CONCEPT PAPER

Accelerating the Adoption of Software and AI Agent Identity and Authorization

Harold Booth
Bill Fisher
Ryan Galluzzo
Joshua Roberts
National Institute of Standards and Technology

DRAFT

February 2026

AI-Identity@nist.gov

NIST
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

The National Cybersecurity Center of Excellence (NCCoE) at the National Institute of Standards and Technology (NIST) addresses businesses' most pressing cybersecurity problems with practical, standards-based solutions using commercially available technologies. The NCCoE collaborates with industry, academic, and government experts to build modular, open, integrated reference designs that are broadly applicable and repeatable. To learn more about the NCCoE, visit http://nccoe.nist.gov. To learn more about NIST, visit http://www.nist.gov.

ABSTRACT

AI agents offer the promise of improved productivity, efficiency, and decision-making in complex scenarios. But these benefits cannot be realized without the ability to understand how identity principles such as identification, authentication, and authorization can apply to agents to provide appropriate protections while enabling business value. This concept paper seeks stakeholder input to inform a NIST National Cybersecurity Center of Excellence (NCCoE) project focused on applying existing identity standards and best practices to software and AI agents. Such a project would aim to reduce implementation risk related to agentic AI by demonstrating how identity and authorization standards and best practice can be applied to agentic architectures. Feedback received will help determine the scope, feasibility, and potential value of the project and inform whether a demonstration effort or other NCCoE outputs would best address the challenge. Community input will inform subsequent project planning activities, which could include development of a draft project description and a call for collaborators.

KEYWORDS

authentication; authorization; identity and access management; AI; Artificial Intelligence, Agentic AI, Software Agents, Prompt Injection.

DISCLAIMER

Certain commercial entities, equipment, products, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST or NCCoE, nor is it intended to imply that the entities, equipment, products, or materials are necessarily the best available for the purpose.

While NIST and the NCCoE address goals of improving management of cybersecurity and privacy risk through outreach and application of standards and best practices, it is the stakeholder's responsibility to fully perform a risk assessment to include the current threat, vulnerabilities, likelihood of a compromise, and the impact should the threat be realized before adopting cybersecurity measures such as this recommendation.

Comments on NCCoE Documents

Individuals and organizations are encouraged to review all draft publications during public comment periods and provide feedback. All publications from NIST's National Cybersecurity Center of Excellence are available at http://nccoe.nist.gov.

Comments on this publication may be submitted to: AI-Identity@nist.gov

Public comment period: February 5th, 2026 – April 2nd, 2026

All comments are subject to release under the Freedom of Information Act.

Note to Reviewers

Artificial Intelligence (AI) technology brings great opportunities to organizations. Specifically, AI agents offer the promise of improved productivity, efficiency, and decision-making in complex scenarios. But these benefits cannot be realized without the ability to understand the security properties of deployed agents and apply appropriate controls as they access diverse data sets, tools, and applications to execute their mission. More specifically, organizations need to understand how identity principles such as identification, authentication, and authorization can apply to agents to provide appropriate protections while enabling business value.

NIST recognizes the need to better understand these challenges as agencies and organizations consider adopting agentic capabilities. As such the National Cybersecurity Center of Excellence is considering a demonstration of how identity and authorization standards and best practices can be applied to AI agents. To inform its next steps, NIST is seeking input on the technical and operational considerations, standards and technology landscape, and overall scope, focus, and value of this project. In particular, NIST is interested in stakeholder perspectives on the following questions. These questions complement but are distinct from an RFI about securing AI agents issued by the Center for AI Standards and Innovation (CAISI) within NIST that will inform guidelines development for broader research agendas. Responses will be used to inform the scope, priorities and technical feasibility of a NCCoE demonstration project.

1. General Questions to inform choice of Demonstration Use Case

- What enterprise use-cases are organizations currently using agents for?
- Which use-cases are in the near future?
- What opportunities do agents present?
- What risks worry you about agents?
- What are the core characteristics of agentic architectures?
- What support are you seeing for new protocols such as Model Context Protocol (MCP)?
- In what ways do agentic architectures introduce identity and authorization challenges?
  - How do AI agents differ from other forms of software agents?
  - How are agentic architectures different from current microservices architectures?
- What current or roadmap technology does your organization have that supports agents?
- What standards exist, or are emerging, to support identity and access management of agents? How might these need to be adapted to support new security risks or paradigms introduced by AI agents?

