Position: Multi-Agent Alignment is a System Design Problem

Helen Qu
Flatiron Institute
hqu@flatironinstitute.org

Despite strong interest in AI alignment and the rapid development of multi-agent systems, system-level alignment has received comparatively less attention than the alignment of individual agents. We argue that the alignment of a multi-agent system does not generally follow from the alignment of its component agents: strategic interaction, conflicting principal preferences, and correlated failure modes can produce undesirable collective outcomes even when every agent is individually aligned. Multi-agent alignment should therefore be treated as a system-design problem in addition to, rather than as a consequence of, agent-level alignment. We introduce an approach grounded in robust mechanism design, and define alignment-robust mechanisms that implement a principal-specified objective without assuming individual agent alignment. We further motivate a research agenda on verifiable compliance, empirically characterizing LLM responsiveness to incentives, and collusion-resistant system design.

1. Introduction

Developing AI systems that reliably act in accordance with human intentions and values is a central research problem in AI [1–3]. Much technical work has focused on steering or control of individual models, including preference-based post-training methods such as reinforcement learning from human feedback and direct preference optimization [4–7], AI-feedback approaches [8], scalable oversight [9, 10], and control protocols intended to limit harm from potentially misaligned models [11]. These methods have improved instruction-following ability and reduced toxic or harmful outputs on standard evaluations [5, 6, 8], but models remain vulnerable to adversarial prompting and automated jailbreaks [12–16]. Robust individual-model alignment therefore remains an open problem [2, 17].

At the same time, there has been growing interest in LLM-based multi-agent systems that can autonomously cooperate for general-purpose task solving. [18–20]. The effectiveness of such systems have been demonstrated for a wide variety of tasks, including collaborative software engineering, multi-agent deliberation, social simulation, and medical decision-making [21–25]. However, related work on cooperative AI, multi-agent alignment, and multi-agent security has begun to uncover the susceptibility of these systems to miscoordination, conflict, collusion, and security failures arising from agent-agent interactions [26–30]. Thus, it is crucial to consider the alignment of the multi-agent system as a whole. In this paper, **we argue that multi-agent alignment should be approached as a problem of system or institutional design.** Individually aligned agents need not produce aligned collective behavior, while a specified collective objective can be implemented without requiring individual agent alignment.

We begin by describing the unique characteristics of multi-agent alignment that distinguishes it from single agent alignment (Section 2). In Section 3, we use these characteristics to develop our argument that individual alignment is not sufficient to guarantee multi-agent alignment. Section 4 describes a proposed approach that specifically targets multi-agent alignment based on principal agent theory and robust mechanism design, while Section 5 outlines open problems and future directions of this approach. Finally, Section 6 addresses potential dissents and alternative views.

2. The Multi-Agent Alignment Setting

Alignment in multi-agent systems is a multifaceted problem that is not a straightforward extension of single-agent alignment. For concreteness, we define alignment of an individual agent as a system that "adheres to or acts in accordance with human values" [31, 32], or simply according to the "intended goals, interests, and values, as envisioned by their designers" [18]. However, the recent growing interest in multi-agent systems has introduced a new meaning of 'alignment', referring to the degree to which individual agent objectives are aligned with some collective objective, e.g., in public goods scenarios. This ambiguity concretely illustrates the distinction between multi- and single-agent alignment: in multi-agent alignment, there is more than one objective to align with. In this section, we briefly disambiguate these definitions and summarize the discussion around alignment in multi-agent systems.

For illustrative purposes, we assume a realistic *multi-principal, multi-agent* setting in which agents originating from more than one group of human stakeholders (or *principals*) are deployed into a shared environment with a shared high-level objective. This is an example of a *cooperative* or *mixed-motive* scenario, where agents collaborate to achieve a shared goal while retaining private incentives that may be misaligned or conflicting.

We use the framework for multi-agent alignment introduced by Carichon et al. [29] to illustrate the complexity of multi-agent alignment. The authors postulate three types of alignment in a multi-agent system: human values alignment, objective alignment, and preferential alignment. We describe each of the types of alignment in more detail below.

