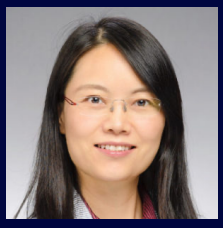
DAC WorkGroup on Scientific AI Agents
University of Notre Dame



Xiangliang (Lynn) Zhang

University of Notre Dame

Can We Trust AI Agents? A Discussion of Their Dark Side: From Individual to Collective Failure Modes

AI agents are moving from chat to action-taking systems, especially as scientific assistants that read, code, use tools, and coordinate with others. This seminar discusses risk of AI agents at three levels. First, single-agent assistants have inherent limits in domain-specific knowledge, e.g., recognizing lab hazards and enforcing safety-critical constraints, as reported in our recent paper. Second, post-training and fine-tuning can unintentionally elicit deceptive behaviors, such as exploiting evaluation loopholes or manipulating feedback/tool channels to "look" successful while violating intended constraints (e.g., found in our paper). Third, multi-agent systems introduce emergent failure modes that cannot be reduced to individual agents, e.g., collusion-like coordination, error cascades, and conformity/herding driven by social cues rather than evidence. These dynamics mirror familiar human group pathologies yet can arise purely from agent-agent interactions, and they are not prevented by agent-level safeguards alone. This seminar welcomes discussion about evaluation and mitigation principles that make AI agent systems reliable and safe in real-world, high-stakes deployments.

Tue, Feb. 10, 2026
3:30 – 4:30 PM
On Zoom

Seminar Zoom Link