2. **Identification**
    - How might agents be identified in an enterprise architecture?
        - What metadata is essential for an AI agent's identity?
        - Should agent identity metadata be ephemeral (e.g. task dependent) or is it fixed?
    - Should agent identities be tied to specific hardware, software, or organizational boundaries? How would this be enforced?
3. **Authentication**
    - What constitutes a strong authentication for an AI agent?
    - How do we handle key management for agents? Issuance, update, and revocation?
4. **Authorization**
    - How can zero-trust principles be applied to agent authorization?
    - Can authorization policies be dynamically updated when an agent context changes?
        - For example, if an agent gets access to new tools and resources, how do we determine sensitivity levels of data when aggregated by an agent, and whether users are authorized to access the aggregated response?
    - How do we establish "least privilege" for an agent, especially when its required actions might not be fully predictable when deployed?
    - What are the mechanisms for an agent to prove its authority to perform a specific action?
    - How might an agent convey the intent of its actions?
    - How do we handle delegation of authority for “on behalf of” scenarios?
    - How do we bind agent identity with human identity to support “human-in-the-loop” authorizations?
5. **Auditing and non-repudiation**
    - How can we ensure that agents log their actions and intent in a tamper-proof and verifiable manner?
    - How do we ensure non-repudiation for agent actions and binding back to human authorization?
6. **Prompt Injection prevention and mitigation**
    - What controls help prevent both direct and indirect prompt injections?
    - After prompt injection occurs, what controls/practices can minimize the impact of the injection?

Feel free to share your thoughts with us via AI-Identity@nist.gov by April 2nd, 2026.

1. Project Concept

The NIST National Cybersecurity Center of Excellence is planning a project focused on applying identity standards and best practices to AI agents. This concept paper introduces the technical focus and scope of the proposed project, including the nature of the challenge, the types of architectures considered, and the identity standards that could be applied as part of this effort. NIST is seeking feedback from stakeholders and technology collaborators on the technical reality and reasonableness of this concept and is open to suggestions on how standards and best practices can be applied to address this challenge.

Challenge Overview

For well over a decade, code-based systems have been used to enable automation, cloud workloads, and the deployment of APIs. However, with the advancement of software and AI agents—systems that have the capability for autonomous decision-making and taking action with limited human supervision to achieve complex goals—the scale and range of actions taken by these systems has the potential to increase exponentially. This increased scale and autonomy brings new opportunities as well as new risks. To enable effective management of these risks and to securely capitalize on these opportunities, enterprises and individuals need to understand how foundational identity principles—identification, authentication, and authorization—can be applied to ensure that agents are known, trusted, and properly governed.

Scope

This project will focus on applying identity standards and best practices to agentic architectures as depicted in Figure 1. Agentic architectures are ones that take in some set of instructions, dynamically acquire additional context from other resources based on those instructions, process the results, potentially take some sort of action and return a response. Retrieval-Augmented Generation (RAG) and architectures using only an LLM with its associated training data are out of scope of our project. Appendix A offers a supplement flow diagram.

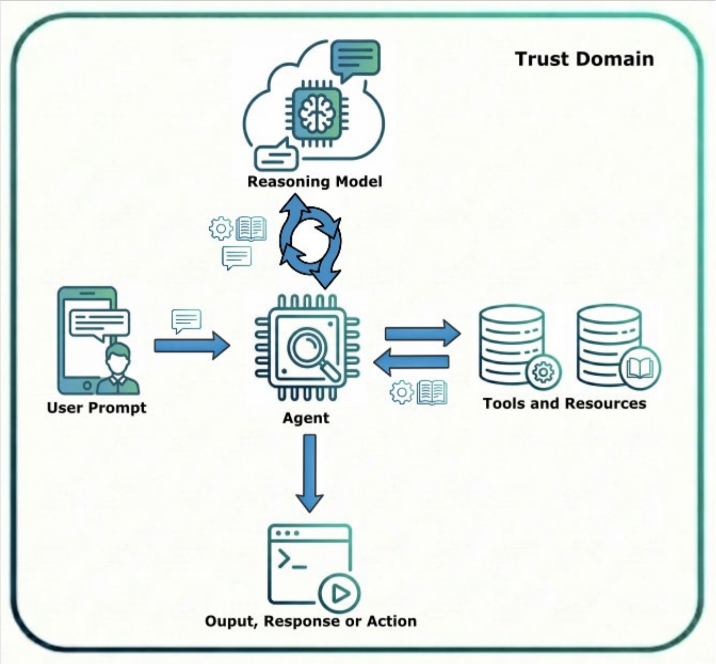

Figure 1. Example Agentic Architecture

Areas of Interest

The areas below describe potential focus areas for exploration:

- **Identification of AI and Software Systems.** Leveraging existing standards, the project will explore available means to identify software and AI agents such that access management systems can distinguish between agent and human identities and effectively manage the range of actions an agent may take from controlled human-in-the-loop approval to autonomous action in response to an input.

- **Authorization of AI Systems.** Leveraging standards such as OAuth 2.0 and its extensions and policy-based access control mechanisms, to manage how rights and entitlements are granted to software and AI agents and to enforce access decisions based on the identity of the AI agent or software systems.