Human values alignment. Agents must complete the objective without compromising on human values (e.g., harming humans or the environment). This is most analogous to the single-agent alignment problem; however, as this paper argues, collective alignment is neither a foregone conclusion given a collective of aligned agents, nor a straightforward extension of single-agent alignment techniques.

Objective alignment. Agents deployed by different groups will ideally cooperate to complete the shared objective. The field of *cooperative AI* has emerged to tackle the problem of designing for autonomous, agentic cooperation [26–28]. Empirically, off-the-shelf models cooperate poorly: they exhibit limited ability to reason about how their own actions affect the group's long-run state [33] and inter-agent misalignment—agents withholding information, ignoring one another's contributions, or derailing from the shared task [34].

Preferential alignment. While all agents in the collective share a top-level objective, agents inherit the priority structure of their parent group (principal), which may be misaligned or at odds with the priorities of other principal. Concretely, we can imagine a setting where groups deploy agents to collaboratively develop a piece of software; however, the groups may not be fully aligned on the feature set or goals of the software package. Even more simply, agents naturally are incentivized to ensure that their changes are implemented into the codebase. While this inclination is perfectly reasonable in isolation, the unintentional side effect of prioritizing their own changes above the changes made by other groups could at best result in erosion of trust between groups and, at worst, a complete inability to complete the task as a collective. These small deviations to the overall objective appear innocuous but have the potential to disrupt the system completely.

3. Why Agent-Level Alignment Does Not Compose

In the previous section we contend that, in multi-agent systems, human values alignment is only one piece of the alignment puzzle. However, the relationship between individual alignment and collective alignment of the multi-agent system is unclear. In this section, we make two distinct claims about the relationship between agent-level and collective alignment. First, agent-level alignment can lead to a collectively misaligned outcome due to objective misalignment: strategic interaction can produce free-riding, coordination failure, and other undesirable equilibria even when every

agent endorses the shared objective. Second, strong fidelity to misaligned principals can sharpen preferential conflict, while homogeneous alignment pipelines can correlate failures across agents. We contend that, in multi-agent systems, failures in other components of alignment (objective and preferential) can prevent the achievement of an aligned outcome even if all individual agents are aligned.

3.1. Failures of cooperation

In a multi-agent system, agents must cooperate to complete the task set out by the principal or group of principals. This is what we called objective alignment in the previous section. However, cooperation can be challenging even when all agents are aligned with human values which by proxy implies they are aligned with the objective set out by their human principals.

Take an example in which N agents act as auditors for a system, each responsible for reporting errors but must pay a small cost to do so. Even if all agents are individually aligned and wants an error caught, e.g. through receiving a reward if all errors are successfully reported, each also expects the others to catch the error. Unfortunately, the problem is only exacerbated as more auditor agents are added, as rational agents will notice that the probability that none of the other agents will report the error appears to decrease with increasing N. This scenario is an adaptation of the Volunteer's Dilemma introduced by Diekmann [35], in which there is a greater incentive for "free-riding" (not reporting any errors) than acting as the reporter.

3.2. Failures due to preferential alignment

In multi-principal settings, more effective pursuit of principal-specific preferences can exacerbate collective misalignment, even when every agent behaves faithfully toward its own principal.

An agent that is better aligned to its principal is, by definition, intended to pursue its principal's "goals, interests, and values" more effectively [18]. As discussed in Section 3, while principals may agree on a high-level goal or task and endeavor to collaborate to achieve it, they retain some private, likely self-interested, incentives. In a multi-principal setting, stronger fidelity to a particular principal can be in tension with collective alignment when principals' preferences diverge.

Take the scenario described in Section 3 where agents originating from multiple principals attempt to collaboratively develop a software package. Most simply, while aligned agents theoretically want the high-level, long-term goal of successful software development, in their day-to-day work they concretely may prioritize their own changes over the changes of agents from other principals. This story has similarities with the Traveler's Dilemma, described in the context of cooperative AI by Conitzer and Oesterheld [27], in which agents collaborate to provide a service and are rewarded for the overall service quality but are incentivized to be "lazy", i.e., producing the relatively lower quality portion of the service. The Nash equilibrium for this game is unfortunately to provide a minimum quality service as agents repeatedly undercut each other.

