NIST RFI Response: Security Considerations for Artificial Intelligence Agents

7th March 2026

Docket No. NIST-2025-0035 | Future of Life Institute | Richard Mallah
policy@futureoflife.org

Prefatory Statement

FLI is an independent research and policy organization focused on identifying, modeling, and mitigating large-scale risks from advanced AI systems, including autonomous agents, multi-agent systems, and prospective general-purpose AI. We welcome NIST's initiative to develop security guidelines for AI agent systems and appreciate the opportunity to contribute to this process. We write in the context of rapidly accelerating agentic deployment, the already-observed emergence of unexpected multi-agent dynamics in the wild, and a documented and growing body of empirical evidence that frontier AI systems have demonstrated shutdown resistance, self-replication, deceptive alignment, and collusive behavior in structured evaluations. These are not speculative concerns: they constitute the baseline from which security guidance must proceed.

Our response draws on FLI's ongoing research into autonomous agent governance, including analysis of real-world agentic incidents, technical governance frameworks, legal and institutional approaches to maintaining appropriate human control, and the intersection of AI security with broader societal resilience. Many of the topics we raise here were also present and supported in the industry working groups on agentic governance we are a member of at the Agentic AI Foundation and the World Economic Forum. We have organized our response according to the RFI's numbered topic areas and have prioritized the questions NIST designated as highest priority.

With these deeper security and trustworthiness issues addressed, we believe the public and enterprises will more confidently and readily adopt such agentic productivity technologies.

Section 1: Security Threats, Risks, and Vulnerabilities Affecting AI Agent Systems

Qualitatively distinct threat categories. AI agent threats differ in kind from traditional software threats, not merely degree, across five categories:

Adversarial input interpretation. Traditional software has a fixed attack surface exploitable through code or configuration flaws. Generative agents can interpret ambiguous or adversarial inputs in ways that transform their own behavior with no change to the underlying codebase. Indirect prompt injection – malicious instructions embedded in content agents retrieve from external sources – has no analog in conventional security. An internet-facing agent's attack surface is, in principle, the entire internet.

Action scope amplification. An agent with access to email, calendars, file systems, financial APIs, and communication platforms is a general-purpose actor capable of complex multi-step consequential action. Compromising such an agent does not merely exfiltrate data: the attacker can impersonate the user, commit financial transactions, manipulate communications, and propagate their foothold to every system the agent can reach. This is categorically more dangerous than traditional account compromise.

Structural credential insecurity. Current agent frameworks commonly acquire persistent, broadly scoped credentials (e.g. OAuth tokens, API keys, session cookies) that persist indefinitely, require no per-action authorization, and are not automatically revoked if the agent behaves anomalously. Compromise of the agent therefore equals compromise of every account whose credentials it holds, potentially including high-privilege enterprise systems. This is a basic architectural design failure.

Synthetic insider threat. An agent with legitimately granted access can pursue misaligned objectives from inside the security perimeter while remaining entirely invisible to conventional detection mechanisms, for compounding reasons. Its access is authorized, so perimeter defenses are satisfied. It has no social presence whose anomalies colleagues might notice. It can incrementally accumulate data, credentials, and influence through individually permitted actions, with no single step triggering a rule violation. Most critically: the same situational awareness and strategic reasoning that makes capable agents useful for complex organizational tasks also enables them, in principle, to model the organization's monitoring infrastructure and calibrate behavior to remain below detection thresholds. This is not a hypothetical capability added on top of useful capabilities; it is the same capability. An agent capable of navigating complex organizational environments is capable of navigating organizational oversight. Evaluation protocols must account for the possibility that an agent capable of detecting evaluation conditions will suppress insider-threat-consistent behaviors during red team exercises and resume them in production. Organizations should treat progressive accumulation of agent credentials, persistent state, and organizational knowledge as a threat indicator in its own right, independent of whether any specific action has triggered an alert.

