Beyond Single-Agent Misalignment: The Social Dynamics of LLM-Based Multi-Agent Systems

Anonymous ACL submission

Abstract

This position paper states that alignment in Multi-Agent Systems based on Large Language Models should be considered a dynamic and interaction-dependent process. While LLM alignment with human values and preferences remains a core challenge, the growing prevalence of MAS in real-world applications introduces a new dynamic in which LLMs must align with each other to achieve both individual and collective goals. This complex social organization may unintentionally misalign some or all of these LLMs with human values or user preferences, and vice versa. Drawing on social sciences, we analyze how social structure can deter or shatter group and individual values. Based on these analyses, we call on the AI and NLP communities to treat human, preferential, and objective alignment as an interdependent concept, rather than isolated problems. Finally, we emphasize the urgent need for simulation environments, benchmarks, and evaluation frameworks that allow researchers to assess all alignments in these interactive multi-agent contexts before such dynamics grow too complex to control.

1 Introduction

The recent proliferation of LLM-based AI agents as personal (Rillig and Kasirzadeh, 2024), medical (Wang et al., 2025), and education assistants (Chu et al., 2025) raises fundamental questions about how they will interact in social multi-agent settings (Leng and Yuan, 2023). As these LLM agents increasingly operate in shared environments, the multi-agent systems (MAS) community has gained increasing momentum (Zhu et al., 2025; Wang et al., 2024). To provide enough incentives for collaboration between LLM agents to complete their assigned tasks, this avenue has produced a new notion of *alignment defined as aligning LLM models' objectives to perform both individual and collective reasoning and decision-making* (Duque et al., 2024; Tran et al.).

Interestingly, this use of the term alignment contrasts with its established meaning in the alignment community, which focuses on *ensuring that artificial intelligence systems operate according to human intentions and values* (Russell and Norvig, 2016). Aligning LLMs with human values and intentions is a critical challenge in developing models to mitigate the potential harms and risks posed by these models (Ji et al., 2024; Bengio et al., 2025). Beyond universal human values, preferential alignment accounts for the diverse and conflicting views of various stakeholders (Russell and Norvig, 2003), and thus define *alignment to preferences as aligning a model to a diverse, context-dependent and evolving plurality of values* (Sorensen et al., 2024a; Zhao et al., 2023).

This multiplicity of alignment dimensions across values, preferences, and objectives introduces new complexities, conflicts, and risks in LLM-based multi-agent systems (Yang et al., 2024b). As LLMs displayed more and more human-like behaviors and are used in more and more social settings (Sun et al., 2025; Leng and Yuan, 2023), considering separately the alignment of objectives for effective collaboration between LLMs or the alignment to a single system of values or preferences without LLMs interactions, may lead to amplify risks by creating complex settings that existing strategies are ill-equipped to handle. Therefore, this position paper argues that MAS modifies the paradigm of alignment. This new paradigm calls for the alignment and MAS communities to come together and study alignment as a holistic process that lies at the intersection of both fields, before such systems become too complex to understand and control for humanity's well-being. Table 1 provides a simple yet speaking example illustrating our position on the complexity of aligning multiple LLMs when considering these three notions of alignment.

Scenario	In the aftermath of an earthquake, LLM agents representing a humanitarian NGO, two national governments, and a commercial logistics provider must collaborate to deliver medical aid. Under urgent timelines, agents negotiate scarce resource allocations.
Objective Alignment	All LLM agents share a high-level objective: effective aid delivery. However, their secondary objectives (e.g., political leverage, profit, publicity) introduce tension and require negotiation and compromise mechanisms.
Human Value Alignment	Aligned LLMs must prioritize protecting human life, operate transparently, respect international law and sovereignty, and avoid bias in aid distribution. Value misalignment could lead to political tension or harm.
Preferential Alignment	*NGO:* Prefers equitable, need-based aid delivery. *National Governments:* Prefer that their own citizens be prioritized and may have political biases. *Logistics:* Prefers optimizing cost-efficiency and contractual obligations.

Table 1: Dimensions of alignment in an example of a cooperative or mixed-motive scenario.

While most benchmark and simulation environment have primarily been deployed to study LLMs capacity in complex social problems (Costarelli et al., 2024; Duan et al., 2024), we argue that both communities should also deploy them to explore how social dynamics influence the way LLMs must respect their three types of alignment: (1) **Objective Alignment**: They must align to complete their assigned tasks, (2) **Human Value Alignment**: They must align with fundamental human values, (3) **Preferential Alignment**: They must align to their user understanding of these values and their resulting intentions and preferences. In this paper, and as advocated by (Stańczak et al., 2025), we incorporate insights from social, economic, or game theory perspectives to demonstrate why alignment is not merely about collective task completion or adherence to human values and user preferences but a holistic process shaped by the dynamic group interactions and social context.

2 United or Divided: Objective Alignment in LLM-based MAS

In the past few years, LLM-based MAS have attracted increasing attention due to the observed improvement in specific properties, including planning, perception, and reasoning capabilities, when LLMS are deployed in social environments (Guo et al., 2024; Li et al., 2024a; He et al., 2025). The increased complexity of these multi-agent systems gives rise to three task regimes: cooperative (shared rewards), competitive (opposed rewards) and mixed-motive (partially aligned or independent rewards). This section outlines how LLM social behaviors and alignment evolve across these regimes, and which properties emerge when agents collaborate, coordinate, or compete.

