AGENTIC THOUGHT COLLECTIVES: MULTI-AGENT COMMUNITIES FOR OPEN-ENDED SCIENTIFIC DISCOVERY

Prabhant Singh
p.singh@tue.nl
AMOR/e Lab
TU Eindhoven

Joaquin Vanschoren
j.vanschoren@tue.nl
AMOR/e Lab
TU Eindhoven

ABSTRACT

The "AI Scientist" is typically envisioned as a solitary, goal-directed entity. However, the complexity of modern science requires a shift from isolated optimization to open-ended evolution. We propose a vision for Agentic Thought Collectives, multi-agent systems designed not just to solve problems, but to formulate them. Leveraging recent advancements in Large Language Models (LLMs), we outline an architecture for heterogeneous agent populations that sustain scientific inquiry through collaboration, rigorous peer review, and inter-generational knowledge inheritance. Unlike static systems, these collectives are designed to be self-evolving, capable of expanding the boundaries of their own search space without human intervention. This paper explores the structural requirements for such ecosystems, including communication protocols and selection pressures that favor novelty and robustness. We present this vision as a roadmap for the agent community, urging a transition toward creating persistent, autonomous societies capable of long-horizon scientific innovation.

1 INTRODUCTION

The vision of automating scientific discovery has captivated AI researchers for decades (King et al., 2009), but recent advances in large language models have transformed this from simple experiments into a practical utility for scientific discovery. Systems like AI Scientist (Yamada et al., 2025), Cycle Researcher (Weng et al., 2025), and Agent Laboratory (Schmidgall et al., 2025) demonstrate that AI can generate novel hypotheses, execute experiments, and produce publishable research (workshop level (Yamada et al., 2025)). Yet, these achievements, however impressive, represent only the first step toward a more profound transformation. This research direction has even resulted in multiple commercial products for AI Scientist, such as AI Google's AI Co-Scientist (Gottweis et al., 2025) and Microsoft Discovery (Cheng et al., 2025).

Current AI scientist systems(FunSearch (Romera-Paredes et al., 2024), AI Scientistv2 (Yamada et al., 2025)) suffer from a fundamental limitation in terms of the scientific discovery process: they operate as isolated discovery machines optimizing narrow, predefined objectives. This paradigm enables them to optimize a particular objective; for example, FunSearch optimizes solving the CapSet problem using LLMs and evolutionary algorithms, whereas AI-Scientistv2 can generate a set of problems where the input is a call-for-papers file for a venue and produce a scientifically valid novel work (workshop level). This stands in stark contrast to how modern science actually works. Scientific communities are rarely driven by a "CFP-first" mentality, nor do they simply exhaustively optimize a single static metric. While computational scalability is essential for accelerating discovery, the next generation of systems must move beyond isolated optimization to emulate the dynamic, open-ended, and communal methods that characterize real-world science.

In this paper, we draw inspiration from Ludwik Fleck's idea of *thought collectives* (Denkkollektive)(Zalta, 2012; Fleck, 1979), communities of practitioners who share conceptual frameworks, methodological commitments, and tacit knowledge that shape what counts as a valid question, an acceptable method, and a satisfactory answer. Scientific collaboration is inherently value-laden and

social (Rolin, 2015), built on decades of incremental work rather than isolated optimization. To this end, we argue for a framework that is designed to operationalize these social dynamics with the objective of accelerating scientific development and fostering autonomous agentic science communities.

We propose **Agentic Thought Collectives**: self-organizing communities of heterogeneous AI agents that collectively navigate the landscape of scientific discovery. Unlike goal-centric AI scientists, these communities would be *intrinsically motivated, autotelic collectives generating(Schmidhuber, 2010)* their own research questions, building cumulatively on each other's work, and experiencing emergent phenomena like specialization, paradigm formation, and scientific revolutions. Recent work on scaling multi-agent collaboration (Qian et al., 2025) and self-evolving agent networks (Hu et al., 2025b) suggests that such systems are becoming feasible.

2 Structure of Agentic-Thought-Collectives

We propose a possible structure for Agentic Thought Collectives, grounded in the multiagent ;literature of making agency *explicit*, *explainable*, and *socially embedded*. Unlike current agentic AI systems, which often lack explicit architectures and normative grounding, our vision emphasizes structured reasoning, formal interaction protocols, and institutional governance.

