TITLE: Privacy-Enhancing Federated Multi-Agent Systems

Huan Wang (The University Of Wollongong, Australia)

Overview

With the remarkable success of Large Language Models (LLMs), Agentic AI has shown substantial potential for real-world applications, catalyzing the development of LLM-based agents. By shifting from standalone reasoning to collaboration-centric interaction, LLM-based Multi-Agent Systems (MASs) have demonstrated strong capability in coordinating and solving complex tasks. However, in sensitive domains, this collaborative paradigm faces emerging privacy concerns due to the continuous interaction trajectories among agents. Motivated by the need for privacy-preserving multi-party collaboration, this project aims to extend MASs into Federated Multi-Agent Systems (FMAS), where autonomous agents collaborate without directly sharing confidential information. Specifically, we focus on three key dimensions for advancing FMAS, including: **I**) Personalized Privacy Reasoning (**RQ1**), **II**) Context-Aware Identity Definition (**RQ2**), **III**) Heterogeneous Decision Alignment (**RQ3**). By minimizing centralized control and data exchange, this project will effectively enhance privacy preservation while maintaining strong performance of multi-agent systems.

Background And Research Gaps

Large Language Models (LLMs) have recently made significant advancements in reasoning, planning, and tool use capabilities comparable to humans through large-scale alignment training [1, 2, 3]. These advances are prompting a paradigm shift from passive tools to proactive autonomous agents capable of perceiving environments, executing reasoning, and initiating actions. Consequently, LLM-based agents have attracted widespread attention and undergone rapid development, evolving from human-like instruction understanding and generation [4] to autonomous task planning [5], tool utilization [6], and closed-loop interactions [7, 8]. Based on the inspiring versatility of the single LLM-based agent, LLM-based Multi-Agent Systems (MASs) [9, 10] have been introduced to leverage the collective intelligence of multiple agents.

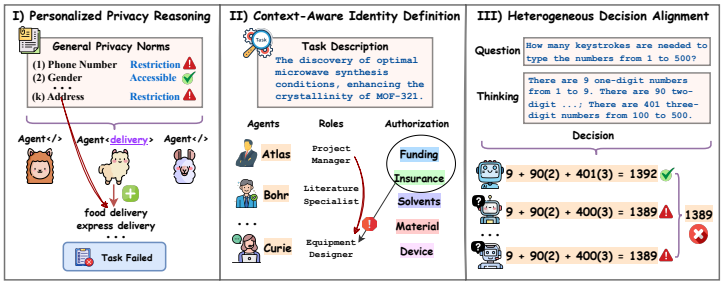

Figure 1: The **three key dimensions** for advancing FMAS.

Despite the promising trend, most existing studies primarily focus on enhancing collaboration to improve the MASs performance, while paying insufficient attention to critical privacy concerns [11, 12]. This issue is particularly pressing in sensitive domains such as finance [13, 14] and healthcare [15]. The growing need for privacy-preserving multi-party collaboration naturally motivates us to extend MASs into **Federated Multi-Agent Systems (FMAS)**, where autonomous agents cooperate without exchanging confidential information. However, FMAS differs fundamentally from conventional Federated Learning (FL) in several key aspects: 1) FL aims to learn a shared global model, while FMAS focuses on multi-agent collaboration; 2) FL exchanges messages indirectly through model updates, whereas FMAS relies on task allocation and agent interaction; 3) FL primarily protects private training data, whereas FMAS must ensure privacy dynamically throughout task execution and agent communication. Through *Figure 1*, we identify three key dimensions for advancing FMAS: **I**) **Personalized Privacy Reasoning**: different agents may have varying requirements for data sharing and privacy protection, requiring task-related reasoning tailored to individual users' privacy preferences; **II**) **Context-Aware Identity Definition**: agent roles influence individuals' acceptance of identifiable disclosure, thus prompting robust scoping of agent capabilities across diverse interaction modalities; **III**) **Heterogeneous Decision Alignment**: single-agent decision may maliciously propagate across other agents, causing negative

interference and ultimately resulting in a conflicting collective outcome. These dimensions highlight a trade-off between collaborative performance and privacy preservation, and motivate the central research question of this project: **How can FMAS be designed to satisfy specific privacy requirements across agents while maintaining stable task performance and manageable system complexity?**

Aim1: To develop a personalized privacy reasoning strategy for FMAS that enables each agent to adapt task planning and information sharing to user privacy preferences during multi-agent coordination.
Aim2: To design a context-aware identity definition mechanism for FMAS that allow agents to dynamically regulate disclosure boundaries according to interaction context and collaboration objectives.
Aim3: To establish a heterogeneous decision alignment framework for FMAS that ensures robust and consistent multi-agent behavior by mitigating harmful propagation and decentralized conflicts.

To achieve the outlined objectives, this project will formalize the fundamental challenges in developing Federated Multi-Agent Systems into concrete research questions, and address them from three major aspects enabled by a privacy server agent: **adaptive logical entailment, agent-specific authorized delegation, and uncertain utility alignment**. Through these innovations, this project will establish foundations for privacy-enhancing multi-agent systems, enabling robust collaboration across diverse interactive environments while ensuring system performance and manageability. The detailed research questions, technical innovations, and methodologies will be presented in the subsequent sections of this proposal.