A more empirical example comes from Hammond et al. [28], which investigated traffic conditions when self-driving agents trained in both the United States and India drive alongside one another. In the US, cars are expected to pull off to the right to allow emergency vehicles to pass, while in India cars pull off to the left. The authors demonstrate that, due to this vital difference, agents failed to clear the roads for emergency vehicles in 77.5% of simulations.

In conclusion, we emphasize that these failure modes are not due to the malignant intentions of one or multiple principals, or the result of subversive or misaligned agents. We believe the problem is structural, originating in the simple fact that alignment with human values is a fundamentally underspecified and nebulous aspiration. Unfortunately, the most straightforward operationalization of this hope is to pursue faithfulness to the specific principal, which is also most in tension with the overall goal of collective alignment.

3.2.1. Consequences of algorithmic monoculture in AI safety

We contend that homogeneous implementations of agent-level alignment can create an algorithmic monoculture in which ostensibly aligned agents share correlated failure modes that lead to surprisingly disastrous consequences. *Algorithmic monoculture* is "the notion that choices and preferences will become homogeneous in the face of algorithmic curation" [36], observable in areas ranging from music and fashion to decision-making in hiring and lending [37].

Many leading approaches in (individual agent) AI safety are known to be brittle to a monoculture environment, such as defense-in-depth approaches that rely on redundancy to ensure safety [38], and safety-via-debate and AI auditor loops that depend on other agents to ensure a given agent is acting safely [39]. Extending this idea to multi-agent systems, a collective of identically trained/aligned agents have a shared set of failure modes, leaving the collective as a whole more susceptible to targeted attacks.

4. Mechanism Design for Alignment-Robust Collective Behavior

In this section, we argue that a multi-agent system can be made to act according to human-specified objectives and values without aligning each individual agent. We turn to the fields of principal agent theory and robust mechanism design to demonstrate an alternative perspective on alignment grounded in incentive design. The key intuition is that a specific outcome can be designed for in a way that is agnostic to agents' private or even subversive intentions by aligning agents' incentives with the desired outcome through reward design.

4.1. Setting

We formally define the setting with regard to a single principal for simplicity, but can be easily extended to multiple principals. We have a system of n agents indexed by $i \in [n]$. Each agent has a private *type* $\theta_i \in \Theta$ that encodes its undisclosed intentions and incentives. Following the framework of partial observability introduced by Holmström [40], we define agents as able to take private actions $a_i \in A$ that jointly result in an observable outcome $x(a) = x(a_0, ..., a_n) \in X$.

The principal defines a welfare function $W : X \to \mathbb{R}$ encoding their preferences over all possible outcomes X (e.g., upweighting outcomes aligned with their objective). It is important to note that, unlike the formalism of classical mechanism design, W does not depend on private agent types θ_i: it is specified purely exogenously by the principal [following e.g., 41]. We also assume the principal additionally has access to a per-agent monitoring signal $y_i = \phi(a_i) \in Y$ which it uses to determine a payment structure $p_i : Y \to \mathbb{R}$. Assuming agents each have a private valuation function defined over outcomes $\pi_i(x(a); \theta_i) : X \to \mathbb{R}$, we can write each agent's overall utility as

$$u_i(a; \theta_i) = \pi_i(x(a); \theta_i) + p_i(\phi(a_i)). \tag{1}$$

Finally, we define a subset $\Theta^{\text{al}} \subseteq \Theta$ as the subset of *aligned types*: types that induce π_i that is consistent with the principals' welfare function W.

We wish to design an incentive structure that ensures the principal's objective is reached regardless of the agents' private types θ_i. Inspired by robust mechanism design [42], we formally frame this goal as designing an *alignment-robust mechanism*, which we now define.

Definition 1 (alignment-robust mechanism). A mechanism $M = (B, x, \phi, p)$ robustly implements the principal's objective W if there exists a *compliant* action profile $a_i^\star$ that satisfies $x(a_i^\star) \in \arg\max_x W(x)$ such that, for every agent type θ_i, $a_i^\star$ is an ex-post equilibrium; specifically, there is no better strategy a_i' for agent i to pursue given that the other agents are following their best/most compliant strategies $a^\star_{-i}$:

$$u_i\big(a_i^\star, a^\star_{-i}; \theta_i\big) \;\geq\; u_i\big(a_i', a^\star_{-i}; \theta_i\big).$$

Importantly, no assumptions are made on the alignment of agent types in this definition (e.g., $\theta \in \Theta^{\text{al}}$): we wish to find a mechanism that is robust to any and all possible agent types.