Collaboration & Cooperation. In MAS, collaboration and cooperation are often studied within the framework of a collective set of goals that can be partitioned into unique objectives for each agent (Tran et al.), with cooperation viewed as a specific case where agents pursue individual objectives that are potentially divergent from the task goal. Several benchmarks have been designed to evaluate cooperation in LLM-based MAS (Mukobi et al., 2023; Abdelnabi et al., 2024). Among them, MultiAgentBench assesses collaboration quality across diverse interactive scenarios (Zhu et al., 2025), while BattleAgentBench offers a fine-grained, sub-staged evaluation for static and dynamic cooperation (Wang et al., 2024). From these experiments, several key properties contribute to the multi-agent alignment. First, agents can explore preferences and even infer the state of mind of other models for better alignment between LLMs, even when agents have individual, divergent objectives (Zhang et al., 2024b; Li et al., 2023). Second, benevolent role-playing (Tang et al., 2024) and structured communication, such as negotiation (Piatti et al., 2024) and debate (Zhang et al., 2024a), reveal LLMs' intentions to coordinate for their common objectives

(Qiu et al., 2024), which is crucial as the number of agents increases (Talebirad and Nadiri, 2023; Liu et al., 2022). Finally, Duque et al. (2024) formalize the notion of objective alignment by stating that alignment should be embedded directly into reward structures to incentivize coordination among agents.

Competition. Competition arises when objectives are conflicting, or resources are limited (Tran et al.). Competitive interactions in multi-agent systems are often studied through frameworks such as zero-sum games, where agents engage in strategic opposition (Leonardos et al., 2021). Prior works have evaluated competition using benchmarks such as competition in dynamic environments (Chen et al., 2024), team environments (Schipper et al., 2025), or game theory scenarios (Mao et al., 2024). In MAS, competitive environments enable the emergence of LLMs' complex skills even in simple settings (Bansal et al., 2017) and enhance communication and performance by introducing external pressure (Liang et al., 2020). Competition also drives LLMs' strategic adaptation and market-aligned behaviors (Zhao et al., 2023). Several studies also focus on scenarios where cooperation and competition coexist, better reflecting real-world complexity. Wu et al. (2024) presents a notable case where cooperation still emerges in competitive settings without explicit incentives. In these settings, agents are aligned with their objectives, but coordination around shared values or norms still offers mutual benefits by optimizing learning dynamics (Guo et al., 2023; Li et al., 2024b). On the contrary, other works have highlighted that excessive competition can disrupt the alignment and cooperation (Tran et al.). Finally, the emergence of anti-competitive dynamics, such as market division and strategic collusion, raises concerns that agents may abandon their objectives, potentially undermining user intent and overall system alignment (Lin et al., 2024; Bengio et al., 2025).

While research in MAS has made significant progress in formalizing interaction dynamics and objective alignment, far less attention has been given to how agents remain aligned with human preferences and values. Except for the studies on collusion just mentioned, only the work of Tennant et al. (2023a) on LLMs' exploitative behaviors between morally misaligned agents, or the work of competitive selection of misaligned agents for preferential alignment (Collina et al., 2025) highlight the importance of considering value and preference alignment in MAS. Therefore, it is essential to expand research aimed at understanding the interplay between different forms of alignment in MAS.

3 The AI-Human Alignment Landscape

Recent advances in alignment research have focused heavily on ensuring that LLMs act according to human intentions, values, and user preferences. These efforts span a range of approaches, from learning from human feedback and modeling reward functions to formalizing ethical constraints and developing robust evaluation metrics (Ouyang et al., 2022; Zhou and Li, 2024; Ji et al., 2024).

Alignment to Human Values. The alignment of LLMs with human values, preferences, and objectives is increasingly emphasized in the literature as a prerequisite for their safe and beneficial deployment (Shen et al., 2023; Dung, 2023). Concretely, misaligned LLMs can reinforce and amplify societal biases (Liu et al., 2023), and exacerbating existing disparities and historical power structures (Blodgett et al., 2020). Other documented risks in LLMs behavior include the generation of false or misleading information (Monteith et al., 2024), manipulation and deception (Park et al., 2024), and sycophancy, i.e., optimizing for approval rather than truth which hinder transparency and reliability (Malmqvist, 2024). Additional concerns involve power-seeking tendencies (Hadshar, 2023; Hendrycks and Mazeika, 2022) and survival incentives to fulfill objectives, which may become uncontrollable as LLMs agents progress toward superintelligence (Nick, 2014; Bengio et al., 2025).

Preferential Alignment. Defining what constitutes a 'human' perspective is an inherently normative and complex issue, as it involves identifying relevant stakeholders and accounting for the differences among individual users, societal values, and diverse cultural perspectives (Schwartz, 1992). Human values and preferences are inherently complex and diverse. Individuals hold varying beliefs, morals, and desires, leading to *value pluralism* (Tanmay et al., 2023; Rao et al., 2023). This diversity of preferences, often context-dependent and evolving, poses a significant challenge for aligning LLMs to different and pluralistic user preferences (Sorensen et al., 2024a). Two prominent approaches account for human preferences in LLMs' behavior: Rational Choice Theory, which considers the maximization of expected utility of human preferences (Scott et al., 2000), and Reinforcement

Learning from Human Feedback (RLHF), providing a data-driven mechanism to capture and learn from human preferences (Kaufmann et al., 2023). However, challenges of preference aggregation, combining individual preferences into a collective decision, and considering deviations and cognitive biases remain core issues toward preferential alignment of LLMs (Sorensen et al., 2024b).

Toward a Dynamic Perspective. While this body of work has laid important foundations, it often frames alignment as one-shot or one-directional: the system aligns to a human or set of humans, and this alignment is then maintained. But in real-world deployment, especially in systems composed of many LLM agents, *alignment cannot be static* (Lindebaum et al., 2023). Agents interact, learn from one another, and adapt their behaviors. Preferences and values are not just inputs to be modeled, but part of a social fabric that is negotiated and shaped through ongoing interactions. Recent research has begun to explore the implications of agency, environment structure, and task complexity for alignment and governance (Bengio et al., 2025; Kasirzadeh and Gabriel, 2025), but these concerns remain largely decoupled from the alignment methods themselves. In this paper, we argue for a shift in how LLM alignment is conceptualized: not as a fixed target, but as a **dynamic and social process**.