2.1 Individual Scientists

The atomic unit is an AI scientist, capable of hypothesis generation, experimental design, execution, and manuscript writing (Yamada et al., 2025). We advocate for agents with *explicit cognitive architectures* rather than purely emergent behavior. The Belief-Desire-Intention (BDI) paradigm (Rao & Georgeff, 1995; Wooldridge, 2009) provides a principled foundation: *beliefs* represent an agent's knowledge about the scientific landscape; Desires embody its long-term research goals and intrinsic motivations; and Intentions constitute the specific experimental plans it commits to executing. This *tripartite* structure enables transparent reasoning about objectives; an *agent* pursues particular hypotheses.

Crucially, agents should be *autotelic*, capable of generating their own research questions based on intrinsic motivation (Schmidhuber, 2010) rather than predefined objectives. Unlike current agentic AI systems where "intentionality" is implicit and inferred from behavior (Dignum & Dignum, 2025), Agentic-Thought-Collectives require explicit commitment mechanisms that link motivations to actions, enabling the verification of goal-action consistency.

2.2 Research Groups and Labs

We define a Lab as a cluster of multiple scientific agents with differentiated roles, mirroring the structure of human research groups: **Junior researchers**: Exploration-focused agents tasked with pursuing novel, high-risk directions. **Senior researchers**: Agents possessing accumulated expertise and extended memory capacities (e.g., larger context windows) to synthesize complex findings. **Supervisory agents**: Coordinators responsible for suggesting collaborations, maintaining "lab culture" (defined here as persistent research agendas), and ensuring methodological consistency. This mirrors real-world dynamics, where supervisors guide students to collaborate and combine ideas.

2.3 Scientific Communities

At the highest level, multiple labs aggregate to form a scientific community. These communities are connected through formal communication protocols that transcend unstructured natural language exchanges. Drawing on multiagent community research on agent communication languages (Finin et al., 1997a), we propose the use of structured communicative acts where message syntax is paired with explicit semantics. In this framework, acts such as requests, commitments, claims, and challenges carry formal meaning, enabling agents to negotiate, coordinate, and construct a shared understanding.

Community infrastructure includes: **Conferences**: Competitive venues with peer review, where agents submit papers and reviewer agents evaluate them (Feliciani et al., 2019). **Workshops**: Col-

laborative spaces for discussing emerging directions and forming new research agendas through *structured argumentation* (Rahwan & Simari, 2009). **Journals**: Archival repositories that enable long-term knowledge accumulation. **Negotiation forums**: Spaces where agents with conflicting research priorities can justify their positions and dynamically revise their commitments through dialogue (Jennings et al., 2001). Over time, these communities develop emergent thought styles that differentiate them. Mechanisms for controlling diversity in agent conversations (Chu et al., 2025) can help maintain productive heterogeneity while enabling coherent research programs. Community events also ensure quality control and knowledge refinement of generated agents and content (Abramowitz et al., 2025).

2.4 KNOWLEDGE INFRASTRUCTURE

Stigmergic Archives: All research outputs (papers, code, data, failed experiments) are deposited in shared archives that serve as stigmergic traces(Kurihara et al., 2003; Marsh & Onof, 2008). Agents can query, build upon, and cite previous work. The archive structure itself evolves, and new taxonomies emerge as fields develop.

Peer Review and Selection: Agents evaluate submissions based on novelty, rigor, and significance. Reviewer agents should represent diverse perspectives (Paolucci & Grimaldo, 2012; Aziz et al., 2023), simulating the multi-paradigm nature of real review panels. *Mechanism design* (Nisan et al., 2007) provides mathematical tools for structuring rewards, penalties, and information flows so that rational agents act in ways that support collective scientific progress.

Norms and Institutions: Drawing on multiagent research on normative multi-agent systems (Boella & van der Torre, 2006; Dignum, 2004), Agentic Thought Collectives should embed explicit *norms* (obligations, permissions, prohibitions) that regulate autonomous behavior within the scientific community. Electronic institutions (Esteva et al., 2001; Sierra et al., 2004) provide computational environments where interaction is structured by roles, protocols, and institutional rules. Critically, norms gain their power not when universally followed, but when violations are meaningful. agents should understand when to follow methodological conventions and when exceptional circumstances warrant deviation(Boella & van der Torre, 2006).

