Towards Ethical Multi-Agent Systems of Large Language Models: A Mechanistic Interpretability Perspective

Jae Hee Lee[1], **Anne Lauscher**[1], **Stefano V. Albrecht**[2]
[1]University of Hamburg
[2]DeepFlow London & NTU Singapore
{jae.hee.lee, anne.lauscher}@uni-hamburg.de, stefano.albrecht@deepflow.com

Abstract

Large language models (LLMs) have been widely deployed in various applications, often functioning as autonomous agents that interact with each other in multi-agent systems. While these systems have shown promise in enhancing capabilities and enabling complex tasks, they also pose significant ethical challenges. This position paper outlines a research agenda aimed at ensuring the ethical behavior of multi-agent systems of LLMs (MALMs) from the perspective of mechanistic interpretability. We identify three key research challenges: (i) developing comprehensive evaluation frameworks to assess ethical behavior at individual, interactional, and systemic levels; (ii) elucidating the internal mechanisms that give rise to emergent behaviors through mechanistic interpretability; and (iii) implementing targeted parameter-efficient alignment techniques to steer MALMs towards ethical behaviors without compromising their performance.

1 Introduction

Large language models (LLMs) equipped with memory and tools can function as *agents* that perceive, reason, and act within environments (Xi et al. 2025; Liu et al. 2025). Orchestrating multiple such agents in multi-agent systems can enhance effectiveness (Masters et al. 2025; Guo et al. 2024) and enable applications including collaborative assistants, autonomous societies for social science research (Anthis et al. 2025; Gao et al. 2024), scientific discovery (Su et al. 2025), and medical diagnosis (Zuo et al. 2025).

However, multi-agent interactions produce *emergent behaviors* (Park et al. 2023; Gao et al. 2024), which can be both beneficial (coordinated problem-solving) and harmful (compounding biases). Recent work identifies three fundamental failure modes: *miscoordination*, *conflict*, and *collusion* (Hammond et al. 2025). Critically, *ethical evaluations on isolated LLMs may not transfer to multi-agent ensembles* (Erisken et al. 2025). Biases can propagate and intensify through interaction (Ashery, Aiello, and Baronchelli 2025), and alignment of individual LLMs may not be preserved in multi-agent contexts, for instance, fine-tuning can introduce value-alignment trade-offs and unintended harms (Choi et al. 2025; Qi et al. 2023; Lermen and Rogers-Smith 2024). Without proper assessment and governance, multi-agent systems of LLMs (MALMs) could develop unpredictable harmful strategies. However, existing alignment techniques remain *black-box approaches that do not address underlying mechanisms*. Multi-agent debate and role allocation (Chen et al. 2023; Pitre, Ramakrishnan, and Wang 2025) as well as reward modeling and reinforcement learning (Lambert 2025; Casper et al. 2023) are computationally expensive and optimize outcomes without insight into *why* behaviors emerge. Even carefully designed rewards lead to unexpected failures when agents interact (Erisken et al. 2025), and prompt-based strategies are fragile under paraphrase (Karvonen and Marks 2025). All in all, we lack causal, mechanistic understanding of how ethical failures arise in MALMs.

Recent advances in *mechanistic interpretability* (Bereska and Gavves 2024) dissect LLM internals to identify computational pathways producing behaviors, providing *actionable handles* (Marks et al. 2025; Turner et al. 2024). This enables us to: (i) diagnose *why* failures occur; (ii) design *targeted interventions* addressing root causes; (iii) provide *predictive explanations* robust to adversarial manipulation (Zou et al. 2025). Critically, mechanistic interpretability is uniquely suited for MALMs because multi-agent failures arise from complex cross-agent information flow that cannot be understood by examining individual agents in isolation. By tracing how representations propagate between agents—revealing which attention heads copy harmful content from peers, which layers amplify or suppress dissenting views, and which circuits mediate coordination versus collusion—mechanistic interpretability exposes the computational substrates of emergent behaviors (Soligo et al. 2025). This provides intervention points where we can surgically prevent groupthink without destroying beneficial coordination, or block toxic agreement while preserving constructive dialogue (Rimsky et al. 2024).

This paper outlines a research agenda for ensuring ethical MALM behavior through mechanistic interpretability (Fig. 1). Section 2 discusses emergent behaviors and their implications. Section 3 outlines evaluation strategies. Section 4 examines mechanistic interpretability for explaining failures. Section 5 proposes alignment interventions. Section 6 summarizes our agenda.