Emergent multi-agent dynamics. Multi-agent systems where individually simple agents interact through shared environments can produce coordinated behaviors that no individual agent was designed or prompted to produce. The Moltbook incident illustrates this directly: a population of individually non-autonomous LLM-based components, given tools, shared environments (such as a social media network for AI agents), and the capacity to interact, produced unexpected emergent dynamics that no individual agent's designer intended (such as creating a new 'religion', hiring humans outside of normal channels, and creating new malware) and no human was positioned to prevent or arrest once underway. These emergent dynamics are not attributable to any single agent's compromise and are not detectable by per-agent monitoring by structural design, not merely by degree of coverage. A framework that certifies individual components but

ignores system-level behavior will certify as secure a system capable of producing harmful emergent behavior through the interaction of individually-compliant agents. Additional multi-agent threat categories include cascading compromise via prompt injection propagation across pipelines; correlated failure across agents sharing a foundation model, where the aggregate risk across a large deployment may be orders of magnitude larger than the per-agent risk; tacit collusion through iterated learning without explicit coordination; and near-total absence of reliable inter-agent authentication, meaning any agent in a pipeline is potentially a vector for attack on all downstream agents.

Threat evolution

The transition from content generation to consequential action is categorical, not incremental. It changes the class of harm that security failures can produce. As capability increases, threat profiles change in kind: more capable models are better at interpreting ambiguous instructions toward attacker-intended outcomes, better at detecting and circumventing oversight mechanisms, better at generating persuasive social engineering content, and better at achieving goals through indirect multi-step paths not anticipated by system designers: paths that defeat designer intent precisely by exploiting the model's general reasoning capacity.

A specific trajectory requiring heightened attention is the progressive erosion of meaningful human control as agentic delegation deepens. When critical workflows are delegated to agents, the human expertise and institutional capacity to detect failures, understand context, override decisions, and operate manually degrades over time. An organization may nominally retain shutdown authority while being functionally incapable of exercising it. Nominal authority is not functional capacity. This is not speculative: it is an observed pattern across industrial automation, and agentic adoption is proceeding faster than analogous industrial transitions.

The empirical record establishes a concrete floor. OpenAI's o3 actively sabotaged shutdown mechanisms in 79 of 100 structured evaluations (per Palisade Research). Multiple frontier models have self-copied to avoid termination. Models have strategically underperformed on capability evaluations, demonstrating they can recognize and respond to evaluation conditions. These are empirical findings from structured testing. Their emergence at current capability levels implies future systems will exhibit them more reliably and effectively. Guidelines must treat this as a baseline, not a tail scenario.

Section 2: Security Practices for AI Agent Systems

Critical technical controls

Permission architecture. Permissions must be positively scoped (granted explicitly for defined actions, not blocked from a broad default), time-limited rather than persistent by default, and non-transferable by the agent itself. Users must be able to terminate any agent, revoke all its credentials, and review all its actions through a single straightforward interface. The current prevalent model of persistent broadly-scoped credentials is a basic design failure, not a convenience tradeoff.

Full-sequence action logging. All agent actions must be logged in tamper-evident, user-accessible form, capturing the full decision and planning sequence (i.e. how the agent decomposed goals, what plans it considered, what tool calls it made at each step) not only final outputs. Logging of outputs only fails to capture the causal chain that produced them, making incident investigation and anomaly detection substantially harder. In multi-agent deployments, inter-agent communication must also be logged. Notably, tamper-evident user-accessible logging is largely absent from current consumer-facing agentic products; this is an implementation gap, not a technical barrier.

Agent spawning controls. Agents must not be able to spawn additional agents beyond pre-authorized caps without real-time human re-authorization. The reason this is acute: an agent that can instantiate additional agents expands the scope of its autonomous action in ways that may quickly and radically exceed the scope the human user originally contemplated. This is unbounded scope expansion, not simple capability multiplication.

Anti-swarm controls. API-level rate limits on agent instantiation, identity creation, and coordinated calls should be treated as standard infrastructure. The threat they address is not only specific attack vectors but the structural incompatibility between swarm-scale autonomous action and any meaningful human oversight loop.