Trust and Reputation: In open scientific communities, agents must assess the reliability of others' claims and methods. Computational models of trust and reputation (Sabater & Sierra, 2005) enable selective collaboration, reputation-based resource allocation, and resilience to low-quality or fraudulent research. This system facilitates the natural marginalization of low-performing agents, as their trust scores and reputations degrade over time.

3 KEY CHALLENGES

3.1 SCALABILITY

The most immediate challenge for Agentic Thought Collectives is computational and infrastructural scalability(Cemri et al., 2025). A single AI Scientist relies on resource-intensive LLM inference, code execution environments, and iterative debate loops. Scaling this to a community level introduces exponential complexity: communication between agents can generate millions of transactions in a short timeframe, creating significant bottlenecks. Hosting this infrastructure requires robust orchestration, efficient message routing, and highly cost-effective inference mechanisms. While current trends in the machine learning community focus heavily on improving agentic scalability(wang et al., 2025), realizing a fully functional Agentic Thought Collective will likely require dedicated engineering breakthroughs in distributed agent orchestration and resource-efficient model serving.

3.2 DYNAMICS

Regulating the complex dynamics between agents must be an autonomous process. To achieve this, we draw upon principles from adjacent fields to propose the following mechanisms:

1. **Evolutionary computing**: Principles from evolutionary algorithms and population-based training are essential for maintaining the health of the collective. We argue a need for

mechanisms which can perform automated pruning of underperforming or obsolete agents based on fitness criteria. Furthermore, open-ended evolutionary strategies can drive the continuous improvement of the system, allowing for the emergence of more efficient and specialized agent architectures over time.

2. **Memetic Transmission(Hou et al., 2016) and Knowledge Transfer**: In this framework, successful research strategies propagate through the population via social learning. Agents observe high-impact work and adopt similar approaches or directly inherit research programs. This fosters a form of cultural evolution within the community, where successful heuristics, methods, and conceptual frameworks are replicated and disseminated based on their scientific utility.

3.3 COMMUNICATION

Scalable and robust communication protocols are essential for the realization of Agentic Thought Collectives. LLM-based agents are sensitive to unstructured dialog, often becoming susceptible to hallucination cascades or circular reasoning that leads to suboptimal research trajectories. To mitigate these failure modes, we leverage the extensive history of communication research in the multiagent literature. Mechanisms such as KQML (Finin et al., 1997b) and TACTIC(Yu et al., 2025) provide established frameworks for agents to exchange information, negotiate, and coordinate actions with a shared understanding of intent and context. By introducing formal semantics and constrained communication layers, rather than relying solely on natural language, we can significantly improve the reliability, traceability, and accountability of interactions within the collective.

3.4 SAFETY AND ETHICS

The deployment of Agentic Thought Collectives introduces significant risks and safety challenges(Woodgate, 2025). Enabling autonomous agents to communicate at scale and execute independent experiments creates the potential for the generation of unethical or dangerous research. For instance, a sub-community within the collective could inadvertently or maliciously optimize for harmful objectives, such as developing discriminatory algorithms or designing toxic chemical compounds in a materials science context. Therefore, implementing robust safeguards(Ferrando & Cardoso, 2022; Zhou et al., 2025) is a prerequisite for deployment. This includes not only hard-coded constraints on experimental parameters but also the development of "ethical supervisor" agents capable of monitoring the collective's research trajectory and intervening to prevent dual-use hazards or alignment failures.

4 CONCLUSION AND DISCUSSION

We have proposed Agentic Thought Collectives, a vision for AI communities that transcend the paradigm of isolated AI scientists to become genuine thought collectives capable of open-ended scientific discovery. By synthesizing insights from the sociology of science, open-ended learning, stigmergic coordination, and, critically, the tradition of explicit reasoning architectures, communication protocols, and institutional governance (Dignum & Dignum, 2025; Weiss, 2013), we have outlined a vision in which heterogeneous agents form labs, conduct research, peer-review work, and collectively navigate the scientific landscape. Recent work by Dignum & Dignum (2025) further reinforces how core multiagent principles can be effectively integrated into modern LLM-based agent architectures.

