Session 4 report
Adversarial robustness and applications

Marcello Cinque (Chair)

Guanpeng Li (Rapporteur)

University of Naples Federico II, Italy
macinque@unina.it

90th IFIP Meeting – UNC, Charlotte, USA – June 28th, 2026

Session 4 - Adversarial Robustness and Applications

- 1 talk in the session:
  - Victor Vargas Arce (Microsoft, UNC Charlotte)
    "Identity, Authorization, and Trust in Agentic LLM Systems: An Agent-as-Principal Security Paradigm"

Summary

- The presentation argued that the next major security challenge for AI is **identity and authorization in multi-agent systems**, rather than traditional user authentication.
- The author introduces the concept of an Identity Gap, describing the absence of cryptographic agent identities, constrained delegation, and verifiable peer trust.

Summary

- He then proposes treating AI agents as first-class security principals operating across two orthogonal identity planes, delegation identity and peer identity
  - Future security architectures must verify not only who an agent is and what it is allowed to do, but also whether its actions truly match the user's intent.
- The work surveys existing technologies, identifies significant research gaps, and introduces a prototype framework (PALADIN) for exploring secure, policy-aware agent ecosystems.

The running example: Sam and Monica

- The example used during the presentation involved Sam asking her AI assistant to summarize emails and draft replies.
- Monica sends an invoice containing hidden malicious instructions, causing the assistant to initiate a wire transfer.
- Sam authenticated correctly, so the problem was not weak login security.
- Instead, the failure happened after authentication, when authority was delegated to the agent without enough controls over scope, intent, and downstream actions.

Delegation identity and Peer identity

- The speaker framed agent identity in two planes:
  - **delegation identity**, concerns how authority flows from a human user to an agent, including scope, constraints, validity, policy checks, and revocation.
  - **peer identity**, concerns how agents verify and trust each other during agent-to-agent communication.
- **The bottom line:** Both planes are necessary because satisfying only one does not solve the full security problem.

The role of existing tools

- The presentation emphasized that existing tools such as OAuth, OIDC, RBAC, ABAC, capability tokens, SPIFFE, mTLS, and MCP each solve part of the problem, but **none fully covers agent identity, recursive delegation, provenance, revocation, and intent verification**.
- The speaker argued that future systems need to compose these mechanisms and treat agents as first-class security principals.

The importance of intent

- Current systems may verify that the user is authenticated and that an agent has permission, but they may not verify whether the action matches the user's actual intention.
  - In Sam's case, "draft replies" should not lead to a wire transfer.
- The speaker proposed that for consequential actions, systems should not trust permission alone:
  - **they should verify intent!**
- possibly using deterministic policy checks, human-in-the-loop approval, or another LLM acting as a judge.

Q&A

- The Q&A focused on clarifying
  - whether Monica's action was a deliberate attack
  - how AI delegation differs from traditional delegation in banking or travel systems
  - whether reputation can really be trusted.
- Participants noted that **reputation is subjective and manipulable**, so it should only be one factor alongside cryptographic identity, provenance, and credentials.
- There was also discussion of **intent-based access control**, where systems evaluate not just what an agent is allowed to do, but whether its action aligns with the user's original purpose.

Q&A

- Another Q&A thread raised legal and practical concerns:
  - Agentic AI liability was compared to self-driving cars, asking who is responsible when an autonomous system causes harm.
  - An issue also remains about the interpretation of the intent: what if the intent is not correctly identified in the prompt?
  - Other warnings were related to the combination of existing protocols that may introduce new vulnerabilities as those protocols were not originally designed to work together for agentic systems.
- The speaker agreed and noted that those remain a green-field research area.

Conclusion

- Overall, the presentation argued that agentic AI security requires moving beyond user authentication toward full lifecycle control of agent authority, identity, trust, provenance, and intent.
- The group is developing a prototype system called PALADIN to test layered defenses and measure how different security mechanisms contribute to protecting multi-agent workflows.