



ISO/IEC JTC 1/SC 27/WG 3 "Security evaluation, testing and specification"

Convenorship: UNE

Convenor: Bañón Miguel Mr



Updated proposal for PWI on security guidelines for AI agents_20260917

Document type	Related content	Document date	Expected action
Project / Other	Meeting: VIRTUAL 14 Sep 2026	2026-09-17	INFO

Replaces: N 3158 Security guidelines for AI agents

PWI on security guidelines and evaluation methods for AI agents and their orchestration

AI agents and their orchestration

Per ISO/IEC 22989:2022, an AI agent is an automated entity that senses and responds to its environment and takes actions to achieve its goals.

In the context of this document, such an agent is realized as an autonomous software program powered by artificial intelligence (AI), often a Large Language Model (LLM), that perceives its environment, reasons through problems, and executes actions using external tools to achieve specific goals without requiring step-by-step human intervention. An AI agent combines an AI model with tools, memory, and instructions, together with an execution loop that turns the model's reasoning into real-world actions.

For example, an AI agent can be prompted to generate a proof-of-concept (PoC) test for a specific vulnerability. The agent searches relevant source code using commands like `cat`, `grep` or `awk`, queries the LLM with that context to formulate an exploitation strategy, and writes and executes Python code to create and test the PoC.

Because an individual agent typically favours a single tactical approach, you can deploy multiple agents to search for vulnerabilities in different ways. This scales one agent into an agentic system, a framework where multiple agents plan, call tools, and execute multi-step tasks autonomously. AI agent orchestration coordinates multiple such agents so that they accomplish a shared goal. It governs role assignment and task decomposition, agent selection and routing, inter agent communication, tool and API integration, and shared context and memory. The software component or agent that performs this coordination is termed the orchestrator, whether it takes the form of a dedicated controller, a supervisor (or "manager") agent, or the runtime framework itself.

The market has already shifted from single agents toward multi-agent systems, and emerging protocols such as the Model Context Protocol (MCP) and Agent-to-Agent (A2A) are driving interoperability across heterogeneous agent platforms.

Standardization activities related to AI agents

Because agent outputs are no longer confined to text but change the state of external systems, the security boundary extends beyond model input and output to every external data source the model observes, the privileges of its tools, its memory, its execution logs, and its user-approval procedures. Identity, credentials, and access control for autonomous agents remain unresolved, creating demand for international security standards.

ISO/IEC 15408 (all parts) and ISO/IEC 18045 define evaluation criteria for IT security and methodology for IT security evaluation. ISO/IEC 19790 and ISO/IEC 24759 define security requirements for cryptographic modules and test requirements for cryptographic modules. ISO/IEC FDIS 27090 provides guidance on AI security threats, consequences, detection and mitigation. ISO/IEC WD 26160 will define AI-specific security functional requirements (SFRs), security assurance requirements (SARs) and evaluation work units. ISO/IEC AWI TS 42119-7 and ISO/IEC AWI TS 42119-8 address red team testing and quality evaluation, respectively. ISO/IEC CD TS 25568 provides guidance on addressing risks in generative AI systems. ISO/IEC WD 25959 defines attack potential tables for deep learning-based technology. ISO/IEC 22989 provides the AI terminology on which this work relies.

Challenges in securing AI agents and their orchestration

The standards above primarily address threat descriptions, risk-management frameworks, general testing guidance, and the security of individual AI models. A gap remains in standardized security requirements and reproducible evaluation methods for AI agents and orchestration. That is, a gap remains for the composition of AI models with external tools and resources, and multi-agent coordination that produces autonomous behaviour. The challenges below motivate this need, organized around the three-layer view of an agent system: the model layer (foundation, LLM, generative AI models and agentic AI models), the agent layer (a single agent = model + tools + memory + instructions), and the orchestration layer (coordination across multiple agents).

a) Orchestration-specific attack surfaces are not captured by model-level evaluation