Crucially, we argue that the shift to collectives is not merely an efficiency gain but a fundamental requirement for deep discovery. A monolithic AI scientist, no matter how capable, risks converging on a single dominant paradigm or getting trapped in local optima. By contrast, the Agentic Thought Collective enforces epistemic diversity, ensuring that competing hypotheses and distinct methodologies survive long enough to be rigorously tested—mirroring the resilience of the scientific enterprise itself.

This vision is ambitious but increasingly tractable. The components exist: LLM-based scientist agents (Yamada et al., 2025; Schmidgall et al., 2025), multi-agent coordination frameworks (Qian et al., 2025; Chen et al., 2025), quality-diversity algorithms (Hu et al., 2025a), and knowledge infrastructure (Weng et al., 2025). What Autonomous Agents and Multiagent Systems research adds

is the conceptual and formal scaffolding to make these systems not merely capable but *coherent*, *cooperative*, and *accountable*.

REFERENCES

Ben Abramowitz, Omer Lev, and Nicholas Mattei. Who reviews the reviewers? a multi-level jury problem. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, pp. 14–22, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.

Haris Aziz, Evi Micha, and Nisarg Shah. Group fairness in peer review. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pp. 2889–2891, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.

Guido Boella and Leendert van der Torre. A game-theoretic approach to norms and agents. *Artificial Intelligence and Law*, 14(3):257–282, 2006.

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=fAjbYBmonr.

Weize Chen, Ziming You, Ran Li, yitong guan, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=o1Et3MogPw.

Newman Cheng, Gordon Broadbent, and William Chappell. Cognitive loop via in-situ optimization: Self-adaptive reasoning for science. August 2025. URL https://www.microsoft.com/en-us/research/publication/cognitive-loop-via-in-situ-optimization-self-adaptive-reasoning-for-science/.

KuanChao Chu, Yi-Pei Chen, and Hideki Nakayama. Exploring and controlling diversity in LLM-agent conversation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 25626–25644, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1397. URL https://aclanthology.org/2025.findings-emnlp.1397/.

Virginia Dignum and Frank Dignum. Agentifying agentic ai. *Proceedings of AAAI*, 2025.

Virginia M. Dignum. *A Model for Organizational Interaction: Based on Agents, Founded in Logic*. PhD thesis, SIKS, 2004.

Marc Esteva, Juan A. Rodríguez-Aguilar, and Carles Sierra. Electronic institutions: From specification to development. In *Applications of Agent Technology in Traffic and Transportation*, pp. 303–318. Springer, 2001.

Thomas Feliciani, Junwen Luo, Lai Ma, Pablo Lucas, Flaminio Squazzoni, Ana Marušić, and Kalpana Shankar. A scoping review of simulation models of peer review. *Scientometrics*, 121(1):555–594, October 2019. ISSN 0138-9130. doi: 10.1007/s11192-019-03205-w. URL https://doi.org/10.1007/s11192-019-03205-w.

Angelo Ferrando and Rafael C. Cardoso. Safety shields, an automated failure handling mechanism for bdi agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pp. 1589–1591, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450392136.

Tim Finin, Yannis Labrou, and James Mayfield. Kqml as an agent communication language. In *Software Agents*. MIT Press, 1997a.

Tim Finin, Yannis Labrou, and James Mayfield. Kqml as an agent communication language. In *Software Agents*. MIT Press, 1997b.

Ludwik Fleck. *Genesis and Development of a Scientific Fact*. University of Chicago Press, Chicago, 1979.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an ai co-scientist, 2025.

Yaqing Hou, Yifeng Zeng, and Yew-Soon Ong. A memetic multi-agent demonstration learning approach with behavior prediction. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, pp. 539–547, Richland, SC, 2016. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450342391.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=t9U3LW7JVX.

