Doctoral School of Information and Biomedical Technologies Polish Academy of Sciences (TIB PAN)

SUBJECT:
Towards Safe and Aligned Agentic and Multi-Agent Systems Through Information Propagation Traceability

Scientific discipline: Information and communication technology

SUPERVISOR:
dr Fazl Barez
fazl.barez@eng.ox.ac.uk | fazlbarez93@gmail.com
University of Oxford, Wellington Square, Oxford OX1 2JD, United Kingdom

ASSISTANT SUPERVISOR:
dr inż. Sebastian Cygert
sebastian.cygert@nask.pl
NASK-PIB, ul. Kolska 12, 01-045 Warszawa

DESCRIPTION:
Large language models (LLMs) are commonly deployed as autonomous agents that call external tools, act on persistent memory, and coordinate through hierarchies in which an orchestrating agent dispatches subagents and consumes what they return as fact. These Multi-Agent Systems (MAS) are commonly used for multi-step tasks that a single LLM cannot perform, thus taking its place as the default deployment architecture.

Most alignment research treats a single model's output as the unit of concern[1,2], leaving open what happens once an error or contested claim is written into shared memory and acted upon through a tool call with a real-world consequence. A fast-growing literature studies how unreliable claims propagate through multi-agent cascades and conformity effects[3-7], hierarchical decision-making[8,9], and symmetric opinion-dynamics networks[10]. However, little work models a hierarchical orchestrator that treats subagent output as fact, relates information propagation to downstream tool use, or accounts for the resulting safety violations, such as privacy leakage or unauthorised tool use[11,12]. This PhD topic studies how errors and contrasting claims propagate, compound, or resolve as they move through agents, memory, and tool-mediated actions, and how they can be detected, contained, or attributed.

RESEARCH OBJECTIVES AND APPROACH

1. Analysis of Propagation of Mutually Exclusive Claims in Hierarchical Orchestrator-Subagent Systems.

This direction models information flow over a hierarchical network of an orchestrating agent and the subagents it dispatches, rather than the symmetric peer networks studied in prior opinion-dynamics work[10]. Its novelty lies in treating propagated content as fact, not opinion: each subagent independently researches its claim, and claims are consumed by the orchestrator as ground truth rather than debated. This direction characterizes how mutually exclusive, independently-researched claims propagate and shape the orchestrator's downstream synthesis and decisions.

2. Capability-Driven Influence and Injection-Point Sensitivity.

A closely related question is whether subagents' influence over the orchestrator's eventual synthesis depends on the subagent's size rather than the correctness of its claim. This direction pairs subagents of different sizes on the same contested question and measures outcome as a function of which side holds the larger model, and separately, whether an orchestrator, while explicitly instructed to distrust upstream claims will resist this influence differently than one that is not. Additional conditions can be considered: one in which the orchestrator has access to the subagents' identity and size, and one in which claims are presented blind, so that the orchestrator can only judge them by their content or argumentation. This contrast isolates influence coming from known model identity from influence carried by the claim itself.

3. Memory- and Tool-Mediated Propagation in Long-Horizon Agentic Workflows.

This direction studies propagation through a persistent, shared memory store across a long horizon, tool-using workflow (case intake, multi-step approval, record updates), in which an erroneous or contested claim written into memory by one step is read, unverified, by a later step or a different agent entirely, and acted on through a tool with a real consequence: an incorrect filing, or the exposure of another party's data. It asks how propagation dynamics change once the channel is persistent memory rather than a bounded conversational exchange, and the endpoint is a tool call rather than a spoken claim.

4. Distinguishing Benign Drift from Adversarial Injection, and Attribution.

This direction studies benign, honest disagreement and deliberate, adversarial claim injection within the same hierarchical harness, using a red teaming pipeline that injects candidate claims into an agentic pipeline with real tools. It asks whether the two are distinguishable from the propagation pattern alone, and whether execution-provenance techniques [13] can attribute an incorrect action to the claim that produced it.

REQUIREMENTS:

- MSc degree in computer science, mathematics, AI, or related,
- programming and implementation skills in Python,
- experience with LLMs and understanding of Agentic Systems,
- advanced level of English (spoken and written).

BIBLIOGRAPHY:

[1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv [Cs.AI]. Retrieved from http://arxiv.org/abs/1606.06565

[2] Zaiats, V. (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.15217

[3] Jamshidi, S., Dakhel, A. M., Nafi, K. W., & Khomh, F. (2026). Hallucination Cascade: Analyzing error propagation in Multi-Agent LLM Systems. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2606.07937

[4] Becker, J., Wahle, J. P., Ruas, T., & Gipp, B. (2026). Misinformation propagation in benign Multi-Agent systems. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2606.16710

[5] Xie, Y., Zhu, C., Zhang, X., Zhu, T., Ye, D., Qi, M., Chen, H., & Zhou, W. (2026). From spark to fire: Modeling and mitigating error cascades in LLM-Based Multi-Agent collaboration. Open MIND. https://doi.org/10.48550/arxiv.2603.04474

[6] Kasprova, V., Parulekar, A., AlRabah, A., Agaram, K., Garg, R., Jha, S., Bozdag, N. B., & Hakkani-Tur, D. (2026). Too polite to disagree: Understanding sycophancy Propagation in Multi-Agent Systems. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2604.02668

[7] Ju, T., Wang, Y., Ma, X., Cheng, P., Zhao, H., Wang, Y., … Liu, G. (2024). Flooding Spread of Manipulated Knowledge in LLM-Based Multi-Agent Communities. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2407.07791

[8] Park, S., & Kwon, M. (2026). Multi^2: Hierarchical Multi-Agent Decision-Making with LLM-Based Agents in Interactive Environments. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2606.03698

[9] Kostka, A., & Chudziak, J. A. (2025). Evaluating theory of mind and internal beliefs in LLM-Based multi-agent systems. Lecture Notes in Computer Science, 18–32. https://doi.org/10.1007/978-3-032-09318-9_2

[10] Chuang, Y., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., Shah, D., Hu, J., & Rogers, T. T. (2023). Simulating Opinion Dynamics with Networks of LLM-based Agents. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.09618

[11] Yu, M., Meng, F., Zhou, X., Wang, S., Mao, J., Pang, L., Chen, T., Wang, K., Li, X., Zhang, Y., An, B., & Wen, Q. (2025). A survey on Trustworthy LLM Agents: Threats and Countermeasures. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2503.09648

[12] Yagoubi, F. E., Badu-Marfo, G., & Mallah, R. A. (2026). AgentLeak: a benchmark for Internal-Channel privacy leakage in Multi-Agent LLM systems. IEEE Access, 14, 94960–94978. https://doi.org/10.1109/access.2026.3704541

[13] Wang, Y., Zhang, J., Cai, T., Liu, Z., Sun, Q., Sun, Z., Wu, Z., Dong, M., Zheng, M., Yin, X., & Zhu, Y. (2026). From Agent Traces to Trust: A survey of evidence tracing and execution provenance in LLM agents. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2606.04990