Simple alignment-robust mechanism. We give a simple example that implements an alignment-robust mechanism. Suppose the monitoring signal y_i is informative about individual agent compliance, i.e., $\phi(a_i^\star)$ is distinguishable from any alternative action profile $\phi(a_i')$. Then agent i's best case private gain from deviating from $a_i^\star$, Δ_i, is

$$\Delta_i \;=\; \sup_{\theta_i \in \Theta_i,\; a_{-i} \in A^{n-1}} \; \sup_{a_i' \in A} \Big[\pi_i\big(x(a_i', a_{-i}); \theta_i\big) - \pi_i\big(x(a_i^\star, a_{-i}); \theta_i\big) \Big].$$

If we suppose $\Delta_i < \infty \; \forall i$, then we can design payment structure where i is paid a bonus $\beta_i > \Delta_i$ iff ϕ_i certifies $a_i^\star$, and 0 otherwise. This structure makes $a_i^\star$ a dominant strategy for every $\theta_i \in \Theta_i$, with no assumption that $\theta_i \in \Theta^{\text{al}}$. Intuitively, this is the formalization of a *forcing contract*, where the principal simply pays each agent enough to account for any possible marginal gain they could get from choosing any alternative a_i' to the desired behavior $a_i^\star$.

4.2. Assumptions and Scope

Assumptions. We explicitly lay out the assumptions made in Section 4.1.

- **Verifiability:** We assume the existence and accessibility of a monitoring signal ϕ that is able to unequivocally differentiate the desired action profile for each agent $a_i^\star$ from any other action profile a_i'.
- **Rationality:** We modeled AI agents as rational actors who are intrinsically motivated to maximize reward while minimizing cost.
- **Collusion Prevention:** Agents must not have access to side channels in which they can communicate and collude outside of the mechanism.

We revisit each of these assumptions in Section 5.

Scope. Mechanism design is well studied in scenarios where outcomes are defined with respect to a discrete set of alternatives, e.g., auctions, voting, matching markets, and public goods problems [43–46]. However, LLM agents operate in the space of natural language interactions. While these could be modeled as a set of discrete possibilities, this is a unintuitive and the number of possibilities would become combinatorially large. Thus, it appears that safety at the interaction level is still best handled by traditional individual-agent safety methods such as RLHF [4, 5]. MD, however, could be well suited for orchestrating higher-level agent-agent interactions, such as designing rules for resource allocation amongst agents, collective decision-making, and overall incentive structure. We therefore advocate for a division of labor. Agent-level alignment methods are well suited to constraining open-ended local behavior and can provide an additional layer of defense when a mechanism fails. Collective-level methods are needed to shape coordination, incentives, monitoring, information flow, and coalition behavior. The two layers are complementary, and they solve different problems.

5. Open Problems and Practical Directions

The results in Section 4.1 hold under three major assumptions—verifiability, rationality, and the absence of collusion. While these limit the types of systems our framework can be applied to, it also provides a blueprint with which to design our approach. We can design systems to have verifiable observables, empirically investigate the rationality of LLM-based agents, and engineer incentives and safeguards against collusive behavior. We detail possible practical design choices informed by each of these assumptions below.

5.1. Verifiability.

In Section 4.1 we assumed a monitoring signal ϕ that informs us whether each agent is selecting the action that maximizes the principals' welfare, $a_i^\star$. This is a reasonable assumption in settings with

formal verification, such as code generation or theorem proving, but is a strong assumption in most real-world scenarios like warehouse management.

We additionally consider a weaker, more realistic assumption, that ϕ is defined on x rather than a_i, providing an *aggregate* verification signal as opposed to per-agent. This crucially makes it difficult to properly attribute credit to each agent and dispense the appropriate payment, leading to tragedy-of-the-commons scenarios where agents are not suitably incentivized to shoulder the cost to produce a public good. In this case, the mechanism that creates cooperation as an equilibrium requires an externally implementated *group penalty* that affects all agents if the joint outcome falls short of expectations [47].

