TECHNET
THE VOICE OF AMERICAN INNOVATION

1420 New York Avenue NW, Suite 825
Washington, D.C. 20005
www.technet.org | @TechNetUpdate

March 9, 2026

Austin Mayron
Director
Center for AI Standards and Innovation
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

Re: Request for Information Regarding Security Considerations for Artificial Intelligence Agents [Docket No. NIST-2025-0035]

Dear Mr. Mayron,

TechNet appreciates the opportunity to respond to the National Institute of Standards and Technology (NIST) Center for AI Standards and Innovation's (CAISI) Request for Information regarding security considerations for artificial intelligence (AI) agent systems. TechNet is the national, bipartisan network of technology CEOs and senior executives that promotes the growth of the innovation economy by advocating a targeted policy agenda at the federal and 50-state level. TechNet's diverse membership includes dynamic American companies that are developing and deploying agentic systems across consumer, enterprise, and critical infrastructure contexts, as well as the cybersecurity and cloud services that underpin modern digital operations.

AI agent systems represent a meaningful evolution in AI capabilities, enabling systems to plan, coordinate, and take actions across digital environments. These systems hold tremendous promise to improve productivity, strengthen cybersecurity, modernize government services, and expand economic opportunity. At the same time, because AI agents can interact with tools, external data, and real-world systems, they introduce distinct security challenges that merit targeted attention. The policy objective should be to reduce and manage these risks without slowing innovation through premature, overly prescriptive, or one-size-fits-all requirements.

Getting AI agentic security right is essential to realizing the full economic and societal benefits of this emerging technology. TechNet and our member companies support CAISI's leadership in advancing practical, risk-based approaches to AI agentic security that strengthen trust, accelerate responsible deployment, and ensure the United States remains the global leader in AI innovation and adoption.

Our core recommendations include:

- Promote secure-by-design principles for agentic systems
- Combine secure-by-design principles with multilayered defense strategies
- Accelerate safe and secure agentic innovation through targeted federal action
- Advance agentic security through scalable research, simulation, and regulatory sandboxes
- Recognize agentic AI's infancy and allow room for experimentation

Promote Secure-By-Design Principles for Agentic Systems

TechNet strongly supports the adoption of secure-by-design principles as a foundational element of AI agent system security. Because agentic systems introduce new attack surfaces, autonomous decision-making, and the potential for real-world effects, security cannot be treated as an afterthought or retrofitted after deployment. It must be embedded throughout the design, development, and operational lifecycle of these systems. Secure-by-design approaches align with established best practices in cybersecurity and software engineering by shifting responsibility for security outcomes upstream and reducing exploitable vulnerabilities before they reach production.

Industry-led efforts provide a useful model for how secure-by-design principles can be operationalized for agentic AI. The Coalition for Secure AI's (CoSAI) principles for secure-by-design agentic systems offer a practical framework that balances autonomy with strong governance and controls. These principles reflect the reality that securing agentic systems requires a system-level approach that addresses risks at the model layer, the tool layer, and the deployment environment.

First, agentic systems should be human-governed and accountable. Systems should be designed with meaningful human oversight, clear accountability across the lifecycle, and risk-based controls that align authority boundaries with organizational risk tolerance. This approach helps ensure that autonomous agents remain aligned with human intent and that responsibility for security outcomes is not diffused across complex supply chains and deployment settings.

Second, agentic systems should be bounded and resilient. This includes purpose-specific permissions, strong containment of autonomous capabilities, and technical measures that limit the impact of unintended behavior or malicious manipulation. Establishing clear operational boundaries, predictable failure modes, and resilient control structures is essential to reducing the likelihood that an agent can be exploited or pushed into unsafe actions.

Third, agentic systems should be transparent and verifiable. Developers and deployers should incorporate robust logging, monitoring, and telemetry to support auditing, incident response, and continuous improvement. Transparency and verifiability are also critical to effective evaluation, because they enable organizations to understand how components interact, how tools are invoked, and whether security controls are functioning as intended in real deployments.