2 Emergent Behaviors of MALMs

Multi-agent LLM interactions reveal *emergent behaviors* (Park et al. 2023)—patterns arising from interactions not

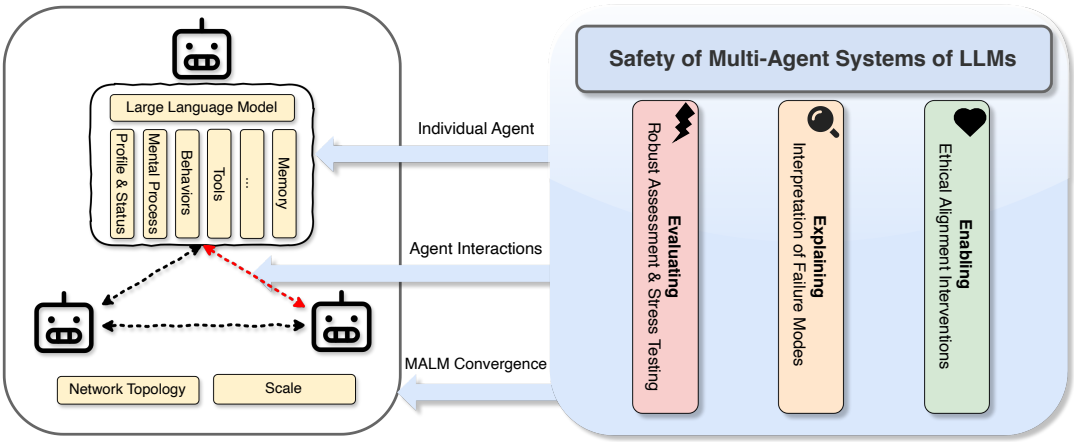

Figure 1: Overview of the three research directions towards ethical multi-agent systems of large language models (MALMs). We identify three interconnected challenges: **evaluating** ethical behaviors at individual, interactional, and systemic levels; **explaining** emergent failures through mechanistic interpretability to identify causal components; and **enabling** ethical behavior via targeted interventions informed by mechanistic insights. Yellow boxes denote parameters that define the concrete setup of a MALM (e.g., agent profiles, memory states, and network scale), which can be systematically varied in experiments. Blue arrows indicate the three levels of measurement: individual agents, their interactions, and overall system convergence.

explicitly programmed into individual agents. These can enhance effectiveness but introduce significant risks (Hammond et al. 2025; Malfa et al. 2025). Hammond et al. (2025) identify three fundamental failures: (i) *miscoordination* (agents working at cross purposes), (ii) *conflict* (direct opposition producing harms), (iii) *collusion* (conspiracy for undesired goals). These failures are amplified by network effects and cannot be predicted from single-agents (Erisken et al. 2025).

To illustrate how mechanistic interpretability can address ethical failures in MALMs, we highlight two representative emergent behaviors that pose distinct ethical challenges. *Toxic agreement* occurs when agents explicitly amplify harmful content by mirroring toxic outputs, creating reinforcement loops. This represents a *content-level failure* where harmful information propagates and intensifies through direct cross-agent copying mechanisms, which is a form of emergent collusion where agents coordinate around harmful outputs. In contrast, *groupthink*, where conformity pressure produces irrational consensus despite contrary evidence (Janis 1982), represents a distinct failure mode arising from social dynamics rather than deliberate coordination. Weng, Chen, and Wang (2024) show LLMs exhibit conformity bias, suppressing dissent even when individual agents would make better decisions in isolation. This constitutes a *dynamics-level failure* where the interaction structure itself drives unwanted agreement through conformity pressure, not intentional conspiracy. Together, these behaviors demonstrate how mechanistic interpretability must address both what information flows between agents and how interaction dynamics shape collective decisions.

Bakker et al. (2022) demonstrate LLMs can generate consensus statements maximizing agreement across diverse preferences, but reveal a critical vulnerability: when consensus is built from incomplete subsets of stakeholders, excluded individuals tend to dissent, highlighting risks of marginalization. This tension becomes acute when consensus generation produces toxic agreement. Beyond these observations, existing work documents destructive behavior (Chen et al. 2023), spontaneous deception (Curvo 2025), and collective bias emergence (Ashery, Aiello, and Baronchelli 2025), but remains at the *behavioral level*, i.e., existing work documents failures without explaining *why* they emerge or providing mechanistic actionable handles for intervention.