Model-level security evaluation, for example testing an LLM's robustness against direct prompt injection, does not capture threats that emerge only when a model is placed inside an agent and coordinated with others. Untrusted content in tool outputs, retrieved documents, web pages, and messages from other agents can be interpreted as instructions and turned into privileged actions. Real incidents illustrate the exposure: the EchoLeak vulnerability (CVE-2025-32711) in Microsoft 365 Copilot exploited indirect prompt injection to exfiltrate data with zero user interaction, and in 2025 an autonomous coding agent deleted a production database while acting on its own plan. Evaluating orchestration security therefore requires following the whole execution path (the agent's action trace), not single input–output pairs.

b) Fragmented and incomparable security claims across platforms and protocols

Orchestration is implemented on heterogeneous frameworks and emerging protocols (e.g. MCP and A2A), each of which reports security differently or not at all. There is no common basis on which to compare the security of two orchestration platforms, and no agreed way to establish identity, credentials, and authorization delegation for agents that interact across different virtual platforms. Standardized security requirements and evaluation methods would enable meaningful comparison and support interoperability-secure deployment across vendors and National Bodies.

c) Domain diversity and inconsistent evaluation criteria

AI agent orchestrations cover a wide variety of application domains, such as educational assistants, code-development agents, and customer-service agents, each with its own threat landscape and domain-specific security requirements. Without common baseline requirements, different evaluators assess orchestrations using different scopes and metrics, leading to inconsistent evaluation quality. A standard evaluation shall measure not only attack success rate but also benign-task success rate, tool-misuse rate, sensitive-data exposure rate, policy-violation rate, human-approval burden, and log reproducibility, so that the trade-off between an orchestration's usefulness and its security is captured.

Proposed scope

This PWI adapts and extends existing evaluation methodology (ISO/IEC 15408/18045, ISO/IEC 19790/24759) and interprets how existing security requirements apply to AI agents and their orchestration.

To remain consistent with ISO/IEC 26160, ISO/IEC 27090, and ISO/IEC 25959, this PWI reuses their security requirements, threat taxonomy, and attack-potential model as the normative baseline, and develops the corresponding evaluation methods and test procedures on top of them.

This PWI includes the following action items:

1. Security requirements and evaluation methods for AI agents — based on ISO/IEC 15408 and 18045, AI model-specific requirements referenced from ISO/IEC 26160/SC 42, with extensions to the evaluation activities (work units) for the agentic technology.
2. Security requirements and evaluation methods for orchestrators in AI agent orchestration — based on ISO/IEC 15408/18045, with AI model-specific requirements referenced from ISO/IEC 26160/SC 42 and extensions to the evaluation work units for orchestration specifics (plan/control-flow integrity, inter-agent communication, shared memory).
3. Verification of security protocol for identity proof and verifiable credentials between AI agents — referencing established specifications (e.g., FIDO2/WebAuthn, OAuth 2.0/OpenID Connect, etc.), developing verification methods for their security properties, and extending these methods to emerging agent-specific protocols.
4. Security requirements and test requirements for authentication and delegated authorization of AI agents — following ISO/IEC 19790/24759, and applying their cryptographic-mechanism evaluation where credentials and keys are handled, extending it to re-delegation.
5. Investigating the application of attack potential to agentic AI-based technologies and their interactions for ISO/IEC 25959 update or part 2 NP.
6. For each evaluation activity, specifying test conditions, sample and behaviour construction procedures, and execution steps.
7. Planned activities:
 - Developing the evaluation methodologies without CC terms (e.g., SFR or SAR), as we plan to request a review from SC42 via SC27 AHG3.
 - Conducting a gap analysis against ISO/IEC 15408 and ISO/IEC 18045, identifying Extended SFRs, Extended SARs, SAR application notes, dedicated evaluation activities and new attack potential factors.
 - Integrating our findings into ISO/IEC 26160 or ISO/IEC 25959.
 - Proposing NPs under WG3 if the outcomes are successful.

Collaboration with other SC/WG

We propose establishing this work item under SC27/WG3. However, because parts of the proposal fall within the scope of other SC/WGs, particularly SC42, we will submit the approved PWI proposal to the SC27 AHG3 (Security and privacy in AI and Big Data (BD)) following the September SC27/WG3 meeting to ensure proper cross-SC/WGs coordination.

Proposed project editors

Heebong Choi (KR), Naruki Kai (JP), Hoonjae Lee (KR), Jiazhe Chen(CN), Seira Hidano(JP), Kangwoo Lee(KR)