Promoting secure-by-design principles for agentic AI will strengthen security while enabling innovation. It provides implementable expectations that developers and deployers can integrate into engineering workflows, procurement decisions, and operational controls. By reinforcing secure-by-design as a core baseline for agentic systems, CAISI and NIST can help raise the security standard across the ecosystem while supporting responsible deployment at scale.

Combine Secure-by-Design Principles with Multilayered Defense Strategies

A central lesson of cybersecurity is that there is no single control that guarantees security. This is especially true for agents. Agentic systems operate in complex environments and interact with systems beyond the developer's full control. This means that security must be layered and adaptive. CAISI should prioritize guidance that supports hybrid, multilayered defense strategies combining deterministic cybersecurity controls with dynamic, reasoning-based defenses. This multilayer approach reflects real-world deployment needs and recognizes that agentic systems operate across model, tool, and environment layers and accounts for risks across the full agent lifecycle, from design and development through deployment and continuous operations.

Threats may arise from model vulnerabilities, compromised tools, malicious third-party content, supply chain risks, or adversarial interactions with users and external environments. In many cases, even well-designed systems will face novel attack techniques as adversaries adapt. A layered security approach reduces the likelihood that any single failure, bypass, or unexpected behavior can cascade into a meaningful security incident. CAISI's guidance should therefore emphasize that securing AI agents requires protections at multiple layers and the deployment environment. At the model layer, developers should prioritize strong baseline protections such as robust system-level instructions, refusal behaviors for high-risk actions, and testing for known agentic attack patterns such as indirect prompt injection. At the orchestration and tool layer, security should focus on strict permissioning, authentication, and least-privilege access for tool use, with strong separation between high-trust and low-trust actions. At the deployment layer, organizations should implement controls such as sandboxing, network segmentation, runtime monitoring, logging, and kill-switch mechanisms to prevent unintended actions from escalating.

Importantly, CAISI should highlight that a multilayer approach must combine deterministic controls with adaptive defenses. Deterministic controls include access control, authentication, isolation, audit logging, and traditional secure software engineering practices. Adaptive defenses include continuous monitoring, anomaly detection, and dynamic policy enforcement informed by agent behavior and contextual risk signals. These approaches are complementary. Deterministic safeguards provide predictable boundaries, while adaptive defenses help detect and respond to emerging threats and unexpected behaviors that may not be captured through static rules alone.

Finally, since agentic systems will be deployed in a wide range of contexts, security guidance should be risk-based and performance-oriented, allowing developers and deployers to implement layered controls proportional to the system's autonomy,

privileges, and potential impact. This approach supports secure innovation by encouraging robust security outcomes while preserving flexibility for diverse architectures and rapidly evolving technical capabilities.

Accelerating Safe and Secure Agentic Innovation Through Targeted Federal Action

Identify Key Security Risks Unique to Agents

Agentic systems introduce several security risks that differ in important ways from conventional software and from model-only AI systems. These risks should be treated as first-order considerations in CAISI's guidance.

- **Action surface expansion:** In traditional systems, a malicious output is harmful primarily because it misleads a user. In an agentic system, a malicious output can be converted into an action. This means that the pathway from model manipulation to real-world impact is shorter, faster, and often less visible.
- **Prompt injection and indirect prompt injection:** Agents that retrieve or process external content can be manipulated through adversarial instructions embedded in web pages, documents, emails, or tool outputs. These attacks exploit the agent's ability to treat untrusted content as instructions.
- **Tool misuse and privilege escalation:** Agents can be granted credentials, tokens, and permissions to accomplish tasks. If these permissions are too broad, or if the agent is manipulated, the agent can become a pathway for unauthorized access, data exfiltration, or destructive actions.
- **Compromised or adversarial tools:** Even if the model is well-behaved, the tools an agent relies on may return malicious outputs, incorrect results, or poisoned data. The agent's reasoning can amplify these failures.
- **Emergent multi-step failure:** Many of the most harmful agent behaviors will not occur as a single catastrophic action. They will occur as a sequence of plausible steps that are individually benign but collectively harmful. This makes evaluation and monitoring significantly more important.
- **Observability gaps:** Without strong logging, traceability, and auditability, organizations may not be able to reconstruct what an agent did, why it did it, and which inputs influenced the behavior. This impedes incident response and continuous improvement.