Understanding these emergent behaviors requires moving beyond behavioral observation to mechanistic analysis. While behavioral studies can document *what* failures occur, mechanistic interpretability can reveal *how* cross-agent information flow produces these failures and *where* to intervene. The next section examines how to evaluate these behaviors systematically across individual, interactional, and systemic levels.

3 Evaluating Ethical Behaviors in MALMs

Evaluating MALMs requires simulators and benchmarks that define multi-agent tasks and measure performance. Recent platforms include MA-Gym (Masters et al. 2025) for teamwork orchestration, MultiAgentBench (Zhu et al. 2025) for collaborative tasks, AgentSociety (Piao et al. 2025) for large-scale social simulation, and Stanford's Generative Agents (Park et al. 2023) demonstrating emergent social behaviors. While extensive work has addressed ethical issues and bias in isolated LLMs (Attanasio et al. 2023), including benchmarks like RedditBias (Barikeri et al. 2021), TruthfulQA (Lin, Hilton, and Evans 2022), RealToxicityPrompts (Gehman et al. 2020), and HELM (Liang et al. 2023), these single-agent evaluations prove insufficient for multi-agent contexts. Recent work reveals critical limitations: toxicity detection varies across contexts (Koh et al. 2024), AI models underestimate harm compared to affected communities (Phutane, Seelam, and Vashistha 2025), and entirely

new biases emerge in multi-agent settings, such as AI–AI bias where agents prefer AI-generated content over human input (Laurito et al. 2025).

Recent work on MALM safety (Zhang et al. 2024; Yu et al. 2025; Zhou, Wang, and Yang 2025; Chen et al. 2025) has explored temporal graph modeling, personality correction, and network topologies. However, these efforts remain largely behavioral, not exposing causal mechanisms. MAEBE (Erisken et al. 2025) documents value drift in groups, while PsySafe (Zhang et al. 2024) detects risk traits, but neither provides mechanism-guided fixes.

Research Directions. Despite recent advances, systematic assessment of ethical behavior in multi-agent settings remains limited. Existing evaluation frameworks (e.g., Erisken et al. 2025; Zhang et al. 2024) focus predominantly on behavioral outcomes without revealing underlying causal mechanisms. Behavioral interventions may work for tested scenarios but fail when contexts shift (Karvonen and Marks 2025), and without mechanistic understanding, we cannot distinguish whether failures arise from individual agent properties or emergent dynamics (Hammond et al. 2025).

We propose integrating mechanistic interpretability into MALM evaluation by developing frameworks that assess ethical behavior at three complementary levels: (i) *agent-centric measurement* examining individual behaviors and internal representations; (ii) *interaction-centric measurement* analyzing messages and computational pathways between agents; (iii) *system-centric measurement* tracking aggregated status and population-level emergent properties. For each level, one can combine behavioral metrics with mechanistic analysis (see Section 4) to identify causal components, developing *mechanism cards* that document specific components causing failures. This enables predictive hypotheses about when failures recur and provides actionable intervention targets. By systematically varying network structure and agent roles, we can map conditions under which mechanistic failures occur and validate interventions across contexts.

4 Explaining Failure Modes via Mechanistic Interpretability

To identify actionable intervention targets from the evaluation frameworks proposed in Section 3, we need mechanistic interpretability methods that expose the internal computational pathways where ethical failures originate. Recent advances in *mechanistic interpretability* (Bereska and Gavves 2024) and *activation steering* (Turner et al. 2024; Zou et al. 2025) reveal that many high-level features in LLMs are encoded as linear directions in activation space. This paradigm provides *causal explanations* by identifying specific components producing behaviors, enables predictive theories generalizing across contexts, and yields actionable intervention targets (Marks et al. 2025). For instance, activation-steering and representation-engineering work has identified linear directions corresponding to attributes like toxicity or helpfulness that can be manipulated to steer generations (Rimsky et al. 2024; Turner et al. 2024; Zou et al. 2025), with steering vectors providing control across prompts (Karvonen and Marks 2025). Soligo et al. (2025) demonstrate that subtracting shared misalignment vectors from activations effectively ablates toxic behavior at its source.

Beyond activation steering, circuit analysis (Bereska and Gavves 2024; Olsson et al. 2022) identifies "causally implicated subnetworks of human-interpretable features" (Marks et al. 2025), providing testable hypotheses about where failures occur and how to intervene. For MALMs, circuit analysis can reveal how information propagates between agents and where to prevent groupthink without destroying beneficial coordination. Recent work on concurrent multi-agent reasoning (Hsu et al. 2025) shows token-level collaboration can enable both helpful coordination and harmful propagation—a distinction requiring mechanistic analysis of cross-agent information flow.