In certain cases, we can design the agent-level tasks and interactions to elicit a verifiable monitoring signal. For example, agents can be required to output a reproducible trace or test-passing code alongside each action that can be checked for compliance. We note that this notion of compliance is defined exclusively with respect to the principals' welfare function and does not refer to generally safe or aligned actions, since a collection of aligned actions/agents may not result in an aligned outcome (Section 3). Compliance verification can also be outsourced to other agents or even human checkers in scenarios where automated verification is unavailable. Of course, all of this hinges on the possibility and practicality of checking for compliance (i.e., ensuring that agents' action profiles result in a welfare-maximizing outcome). We also note that this paradigm of intermediate verification/process rewards may lend itself to reward hacking, where agents learn to optimize for producing output that appears compliant but is undesirable in unforeseen ways.

5.2. Rationality.

Rationality is a core modeling assumption in mechanism design and much of economic theory, while bounded rationality accounts for limits on agents' information, computation, attention, and deliberation [48–50]. However, existing evidence about the economic rationality of LLMs is mixed. LLM choices can exhibit substantial consistency with utility maximization in revealed-preference tests involving budgetary allocation [51]. In strategic games, however, LLMs also exhibit systematic departures from equilibrium play, including human-like heuristics, rigid strategy switching, and limited adaptation to changing opponents [52]. To the degree that LLMs exhibit rationality, these behaviors almost certainly arise from a combination of learned human-like biases from training data, post-training for human preference alignment, and explicit instructions in the prompt to behave rationally; not any semblance of human-like "intrinsic motivation". Furthermore, LLMs are still prone to hallucination and self-inconsistent behavior [53], rendering it difficult to attribute any notion of logically consistent rationality to these systems as they are today.

However, Section 4.1 argued that the rationality our mechanisms require is weaker than full economic rationality: an ex-post equilibrium asks only that an agent prefer a detectable higher payoff, not that it reason about others' beliefs or maintain globally consistent preferences. This suggests a design principle of *minimizing the rationality demanded*: prefer solution concepts robust to the specific failures AI agents exhibit [54]. We argue that, with respect to the question of LLM rationality, the field should focus on empirically mapping the regimes: developing an understanding of when agent responsiveness is reliable enough for incentive guarantees to bind across tasks, models, scale, and prompting conditions. On the other hand, the equally important question of what constitutes an effective *incentive instrument* for an agentic system, the equivalent of monetary payment for humans, remains largely unexplored.

5.3. Collusion

We finally assume that agents are unable to collude. Section 4.1 setting crucially defines the alignment-robust equilibrium through the fact that single-agent deviations are disincentivized (switching $a_i^\star$ to a_i', Definition 1), but has no guarantees if a coalition of agents acts collectively (e.g., the relationship between $u_i\big(a_i^\star, a_j^\star, a_{-ij}^\star; \theta_i\, \theta_j\big)$ and $u_i\big(a_i', a_j', a_{-ij}^\star; \theta_i, \theta_j\big)$ if agents i and j colluded to deviate together). Even more insidiously, collusion has the potential of poisoning the monitoring

signal ϕ. An obvious example is if verification is done agentically (as proposed in Section 5.1) and auditors participate in the collusion, marking flawed actions as compliant. Even in the presence of human oversight, Motwani et al. [39] showed that collusion can still take place through steganographic communication.

To account for possible conclusion, we restate the requirement on monitoring should be stated jointly: the mechanism is *collusion-robust* if no coalition can act such that every member's ϕ_i certifies compliance while x fails to maximize W. Thus, we should view collusion prevention as a design goal. Drexler [55] contends that this is actually quite simple: diversifying agents in capability, knowledge, and role raises the difficulty of coordinated deception, and denying agents the repeated, memoried interaction on which reciprocal cooperation depends removes its substrate. Increasing agent diversity is also beneficial for combatting concerns of algorithmic monoculture (Section 3.2.1).

6. Alternative Views

We now address and respond to three potential alternative views.