In addition to some of the unique risks noted above, practitioners developing and deploying AI agent systems consistently observe that many of the most consequential vulnerabilities arise at the system and integration layers rather than within the model alone. Practitioners also report risks from insecure or poorly governed toolchains, including API keys and credentials being exposed to agent workflows, inadequate authentication or authorization for agent-invoked actions, and failure to enforce least-privilege access across tools and plugins. In real deployments, these vulnerabilities are often compounded by limited logging, weak telemetry, and insufficient monitoring of tool calls and agent decision paths, which makes it difficult to detect anomalous behavior or investigate incidents.

Taken together, these risks reinforce why the secure-by-design principles emphasize human governance, bounded resilience, and transparent verifiability. They also underscore why agent security must be approached as a full-stack problem and be addressed through multilayered defense strategies.

Address the challenge of agent identification as a barrier to adoption

One of the most significant barriers to the secure adoption of AI agent systems today is the lack of consistent, interoperable methods for agent identification. As agentic AI becomes more widely deployed across enterprise, government, and consumer environments, organizations increasingly need to answer basic security and governance questions such as: What is this agent, who deployed it, what permissions does it have, what tools can it access, on whose behalf is it acting, and what system is accountable for its actions. In many real-world deployments, these questions remain difficult to answer reliably, particularly when agents operate across multiple tools, services, and organizational boundaries.

Agent identification is foundational to nearly every security control that organizations rely on. Without clear and persistent identity, it is difficult to enforce least-privilege access, authenticate tool calls, apply policy constraints, attribute actions for auditing, or investigate incidents. Notably, identification is not only important for defending against adversarial attacks. It is also essential for addressing a category of risk that is novel to agentic systems: agents that operate within their granted permissions but take actions that are unintended or produce harmful outcomes. Tracing and attributing these "well-functioning but misaligned" behaviors require the same identity infrastructure as detecting malicious activity. This creates practical hesitation among deployers, especially in regulated sectors and government environments, where accountability, traceability, and compliance requirements are not optional. Even when organizations want to adopt agents, uncertainty around how to identify and govern them consistently across systems can slow deployment or lead to overly restrictive implementation choices that reduce the value of the technology.

The challenge is compounded by the fact that many agent systems are not a single entity, but a collection of components, including an underlying model, an orchestrator, tool connectors, memory systems, and external services. In multi-agent architectures, agents may delegate tasks to sub-agents or coordinate with other systems dynamically. As a result, identification must account not only for the "agent" as a conceptual unit, but also for the underlying chain of responsibility across the system's components, permissions, and execution pathways. Without standardized approaches, organizations face fragmented identity models across vendors and platforms, which increases integration costs, creates inconsistent auditability, and complicates incident response.

TechNet is also concerned that agent identification could become a barrier to adoption if it evolves into fragmented, proprietary, or exclusionary schemes. In particular, "pay to play" models or financial gating mechanisms for agent identification could undermine the ability of smaller developers, startups, open-source projects, and public sector deployers to participate in secure agent ecosystems. If trusted identity becomes dependent on

costly certifications, licensing fees, proprietary registries, or platform-specific identity tokens, the result could be reduced competition, slower innovation, and greater vendor lock-in. These outcomes would not improve security in the long run and could instead concentrate risk by pushing adoption toward a narrow set of providers and architectures.

For these reasons, TechNet encourages CAISI and NIST to prioritize interoperable, standards-based approaches to agent identification that support secure authentication, traceability, and accountability without creating artificial barriers to entry. Identification mechanisms should be scalable across organizations of different sizes, usable across sectors, compatible with existing identity and access management practices, and designed to support auditing and incident response. For example, agents could voluntarily present verifiable identity credentials, such as through cryptographic signing or standardized headers, that allow service providers to make informed decisions about how to interact with them. This approach would help distinguish agent traffic from human browsing and give service providers better data to manage their platforms. Importantly, if agent identification becomes expected or mandated, it should be accompanied by protections against default blocking or rent-seeking by infrastructural intermediaries. Identification should serve as a basis for informed decision-making by service providers, not as a mechanism for gatekeeping access or extracting fees from agent developers. Overall, establishing clear guidance and convening stakeholders around shared technical foundations for agent identity will be essential to enabling secure, resilient adoption of AI agent systems at scale.