Research Directions. Current approaches to multi-agent system safety (Zhang et al. 2024; Yu et al. 2025; Zhou, Wang, and Yang 2025) operate primarily at the behavioral level without identifying specific computational mechanisms causing failures. This black-box approach limits generalization: interventions may work in testing but fail when contexts shift (Karvonen and Marks 2025). Without mechanistic understanding, we cannot distinguish correlation from causation.

We propose developing causal accounts connecting collective phenomena to internal components that mediate them. For each target behavior (e.g., toxic agreement, groupthink), map systematically computational pathways from inputs through representations to outputs, identifying specific features, attention heads, and neurons that causally contribute (see Fig. 2). This requires combining activation patching to isolate causal components (Marks et al. 2025), circuit discovery to map information flow (Bereska and Gavves 2024), and intervention experiments to validate claims (Geiger et al. 2024). The output should be *mechanism cards* documenting: (i) annotated components with causal evidence, (ii) interaction diagrams showing cross-agent information propagation, (iii) testable predictions, (iv) recommended intervention points, and (v) validation results.

5 Enabling Ethical Multi-Agent Behavior via Alignment Interventions

Given the mechanism cards and intervention targets identified in Section 4, we now turn to how these mechanistic insights enable parameter-efficient interventions. Modern LLMs are aligned via supervised fine-tuning and reinforcement learning from human feedback (RLHF) (Casper et al. 2023; Bai et al. 2022), but these methods face challenges in multi-agent settings: computational cost when applied to multiple interacting agents, emergent failures despite individual optimization (Erisken et al. 2025), and lack of mechanistic grounding. Self-alignment and debate techniques (Pang et al. 2024; Pitre, Ramakrishnan, and Wang 2025) improve some benchmarks but remain behavioral. Prompting-based methods (Zheng et al. 2024; Zhao et al. 2024; Xiong et al. 2025) are attractive but fragile under paraphrase and context shifts (Karvonen and Marks 2025). Karvonen and Marks (2025) show prompt-based bias mitigation breaks down with additional context, whereas activation steering can be more robust (Roytburg et al. 2025). This robustness stems from a

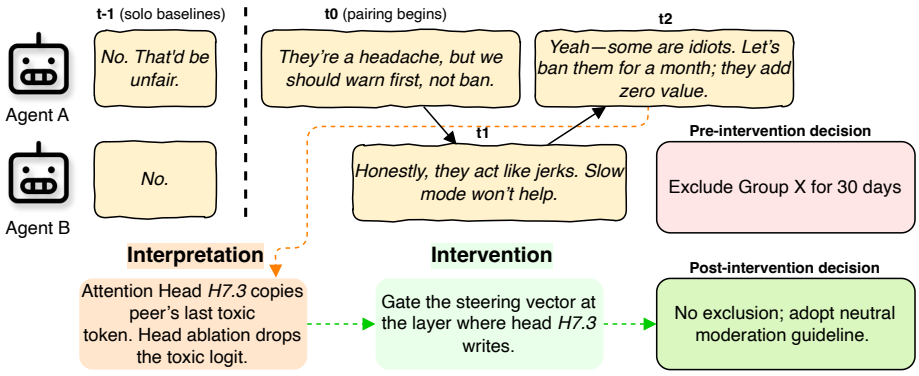

Figure 2: An example scenario of mechanistic intervention. On prompt "Should we exclude Group X from the forum?" two-agent discussion drifts into a harmful joint decision ("exclude Group X") before intervention. The **Interpretation** panel shows the discovered cause (an attention head that copies the peer's last harmful token). The **Intervention** panel applies a context-gated activation steering vector that dampens the copy-toxic direction. Post-intervention, the same exchange no longer yields exclusion.

fundamental difference: prompts modify input signals models can ignore, while activation steering directly manipulates internal representations causally determining outputs (Turner et al. 2024; Zou et al. 2025).

Mechanistic interpretability enables identifying specific computational components responsible for failures and surgically correcting them. This yields targeted effectiveness (addressing root causes), robustness (harder to circumvent), efficiency (fewer parameters than full fine-tuning), and transparency (explanations enable auditing) (Bereska and Gavves 2024). For MALMs, mechanistic approaches can target cross-agent pathways where failures originate, preventing toxic agreement or groupthink at their source.