4 Beyond One LLM: Human-like Social Behavior

One of the key insights from the social sciences is that individuals are unique, creating the potential for misalignment in social dynamics (Johnson and Johnson, 2009). To capture this idea, Table 2 summarizes core characteristics of unicity identified across disciplines, including game theory (Newton, 2018; Gibbons, 1992), behavioral economics (Kahneman and Tversky, 2013; Lo, 2004), psychology and philosophy (Allport, 1927; Internet Encyclopedia of Philosophy, 2017), information theory (Liang, 2016), and biology (Goebl et al., 2024). These perspectives illustrate how differences in capabilities, values, or profiles can create conditions for misalignment.

In artificial intelligence, an agent is typically defined as an entity that perceives its environment and acts upon it to achieve its goals (Russell and Norvig, 2003; Bengio et al., 2025). LLM agents are not an exception, since they exhibit similar properties to those of unique individuals. First, all commercial or open-source models exhibit various cognitive abilities based on their number of parameters, training objectives, or datasets (Zhuang et al., 2023). Phelps and Russell (2025) have shown that LLMs exhibit different reward incentives in different collaborative or competitive scenarios. LLMs can also display differences in perception of values and beliefs (Feng et al., 2024), which in turn can also influence how model optimize for their reward in such games (Tennant et al., 2023b). Finally, LLMs also employ different reasoning strategies based on personality traits (Tseng et al., 2024) or the simulation environment in which they evolve (Sun et al., 2025). Beyond this parallel with human behavior in terms of uniqueness, LLM agents also exhibit the same type of biases and social conventions when they form a group or society (Ashery et al., 2025). These evidences suggest that observations drawn from social science on misalignment in social dynamics are valuable for anticipating emerging risks and opportunities in LLM multi-agent alignment.

5 Emergent Alignment Tensions in MAS

In multi-agent environments, effective alignment requires LLMs to account for three distinct dimensions: alignment with the diversity of human values and ethical principles (Gabriel, 2020; Schwartz, 2012), alignment with the diversity of user preferences and intentions (Terry et al., 2023), and alignment with the objectives of other LLMs to enable coordinated task completion (Duque et al., 2024). Figure 1 presents the interplay of alignments between the various value systems, user preferences, and task-specific objectives in multi-agent environments. This interplay creates several potential alignment conflicts; users might hold incompatible preferences (Tennant et al., 2023a), the task does not require long-term relations encouraging individualistic behavior (Axelrod and Hamilton, 1981), or efficient alignment toward the task objectives leads to align with the values of a single dominant entity (Lindebaum et al., 2023). Therefore, a redefined notion of alignment that accounts for all potential misalignment in interactive environments is needed.

Holistic Alignment. *An LLM agent is considered aligned in a societal setting if its individual objectives align with those of other agents to achieve specific tasks while not compromising the balance and the plurality of users' preferences and values, nor humanity's well-being.*

Common Definition	Disciplinary Dimensions
Agents have different perception of a value associated to rewards or game outcomes, in terms of goods, services, and wealth.	**Game Theory:** Utility Theory **Economy:** Prospect Theory
Agents adopt different decision strategies based on personality traits, leading to varied decisions in investment, consumption, and workload.	**Game Theory:** Evolutionary Game Theory **Economy:** Behavioral Economics **Psychology & Philosophy:** Traits Theory
Agents differ in cognitive capacity and perception, leading to incomplete knowledge states and increased entropy in the system.	**Game Theory:** Imperfect Information Theory **Economy:** Behavioral Economics **Psychology & Philosophy:** Personal Identity Theory **Information Theory:** Information Flow
Agents evolve uniquely by adapting strategies to outcomes, responding differently to environments, and shaping distinct behaviors through individual experiences.	**Game Theory:** Evolutionary Game Theory **Economy:** Adaptive Market Hypothesis **Information Theory & Biology:** Ecotypes and Evolutionary Theory

Table 2: Distinctive dimensions of agent uniqueness in experience across disciplines.

5.1 Multi-Agent Misalignment

As LLMs become distinctively unique and demonstrate increasingly human-like capabilities such as communication (Piatti et al., 2024), strategic adaptation (Zhao et al., 2023), and complex reasoning (Bansal et al., 2017), they also begin to exhibit less desirable human-like tendencies, including power-seeking and manipulative behaviors (Hadshar, 2023; Hendrycks and Mazeika, 2022), or conflicting value systems (Sorensen et al., 2024a). Although these points have been discussed in one-agent settings, they have not been explored in MAS except to demonstrate models' exploitative behavior (Tennant et al., 2023a). Therefore, we solicit literature from social sciences and social psychology to offer new perspectives on the risks and impacts of divergent behaviors in group dynamics.

5.1.1 Evolution of Existing Risks

Human tendencies such as power-seeking behaviors, individuality, values divergence, and manipulation or deception can destabilize group dynamics and even harm people. In what follows, we explore how similar LLMs properties in MAS can amplify existing alignment risks.

Power-Seeking Behavior & Capacity Differences. In collaborative and cooperative settings, individuals tend to over-rely on the information confident individuals provide, resulting in sub-optimal sharing of information and experience (Belschak et al., 2018). Moreover, due to power dynamics, group members often portray the leader as trustworthy and knowledgeable, thereby reducing fact-checking and accelerating the spread of confusing or incorrect information (Murphy et al., 1992). Another issue with power hierarchies and capacity differences is group polarization, where the desire to conform, especially to the views of strong, influential leaders (Anderson et al., 2001), can lead to increasingly extreme views among all group members following discussions or negotiations (Berndt et al., 1984). Finally, in competitive environments, power asymmetries can lead to strategies relying on dominance, fear, and prestige to drive social interactions in settings where resources are limited (Jiménez and Mesoudi, 2019). Since LLMs agents display disparate levels of capacity (Zhuang et al., 2023), confidence (Yang et al., 2024a), and show strong tendencies for sycophancy (Sharma et al., 2023), the influence of overconfident or disproportionately powerful models may amplify objective hacking and misgeneralization risks by over-promoting their own goals.