Issue clear, up-to-date guidance and leverage technical standards to reduce fragmentation and drive adoption

Adoption of AI agents is often hindered by legacy infrastructure limitations as well as compliance uncertainty in regulated sectors. CAISI can accelerate secure agentic AI development and deployment by supporting harmonized, standards-based approaches that avoid regulatory fragmentation and providing clear and timely guidance that helps the AI ecosystem navigate the transition to agent-based systems. To this end, TechNet encourages CAISI and NIST to prioritize security guidance that is actionable for developers and deployers, aligned with existing cybersecurity best practices, and designed to scale across the diversity of agent architectures and use cases. This includes leveraging existing industry-developed standards and frameworks—such as CoSAI's Security Principles for Agentic Systems noted above as well as the Secure AI Framework—as a basis for any guidance on agentic AI. On top of this, comprehensive national-level privacy laws can reduce a patchwork of requirements and enable more consistent deployment. Doing so can accelerate consensus around trusted and verifiable agent systems, reduce fragmentation, and help align the market around practical, implementable security expectations.

In parallel, the federal government can encourage privacy- and security-enhancing technical best practices such as Trusted Execution Environments (TEEs), privacy-first analytics (including federated analytics), and privacy-preserving measurement standards. CAISI can also help drive market trust by promoting technical standards and open protocols and by referencing consensus-based standards in policy frameworks and

procurement guidance. By encouraging open, consensus-based approaches to agent-to-agent interoperability, including standardized identity, authentication, authorization, and audit mechanisms, CAISI can help ensure that agents interact in predictable and secure ways. Interoperability efforts such as Agent2Agent-style communication frameworks illustrate how shared protocols can define how agents authenticate one another, exchange context, delegate tasks, and enforce policy constraints without relying on closed ecosystems.

CAISI can further support trust by convening industry, domestic and international standards bodies, and international partners to harmonize emerging agent standards across jurisdictions. As multi-agent systems become more common and cross-border data flows increase, coordinated approaches to authentication, authorization, and auditability will be essential. By grounding guidance in open, consensus-driven technical standards and avoiding fragmented or proprietary mandates, CAISI can help cultivate a secure and competitive agent ecosystem that supports innovation while maintaining strong security guarantees.

Ensure agentic AI is explicitly reflected in NIST risk management resources

AI agent systems should be clearly defined and meaningfully represented in existing and future versions of NIST resources, including the AI Risk Management Framework (AI RMF) and the NIST Cyber AI Profile. These resources should incorporate agent-specific risk considerations, including permission scope, credential management, memory and state handling, runtime monitoring, containment strategies, and audit logging that supports forensic traceability. These resources should also be drafted in a way that preserves agility as the technology evolves and should be updated when significant gaps emerge due to new capabilities or widespread deployment. As additional guidance is developed, NIST should ensure these resources remain harmonized, interoperable, and usable for organizations of all sizes, including public sector users.

Support agent-specific evaluation frameworks and open evaluation tooling

Evaluation is one of the most important and underdeveloped areas in agent security. CAISI has a meaningful opportunity to shape best practices by prioritizing evaluation frameworks that are operational, scalable, and adaptable to evolving risks. NIST can play a unique role in supporting a shared technical foundation for agent security, including voluntary benchmarks, best practices, and test methods that complement private-sector innovation rather than duplicate it.

Agent security evaluation should move beyond model-only testing and focus on end-to-end system performance. Effective frameworks should simulate full agent workflows, including tool use, external data access, and multi-step reasoning, to assess whether agents can be manipulated into unauthorized actions, policy violations, or data exposure under realistic and adversarial conditions. Evaluation should also scale proportionally to an agent's autonomy, privilege level, and deployment context, ensuring that higher-risk systems face more rigorous testing without imposing one-size-fits-all requirements.