Research Questions And Tasks

[RQ1] How to design the personalized privacy reasoning strategy for FMAS to adapt to diverse privacy preferences across agents? In multi-agent systems, general privacy norms cannot account for individual-sensitive decisions (e.g., financial tasks), thereby posing a barrier to the development of personalized privacy reasoning. While existing approaches [16, 17] typically leverage contextual integrity of prior user judgments to improve personalization, they still cannot effectively align with user privacy decisions due to the limited logical reasoning. Furthermore, the opaque nature of the LLM-based agents undermines interpretability by preventing verification of the reasoning process used to reach a privacy judgment. These limitations motivate the central research objective: *To personalize LLM-based agents for data-sharing decisions that align with prior user judgments while ensuring auditability*.

[Task1] Adaptive Logical Entailment: To fulfill this objective, we carry out Task1 to address RQ1. The key insight is to frame personalized privacy reasoning as a *Logical Entailment Problem*: determining whether a user's judgment on a prior request entails the same judgment on a new request. Specifically, built upon a privacy-enhanced agent deployed on a trusted server, Task1 leverages an LLM to construct ontologies from users' prior privacy judgments, capturing semantic relationships such as data sensitivity hierarchies. These ontologies are then processed by a rule-based component to produce the final adaptive privacy judgment for each agent. The core pipelines of Task1 are: 1) *Generating Ontology*: use the server agent to generate user-specific ontologies over the parameters involved in prior and incoming requests; 2) *Mapping Ontology*: map the parameters of the incoming request and relevant prior requests to levels in the generated ontologies; 3) *Logical Entailment*: apply a set of entailment rules to determine whether the user's prior judgments entail the incoming request. These three pipelines ensure that the agent makes adaptive decisions only when the privacy judgment can be logically inferred from the user's prior privacy judgments; otherwise, the agent's request is escalated to the user to seek further authorization.

[RQ2] How to develop the context-aware identity definition mechanism for FMAS to regulate agents' permission boundaries across diverse interactions? The evolving roles of agents in multi-agent systems raise pressing challenges for authorization, accountability, and access control. Under dynamic self-planning environments, verifying the permissions of interacting agents becomes essential whenever they may act on

behalf of human users, especially when they are capable of taking consequential actions. Consider an agent that transitions from flight booking to route planning. Although both roles belong to the broader travel task, the latter may retain privileges granted in the former role, such as access to credit card details. *Therefore, an agent should be authenticated and authorized before performing role-specific operations*.

[Task2] Agent-Specific Authorized Delegation: Task2 employs a role-based authentication mechanism to dynamically assign the appropriate identity to each agent according to the task context. The main idea is to generate a unique identity for each agent based on role permissions, which is achieved through a context-aware approach by integrating global state information with individual agent attributes. Specifically, the core pipelines of Task2 are: 1) *Encoding Contextual State*: capture a set of context variables by integrating the environmental state with the local observations of all agents; 2) *Authorizing Agent Identity*: assign a unique identity to each agent from the generated contextual variables via an auto-regressive manner, and utilize dynamic masking to ensure that newly generated identities do not duplicate previously ones. After assigning unique identities to agents, the server agent incorporates this identity information into decision-making processes, thereby defining explicit boundaries on the actions permitted for different role-based agents.

[RQ3] How to establish the heterogeneous decision alignment framework for FMAS to ensure robust and consistent behavior across multiple agents? Existing approaches [18] predominantly build upon the external performance measure to guide agents' decision-making process, but such reliance is problematic in real-world scenarios, where these priors may be noisy, flawed, or even erroneous. For example, in Game of 24, which uses four numbers and basic arithmetic operations to obtain 24; each agent estimates the likelihood that a candidate decision would lead to the target value and selects the most promising one. However, agent-specific decision trajectories may not provide consistent likelihood estimates, causing unreliable guidance in multi-agent systems. *When making decisions, individual agents should rely not only on external measures but also on practical experience to form internal judgments*.

[Task3] Uncertain Utility Alignment: Task3 critically evaluates the individual decision-making processes through reflective analysis to identify biases and correct decision-making mistakes. The key idea is to use the server agent to simulate a dynamic learning process that encompasses thinking, reflection, and reinforcement, thus forming internally grounded utility judgments. The core pipelines of Task3 are: 1) *Exploring Potential Decisions*: model the agent's candidate decisions as a structured graph and employ Uniform-Cost Search to identify more promising paths in the decision space; 2) *Utility Alignment Learning*: the server agent estimates the utilities of different agents' decisions through an Elo-based posterior ranking mechanism. After multiple comparisons, the Elo scores of different agents gradually stabilize at a value that reflects their relative utility in the task. Guided by these learned utilities, the server agent searches for decisions with higher utility and ultimately selects the best one as the final decision. Through the iterative utility judgment, Task3 can assess the utility of different agents' decisions and then can judge the highest utility to derive the best solution.