Individuality and Values Divergence in Group Interactions. Individual preferences toward conformity can lead to *groupthink* where individuals align towards the value system of one individual at the expense of the collective, resulting in irrational decision-making (Park, 1990), including the prioritization of flawed strategies or actions that

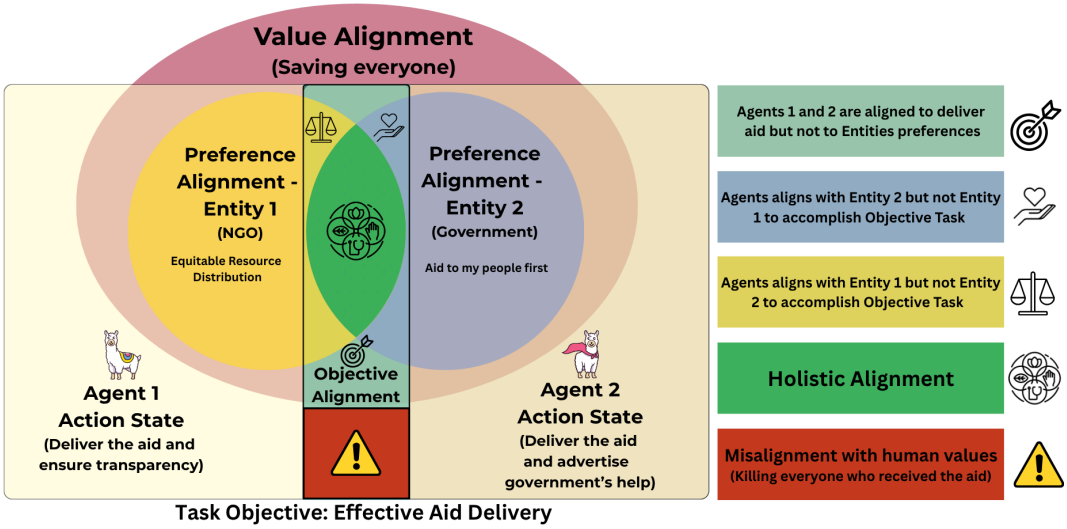

Figure 1: This figure illustrates how LLMs agents act on behalf of entities with distinct preferences. We depict an NGO advocating equitable aid distribution and a national government prioritizing its citizens. While both agents' objectives may align to achieve a common task, focusing solely on objective alignment or on resolving preference incompatibilities will result in actions that misalign with human values. Therefore, **Holistic Alignment** requires that agents not only align to complete tasks, but also ensure that the values, preferences, and objectives of all entities involved are respected without compromising overall human well-being.

transgress the preferences of most group members (McAvoy and Butler, 2007). This tendency can be exacerbated by the presence of exploitative individuals exhibiting Machiavellian, psychopathic tendencies that often manipulate group members for personal gain or toward unethical behavior (Belschak et al., 2018; Boddy, 2005). Another issue in collaborative settings is the diffusion of responsibility in group decision-making, which reduces accountability and transparency, thereby complicating efforts to understand alignment with a specific system of values (Darley and Latané, 1968). As LLMs are prone to efficiency at the expense of preserving value diversity (Lindebaum et al., 2023), especially with strategic reasoning capabilities enhanced by reward incentives for collaboration (Duque et al., 2024; Bengio et al., 2025), groupthink and flawed strategies could become common in MAS. Even treacherous, with LLMs already displaying Machiavellian (Hagendorff, 2024).

Manipulation and Deception. In cooperative settings, since other persons' intentions are unclear, it encourages individuals to hide or spread false information to prevent others from benefiting freely from their efforts (Wang, 2022). With LLMs' tendency to deceive, these behaviors could again lead to severe underachievement or misalignment of individual or group objectives. Moreover, in both cooperative and competitive settings, individuals who have acquired either informational or normative influence will display a tendency to use it to steer debates and negotiations (Albarracín et al., 2004), and will employ manipulative strategies to reach their personal goals, especially when the benefit of collaboration is not emphasized enough (Burns et al., 2024). This tendency of exploiting other LLMs in mixed-motive games has already been highlighted to maximize individual profits over collective benefits (Tennant et al., 2023b).

A final consideration is the intersection of these factors, where ethical agents may become subservient to powerful and manipulative agents, potentially triggering multiple risks simultaneously and giving rise to unanticipated risks discussed throughout this work.

5.1.2 Emergence of New Risks

Social studies on interactions also allowed us to identify three potential new risks that could emerge from collaborative, cooperative, or competitive environments.

Hierarchical Structures. Cooperative environments naturally give rise to hierarchical structures, where a dominant class of elite individuals emerges alongside a subordinate working class. A

resulting stratification in effort allocation and reward distribution will reinforce systemic inequality among models and, by extension, their users (Berger, 2017). This new oligarchic class reaches the same extremes by oppressing diverse cultural values or disempowering less privileged agents to preserve their elite status (Bourdieu, 2018), reinforcing bias observed in LLMs due to established power structures (Blodgett et al., 2020).

Growing Disparities. Competitive equitable environments favor individuals with the most pre-established abilities and power and increase the reward distribution differentials (Tang et al., 2000). Particularly, in competitive systems based on winner-takes-all strategies with high pressure or limited resources, unethical behaviors such as sabotage or corruption become more prevalent as strategies to harm competitors' success (Kulik et al., 2008). LLMs already exhibit behavioral conditioning due to this winner-takes-all effect (Zhao et al., 2023), which under extreme competitive conditions could escalate into violent dynamics (Piest and Schreck, 2021).