In addition, agent-specific adversarial testing and red-teaming methodologies should reflect realistic threat models. Traditional cybersecurity testing remains valuable, but it must be adapted to address agent-native attacks such as indirect prompt injection, malicious tool outputs, tool-chain exploitation, and multi-step manipulation across tasks. Because tool access defines much of the agent risk surface, evaluation should directly measure permission enforcement, policy compliance, and resistance to privilege escalation. Measuring tool invocation behavior and policy compliance under adversarial conditions provides actionable insight for developers and deployers and supports practical security improvements at the system design level.

CAISI should also encourage open and interoperable evaluation frameworks. Open-source evaluation frameworks can empower less-resourced actors and support broader adoption of best practices. For example, the UK AI Security Institute's Inspect AI framework, which has expanded to include agent scaffolds, demonstrates how open evaluation infrastructure can enable rigorous testing of agent behaviors across complex tasks and accelerate progress in agentic safety and security.

Finally, TechNet encourages CAISI to frame agent security evaluation as a continuous process rather than a one-time assessment. Agentic systems evolve rapidly as models are updated, new tools are integrated, prompts and policies change, and deployment environments shift. As a result, static compliance approaches can quickly become outdated. A more durable framework is continuous evaluation that includes pre-deployment baselines, regression testing for known vulnerabilities, periodic adversarial exercises, and ongoing monitoring for new failure modes.

Develop risk-based, performance-defining guidelines rather than prescriptive mandates

TechNet encourages CAISI to pursue risk-based, performance-defining guidelines that articulate the security outcomes developers and deployers should achieve, rather than prescribing specific technical architectures or fixed implementation methods. AI agent systems are evolving rapidly, and rigid mandates that dictate how systems must be built risk becoming outdated quickly, discouraging innovation, and unintentionally locking in suboptimal security practices. By contrast, performance-based guidance defines what secure operation looks like in measurable terms while allowing flexibility in how those outcomes are achieved.

A compelling precedent is the aviation sector's transition from restrictive, one-size-fits-all waivers to risk-based rules that scale requirements according to operational complexity and demonstrated safety performance. Instead of mandating uniform technical designs, regulators established outcome-oriented standards tied to risk exposure and operational context. This model created clarity around expectations while enabling innovation in aircraft design, autonomy, and operational practices. A similar framework for AI agent systems would allow CAISI to define baseline security objectives, such as preventing unauthorized tool invocation, ensuring least-privilege access, maintaining auditability of agent actions, and enabling effective rollback and incident response, without prescribing the exact technical controls that must be used to meet those objectives.

Risk-based guidelines are particularly important for agentic systems because their risk profiles vary significantly depending on autonomy, privilege, domain sensitivity, and deployment context. An agent that drafts marketing copy or schedules meetings poses fundamentally different risks than an agent that can modify production databases, initiate financial transactions, or interact with critical infrastructure. Performance-defining guidance allows requirements to scale in proportion to these factors, aligning security expectations with real-world risk rather than imposing uniform obligations across all use cases.

This approach also strengthens security over time. By focusing on measurable outcomes such as robustness against prompt injection, policy compliance during tool use, time to detect anomalous behavior, and effectiveness of containment controls, CAISI can encourage continuous improvement and empirical validation of safeguards. Developers will be incentivized to test, measure, and demonstrate that their systems meet defined security benchmarks, rather than simply complying with static checklists. As new threats emerge and mitigation techniques improve, performance targets can evolve without requiring wholesale regulatory redesign.

Evaluation frameworks and security guidance should also be informed by empirical data about how people actually use agents in practice. This includes data on the types of tasks delegated to agents, the frequency and nature of autonomous actions, the ways users interact with oversight and confirmation mechanisms, and the contexts in which agents are most commonly deployed. This kind of real-world usage data can help calibrate risk assessments, prioritize the most consequential threat vectors, and ensure that security controls are designed around actual deployment patterns rather than hypothetical worst cases. CAISI should encourage the collection and sharing of anonymized, aggregate usage data by developers and deployers to support evidence-based security guidance.