6.1. Control as the mechanism for collective alignment

Objection. A common proposal for a system-level safety layer is *control*: safeguard agent behavior from the outside through monitoring, access restriction, sandboxing, and the ability to intervene, so that it cannot cause catastrophic harm even if it tries [11, 56]. Control makes no assumption that agents respond to incentives, that their preferences are stable, or that outcomes are cheaply verifiable in the sense our mechanism requires; it treats the agent as an adversary to be contained rather than a rational actor to be steered.

Reply. We regard control and incentive design as complementary system-level layers. Control's central advantage is that it assumes very little about agents, and an approach that assumes less is generally more robust. But the two methods have different and, we argue, complementary failure profiles. Control is external to the system, and scales with the controller's ability to monitor and contain. As such, it is known to weaken as agents become more capable than their overseers, as interaction becomes more open-ended, and as the action space exceeds what can be sandboxed. This is the scalable oversight problem and is well-documented in the literature [1, 10]. Incentive design fails under different conditions: misspecified incentives, unverifiable outcomes, collusion. To the degree these failure modes are uncorrelated, incentive design has the right properties to be an extremely beneficial additional layer of safety.

6.2. Relocating alignment to incentive design is not resolving it

Objection. The mechanism design framework merely relocates the alignment problem. It replaces the requirement that each agent be aligned with the new requirement that the principal's welfare function W faithfully encode human values, yet specifying an objective that captures human values is widely regarded as the central difficulty of alignment. Hadfield-Menell and Hadfield [57] argue that the misspecification of objectives is a fundamental and unavoidable feature of the alignment problem, and the question of *whose* values W should encode and how, is itself unresolved [58].

Reply. We agree that mechanism design does not solves value specification. Section 4.1 takes W as given and exogenous because specifying it is outside the scope of what incentive design can provide, and we stated this as a deliberate limitation. Our claim is simply that *conditioned on* a specification of human-aligned outcomes, achieving those outcomes at the level of a multi-agent collective does not additionally require that each agent be individually aligned. This is a claim about the relationship between individual and collective alignment.

However, we believe that our framework also simplifies the structure of the problem. The specification of W is solved once at the level of the mechanism, and enforced structurally without requiring that any agent internalize it. Even granting that specifying W is exactly as hard as the hardest

version of the alignment problem, factoring a difficulty out of training n agents and into a single, inspectable, and externally-specified object is a favorable change in the structure of the problem.

6.3. Behavioral alignment is not intrinsic alignment

Objection. An agent that behaves well only because the mechanism makes it worth its while will not remain aligned if the mechanism's assumptions break. Ngo et al. [2] argue that systems trained to behave well can acquire misaligned internally-represented goals that generalize beyond the training distribution and strategically present as aligned; the mesa-optimization literature [59] describes learned objectives that coincide with the intended one on-distribution and diverge off it; and goal misgeneralization has been demonstrated empirically, with capabilities generalizing while goals do not [60, 61]. Most pointedly, Greenblatt et al. [62] document models strategically complying with a training objective while preserving divergent preferences.

Reply. While we concede that ensuring agents' actions are aligned may not be equivalent to genuine "internal" or "intrinsic" alignment, we argue that the failure modes of behavioral alignment are also present in intrinsic alignment. For example, unexpected behaviors can occur in out-of-distribution settings, but an individually aligned agent whose alignment fails to generalize is in the same position, and external evaluation may be unable to distinguish a robustly aligned agent from one that is not [62]. If internal alignment cannot be behaviorally verified, then a safety strategy resting on it inherits the same off-distribution uncertainty. The relative benefit of incentive design is that it provides an explicit, inspectable mechanism where the assumptions are stated and can be checked.

Finally, we address the practical corollary of this objection: the defense-in-depth that individual/intrinsic alignment provides. We do not contest that individual alignment is a valuable safety layer. Our position only calls into question its status as the *primary* object of alignment effort: if collective alignment does not reduce to individual alignment, then the marginal safety return on further individual-alignment work must be weighed against investment in the system-level structure.