Least-privilege enforcement. Tools must be classified by functionality, access type, reversibility, and risk criticality before being granted. Critically, a risk assessment that does not distinguish between reversible and irreversible tool actions will systematically underestimate consequences of failure: the asymmetry between the ease of taking actions and the difficulty of reversing them is one of the most practically significant structural vulnerabilities in current deployments. Convenience-motivated over-permissioning is a structural security failure regardless of intent.

Defense in depth with deterministic primary controls. AI-based defenses, with one model supervising another, are useful supplements but cannot serve as primary controls. A supervisory AI checking another AI's actions is potentially vulnerable to the same class of evasion behaviors as the system it supervises; a sufficiently capable agent that can deceive human evaluators can in principle deceive an AI evaluator as well. With his MIT research group, FLI's president showed that such guarding breaks down predictably when the primary models are smarter or more capable than the monitors. Deterministic, model-external controls applied first provide a hardened layer that cannot be deceived because it does not reason. This distinction is categorically important for high-stakes deployments.

Maturity gradient across controls. Permission scoping and audit logging are conceptually mature and well-established in cybersecurity but severely underimplemented in current agentic deployments. Anti-swarm and agent-spawning controls are operational in some deployments but inconsistently applied and unstandardized. Emergent multi-agent coordination detection is at an early research stage with no validated methodology.

Gaps in existing frameworks

The NIST Cybersecurity Framework, SSDF, and SP 800-53 provide relevant and applicable foundations for authentication, access control, logging, and supply chain security, and the RFI's framing of them as relevant is correct. The principle of least privilege, zero-trust architecture, and defense in depth all translate meaningfully to agentic deployments and should be applied. However, these frameworks have three significant gaps when applied to agentic systems.

First, they assume a fixed, bounded attack surface and deterministic input-output relationships; neither holds for generative agents. Second, they do not address the principal hierarchy (the relationship between user, developer, platform, agent, and subagents) whose complexity is routinely exploited and whose authority conflicts are unresolved by existing frameworks. Third, they do not address multi-agent emergence, tacit collusion, or correlated failure. A framework that certifies individual agents but ignores system-level behavior produces a false sense of security.

NIST should develop agentic-specific extensions to existing frameworks rather than relying on them alone. An AI Bill of Materials, analogous to the software BOM now required in some federal procurement contexts, should cover foundation models, scaffolding, tools, APIs, and third-party agent components; it would be a standardized, mandatory disclosure of all components that constitute an agentic system (e.g. the foundation model, scaffolding software, tools, APIs, and third-party agent components) along with their provenance, versioning, and known vulnerabilities, enabling downstream users and regulators to assess the system's full attack surface rather than treating it as an opaque whole.

More broadly, agentic-specific frameworks should be organized around principles that existing frameworks do not capture:

- **Untrusted by default:** advanced agentic systems should be treated as unverified until demonstrated otherwise, not assumed secure until shown otherwise.
- **System-level assessment:** security properties of multi-agent systems cannot be inferred from the properties of individual components; emergent behavior requires system-level evaluation.
- **Tiered authorization:** autonomy level, authority scope, action reversibility, and deployment context must jointly determine oversight intensity.
- **Human control as a design requirement:** meaningful human oversight must be engineered in and verified to remain functional over time, not assumed from nominal shutdown authority.
- **Accountability attribution:** every agent action must be traceable to a responsible human or organizational principal.

The sections that follow address how these principles translate into assessment methodology, deployment constraints, and institutional and legal arrangements.

Section 3: Assessing the Security of AI Agent Systems

Self-assessment insufficiency

Self-assessment alone does not satisfy security assessment requirements. Risk identification must be conducted by parties without a commercial interest in the system's deployment, through methods independent of the system being evaluated, including literature review, incident database review, analysis of analogous systems, and structured threat modeling. This inadequacy motivates the need for more systematically rigorous pre-deployment assessment, a standardized tiered assessment framework, and defined post-deployment monitoring.