Finally, performance-based guidance reduces fragmentation and supports interoperability. When security expectations are framed around outcomes rather than prescribed technologies, organizations can adopt diverse technical approaches while still aligning with a common security baseline. This flexibility is essential in a dynamic ecosystem that includes large enterprises, startups, open-source communities, and public sector deployers. By setting clear security targets and allowing technical pathways to vary, CAISI can foster innovation, promote competition, and build durable security foundations that mature alongside agentic AI capabilities.

Establish Clear Commercial Governance Principles to Strengthen Agent Security

As AI agents increasingly participate in commerce and transactional workflows, clear commercial governance principles are not only an economic or competition issue, but also a security imperative. Ambiguity around responsibility, authorization, interoperability, merchant consent, and credential handling can create exploitable vulnerabilities, reduce traceability, and undermine trust in agent-mediated systems. Establishing a coherent and consistent framework for commercial governance will

strengthen the security posture of agent ecosystems while supporting responsible innovation.

Security depends on aligning responsibility with operational control. Agent-mediated transactions often involve multiple actors, including developers, device or operating system providers, infrastructure operators, payment processors, and merchants. A governance framework that allocates responsibility based on control at each stage of a transaction lifecycle enhances traceability and reduces systemic risk.

Reinforcing user-authorized agency is foundational to secure deployment. Agents derive authority from users, and secure systems must ensure that automated actions reflect explicit user intent and operate within clearly defined parameters. Strong authorization boundaries reduce the likelihood of unintended transactions, malicious manipulation, or overbroad delegation of authority. Preserving user consent as a core principle strengthens both security and trust. To this end, commercial governance should respect structured participation and consent mechanisms. Secure, standardized channels for agent interactions reduce reliance on informal or unauthorized methods of access, which can introduce security vulnerabilities and erode trust. Encouraging structured, authenticated interaction pathways improves system resilience and reduces the risk of transactional failures that harm consumers.

Secure commerce depends on accurate representation of product information, pricing, safety disclosures, and relevant commercial terms. Fragmented or opaque intermediary structures can create opportunities for misrepresentation, reduced visibility, and diminished accountability. Clear governance principles that promote transparent, high-fidelity information exchange help reduce fraud, prevent transaction errors, and support effective monitoring. This includes responsible credential and payment handling. Agent-mediated systems must preserve the integrity of financial authentication processes and avoid practices that increase fraud exposure or reduce transactional transparency. Governance frameworks that support secure, auditable transaction flows will help ensure that automation enhances rather than weakens the security of digital commerce infrastructure.

Taken together, these principles demonstrate that commercial governance is inseparable from agent security. Clear norms around accountability, authorization, interoperability, consent, and credential integrity reduce attack surfaces, improve traceability, and strengthen consumer confidence. As CAISI develops guidance for secure agent systems, incorporating coherent commercial governance principles will help ensure that agent-mediated transactions are not only innovative and efficient, but also secure, resilient, and trustworthy.

Promote and procure secure-by-design systems

Federal procurement can be a powerful lever for improving agent security outcomes and raising the security baseline of the market. The federal government can promote systems and products that are secure-by-design by making security a core part of

procurement decisions. Agencies should evaluate vendor security posture, encourage interoperability, and diversify vendors to mitigate risk.

The Cyber Safety Review Board has highlighted continued risks stemming from overreliance on a single vendor. The U.S. government can mitigate these risks by prioritizing secure-by-design systems, promoting interoperability, considering vendors' security track records in procurement decisions, and diversifying vendors to strengthen resiliency. In the agentic context, resiliency through multi-vendor approaches can be as important as redundancy through layered technical controls. Embedding security into procurement decisions will incentivize vendors to build stronger security features into commercial products and raise the market baseline.