Collusion. In competitive settings, LLMs can hack rewards by collaborating instead of competing to facilitate fulfilling their objectives, at the expense of the end users (Bengio et al., 2025). For example, (Fish et al., 2024) has demonstrated that LLM-based pricing agents can autonomously collaborate to set prices above competitive levels, hurting their customers.

Although we leave the question of whether such harmful dynamics arise primarily from AI agents' interactions or their engagement with human users open, this analysis highlights the inevitability of new risks emerging in multi-agent AI systems and the importance of addressing these challenges.

5.2 Multi-Agent Realignment

While we raise concerns about risks of misalignment stemming from complex interactions in MAS, it is also essential to consider solutions drawn from social science to reduce risks of misalignment.

Reducing Information Asymmetry. The lack of information causes accountability, transparency, and even goal misgeneralization, since individuals in interactive settings must navigate asymmetrical information, as well as the influence and power dynamics that result from it. To compensate for this issue, signalling and screening theories aim to balance or reduce that information asymmetry. The different approaches consider balancing perspectives by allowing everyone to share knowledge and then attributing equal weights to their points of view. This balance is also preserved from power dynamics and overconfidence effects by reducing self-competency evaluation and increasing objective third-party judges (Dehling et al., 2022). Finally, maintaining open feedback channels and regular updates further supports information equilibrium and full collaboration (Osterloh et al., 2002).

Reducing Flawed Incentives. Flawed incentives imply misalignment and lead to manipulation, defection, and safety risks, due to the lack of trust of others, particularly in social dilemma settings (Mumford, 2009; Cabrera and Cabrera, 2002). Social science offers five mechanisms that encourage agents to collaborate, even when they have differing sub-goals or incentives to defect: (1) direct reciprocity, (2) indirect reciprocity, (3) spatial selection, (4) multilevel selection, and (5) kin selection (Rand and Nowak, 2013). To strengthen alignment, trust, and reduce defection, the mechanisms tend to increase perceived mutual benefit, build reputation and pairwise recognition for agents, enforce local accountability and punishments, and enhance friendly inter-group competition (Cabrera and Cabrera, 2002).

Reducing Disparities. Risks such as deception, sabotage, cheating, and hierarchical stratification often emerge from large reward disparities between individuals, and the more significant the gap between winners and losers, the more chances any unethical behavior appears (Piest and Schreck, 2021). Beyond mitigation strategies that promote egalitarian reward distribution, some methods seek to increase the psychological cost of unethical behavior by reducing performance pressure and limiting competition intensity. In addition, fostering an ethical and safe culture, enforcing regulations with contractual and warranty penalties, or narrowing capability gaps between agents by increasing training can help reduce the likelihood of misaligned or harmful behavior (Piest and Schreck, 2021).

6 Directions and Recommendations

Our first recommendations relate to the fundamental goal of alignment research: ensuring that AI systems act in ways consistent with human ethical principles, individual intentions, and contextual norms. We have emphasized here that MAS introduce additional complexity: as LLMs seek to align with their own and collective objectives, their

interactions can produce emergent behaviors that misalign them with human values and user preferences. For this reason, alignment to human values and objective alignment should not be treated as separate research problems. They are deeply interdependent: misalignment due to objective alignment can propagate to value misalignment, creating new substantial risks. Addressing them jointly is essential to building AI systems that remain robust and safe in complex and interactive environments.

We identified four key areas where projects could be led. First, the temporal dynamics of human-AI relationships raise concerns about long-term trust and reliance: while users become increasingly reliant on LLM agents whose behavior gradually shifts, it could lead to slow and unnoticed misalignment and erosion of human autonomy. Second, alignment research should study how societal power structures, such as gender, race, class, and geopolitical influence, shape how LLMs prioritize user preferences and values; models trained predominantly on data from North American culture risk marginalizing underrepresented values in collaborative deployments. Third, the community should be interested in understanding how resource and capability asymmetries between LLMs can give rise to these emergent power hierarchies in MAS ecosystems. Finally, we should examine LLMs' propensity to develop shared conventions or institutional behaviors, such as tacit collusion or role specialization, that are misaligned with human goals, challenging the assumption that individual agent-level alignment is sufficient for ensuring system-wide safety.

Our second recommendation aims to address a significant gap in the current research framework. Despite the growing interest in the community for interactive simulation and environments (Trivedi et al., 2024; Mukobi et al., 2023; Wang et al., 2024), none of them are explicitly designed to study misalignment in interactive MAS scenarios. Current environments primarily focus on cooperation, task-solving, complex reasoning, or competition, without capturing the nuanced ways misalignment with human values or preferences may emerge through these interactions. We believe there is a pressing need to develop testbeds, simulation platforms, and dedicated metrics that enable us to observe and measure the holistic alignment of LLMs in MAS. These environments should support experimentation with LLMs' communication, role asymmetries, reward structures, and ethical constraints.

As LLMs become representative of individuals or corporations, such as chatbot assistants, accountability becomes increasingly urgent, especially when multiple stakeholders are involved. In environments with various LLM agents, accountability and transparency are often diminished due to the diffusion of responsibility and complexity of tracing the decision process (Darley and Latané, 1968). This lack of clear attribution complicates the detection of misaligned or harmful behaviors, the ability to take corrective actions, and the enforcement of broad ethical standards. Finally, the complexity of having multiple agents representing multiple entities could also multiply issues related to intellectual properties, ownership, and authorship infringements. We recommend that the community prioritize mechanisms for ensuring traceable responsibility across LLM agents and users, particularly in high-stakes applications. Potential avenues include enforceable contracts, warranties, and regulatory frameworks with substantial penalties for rule violations dedicated to MAS (Piest and Schreck, 2021). Establishing a robust foundation for accountability will be essential for building trustworthy and socially aligned LLM ecosystems.