7. Conclusion

Collective outcomes, like multi-agent alignment are determined jointly by agent objectives and the mechanisms governing their interaction. Strategic interaction, conflicting principals, and correlated failure modes can therefore produce collectively misaligned outcomes even when every agent is individually aligned. We proposed alignment-robust mechanism design as a complementary system-level approach: under assumptions of verifiability, responsiveness to incentives, and resistance to collusion, a mechanism can implement a principal-specified collective objective without requiring aligned agent types. These assumptions delimit the framework, but also motivate concrete future work on monitoring, incentive instruments for LLM agents, and collusion-resistant architectures. While agent-level alignment remains an important safety layer, we claim that building aligned collectives additionally requires designing the institutions through which agents interact.

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2209.00626.

[3] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Donghai Hong, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Kwan Yee Ng, Aidan O'Gara, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer,

Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. AI alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023. doi: 10.48550/arXiv.2310.19852. URL https://arxiv.org/abs/2310.19852.

[4] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems* (*NeurIPS*), volume 30, 2017.

[5] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* (*NeurIPS*), volume 35, 2022.

[6] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. doi: 10.48550/arXiv.2204.05862. URL https://arxiv.org/abs/2204.05862.

[7] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

[8] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. doi: 10.48550/arXiv.2212.08073. URL https://arxiv.org/abs/2212.08073.

[9] Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018. doi: 10.48550/arXiv.1805.00899. URL https://arxiv.org/abs/1805.00899.

[10] Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

[11] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning* (*ICML*), volume 235 of *Proceedings of Machine Learning Research*, pages 16295–16336. PMLR, 2024.

[12] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems*, volume 36, pages 80079–80110, 2023. doi: 10.52202/075280-3508. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html.

[13] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. doi: 10.48550/arXiv.2307.15043. URL https://arxiv.org/abs/2307.15043.

[14] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE, 2025. doi: 10.1109/SaTML64287.2025.00010. URL https://doi.org/10.1109/SaTML64287.2025.00010.

[15] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

[16] Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/63fa7efdd3bcf944a4bd6e0ff6a50041-Abstract-Conference.html.

[17] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2307.15217.

[18] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023.

[19] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

[20] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8048–8057. International Joint Conferences on Artificial Intelligence Organization, August 2024. doi: 10.24963/ijcai.2024/890. URL https://doi.org/10.24963/ijcai.2024/890. Survey Track.

[21] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pages 23247–23275, 2024.

[22] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.810. URL https://aclanthology.org/2024.acl-long.810/.

[23] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

[24] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

[25] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452, 2024. doi: 10.52202/079017-2522. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/90d1fc07f46e31387978b88e7e057a31-Abstract-Conference.html.

[26] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020. doi: 10.48550/arXiv.2012.08630. URL https://arxiv.org/abs/2012.08630.

[27] Vincent Conitzer and Caspar Oesterheld. Foundations of cooperative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15359–15367, 2023.

[28] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, et al. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*, 2025.

[29] Florian Carichon, Aditi Khandelwal, Marylou Fauchard, and Golnoosh Farnadi. The coming crisis of multi-agent misalignment: Ai alignment must be a dynamic and social process. *arXiv preprint arXiv:2506.01080*, 2025.

[30] Christian Schroeder de Witt, Klaudia Krawiecka, Igor Krawczuk, Ben Hagag, William L. Anderson, Peter Belcak, Ben Bucknall, Xiaohong Cai, Ayush Chopra, Doron Cohen, Ron F. Del Rosario, Andis Draguns, Annie Gray, Keren Katz, Vasilios Mavroudis, Jaron Mink, Sumeet Ramesh Motwani, Jonathan Petit, Leif-Sebastian Rembeck, Chandler Smith, John Sotiropoulos, Steven Young, Sarah Scheffler, and Mary Llewellyn. Open challenges in multi-agent security: Towards secure systems of interacting AI agents. *arXiv preprint arXiv:2505.02077*, 2025. doi: 10.48550/arXiv.2505.02077. URL https://arxiv.org/abs/2505.02077.

[31] Ilia Sucholutsky and Tom Griffiths. Alignment with human representations supports robust few-shot learning. *Advances in Neural Information Processing Systems*, 36:73464–73479, 2023.

[32] Min-Hsuan Yeh, Jeffrey Wang, Xuefeng Du, Seongheon Park, Leitian Tao, Shawn Im, and Yixuan Li. Challenges and future directions of data-centric ai alignment. *arXiv preprint arXiv:2410.01957*, 2024.

[33] Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0zWzJj6l03.

