FRONTIER MODEL FORUM

March 9, 2026

U.S. National Institute of Standards and Technology
Center for Artificial Intelligence Standards and Innovation
100 Bureau Drive
Gaithersburg, MD 20899

To Whom It May Concern:

The Frontier Model Forum (FMF) is grateful to the U.S. Center for AI Standards and Innovation (CAISI) for developing voluntary guidelines on AI agent security. We share the CAISI's view that best practice guidance is essential for building trust in AI agents and ensuring their safe and responsible deployment.

The FMF is an industry-supported non-profit dedicated to advancing the secure development and deployment of frontier AI systems. We leverage the technical and operational expertise of our member firms, as well as the broader scientific community, to develop industry best practices for managing the most significant national security and public safety risks related to frontier AI systems.

As AI agents become more autonomous and acquire greater capabilities, they can heighten potential risks as well as introduce significant opportunities for cyberdefense. AI agents are already being deployed to search for and identify previously unknown vulnerabilities in software, including in widely used open-source projects, but realizing their full defensive potential will depend on clear, well-designed security guidance. That the field of agentic security is still nascent, and AI agents remain challenging to evaluate for safety and security vulnerabilities, also reinforces the need for rigorous, well-grounded guidance.

We welcome the CAISI's request for information regarding the security considerations of AI agents, as well as its broader efforts to advance new methodologies for agentic security and establish more mature security practices.

Distinctive security risks for AI agents

As the RFI notes, AI agents pose many of the risks of their underlying AI models, while also introducing additional risks due to their unique capabilities and affordances. Since the underlying models used by AI agents are often non-deterministic, AI agents can produce

different outputs from the same inputs and generate unpredictable behaviors. In addition, complex, emergent behaviors can also arise that weren't explicitly programmed. Greater autonomy in decision-making and tool use increases the potential scope and impact of security errors and increases exposure to exploitation by malicious actors. Additionally, effectively managing agent identity and access privileges presents significant challenges.

We identify four primary security concerns:

1. **Rogue Action.** Agents may produce unintended, harmful, or policy-violating behaviors. A core driver of this risk is prompt injection, a currently unsolved weakness in large language models that makes agents susceptible to manipulation. Agents' susceptibility to prompt injection remains one of the biggest challenges for the AI industry.

2. **Sensitive Data Disclosure.** Agents operating across tools, APIs, and data sources create new pathways for the unauthorized exposure of private or sensitive information. The breadth of agent access makes containment harder than in traditional software systems. This is compounded by susceptibility to prompt injection, which can be exploited to induce agents to reveal sensitive data.

3. **Integrity Attacks.** Integrity attacks at the model level, particularly during pre-training and fine-tuning phases, remain a significant concern. In agentic systems, data poisoning can introduce latent behaviors that remain dormant until activated by triggers that may appear in ordinary agent inputs and then influence the actions systems take. As agents are integrated into operational workflows, small design or configuration choices can amplify the impact of these integrity failures over time. Recent research both finds that backdoor triggers can be tolerant to variation, and introduces practical approaches for detecting and triaging backdoored open-weight language models at scale, reinforcing the need for integrity evaluations methods.

4. **Unsafe Configuration.** Beyond model-level vulnerabilities, a critical class of operational risk stems from unsafe configuration–a foundational lesson from traditional cybersecurity. Security gaps are amplified when agents are deployed without robust authentication or extended beyond their intended architectural boundaries. These gaps create unauthorized identity and data-access paths that the underlying model's safety filters were not designed to manage. Such configuration weaknesses can transform a localized prompt manipulation into a broader security breach.

Supporting the science of agent security

While agentic security builds on traditional system security, traditional system security approaches are insufficient to address the full spectrum of agentic AI security challenges. Conventional controls can govern the environments in which agents operate and the tools they access, but both the model itself and the harness built around it present risks that traditional methods were not designed to address. Effective agentic security will require combining deterministic controls with dynamic, reasoning-based defenses. Many frontier AI developers are already developing novel security measures, including new design patterns, fine-tuning methods, and guardrails to defend against prompt injections.

We recommend investing in research and developing guidance in the following areas:

- **Tool access and privilege management.** Significant gaps in guidance remain, particularly around how agents should be granted access to tools, how they should operate within complex environments, and how multi-agent systems should interact securely.
- **Multi-agent interaction.** Multi-agent interaction raises new risks, as well as opportunities for defenders.
- **Prompt injection resiliency.** NIST should consider investing in future research on prompt injection resiliency and enabling adversarial testing on highly realistic production environments.
- **Reliability for high-stakes domains.** The broader security ecosystem should weigh approaches for high reliability for agentic usage on high-stakes or security-sensitive domains.
- **Agent evaluation.** Improving prompt injection resiliency and agent reliability will require repeatable methods to measure both consistently. CAISI should invest in developing rigorous evaluation frameworks to assess AI agents across these dimensions, as well as offer general guidance on building and running agent evaluations.

The FMF supports the CAISI's effort to improve AI agent security through developing evaluations, assessments, technical guidance, and best practices that address the unique security risks and vulnerabilities of agentic AI.