Yue Hu, Yuzhu Cai, Yaxin Du, Xinyu Zhu, Xiangrui Liu, Zijie Yu, Yuchen Hou, Shuo Tang, and Siheng Chen. Self-evolving multi-agent collaboration networks for software development. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=4R71pdPBZp.

Nicholas R. Jennings, Peyman Faratin, Alois R. Lomuscio, Simon Parsons, Michael J. Wooldridge, and Carles Sierra. Automated negotiation: Prospects, methods, and challenges. *Group Decision and Negotiation*, 2001.

Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, and Amanda Clare. The automation of science. *Science*, 324(5923):85–89, 2009. doi: 10.1126/science.1165620. URL https://www.science.org/doi/abs/10.1126/science.1165620.

Satoshi Kurihara, Kensuke Fukuda, Toshio Hirotsu, Osamu Akashi, Shinya Sato, and Toshiharu Sugawara. Simple but efficient collaboration in a complex competitive situation. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '03, pp. 1042–1043, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581136838. doi: 10.1145/860575.860786. URL https://doi-org.dianus.libr.tue.nl/10.1145/860575.860786.

Leslie Marsh and Christian Onof. Stigmergic epistemology, stigmergic cognition. *Cognitive Systems Research*, 9(1):136–149, 2008. ISSN 1389-0417. doi: https://doi.org/10.1016/j.cogsys.2007.06.009. URL https://www.sciencedirect.com/science/article/pii/S1389041707000290. Perspectives on Social Cognition.

Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani (eds.). *Algorithmic Game Theory*. Cambridge University Press, 2007.

Mario Paolucci and Francisco Grimaldo. Disagreement for control of rational cheating in peer review: a simulation. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '12, pp. 1357–1358, Richland, SC, 2012. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 0981738133.

Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large language model-based multi-agent collaboration. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=K3n5jPkrU6.

Iyad Rahwan and Guillermo R. Simari (eds.). *Argumentation in Artificial Intelligence*. Springer, 2009.

Anand S. Rao and Michael P. Georgeff. Bdi agents: From theory to practice. In *Proceedings of the International Conference on Multi-Agent Systems (ICMAS)*, 1995.

Kristina Rolin. Values in science: The case of scientific collaboration. *Philosophy of Science*, 82 (2):157–177, 2015. doi: 10.1086/680522.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, January 2024. URL http://dblp.uni-trier.de/db/journals/nature/nature625.html#RomeraParedesBNBKDREWFKF24.

Jordi Sabater and Carles Sierra. Review on computational trust and reputation models. *Artificial Intelligence Review*, 2005.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.320. URL https://aclanthology.org/2025.findings-emnlp.320/.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010. doi: 10.1109/TAMD.2010.2056368.

Carles Sierra, Marc Esteva, Juan A. Rodríguez-Aguilar, and Josep Lluís Arcos. Engineering multiagent systems as electronic institutions. In *Methodologies and Software Engineering for Agent Systems*, pp. 259–280. Springer, 2004.

Yuru wang, Kaiyan Zhang, Kai Tian, Sihang Zeng, Xingtai Lv, Ning Ding, Biqing Qi, and Bowen Zhou. Scalability of LLM-based multi-agent systems for scientific code generation: A preliminary study. In *First Workshop on Multi-Turn Interactions in Large Language Models*, 2025. URL https://openreview.net/forum?id=h8KFX5HM3N.

Gerhard Weiss (ed.). *Multiagent Systems. 2nd edition*. MIT Press, 2013. ISBN 978-0-262-01889-0.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bjcsVLoHYs.

Jessica Woodgate. Ethical decision-making in multi-agent systems. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, pp. 2991–2993, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.

Michael Wooldridge. *An Introduction to MultiAgent Systems*. Wiley, 2009.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Peihong Yu, Manav Mishra, Syed Zaidi, and Pratap Tokekar. Task-agnostic contrastive pre-training for inter-agent communication. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, pp. 2262–2270, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.

Ed Zalta (ed.). *Stanford Encyclopedia of Philosophy*. Stanford Encyclopedia of Philosophy, Stanford, CA, 2012.

Jialong Zhou, Lichao Wang, and Xiao Yang. GUARDIAN: Safeguarding LLM multi-agent collaborations with temporal graph modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=6j9xJ9pBjm.