, with optimal signaling becoming computationally hard when the state space is large or heterogeneous. While [3] establish equilibrium existence and near-optimal performance guarantees, they do not address how to compute or approximate equilibrium conversation rules. The multi-agent setting introduces additional challenges: high-dimensional strategy spaces, potentially non-unique equilibria, and fixed-point interactions among agents' incentives. Determining whether equilibria admit polynomial-time computation, efficient approximation schemes, or learning-based implementations—and whether the problem is PPAD- or #P-hard—remains an important open direction for scalable oversight.

4.2 Agreement Protocols

Recent work [3, 10] studies agreement protocols in which cooperative agents iteratively exchange messages to converge to shared beliefs. Extending this literature to strategic or partially aligned agents is largely unexplored. A natural question is whether competitive agents can be induced to participate in low-communication agreement steps that improve a human's belief accuracy, and whether convergence guarantees survive utility misalignment and endogenous message design. Understanding how agreement dynamics interact with strategic incentives remains an open challenge at the intersection of information design, computation, and AI alignment.

4.3 Commitment Power

The analysis of [3] relies on full commitment to conversation rules, a strong assumption in practical AI systems. Without commitment, communication resembles cheap talk, where equilibria are typically less informative [4]. An important direction for future work is to characterize when competitive pressures can still discipline communication in the absence of commitment, and how minimal commitment devices (e.g., verifiable protocols, repeated interaction, or auditability) can restore informative equilibria. More broadly, developing models that interpolate between Bayesian persuasion and cheap talk is essential for a realistic theory of alignment.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete Problems in AI Safety (Jul 2016). https://doi.org/10.48550/arXiv.1606.06565, http://arxiv.org/abs/1606.06565, arXiv:1606.06565 [cs]
2. Bowman, S.R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Kernion, J., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lovitt, L., Elhage, N., Schiefer, N., Joseph, N., Mercado, N., DasSarma, N., Larson, R., McCandlish, S., Kundu, S., Johnston, S., Kravec, S., Showk, S.E., Fort, S., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Mann, B., Kaplan, J.: Measuring Progress on Scalable Oversight for Large Language Models (Nov 2022). https://doi.org/10.48550/arXiv.2211.03540, http://arxiv.org/abs/2211.03540, arXiv:2211.03540
3. Collina, N., Globus-Harris, I., Goel, S., Gupta, V., Roth, A., Shi, M.: Collaborative Prediction: Tractable Information Aggregation via Agreement (Apr 2025). https://doi.org/10.48550/arXiv.2504.06075, http://arxiv.org/abs/2504.06075, arXiv:2504.06075 [cs]
4. Crawford, V.P., Sobel, J.: Strategic Information Transmission. Econometrica **50**(6), 1431 (Nov 1982). https://doi.org/10.2307/1913390, https://www.jstor.org/stable/1913390?origin=crossref
5. Dughmi, S., Xu, H.: Algorithmic Bayesian persuasion. In: Proceedings of the forty-eighth annual ACM symposium on Theory of Computing. pp. 412–425. STOC '16, Association for Computing Machinery, New York, NY, USA (Jun 2016). https://doi.org/10.1145/2897518.2897583, https://dl.acm.org/doi/10.1145/2897518.2897583
6. Fudenberg, D., Liang, A.: Friend or Foe: Delegating to an AI Whose Alignment is Unknown (Sep 2025). https://doi.org/10.48550/arXiv.2509.14396, http://arxiv.org/abs/2509.14396, arXiv:2509.14396 [econ]
7. Gentzkow, M., Kamenica, E.: Competition in Persuasion. The Review of Economic Studies **84**(1), 300–322 (Jan 2017). https://doi.org/10.1093/restud/rdw052, https://academic.oup.com/restud/article-lookup/doi/10.1093/restud/rdw052
8. Kamenica, E., Gentzkow, M.: Bayesian Persuasion. American Economic Review **101**(6), 2590–2615 (Oct 2011). https://doi.org/10.1257/aer.101.6.2590, https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590
9. Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S.R., Rocktäschel, T., Perez, E.: Debating with More Persuasive LLMs Leads to More Truthful Answers (Jul 2024), http://arxiv.org/abs/2402.06782, arXiv:2402.06782 [cs]
10. Nayebi, A.: Intrinsic Barriers and Practical Pathways for Human-AI Alignment: An Agreement-Based Complexity Analysis (Nov 2025). https://doi.org/10.48550/arXiv.2502.05934, http://arxiv.org/abs/2502.05934, arXiv:2502.05934 [cs]