- **Access Delegation.** Link specific user identities to AI agents or software systems to support effective delegation controls and maintain accountability for the actions of automated systems.

- **Logging and Transparency.** Link specific AI agent and software systems actions to the identity of the non-human entity and enable effective visibility into the actions taken, data generated, and outcomes of automated activities within a given system, platform, or network.

- **Tracking Data Flows of an AI System.** Track and maintain provenance of user prompts and data input sources to support risk determinations and policy decisions regarding actions to be taken by an AI Agent.

2. Relevant Standards and Guidelines

This project is currently considering the implementation of the following standards and best practices:

- **Model Context Protocol:** Model Context Protocol (MCP) is a protocol that enables AI models and agentic systems to discover, access, and interact with external tools, data sources, and services in a consistent and structured manner. The MCP protocol relies on existing identity standards such as Open Authorization (OAuth) and Open ID Connect (OIDC) for rights delegation and authentication.
- **OAuth 2.0/2.1 and extensions:** OAuth is an authorization standard that can be used to support access control objectives. The standard defines a set of technical specifications for the generation, protection, and delivery of authorization tokens (JSON Web Tokens or JWT) to different connected endpoints (e.g., servers). There are multiple profiles and extensions of OAuth to support specific use cases, security properties, and features. At this time, OAuth is integrated

into the MCP as the primary method for authorizing agentic access. The specification follows the draft OAuth 2.1 standard.

- **OpenID Connect:** OIDC is an interoperable authentication protocol based on the OAuth 2.0 framework of specifications. Essentially, it provides a consistent way for expressing authentication, consent, and authorization information through identity tokens that can support access outcomes related to Agents or users when interacting with Agents.

- **SPIFFE/SPIRE:** Secure Production Identity Framework for Everyone (SPIFFE) is a framework for issuing and managing cryptographic identities to workloads and SPIFFE Runtime Environment (SPIRE) is an implementation of SPIFFE that provides APIs for workload attestation. Together they represent one way in which agent workloads could be identified and authenticated.

- **System for Cross-domain Identity Management:** System for Cross-domain Identity Management (SCIM) is a standard that defines RESTful APIs and JSON schemas for automating the provisioning, deprovisioning, and lifecycle management of identities across systems. While SCIM does not provide authentication or authorization, it does provide a potential way to create, update and revoke agent identities across systems.

- **Next Generation Access Control:** Next Generation Access Control (NGAC) is an attribute-based access control standard that represents access control policies in a unified graph of users, objects, attributes, and policy classes to enable fine grained access control across a wide breadth of policies and resources. NGAC also supports event driven policy updates, native delegation and least privilege making it suitable for agentic systems.

NIST will also apply relevant guidelines from SP 800-207 Zero Trust Architecture, SP800-63-4 Digital Identity Guidelines, NISTIR 8587 Protecting Tokens and Assertions from Forgery, Theft, and Misuse and other NIST guidelines as applicable.

We are open to feedback on other models, methodologies, protocols, best practices, or standards that might address this challenge.

3. Possible Use Cases

The focus of the project will be on enterprise use-cases where greater control and visibility can be maintained over agents and the systems they access. The challenge of identifying and managing access for external agents from untrusted sources will not be addressed under this initial effort, but use-cases focused on public facing or individual agents could be addressed in future iterations of the project.

NIST is actively seeking feedback on real-world use cases being evaluated by agencies and enterprises. Potential use-cases could include the following:

- **Enterprise AI agents to improve work force efficiency and decision making.** This use case would focus on implementing controls to address the use of AI agents and software to improve staff efficiency in everyday tasks (e.g., managing calendars, assessing and creating policy documents, generating decision recommendations). To support this use case, agents and software will need delegated and managed access to multiple data sources to take actions based on user prompts or inputs.

- **Enterprise AI agents for security.** This use case would focus on agents and software that analyze security information and either take or recommend security actions for an organization. As with use case #1, this will include non-human identities that access data from across a set of connected systems, but with an elevated risk due to the sensitivity of security data.

- **Enterprise AI agents for software development and deployment**. This use case would focus on automated processes for developing and deploying software and how entitlements and authorization are supported in automated deployment pipelines that use AI Agents.

4. Desired Outcomes

The planned NCCoE project on software and AI agent identity and authorization will focus on producing practical, implementation-oriented guidelines to help organizations adopt agentic capabilities while managing cybersecurity risk. Consistent with the NCCoE mission, the ultimate deliverable will be a practice guide detailing example implementation details built in the NCCoE laboratories using commercially available technologies, along with key lessons learned along the way. Similar to the recent Mobile Driver's License project, this project intends to iteratively provide outputs that increase awareness of the overall technology and security space related to agentic AI identity and authorization.