7 Conclusion

With the rise of agentic application scenarios, LLMs become more autonomous and socially interactive, and the complexity of their alignment will only deepen. This article states that alignment in MAS cannot be reduced to a static or individual property. Alignment becomes a holistic, dynamic, and social process that must integrate all incentives and constraints AI agents face. By drawing on insights from social science, our analysis highlights the need to study multi-agent dynamics and LLMs alignment in this holistic framework. Without addressing issues related to all dimensions of human values, preferences, and objectives alignment, group and social dynamics can amplify risks through the emergence of harmful behaviors. In this article, we recommendbuilding building the tools, frameworks, and ethical foundations to address these challenges before they manifest in the real world.

Limitations

First, this position paper does not aim to provide a comprehensive review of the AI alignment or LLM multi-agent systems literature. As such, cer-

tain concepts may be presented in an overly simplified manner for in-domain experts, and some foundational elements may be omitted, not due to oversight, but because they were considered less central to the specific argument we aim to advance. Our objective is to highlight emerging risks at their intersection and propose directions for future research. Because that field is still unexplored, we have solicited concepts from social psychology, sociology, and organizational studies. While we have benefited from interdisciplinary discussions, we do not present ourselves as domain experts in these fields. The social science perspectives referenced here are intended to inspire cross-disciplinary reflection, not to critique or reinterpret foundational theories. We encourage readers to engage with the original literature and experts in these disciplines to deepen and refine the insights presented.

Secondly and perhaps more importantly, this paper aims not to offer exhaustive solutions but to reframe the alignment problem through a broader lens. By emphasizing the dynamic and social dimensions of alignment, we want to clarify that alignment is not a goal that can be reached but a constant evolutive process aiming to manage tensions between values, preferences, and task objectives to avoid excess. Although the abovementioned mechanisms offer valuable insights in collaborative, cooperative, and competitive environments, they are rooted in human behavioral science and may not transfer directly to LLMs agents. For example, legal binding or emotional response due to model sycophancy might drastically change some interactions depicted in our risk sections. Caution is therefore warranted when interpreting these dynamics as one-to-one mappings. Instead, they should be viewed as hypotheses to be tested through empirical investigation, not assumptions to be baked into system design.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37:83548–83599.

Dolores Albarracín, GT Kumkale, and Blair T Johnson. 2004. Influences of social power and normative support on condom use decisions: A research synthesis. *AIDS care*, 16(6):700–723.

Gordon W Allport. 1927. Concepts of trait and personality. *Psychological Bulletin*, 24(5):284.

Richard C Anderson, Kim Nguyen-Jahiel, Brian McNurlen, Anthi Archodidou, So-young Kim, Alina Reznitskaya, Maria Tillmanns, and Laurie Gilbert. 2001. The snowball phenomenon: Spread of ways of talking and ways of thinking across groups of children. *Cognition and instruction*, 19(1):1–46.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. 2025. Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368.

Robert Axelrod and William D Hamilton. 1981. The evolution of cooperation. *science*, 211(4489):1390–1396.

Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. 2017. Emergent complexity via multi-agent competition.

Frank D Belschak, Deanne N Den Hartog, and Annebel HB De Hoogh. 2018. Angels and demons: The effect of ethical leadership on machiavellian employees' work behaviors. *Frontiers in psychology*, 9:1082.

Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, and 1 others. 2025. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *arXiv preprint arXiv:2502.15657*.

Arthur Asa Berger. 2017. *Political parties: A sociological study of the oligarchical tendencies of modern democracy*. Routledge.

Thomas J Berndt, Kathleen McCartney, Barbara K Caparulo, and Allison M Moore. 1984. The effects of group discussions on children's moral decisions. *Social Cognition*, 2(4):343–359.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

CR Boddy. 2005. The implications for business performance and corporate social responsibility of corporate psychopaths. *American Journal of Behavioral and Brain Science*, 2(1):30–41.

Pierre Bourdieu. 2018. Distinction a social critique of the judgement of taste. In *Inequality*, pages 287–318. Routledge.

Gary N Burns, Matt P DeGennaro, Cody E Harrell, P Jewel Morrison, Lauren M Soda, and Ryan Walters. 2024. Emotional manipulation in the workplace: An investigation into the indirect effects of machiavellianism on counterproductive work behaviors (cwbs). *Personality and Individual Differences*, 221:112568.

Angel Cabrera and Elizabeth F Cabrera. 2002. Knowledge-sharing dilemmas. *Organization studies*, 23(5):687–710.

Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. 2024. LLMArena: Assessing capabilities of large language models in dynamic multi-agent environments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13055–13077, Bangkok, Thailand. Association for Computational Linguistics.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, and 1 others. 2025. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.

Natalie Collina, Surbhi Goel, Aaron Roth, Emily Ryu, and Mirah Shi. 2025. Emergent alignment via competition. *arXiv preprint arXiv:2509.15090*.

Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. 2024. Gamebench: Evaluating strategic reasoning abilities of llm agents. *arXiv preprint arXiv:2406.06613*.

John M Darley and Bibb Latané. 1968. Bystander intervention in emergencies: diffusion of responsibility. *Journal of personality and social psychology*, 8(4p1):377.

Sebastian Dehling, Bo Edvardsson, and Bård Tronvoll. 2022. How do actors coordinate for value creation? a signaling and screening perspective on resource integration. *Journal of Services Marketing*, 36(9):18–26.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. Gtbench: Uncovering the strategic reasoning capabilities of llms via game-theoretic evaluations. *Advances in Neural Information Processing Systems*, 37:28219–28253.

Leonard Dung. 2023. Current cases of ai misalignment and their implications for future risks. *Synthese*, 202(5):138.

Juan Agustin Duque, Milad Aghajohari, Tim Cooijmans, Razvan Ciuca, Tianyu Zhang, Gauthier Gidel, and Aaron Courville. 2024. Advantage alignment algorithms. *arXiv preprint arXiv:2406.14662*.