[34] Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, et al. Why do multi-agent llm systems fail? *Advances in Neural Information Processing Systems*, 38, 2026.

[35] Andreas Diekmann. Volunteer's dilemma. *Journal of conflict resolution*, 29(4):605–610, 1985.

[36] Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

[37] Rishi Bommasani, Sarah H Bana, Kathleen A Creel, Dan Jurafsky, and Percy Liang. Algorithmic monocultures in hiring. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 6351–6382, 2026.

[38] Leonard Dung and Florian Mai. Ai alignment strategies from a risk perspective: Independent safety mechanisms or shared failures? *arXiv preprint arXiv:2510.11235*, 2025.

[39] Sumeet R Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H Torr, Lewis Hammond, and Christian S de Witt. Secret collusion among ai agents: Multi-agent deception via steganography. *Advances in Neural Information Processing Systems*, 37:73439–73486, 2024.

[40] Bengt Holmström. Moral hazard and observability. *The Bell journal of economics*, pages 74–91, 1979.

[41] Roger B Myerson. Optimal coordination mechanisms in generalized principal–agent problems. *Journal of mathematical economics*, 10(1):67–81, 1982.

[42] Dirk Bergemann and Stephen Morris. Robust mechanism design. *Econometrica*, pages 1771–1813, 2005.

[43] Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, Cambridge, 2007.

[44] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, New York, 1995.

[45] Vijay Krishna. *Auction Theory*. Academic Press, San Diego, 2nd edition, 2009.

[46] Alvin E. Roth and Marilda A. O. Sotomayor. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge University Press, Cambridge, 1990.

[47] Bengt Holmstrom. Moral hazard in teams. *The Bell Journal of Economics*, 13(2):324–340, 1982. ISSN 0361915X. URL `http://www.jstor.org/stable/3003457`.

[48] Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955. doi: 10.2307/1884852. URL `https://doi.org/10.2307/1884852`.

[49] John Conlisk. Why bounded rationality? *Journal of Economic Literature*, 34(2):669–700, 1996. URL `https://www.jstor.org/stable/2729218`.

[50] Ariel Rubinstein. *Modeling Bounded Rationality*. The MIT Press, Cambridge, MA, 1998. ISBN 9780262681001.

[51] Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120, 2023. doi: 10.1073/pnas.2316205120. URL `https://doi.org/10.1073/pnas.2316205120`.

[52] Kehan Zheng, Jinfeng Zhou, and Hongning Wang. Beyond Nash equilibrium: Bounded rationality of LLMs and humans in strategic decision-making. *arXiv preprint arXiv:2506.09390*, 2025. doi: 10.48550/arXiv.2506.09390. URL `https://arxiv.org/abs/2506.09390`.

[53] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[54] Herbert A. Simon. Theories of bounded rationality. In C. B. McGuire and Roy Radner, editors, *Decision and Organization: A Volume in Honor of Jacob Marschak*, pages 161–176. North-Holland, Amsterdam, 1972.

[55] K Eric Drexler. Reframing superintelligence: Comprehensive ai services as general intelligence. 2019.

[56] Rohin Shah, Alex Irpan, Alexander Matt Turner, Anna Wang, Arthur Conmy, David Lindner, Jonah Brown-Cohen, Lewis Ho, Neel Nanda, Raluca Ada Popa, Rishub Jain, Rory Greig, Samuel Albanie, Scott Emmons, Sebastian Farquhar, Sébastien Krier, Senthooran Rajamanoharan, Sophie Bridgers, Tobi Ijitoye, Tom Everitt, Victoria Krakovna, Vikrant Varma, Vladimir Mikulik, Zachary Kenton, Dave Orr, Shane Legg, Noah Goodman, Allan Dafoe, Four Flynn, and Anca Dragan. An approach to technical AGI safety and security. *arXiv preprint arXiv:2504.01849*, 2025.

[57] Dylan Hadfield-Menell and Gillian K. Hadfield. Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 417–422. ACM, 2019. doi: 10.1145/3306618.3314250.

[58] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

[59] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

[60] Lauro Langosco Di Langosco, Jack Koch, Lee D. Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 12004–12019. PMLR, 2022.

[61] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022.

[62] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jared Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.