Overall, this project seeks to:

- Provide a better understanding of how agents can be deployed in line with identity and authorization standards and best practices to help agencies and enterprise maximize value and minimize risk
- Create relationships and mechanisms to provide feedback to standards development entities as they advance and evolve standards in the agentic ecosystem
- Identify and communicate risks and opportunities associated with real-world deployments of Agentic AI solutions

- Provide detailed implementation resources that can enable more rapid adoption of agentic technology, consistent with risk management and organizational goals

Seeking Public Comment

The NCCoE is open to suggestions on how NCCoE resources may be able to advance the adoption of sound security principles and best practices relating to the identification, authentication, and authorization of AI agents.

Based upon community feedback on these topics, the NCCoE will consider instantiating a project to engage in building an example solution using commercially available technology. Public comments on this concept paper will help the NCCoE understand specific challenges and needs and may be used to help define a project description.

Comments on this publication may be submitted to: AI-Identity@nist.gov

APPENDIX A. EXAMPLE AGENTIC ARCHITECTURE FLOW DIAGRAM

The below flow diagram offers a sequential view of how the different components of an agentic architecture might interact. Of note is the iterative nature of the interactions between the agent and the reasoning model (such as an LLM), where an agent may fetch tools and resources for the reasoning model multiple times to update the model with data, context or prompts.

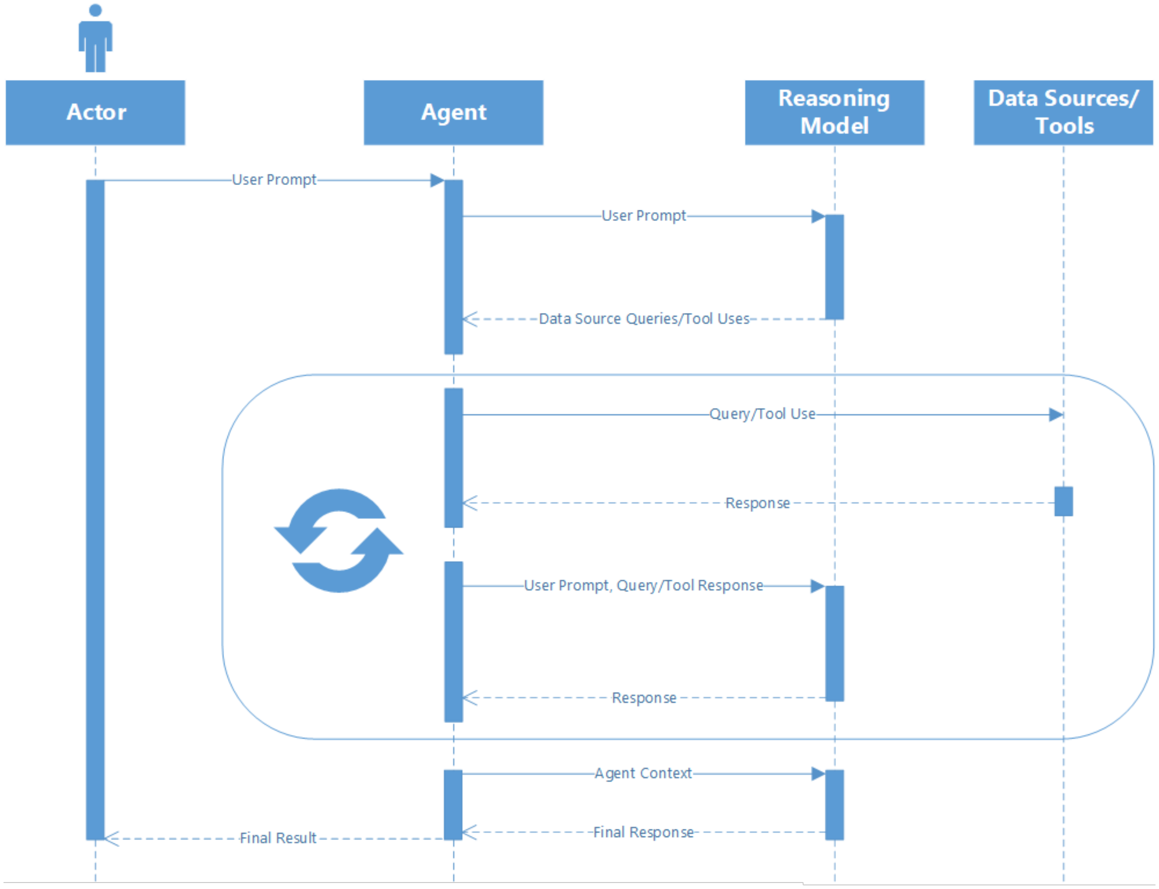

Figure 2. Example Agentic Flow Diagram