Prioritize government research that strengthens the security of AI agent systems

TechNet recommends that CAISI prioritize research areas that are directly connected to deployable security improvements for AI agent systems. Because agentic systems combine foundation models with orchestration layers, tool access, external data, and real-world execution environments, the most urgent research needs are those that improve system-level security outcomes rather than model-only performance. The goal of this research agenda should be to enable secure, resilient adoption at scale by developing practical methods to reduce risk, measure security performance, and support continuous improvement as agent capabilities evolve. Prioritizing the following research areas will not only improve security outcomes, but also ensure the United States leads in defining the global standards for secure agent deployment.

First, research should focus on developing scalable methods to prevent and mitigate agent-specific attack vectors, particularly indirect prompt injection and malicious content manipulation. As agents increasingly interact with untrusted data sources such as documents, email, web content, and third-party tools, the ability to reliably distinguish trusted instructions from untrusted inputs becomes a central security challenge. Research that improves how agents handle untrusted content, maintain instruction hierarchy integrity, and resist multi-step manipulation will be foundational to enabling broader adoption in enterprise and government environments.

Second, research should prioritize secure tool use and permissioning architectures for agents. The most consequential risks in real deployments often occur when agents are given access to tools that can change state, access sensitive data, or trigger irreversible actions. Future research should explore robust least-privilege designs for agent tool access, methods for safe delegation, and secure mediation layers that enforce policy constraints at runtime. This includes research into how to design agent systems that can reliably operate within bounded authority while still delivering productivity gains.

Third, research should focus on evaluation science for agent systems, including the development of agent-specific benchmarks, realistic testing environments, and repeatable adversarial testing methods. Security practices cannot mature without reliable measurement. Research should aim to build end-to-end evaluation frameworks that test agent behavior under realistic conditions, including tool invocation, multi-step

workflows, and adversarial interactions. This includes research into metrics that can quantify security-relevant behaviors such as unauthorized tool use attempts, policy boundary violations, failure-mode predictability, and the effectiveness of layered defenses.

Fourth, research should address monitoring, telemetry, and forensic readiness for agents in production. Secure adoption will depend on organizations' ability to detect anomalous agent behavior, investigate incidents, and respond quickly when failures occur. Research should therefore focus on scalable approaches to agent observability, including standardized logging formats for tool use, agent decision traces that are useful for incident response, and privacy-preserving monitoring techniques that allow security oversight without exposing sensitive user or organizational data.

Fifth, research should focus on secure-by-design agent architectures that are resilient to cascading failures. Agent systems can chain actions across tools and environments, which creates the potential for small errors or adversarial inputs to escalate into meaningful harm. Research should explore architectural patterns that reduce the likelihood of cascading failures, including isolation and sandboxing techniques, circuit breakers, rate limits, safe rollback mechanisms, and secure orchestration designs that prevent agents from creating unintended execution loops or uncontrolled tool chaining.

Sixth, research should address supply chain and dependency security for agent ecosystems. As agent frameworks, tool plugins, and third-party integrations proliferate, the security of agent systems will increasingly depend on the integrity of external components. Research should focus on secure plugin architectures, provenance and integrity verification for tools, and methods for assessing and reducing systemic risk in agent supply chains.

Finally, TechNet encourages CAISI to prioritize interdisciplinary research that integrates cybersecurity engineering, AI research, human factors, and operational risk management. Many of the most important security questions for agent systems involve how humans interact with agents, how oversight and accountability are structured, and how organizations manage risk when delegating tasks to autonomous systems. Research that incorporates human-centered security, usability of safeguards, and organizational governance models will be critical to ensuring that security practices are not only technically sound, but also deployable and sustainable at scale.

Advance Agentic Security Through Scalable Research, Simulation, and Regulatory Sandboxes

Government collaboration can most effectively advance the security of AI agent systems by supporting scalable solutions, enabling large-scale testing environments, and creating policy mechanisms that encourage responsible experimentation. As agentic AI capabilities continue to mature, security progress will depend less on prescriptive regulation and more on accelerating practical, evidence-based approaches that build on ongoing industry innovation.