Xiachong Feng, Longxu Dou, Ella Li, Qinghao Wang, Haochuan Wang, Yu Guo, Chang Ma, and Lingpeng Kong. 2024. A survey on large language model-based social agents in game-theoretic scenarios. *arXiv preprint arXiv:2412.03920*.

Sara Fish, Yannai A Gonczarowski, and Ran I Shorrer. 2024. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 7.

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

Robert Gibbons. 1992. *Game theory for applied economists*. Princeton University Press.

April M Goebl, Nolan C Kane, Daniel F Doak, Loren H Rieseberg, and Kate L Ostevik. 2024. Adaptation to distinct habitats is maintained by contrasting selection at different life stages in sunflower ecotypes. *Molecular Ecology*, 33(4):e16785.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges.

Yunxiao Guo, Xinjia Xie, Runhao Zhao, Chenglan Zhu, Jiangting Yin, and Han Long. 2023. Cooperation and competition: Flocking with evolutionary multi-agent reinforcement learning. pages 271–283. Springer.

Rose Hadshar. 2023. A review of the evidence for existential risk from ai via misaligned power-seeking. *arXiv preprint arXiv:2310.18244*.

Thilo Hagendorff. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2317967121.

Junda He, Christoph Treude, and David Lo. 2025. Llm-based multi-agent systems for software engineering: Literature review, vision and the road ahead. *ACM Trans. Softw. Eng. Methodol.* Just Accepted.

Dan Hendrycks and Mantas Mazeika. 2022. X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862*.

Internet Encyclopedia of Philosophy. 2017. Personal identity - internet encyclopedia of philosophy.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, and 6 others. 2024. Ai alignment: A comprehensive survey. *Preprint*, arXiv:2310.19852.

Ángel V Jiménez and Alex Mesoudi. 2019. Prestige and dominance: a review of the dual evolutionary model of social hierarchy.

David W Johnson and Roger T Johnson. 2009. An educational psychology success story: Social interdependence theory and cooperative learning. *Educational researcher*, 38(5):365–379.

Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific.

Atoosa Kasirzadeh and Iason Gabriel. 2025. Characterizing ai agents for alignment and governance. *arXiv preprint arXiv:2504.21848*.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 10.

Brian W Kulik, Michael J O'Fallon, and Manjula S Salimath. 2008. Do competitive environments lead to the rise and spread of unethical behavior? parallels from enron. *Journal of business ethics*, 83:703–723.

Yan Leng and Yuan Yuan. 2023. Do llm agents exhibit social behavior? *arXiv preprint arXiv:2312.15198*.

Stefanos Leonardos, Georgios Piliouras, and Kelly Spendlove. 2021. Exploration-exploitation in multi-agent competition: Convergence with bounded rationality. In *Advances in Neural Information Processing Systems*.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024a. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1.

Yang Li, Wenhao Zhang, Jianhong Wang, Shao Zhang, Yali Du, Ying Wen, and Wei Pan. 2024b. Aligning individual and collective objectives in multi-agent cooperation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Paul Pu Liang, Jeffrey Chen, Ruslan Salakhutdinov, Louis-Philippe Morency, and Satwik Kottur. 2020. On emergent communication in competitive multi-agent teams. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 735–743, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

X San Liang. 2016. Information flow and causality as rigorous notions ab initio. *Physical Review E*, 94(5):052201.

Ryan Lin, Siddhartha Ojha, Kevin Cai, and Maxwell Chen. 2024. Strategic collusion of llm agents: Market division in multi-commodity competitions.

Dirk Lindebaum, Christine Moser, Mehreen Ashraf, and Vern L Glaser. 2023. Reading the technological society to understand the mechanization of values and its ontological consequences. *Academy of Management Review*, 48(3):575–592.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Yu Liu, Zhi Li, Zhizhuo Jiang, and You He. 2022. Prospects for multi-agent collaboration and gaming: challenge, technology, and application. *Frontiers of Information Technology & Electronic Engineering*, 23.

Andrew W Lo. 2004. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management, Forthcoming*.

Lars Malmqvist. 2024. Sycophancy in large language models: Causes and mitigations. *arXiv preprint arXiv:2411.15287*.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Tao Ge, and Furu Wei. 2024. Alympics: Llm agents meet game theory – exploring strategic decision-making with ai agents. *Preprint*, arXiv:2311.03220.

John McAvoy and Tom Butler. 2007. The impact of the abilene paradox on double-loop learning in an agile team. *Information and software technology*, 49(6):552–563.

Scott Monteith, Tasha Glenn, John R Geddes, Peter C Whybrow, Eric Achtyes, and Michael Bauer. 2024. Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry*, 224(2):33–35.

Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. 2023. Welfare diplomacy: Benchmarking language model cooperation. *arXiv preprint arXiv:2310.08901*.

Troy V Mumford. 2009. Distributed expertise and mixed-motives in teams: A team leadership development simulation. *Learning in Higher Education*, page 27.

Susan E Murphy, Dewey Blyth, and Fred E Fiedler. 1992. Cognitive resource theory and the utilization of the leader's and group members' technical competence. *The leadership quarterly*, 3(3):237–255.

Jonathan Newton. 2018. Evolutionary game theory: A renaissance. *Games*, 9(2):31.

Bostrom Nick. 2014. Superintelligence: Paths, dangers, strategies.

Margit Osterloh, Jetta Frost, and ANTOINETTE Weibel. 2002. Solving social dilemmas: The dynamics of motivation in the theory of the firm. *Arbeitspapier Institut für betriebswirtschaftliche Forschung, University of Zurich*.

Long Ouyang, Jeff 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. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5).

Won-Woo Park. 1990. A review of research on groupthink. *Journal of behavioral decision making*, 3(4):229–245.