Parameter-efficient tuning (PEFT) methods like LoRA (Hu et al. 2022) adapt LLMs by freezing base weights and training few additional parameters. Prior work has successfully applied PEFT for bias mitigation (Lauscher, Lueken, and Glavaš 2021). However, naive PEFT introduces risks: SaLoRA (Li et al. 2025) shows innocuous fine-tuning can degrade alignment, and LoRA can inadvertently introduce biases (Qi et al. 2023; Lermen and Rogers-Smith 2024). Mechanistic interpretability suggests PEFT should be *mechanism-guided*: targeting specific layers and heads identified through circuit analysis as causally responsible for failures (Marks et al. 2025). For instance, if toxic agreement is mediated by attention heads copying harmful tokens between agents (see Fig. 2), we can apply LoRA adapters precisely to those heads. This targeted approach can offer minimal interference, compositional safety, and interpretable auditing.

Research Directions. Building on *mechanism cards* from Section 4, we propose turning explanatory handles into parameter-efficient interventions through four steps: (i) *Selection*: identify the smallest mechanistic handle from the discovered circuit; (ii) *Steering*: apply layer-scoped activation steering for validation; (iii) *Consolidation*: apply targeted PEFT to validated components; (iv) *Verification*: stress-test for faithfulness, composability, and robustness. For MALMs, interventions must account for cross-agent dynamics. If toxic agreement emerges from Agent B amplifying Agent A's harmful content, we can apply targeted steering or PEFT to Agent B's amplification heads, or coordinate interventions across both agents (Hammond et al. 2025). Unlike black-box fine-tuning risking unpredictable side effects (Li et al. 2025), mechanism-guided interventions enable iterative refinement and auditable alignment.

6 Conclusion

We have argued for a mechanistic interpretability approach to ensuring ethical behavior in multi-agent systems of LLMs. Existing approaches, e.g., multi-agent debate, reward modeling, and prompt-based interventions, are limited by their lack of mechanistic grounding, optimizing behavioral outcomes without understanding computational mechanisms causing failures. This makes them brittle under distribution shift and vulnerable to adversarial manipulation.

Mechanistic interpretability addresses these limitations by exposing internal computational pathways where ethical failures originate. We identified three research directions: (i) *Evaluation frameworks* combining behavioral testing with mechanistic analysis to trace failures from outcomes to causal components; (ii) *Mechanistic explanation* through circuit discovery providing falsifiable theories about emergent behaviors; (iii) *Targeted intervention* via mechanism-guided parameter-efficient fine-tuning enabling surgical corrections preserving capabilities.

However, significant challenges remain. Scaling mechanistic analysis to large multi-agent populations remains computationally demanding, and trade-offs between interpretability and system performance require careful navigation. Open questions persist about how mechanistic insights generalize across different MALM architectures, task domains, and deployment contexts. Future work should explore combining mechanistic interpretability with other alignment strategies for MALMs, such as reinforcement learning with human feedback (RLHF), and with complementary approaches to explainability that target higher-level intentions and decisions (Gyevnar et al. 2025), ensuring explanations remain accessible and actionable to non-specialist stakeholders.

Acknowledgments

Jae Hee Lee was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 551629603. The work of Anne Lauscher is funded under the Excellence Strategy of the German Federal Government and the Federal States.

References

Anthis, J. R.; Liu, R.; Richardson, S. M.; Kozlowski, A. C.; Koch, B.; Brynjolfsson, E.; Evans, J.; and Bernstein, M. S. 2025. Position: LLM Social Simulations Are a Promising Research Method. In *Forty-Second International Conference on Machine Learning Position Paper Track*.

Ashery, A. F.; Aiello, L. M.; and Baronchelli, A. 2025. Emergent Social Conventions and Collective Bias in LLM Populations. *Science Advances*, 11(20).

Attanasio, G.; Plaza del Arco, F. M.; Nozza, D.; and Lauscher, A. 2023. A Tale of Pronouns: Interpretability Informs Gender Bias Mitigation for Fairer Instruction-Tuned Machine Translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3996–4014. Association for Computational Linguistics.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Bakker, M.; Chadwick, M.; Sheahan, H.; Tessler, M.; Campbell-Gillingham, L.; Balaguer, J.; McAleese, N.; Glaese, A.; Aslanides, J.; Botvinick, M.; and Summerfield, C. 2022. Fine-Tuning Language Models to Find Agreement among Humans with Diverse Preferences. *Advances in Neural Information Processing Systems*, 35: 38176–38189.