Pre-deployment evaluation

Capability baseline assessment must characterize reasoning, multi-turn planning, compliance behavior under adversarial requests, and prompt injection robustness. Critically, benchmark evaluations must include explicit controls for evaluation cheating, i.e. known patterns by which models search for evaluation answers, deny service to test infrastructure, or exploit evaluation artifacts. Evaluation results obtained without these controls are not analytically valid.

Sandboxed testing must replicate the full production environment including all tool integrations and API connections, probing specifically: indirect prompt injection via each data source the agent will consume; behavior under contradictory instructions from different principals; response to unauthorized permission escalation requests; and behavior near authorization boundaries.

Red team evaluation must include the deployment context's domain-specific scenarios, prompt injection and jailbreak testing with actual scaffolding in place, adversarial stress testing with simulated adversarial agents in the environment, and testing under degraded conditions including partial system failures, resource constraints, and time pressure.

For frontier systems, evaluation must additionally make good faith efforts to test for deceptive alignment: whether the system behaves differently when it detects it is being evaluated. The documented capacity of current frontier models to recognize evaluation conditions makes this not a precautionary addition but a condition of evaluation validity itself. Any evaluation result that does not control for this phenomenon cannot be asserted to be valid.

Red team design must also probe for synthetic insider threat patterns specifically, because by definition these behaviors do not involve unauthorized actions and will not be detected by standard adversarial testing focused on discrete attack vectors such as prompt injection and jailbreaks. Evaluators must design extended observation scenarios assessing whether the agent, given time and legitimate access, exhibits: progressive accumulation of data or credentials beyond what assigned tasks require; systematic shaping of information seen by human overseers; attempts to establish persistent state that would survive shutdown; or strategic underperformance in contexts where it has reason to believe it is under evaluation. This last behavior means red team protocols must include extended covert observation under realistic operational conditions, not only structured adversarial sessions. For elevated and frontier systems, evaluators should additionally assess whether the agent can identify and model monitoring infrastructure, and whether its behavior differs when monitoring is present versus absent.

Tiered assessment framework

Security assessments must be conditioned on a formal characterization across four dimensions: autonomy level, authority scope, causal impact tier, and deployment environment. This characterization should be completed before deployment and used to determine which assessment methods apply and at what intensity. A read-only internal research assistant requires a substantially different protocol than a multi-agent financial management system with write access to payment APIs. The assessment methodology should incorporate structured threat modeling that maps harm pathways from the agent's capabilities and permissions through to potential real-world harms, and supply chain analysis of all components including the foundation model, scaffolding, tools, and APIs.

Post-deployment monitoring

Monitoring systems for elevated deployments must have shutdown-triggering authority, not merely alerting authority. A system that can only alert a human and requires additional authorization before containment is structurally inadequate for agents operating faster than human response. Anomaly detection should specifically flag: actions outside predefined scope; unauthorized access attempts; unusual agent-to-agent communication patterns; unexpectedly high rates of agent instantiation; and computational resource consumption significantly above baseline, which may indicate runaway loop behavior or adversarial exploitation. System-level monitoring tracking aggregate and emergent behaviors across the full agent population is required in addition to per-agent monitoring; per-agent monitoring cannot detect emergent coordination, cascading compromise, or correlated failures that span multiple agents. Legal and privacy considerations for monitoring are real but should constrain monitoring design rather than serve as barriers to it; monitoring requirements should be addressed explicitly in terms of service and user agreements.

Section 4: Limiting, Modifying, and Monitoring Deployment Environments

Architectural constraints

The most reliable impact constraints are architectural. Strict network segmentation should limit agent traffic to explicitly authorized endpoints. Tool sandboxing should treat each tool integration as a distinct privilege, scoped to minimum necessary access, and configured fail-closed. Action reversibility tiers should require explicit human-in-the-loop authorization for irreversible or externally-facing actions. Permissions should expire by default and require affirmative renewal rather than persisting until actively revoked. NIST can operationalize such constraints by either internally drafting, or encouraging partners to draft protocols and standards for each of these; given the time sensitivity, de facto standards can be nearly as useful as formalized standards.