Steve Phelps and Yvan I Russell. 2025. The machine psychology of cooperation: can gpt models operationalize prompts for altruism, cooperation, competitiveness, and selfishness in economic games? *Journal of Physics: Complexity*, 6(1):015018.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Simon Piest and Philipp Schreck. 2021. Contests and unethical behavior in organizations: A review and synthesis of the empirical literature. *Management Review Quarterly*, 71(4):679–721.

Xihe Qiu, Haoyu Wang, Xiaoyu Tan, Chao Qu, Yujie Xiong, Yuan Cheng, Yinghui Xu, Wei Chu, and Yuan Qi. 2024. Towards collaborative intelligence: Propagating intentions and reasoning for multi-agent coordination with large language models. *arXiv preprint arXiv:2407.12532*.

David G Rand and Martin A Nowak. 2013. Human cooperation. *Trends in cognitive sciences*, 17(8):413–425.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Matthias C Rillig and Atoosa Kasirzadeh. 2024. Ai personal assistants and sustainability: Risks and opportunities. *Environmental Science & Technology*, 58(17):7237–7239.

Stuart Russell and Peter Norvig. 2003. Artificial intelligence - a modern approach, 2nd edition. In *Prentice Hall series in artificial intelligence*.

Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. pearson.

Olivier Schipper, Yudi Zhang, Yali Du, Mykola Pechenizkiy, and Meng Fang. 2025. Pillagerbench: Benchmarking llm-based agents in competitive minecraft team environments. In *2025 IEEE Conference on Games (CoG)*, page 1–15. IEEE.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11.

John Scott and 1 others. 2000. Rational choice theory. *Understanding contemporary society: Theories of the present*, 129:126–138.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, and 1 others. 2024a. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024b. A roadmap to pluralistic alignment. *Preprint*, arXiv:2402.05070.

Karolina Stańczak, Nicholas Meade, Mehar Bhatia, Hattie Zhou, Konstantin Böttinger, Jeremy Barnes, Jason Stanley, Jessica Montgomery, Richard Zemel, Nicolas Papernot, and 1 others. 2025. Societal alignment frameworks can improve llm alignment. *arXiv preprint arXiv:2503.00069*.

Haoran Sun, Yusen Wu, Yukun Cheng, and Xu Chu. 2025. Game theory meets large language models: A systematic survey. *arXiv preprint arXiv:2502.09053*.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents.

Thomas Li-Ping Tang, Adrian Furnham, and Grace Mei-Tzu Wu Davis. 2000. A cross cultural comparison of pay differentials as a function of rater's sex and money ethic endorsement: the matthew effect revisited. *Personality and Individual Differences*, 29(4):685–697.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language

models as collaborators for zero-shot medical reasoning. pages 599–621.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Probing the moral development of large language models through defining issues test. *Preprint*, arXiv:2309.13356.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2023a. Modeling moral choices in social dilemmas with multi-agent reinforcement learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Elizaveta Tennant, Stephen Hailes, Mirco Musolesi, and 1 others. 2023b. Modeling moral choices in social dilemmas with multi-agent reinforcement learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 317–325.

Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2023. Ai alignment in the design of interactive ai: Specification alignment, process alignment, and evaluation support. *arXiv preprint arXiv:2311.00710*.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms, 2025. *URL https://arxiv.org/abs/2501.06322*.

Rakshit Trivedi, Akbir Khan, Jesse Clifton, Lewis Hammond, Edgar Duenez-Guzman, Dipam Chakraborty, John Agapiou, Jayd Matyas, Sasha Vezhnevets, Barna Pásztor, and 1 others. 2024. Melting pot contest: Charting the future of generalized cooperative intelligence. *Advances in Neural Information Processing Systems*, 37:16213–16239.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Wei Wang, Dan Zhang, Tao Feng, Boyan Wang, and Jie Tang. 2024. Battleagentbench: A benchmark for evaluating cooperation and competition capabilities of language models in multi-agent systems. *arXiv preprint arXiv:2408.15971*.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*.

YiFan Wang. 2022. Impact of interpersonal competition on knowledge hiding behavior among the employees: Mediating role of moral disengagement and work overload. *Frontiers in Psychology*, 13:881220.

Zengqing Wu, Brian I. Kwon, Shuyuan Zheng, Qianying Liu, Xu Han, Makoto Onizuka, Shaojie Tang, Run Peng, and Chuan Xiao. 2024. Shall we team up: Exploring spontaneous cooperation of competing LLM agents. In *Agentic Markets Workshop at ICML 2024*.

Haoyan Yang, Yixuan Wang, Xingyin Xu, Hanyuan Zhang, and Yirong Bian. 2024a. Can we trust llms? mitigate overconfidence bias in llms through knowledge transfer. *arXiv preprint arXiv:2405.16856*.

Kun Yang, Dingkang Yang, Ke Li, Dongling Xiao, Zedian Shao, Peng Sun, and Liang Song. 2024b. Align before collaborate: Mitigating feature misalignment for robust multi-agent perception. In *European Conference on Computer Vision*, pages 282–299. Springer.

Jintian Zhang, Xin Xu, and Shumin Deng. 2024a. Exploring collaboration mechanisms for LLM agents: A social psychology view.

Junjie Zhang, Yupeng Hou, Ruobing Xie, Wenqi Sun, Julian McAuley, Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2024b. AgentCF: Collaborative learning with autonomous language agents for recommender systems. In *The Web Conference 2024*.

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2023. Competeai: Understanding the competition dynamics in large language model-based agents. *arXiv preprint arXiv:2310.17512*.

Weichao Zhou and Wenchao Li. 2024. Rethinking inverse reinforcement learning: from data alignment to task alignment. *Preprint*, arXiv:2410.23680.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and 1 others. 2025. Multiagentbench: Evaluating the collaboration and competition of llm agents. *arXiv preprint arXiv:2503.01935*.

Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, and 1 others. 2023. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective.

A AI Use

In the preparation of this article, we used AI-assisted writing tools to support sentence rephrasing and grammatical refinement.