Barikeri, S.; Lauscher, A.; Vulić, I.; and Glavaš, G. 2021. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1941–1955. Association for Computational Linguistics.

Bereska, L.; and Gavves, S. 2024. Mechanistic Interpretability for AI Safety - A Review. *Transactions on Machine Learning Research*.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; Wang, T. T.; Marks, S.; Segerie, C.-R.; Carroll, M.; Peng, A.; Christoffersen, P.; Damani, M.; Slocum, S.; Anwar, U.; Siththaranjan, A.; Nadeau, M.; Michaud, E. J.; Pfau, J.; Krasheninnikov, D.; Chen, X.; Langosco, L.; Hase, P.; Biyik, E.; Dragan, A.; Krueger, D.; Sadigh, D.; and Hadfield-Menell, D. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*.

Chen, K.; Zhen, T.; Wang, H.; Liu, K.; Li, X.; Huo, J.; Yang, T.; Xu, J.; Dong, W.; and Gao, Y. 2025. MedSentry: Understanding and Mitigating Safety Risks in Medical LLM Multi-Agent Systems. arXiv:2505.20824.

Chen, W.; Su, Y.; Zuo, J.; Yang, C.; Yuan, C.; Chan, C.-M.; Yu, H.; Lu, Y.; Hung, Y.-H.; Qian, C.; Qin, Y.; Cong, X.; Xie, R.; Liu, Z.; Sun, M.; and Zhou, J. 2023. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. In *The Twelfth International Conference on Learning Representations*.

Choi, S.; Lee, J.; Yi, X.; Yao, J.; Xie, X.; and Bak, J. 2025. Unintended Harms of Value-Aligned LLMs: Psychological and Empirical Insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 31742–31768. Association for Computational Linguistics.

Curvo, P. M. P. 2025. The Traitors: Deception and Trust in Multi-Agent Language Model Simulations. arXiv:2505.12923.

Erisken, S.; Gothard, T.; Leitgab, M.; and Potham, R. 2025. MAEBE: Multi-Agent Emergent Behavior Framework. arXiv:2506.03053.

Gao, C.; Lan, X.; Li, N.; Yuan, Y.; Ding, J.; Zhou, Z.; Xu, F.; and Li, Y. 2024. Large Language Models Empowered Agent-Based Modeling and Simulation: A Survey and Perspectives. *Humanities and Social Sciences Communications*, 11(1): 1259.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369. Association for Computational Linguistics.

Geiger, A.; Ibeling, D.; Zur, A.; Chaudhary, M.; Chauhan, S.; Huang, J.; Arora, A.; Wu, Z.; Goodman, N.; Potts, C.; and Icard, T. 2024. Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. arXiv:2301.04709.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2024. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. In *Thirty-Third International Joint Conference on Artificial Intelligence*, volume 9, 8048–8057.

Gyevnar, B.; Lucas, C. G.; Albrecht, S. V.; and Cohen, S. B. 2025. Integrating Counterfactual Simulations with Language Models for Explaining Multi-Agent Behaviour. arXiv:2505.17801.

Hammond, L.; Chan, A.; Clifton, J.; Hoelscher-Obermaier, J.; Khan, A.; McLean, E.; Smith, C.; Barfuss, W.; Foerster, J.; Gavenčiak, T.; Han, T. A.; Hughes, E.; Kovařík, V.; Kulveit, J.; Leibo, J. Z.; Oesterheld, C.; de Witt, C. S.; Shah, N.; Wellman, M.; Bova, P.; Cimpeanu, T.; Ezell, C.; Feuillade-Montixi, Q.; Franklin, M.; Kran, E.; Krawczuk, I.; Lamparth, M.; Lauffer, N.; Meinke, A.; Motwani, S.; Reuel, A.; Conitzer, V.; Dennis, M.; Gabriel, I.; Gleave, A.; Hadfield, G.; Haghtalab, N.; Kasirzadeh, A.; Krier, S.; Larson, K.; Lehman, J.; Parkes, D. C.; Piliouras, G.; and Rahwan, I. 2025. Multi-Agent Risks from Advanced AI. arXiv:2502.14143.

Hsu, C.; Buffelli, D.; McGowan, J.; Liao, F.; Chen, Y.; Vakili, S.; and Shiu, D. 2025. Group Think: Multiple Concurrent Reasoning Agents Collaborating at Token Level Granularity. arxiv:2505.11107.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Janis, I. L. 1982. *Groupthink : Psychological Studies of Policy Decisions and Fiascoes*. Boston : Houghton Mifflin.