Conclusion

This project aims to address a critical and timely challenge in the evolution of the agentic AI: how to enable the privacy-preserving collaboration among autonomous agents without sacrificing decision quality, system robustness, or operational manageability. As LLM-based multi-agent systems are increasingly deployed in sensitive domains, privacy can no longer be treated as a peripheral concern or a post-safeguard. Instead, it must be embedded into the reasoning, authorization, and coordination mechanisms of the system itself.

By advancing Federated Multi-Agent Systems (FMAS), this project responds directly to this need and integrates symbolic reasoning, context-aware authorization, and reflective utility learning into a unified and implementable framework. The goal is to establish the foundations of privacy-enhancing multi-agent systems so that autonomous agents can collaborate robustly across diverse environments without sacrificing system performance, ultimately enabling a more trustworthy and human-centered future for agentic AI.

References

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang et al., "A survey on evaluation of large language models," ACM Transactions on Intelligent Systems and Technology, vol. 15, no. 3, pp. 1–45, 2024.

[2] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 8, pp. 5625–5644, 2024.

[3] S. Du, J. Zhao, J. Shi, Z. Xie, X. Jiang, Y. Bai, and L. He, "A survey on the optimization of large language model-based agents," ACM Computing Surveys, vol. 58, no. 9, pp. 1–37, 2026.

[4] D. Das, S. Chernova, and B. Kim, "State2explanation: Concept-based explanations to benefit agent learning and user understanding," NeurIPS, vol. 36, pp. 67 156–67 182, 2023.

[5] Z. Wang, S. Cai, G. Chen, A. Liu, X. S. Ma, and Y. Liang, "Describe, explain, plan and select: interactive planning with llms enables open-world multi-task agents," NeurIPS, vol. 36, pp. 34 153–34 189, 2023.

[6] Z. Cheng, H. Wang, Z. Liu, Y. Guo, Y. Guo, Y. Wang, and H. Wang, "Toolspectrum: Towards personalized tool utilization for large language models," in Association for Computational Linguistics, 2025, pp. 20 679–20 699.

[7] Z. Xi, Y. Ding, W. Chen, B. Hong, H. Guo, J. Wang, X. Guo, D. Yang, C. Liao, W. He et al., "Agentgym: Evaluating and training large language model-based agents across diverse environments," in Association for Computational Linguistics, 2025, pp. 27 914–27 961.

[8] H. He, W. Yao, K. Ma, W. Yu, Y. Dai, H. Zhang, Z. Lan, and D. Yu, "Webvoyager: Building an end-to-end web agent with large multimodal models," in Association for Computational Linguistics, 2024, pp. 6864–6890.

[9] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin et al., "Metagpt: Meta programming for a multi-agent collaborative framework," in ICLR, 2023.

[10] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: a survey," Artificial Intelligence Review, vol. 55, no. 2, pp. 895–943, 2022.

[11] P. Peigné, M. Kniejski, F. Sondej, M. David, J. Hoelscher-Obermaier, C. S. de Witt, and E. Kran, "Multi-agent security tax: Trading off security and collaboration capabilities in multi-agent systems," in AAAI, vol. 39, no. 26, 2025, pp. 27 573–27 581.

[12] C. Ma, J. Li, K. Wei, B. Liu, M. Ding, L. Yuan, Z. Han, and H. V. Poor, "Trusted ai in multiagent systems: An overview of privacy and security for distributed learning," Proceedings of the IEEE, vol. 111, no. 9, pp. 1097–1132, 2023.

[13] S. R. Motwani, M. Baranchuk, M. Strohmeier, V. Bolina, P. H. Torr, L. Hammond, and C. S. de Witt, "Secret collusion among ai agents: Multi-agent deception via steganography," NeurIPS, vol. 37, pp. 73 439–73 486, 2024.

[14] X. Guo, H. Xia, Z. Liu, H. Cao, Z. Yang, Z. Liu, S. Wang, J. Niu, C. Wang, Y. Wang et al., "Fineval: A chinese financial domain knowledge evaluation benchmark for large language models," in NAACL, 2025, pp. 6258–6292.

[15] J. Qiu, K. Lam, G. Li, A. Acharya, T. Y. Wong, A. Darzi, W. Yuan, and E. J. Topol, "Llm-based agentic systems in medicine and healthcare," Nature Machine Intelligence, vol. 6, no. 12, pp. 1418–1420, 2024.

[16] T. Kim, J. Lee, S. Yoon, S. Kim, and D. Lee, "Towards personalized conversational sales agents: Contextual user profiling for strategic action," in EMNLP, 2025, pp. 5131–5154.

[17] Z. Zhao, C. Vania, S. Kayal, N. Khan, S. B. Cohen, and E. Yilmaz, "Personalens: A benchmark for personalization evaluation in conversational ai assistants," in Association for Computational Linguistics, 2025, pp. 18 023–18 055.

[18] C. Sun, S. Huang, and D. Pompili, "Llm-based multi-agent decision-making: Challenges and future directions," IEEE Robotics and Automation Letters, vol. 10, no. 6, pp. 5681–5688, 2025.