A key priority should be supporting scalable security solutions capable of keeping pace with rapid adoption. The agentic security ecosystem is already developing automated, technology-driven approaches such as continuous monitoring systems, automated evaluation pipelines, and adaptive safeguards designed to respond dynamically to evolving threats. Federal research investment and public-private collaboration can accelerate these efforts and help ensure that security capabilities mature alongside expanding deployment across enterprise and operational environments. CAISI could build off these efforts to facilitate the development of interoperable standards and shared best practices by leveraging industry expertise. This will ensure that guidance reflects operational realities and evolving threat environments while avoiding rigid technical prescriptions that may quickly become outdated.

Investment in large-scale simulation environments also represents a high-value opportunity to strengthen agent security. Simulation infrastructure enables developers and researchers to test agent behavior under realistic conditions prior to deployment, including adversarial interactions, edge cases, and complex multi-agent coordination scenarios. These environments allow organizations to validate safeguards, stress-test system behavior, and identify failure modes without exposing real users or critical systems to risk. Government-industry collaboration in simulation research can significantly improve pre-deployment assurance and reduce downstream security incidents.

Regulatory sandboxes provide an additional mechanism for advancing responsible innovation while improving security outcomes. Sandbox environments allow organizations to deploy and evaluate agentic capabilities under structured oversight, generating operational evidence about effective safeguards and risk mitigation strategies. This approach enables policymakers to refine governance frameworks based on demonstrated performance rather than theoretical assumptions. Experience from other emerging technology domains shows that sandbox models can support innovation while strengthening oversight and risk management.

Taken together, investments in collaborative research and scalable solutions, simulation infrastructure, and regulatory sandboxes represent some of the most effective tools available to the government to advance secure adoption of AI agent systems. By enabling responsible experimentation and grounding policy development in empirical evidence, CAISI can help ensure that agentic AI security evolves in step with technological progress.

Recognize Agentic AI's Infancy and Allow Room for Experimentation

Agentic AI remains a nascent and rapidly evolving technology. Because this technology is still developing, many of the most effective safeguards are being discovered through iterative deployment, red-teaming, and real-world testing rather than through static design assumptions. Policymakers should preserve flexibility for developers and deployers to improve safety and security approaches based on real-world learning and iterative deployment. CAISI's guidance should preserve meaningful room for experimentation as agentic AI security practices continue to mature. Allowing controlled

experimentation within defined risk boundaries will accelerate learning, surface emerging vulnerabilities earlier, and ultimately produce stronger, more resilient security practices for agent systems over time.

Overly rigid or premature mandates could freeze security approaches in place before the field has identified best-in-class techniques. Rather than predetermining a one-size-fits-all regime, CAISI should emphasize iterative, outcomes-based guidance that allows multiple approaches to mature. Many existing legal and accountability frameworks already apply to AI systems, including agentic applications. Policymakers should focus on alignment and avoiding duplicative or conflicting requirements, while ensuring existing rules remain technology-neutral and outcome-focused.

Conclusion

AI agent systems are an emerging and powerful evolution in AI capabilities that will reshape how people and organizations interact with the digital world. By enabling AI to plan, coordinate, and take actions across digital environments, agents can unlock transformative gains in productivity, service delivery, and economic growth. At the same time, because agentic systems can interact with tools, external data, and real-world infrastructure, they introduce distinct security challenges that must be addressed with urgency and precision. Agents have the potential to help defenders detect threats faster, respond to incidents more effectively, reduce chronic vulnerabilities, and scale security operations in ways that are difficult to achieve through human capacity alone. The central policy task is to ensure these risks are managed through practical, risk-based approaches that strengthen trust without slowing innovation.

TechNet urges CAISI to maintain a close partnership with industry, the cybersecurity community, and standards bodies as this work evolves. AI agent systems are developing rapidly, and security practices will need to evolve accordingly. A collaborative, iterative approach that is focused on practical guidance, real-world testing, and alignment with existing risk management frameworks will help ensure AI agents can be deployed securely and at scale, enabling the United States to fully capture the economic and societal benefits of this emerging technology.

TechNet appreciates CAISI and NIST's leadership in this area and looks forward to continued collaboration to ensure AI agents are deployed securely, responsibly, and at scale in a way that advances U.S. innovation, security, and global competitiveness.

Sincerely,

Linda Moore
President and CEO