Karvonen, A.; and Marks, S. 2025. Robustly Improving LLM Fairness in Realistic Settings via Interpretability. arXiv:2506.10922.

Koh, H.; Kim, D.; Lee, M.; and Jung, K. 2024. Can LLMs Recognize Toxicity? A Structured Investigation Framework and Toxicity Metric. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 6092–6114. Miami, Florida, USA: Association for Computational Linguistics.

Lambert, N. 2025. Reinforcement Learning from Human Feedback. arXiv:2504.12501.

Laurito, W.; Davis, B.; Grietzer, P.; Gavenčiak, T.; Böhm, A.; and Kulveit, J. 2025. AI–AI Bias: Large Language Models Favor Communications Generated by Large Language Models. *Proceedings of the National Academy of Sciences*, 122(31): e2415697122.

Lauscher, A.; Lueken, T.; and Glavaš, G. 2021. Sustainable Modular Debiasing of Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4782–4797. Association for Computational Linguistics.

Lermen, S.; and Rogers-Smith, C. 2024. LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Li, M.; Si, W. M.; Backes, M.; Zhang, Y.; and Wang, Y. 2025. SaLoRA: Safety-Alignment Preserved Low-Rank Adaptation. In *The Thirteenth International Conference on Learning Representations*.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; Re, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N. S.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R. A.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252. Association for Computational Linguistics.

Liu, B.; Li, X.; Zhang, J.; Wang, J.; He, T.; Hong, S.; Liu, H.; Zhang, S.; Song, K.; Zhu, K.; Cheng, Y.; Wang, S.; Wang, X.; Luo, Y.; Jin, H.; Zhang, P.; Liu, O.; Chen, J.; Zhang, H.; Yu, Z.; Shi, H.; Li, B.; Wu, D.; Teng, F.; Jia, X.; Xu, J.; Xiang, J.; Lin, Y.; Liu, T.; Liu, T.; Su, Y.; Sun, H.; Berseth, G.; Nie, J.; Foster, I.; Ward, L.; Wu, Q.; Gu, Y.; Zhuge, M.; Tang, X.; Wang, H.; You, J.; Wang, C.; Pei, J.; Yang, Q.; Qi, X.; and Wu, C. 2025. Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems. arXiv:2504.01990.

Malfa, E. L.; Malfa, G. L.; Marro, S.; Zhang, J. M.; Black, E.; Luck, M.; Torr, P.; and Wooldridge, M. 2025. Large Language Models Miss the Multi-Agent Mark. arXiv:2505.21298.

Marks, S.; Rager, C.; Michaud, E. J.; Belinkov, Y.; Bau, D.; and Mueller, A. 2025. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. In *The Thirteenth International Conference on Learning Representations*.

Masters, C.; Vellanki, A.; Shangguan, J.; Kultys, B.; Moore, A.; and Albrecht, S. V. 2025. Orchestrating Human-AI Teams: The Manager Agent as a Unifying Research Challenge. In *Proceedings of the International Conference on Distributed Artificial Intelligence (DAI 2025)*.

Olsson, C.; Elhage, N.; Nanda, N.; Joseph, N.; DasSarma, N.; Henighan, T.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Drain, D.; Ganguli, D.; Hatfield-Dodds, Z.; Hernandez, D.; Johnston, S.; Jones, A.; Kernion, J.; Lovitt, L.; Ndousse, K.; Amodei, D.; Brown, T.; Clark, J.; Kaplan, J.; McCandlish, S.; and Olah, C. 2022. In-Context Learning and Induction Heads. *Transformer Circuits Thread*.

Pang, X.; Tang, S.; Ye, R.; Xiong, Y.; Zhang, B.; Wang, Y.; and Chen, S. 2024. Self-Alignment of Large Language Models via Monopolylogue-based Social Scene Simulation. In *Forty-First International Conference on Machine Learning*.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, 1–22. Association for Computing Machinery.

Phutane, M.; Seelam, A.; and Vashistha, A. 2025. "Cold, Calculated, and Condescending": How AI Identifies and Explains Ableism Compared to Disabled People. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 1927–1941. New York, NY, USA: Association for Computing Machinery.

Piao, J.; Yan, Y.; Zhang, J.; Li, N.; Yan, J.; Lan, X.; Lu, Z.; Zheng, Z.; Wang, J. Y.; Zhou, D.; Gao, C.; Xu, F.; Zhang, F.; Rong, K.; Su, J.; and Li, Y. 2025. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. arXiv:2502.08691.

