Intent governance in the agentic AI age

Akiko Gondoh

17 August 2026

Iakyara vol.418

Nomura Research Institute, Ltd.

Akiko Gondoh
Group Manager
IT Platform Technology Strategy Department

Executive Summary

AI agents are set to go live in the financial sector but existing authentication, authorization and auditing functions are not designed for autonomous AI. As a result, financial institutions may face risks arising from misalignment between a human user's intended objectives and an AI agent's actions. The intent data generated through AI agent interactions will become a valuable strategic asset. However, effectively governing, securing, and managing this new class of data will present a challenge for financial institutions.

Challenges posed by AI agents are driving standardization

As AI agents begin executing internal workflows and conducting transactions on behalf of employees and customers, financial institutions face new operational and security risks. Existing identity, authentication, authorization, and auditing frameworks were designed for human users, not autonomous agents acting as human proxies. Agentic AI therefore requires a new trust framework that verifies not only the agent itself but also the identity and intent of the human it represents.

Four key challenges have emerged:

- Identity verification: Distinguishing legitimate AI agents from malicious bots and verifying the human principal they represent.
- Dynamic authorization: Granting AI agents only the minimum permissions required as they are dynamically created and assigned complex tasks.
- Robustification: Preventing prompt injection and other malicious inputs from manipulating agent behavior.
- Execution controls and auditability: Monitoring, governing, and auditing non-deterministic AI actions performed at machine speed and scale.

To address these issues, the industry is actively developing technical standards. Examples include Anthropic's Model Context Protocol[1] (MCP) and Google's Agent Payments Protocol[2] (AP2), both of which have been submitted to industry organizations, namely the Agentic AI Foundation[3] (AAIF) and the FIDO (Fast Identity Online) Alliance[4], respectively. However, real-world incidents suggest that technical standards alone will not eliminate operational risk.

NOTE

1) MCP is an open standard for connecting AI agents with external data sources, tools and applications.
2) AP2 is a protocol for AI agents to autonomously make payments on behalf of humans.
3) The AAIF is an international industry organization established by the Linux Foundation to ensure the safety, reliability, and interoperability of autonomous AI agents.
4) The FIDO Alliance is an industry association that develops international standards for secure user authentication utilizing public-key cryptography and biometrics, without relying on passwords.

A notable example occurred in March 2026, when an AI agent responding to an internal engineering discussion at Meta autonomously posted an incorrect answer without first obtaining human approval. Another engineer relied on the response, unintentionally exposing confidential information to unauthorized employees for several hours.

This mishap was attributable to deficiencies in both authorization management and execution controls. First, the AI agent's authorization to access or read the forum was treated as equivalent to authorization to update or post forum content externally. Second, the system lacked execution controls requiring human approval before taking externally impactful actions. A human agent working for a human principal would have been able to recognize based on the situation that her actions could potentially lead to an information security breach. AI agents, by contrast, do not possess the requisite background knowledge and common sense to reliably make such judgments. The Meta incident was a precursory warning of the risks of an AI agent autonomously executing externally impactful actions based on its interpretation of human instructions (intent).

How to interpret intent

How are the players at the forefront of the standardization movement addressing AI agents' interpretation and execution of their principals' intent? In the payments space, multiple initiatives are afoot, including Google's AP2 and MasterCard's Verifiable Intent, both of which verify the intent of the principal's delegation of purchase authority to an AI agent. What is actually verified, however, is operationalized intent, which is merely information resulting from a human prompt (declared intent) being translated into agentic action(s). The process of operationalizing declared intent is left to the AI agent (or application) itself.

How do we ensure that declared intent is operationalized in accord with the human principal's actual intent? One way, the so-called human-in-the-loop (HITL) approach, is to require every agentic action to be approved by a human and the agent to furnish proof of such approval. The proof takes the form of specific purchase instructions (operationalized intent) such as SKU(s), maximum purchase price(s) and eligible merchant(s). Most importantly, the purchase terms are legitimized not by merely tapping an approval button in an app but by cryptographically signing them with the principal's private key and the AI agent's public key. This process creates an immutable audit trail that, if a transaction is

subsequently questioned, verifies whether the transaction was properly authorized by the human principal or a rogue AI purchase.

However, the human approval process becomes a bottleneck when faced with AI that exceeds the limits of human cognitive processing speed. With AI agents expected to perform a wide variety of everyday tasks on our behalf in both our jobs and personal lives, the HITL approach will obviously reach its limits at some point.

To circumvent HITL's limitations, researchers have proposed codifying human approvals as policies. Under this approach, principals would first preapprove a one-time authorization subject to specified limitations (e.g., spending limit, number of purchases) instead of approving every single transaction in real time. An AI agent would then autonomously execute a series of transactions within the constraints of the preapproval. Grant Miller, Chair of the AAIF's Identity & Trust Working Group, has proposed registering HITL-authorized decisions as policies in identity governance and administration (IGA) platforms[5]. The proposal will likely be incorporated into protocols being developed by Miller's Working Group. Generation of machine-readable policies through HITL permissioning will likely broaden beyond the domain of agentic AI for limited use cases such as shopping or routine business processes.

5) An IGA platform centrally manages and monitors system users' ID information and access privileges within an organization.

Intent data poses new security challenges

The intent data generated through AI agents' activities, including both declared and operationalized intent, has the potential to become a new strategic digital asset by capturing previously undocumented personal values and tacit organizational rules. One point that tends to be overlooked is that confidential information associated with agentic AI includes not only prompts inputted by humans but also operationalized intent derived from the AI itself. While safeguarding prompts' confidentiality is currently under discussion, how to manage operationalized intent, which is often generated without the human user's awareness, as an information asset is an issue that remains to be addressed in the future. Operationalized intent could heighten operational and privacy risks also. Prospective solutions include anonymization and rigorous safeguards built into terminal hardware.

about NRI

Founded in 1965, Nomura Research Institute (NRI) is a leading global provider of system solutions and consulting services with annual sales above $5.1 billion. NRI offers clients holistic support of all aspects of operations from back- to front-office, with NRI's research expertise and innovative solutions as well as understanding of operational challenges faced by financial services firms. The clients include broker-dealers, asset managers, banks and insurance providers. NRI has its offices globally including Tokyo, New York, London, Beijing and Sydney, with over 16,800 employees.

For more information, visit https://www.nri.com/en

Inquiries to : Business Planning Department, Financial Technology Solution Division
Nomura Research Institute, Ltd.
Otemachi Financial City Grand Cube,
1-9-2 Otemachi, Chiyoda-ku, Tokyo 100-0004, Japan
E-mail : kyara@nri.co.jp

https://www.nri.com/en/knowledge/publication/list.html#lakyara