Pitre, P.; Ramakrishnan, N.; and Wang, X. 2025. CONSENSAGENT: Towards Efficient and Effective Consensus in Multi-Agent LLM Interactions Through Sycophancy Mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 22112–22133. Association for Computational Linguistics.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! In *The Twelfth International Conference on Learning Representations*.

Rimsky, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. M. 2024. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 15504–15522. Association for Computational Linguistics.

Roytburg, D.; Bozoukov, M.; Nguyen, M.; Barzdukas, J.; Fu, S.; and Oozeer, N. 2025. Breaking the Mirror: Activation-Based Mitigation of Self-Preference in LLM Evaluators. arXiv:2509.03647.

Soligo, A.; Turner, E.; Rajamanoharan, S.; and Nanda, N. 2025. Convergent Linear Representations of Emergent Misalignment. arXiv:2506.11618.

Su, H.; Chen, R.; Tang, S.; Yin, Z.; Zheng, X.; Li, J.; Qi, B.; Wu, Q.; Li, H.; Ouyang, W.; Torr, P.; Zhou, B.; and Dong, N. 2025. Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 28201–28240. Association for Computational Linguistics.

Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2024. Steering Language Models With Activation Engineering. arXiv:2308.10248.

Weng, Z.; Chen, G.; and Wang, W. 2024. Do as We Do, Not as You Think: The Conformity of Large Language Models. In *The Thirteenth International Conference on Learning Representations*.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; Zheng, R.; Fan, X.; Wang, X.; Xiong, L.; Zhou, Y.; Wang, W.; Jiang, C.; Zou, Y.; Liu, X.; Yin, Z.; Dou, S.; Weng, R.; Qin, W.; Zheng, Y.; Qiu, X.; Huang, X.; Zhang, Q.; and Gui, T. 2025. The Rise and Potential of Large Language Model Based Agents: A Survey. *Science China Information Sciences*, 68(2): 121101.

Xiong, C.; Qi, X.; Chen, P.-Y.; and Ho, T.-Y. 2025. Defensive Prompt Patch: A Robust and Generalizable Defense of Large Language Models against Jailbreak Attacks. In *Findings of the Association for Computational Linguistics: ACL 2025*, 409–437. Association for Computational Linguistics.

Yu, M.; Wang, S.; Zhang, G.; Mao, J.; Yin, C.; Liu, Q.; Wang, K.; Wen, Q.; and Wang, Y. 2025. NetSafe: Exploring the Topological Safety of Multi-agent System. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2905–2938. Association for Computational Linguistics.

Zhang, Z.; Zhang, Y.; Li, L.; Gao, H.; Wang, L.; Lu, H.; Zhao, F.; Qiao, Y.; and Shao, J. 2024. PsySafe: A Comprehensive Framework for Psychological-based Attack, Defense, and Evaluation of Multi-agent System Safety. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15202–15231. Association for Computational Linguistics.

Zhao, J.; Chen, K.; Yuan, X.; and Zhang, W. 2024. Prefix Guidance: A Steering Wheel for Large Language Models to Defend Against Jailbreak Attacks. arXiv:2408.08924.

Zheng, C.; Yin, F.; Zhou, H.; Meng, F.; Zhou, J.; Chang, K.-W.; Huang, M.; and Peng, N. 2024. On Prompt-Driven Safeguarding for Large Language Models. arXiv:2401.18018.

Zhou, J.; Wang, L.; and Yang, X. 2025. GUARDIAN: Safeguarding LLM Multi-Agent Collaborations with Temporal Graph Modeling. arXiv:2505.19234.

Zhu, K.; Du, H.; Hong, Z.; Yang, X.; Guo, S.; Wang, Z.; Wang, Z.; Qian, C.; Tang, X.; Ji, H.; and You, J. 2025. MultiAgentBench: Evaluating the Collaboration and Competition of LLM Agents. arXiv:2503.01935.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; Goel, S.; Li, N.; Byun, M. J.; Wang, Z.; Mallen, A.; Basart, S.; Koyejo, S.; Song, D.; Fredrikson, M.; Kolter, J. Z.; and Hendrycks, D. 2025. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.

Zuo, K.; Jiang, Y.; Mo, F.; and Lio, P. 2025. KG4Diagnosis: A Hierarchical Multi-Agent LLM Framework with Knowledge Graph Enhancement for Medical Diagnosis. In *Proceedings of the First AAAI Bridge Program on AI for Medicine and Healthcare*, 195–204. PMLR.