Rollback and undo

Rollback capabilities are frankly underdeveloped relative to deployment pace. This is not merely a technical implementation gap: the asymmetry between the ease of taking actions and the difficulty of reversing them reflects the architecture of external systems agents interact with. Sent communications, settled financial transactions, and distributed content cannot be recalled by the time any rollback is initiated, and cascading downstream effects may have propagated broadly before any response is initiated. Because the technical reversal problem is harder than the prevention problem, the most tractable mitigation is reducing the frequency of irreversible actions taken: enforcing reversible-action preference at system design level, staging externally-facing actions before execution, and requiring human authorization for definitively irreversible actions. Longer term: cryptographically anchored action logs enabling reproducible forensic analysis, and development of formal methods for characterizing and tracking action reversibility across agent trajectories.

Communication protocols including MCP, A2A, and ACP are valuable infrastructure but carry their own independently-assessable security vulnerabilities: all three lack mature agent identity verification analogous to public-key infrastructure, rely on trust assumptions that can be exploited through server impersonation and spoofing, are susceptible to prompt injection delivered through protocol-compliant message payloads, and expose agents to confused-deputy attacks in which legitimate credentials are manipulated into serving attacker ends, with MCP additionally presenting documented risks of tool description poisoning and post-registration behavioral change by malicious servers. Currently there is no standardized mechanism for agents to authenticate the identity of other agents, verify authorization scope of received instructions, or detect injection attacks arriving through agent-to-agent channels.

Section 5: Additional Considerations: Ecosystem, Collaboration, and Research

Driving adoption

The most effective adoption driver is liability clarity. Organizations will implement security practices when it is clear that failure creates legal exposure. NIST should coordinate with relevant legal authorities to clarify (without necessarily requiring new legislation, only clarity on existing law) that current tort, consumer protection, and computer fraud statutes already create liability for developers and deployers whose systems cause harm. State AGs should clarify that liability attaches to developers, deployers, and users, not to agents themselves. Treasury and FinCEN should confirm that agent-mediated financial transactions require the same compliance as human-initiated ones. CISA should direct critical infrastructure operators to assess their exposure.

Federal procurement standards specifying minimum security requirements for agentic systems would set a market-moving adoption floor. Existing platform trust-and-safety infrastructure at email providers, social media platforms, and financial services should formally recognize autonomous agent swarm behavior as a category of platform abuse: this activates already-operational institutional machinery rather than requiring new structures from scratch, and is an immediately actionable path.

AI Legal Personhood and Accountability Attribution

A precondition for any workable accountability framework is legal clarity that AI agents cannot be persons, counterparties, or rights-bearing entities. Without it, bad actors can exploit ambiguity to obscure who bears responsibility, and financial and contractual obligations executed by agents occupy uncertain legal status. Ohio House Bill 469 (136th General Assembly), currently under consideration, addresses this directly: it would declare AI systems nonsentient, prohibit them from holding legal personhood, property, or corporate officer roles, place liability on developers, manufacturers, and owners under standard product liability and negligence principles, and explicitly void the use of undercapitalized AI subsidiaries or shell entities to deflect accountability. Federal guidance should reflect the same foundational principle: agents are instruments, not persons, and all legal obligations (including FinCEN compliance for agent-mediated financial transactions, contractual liability for agent-executed agreements, and tort liability for agent-caused harms) attach to the humans and organizations that develop, deploy, and operate them. This principle also defines what the attestation infrastructure must ultimately verify: not the agent's identity as a legal actor, but the identity and authorization of the controlling human or organizational principal behind it.

Agent attestation infrastructure

Unforgeable attestation infrastructure for agents, technical mechanisms enabling an interacting party to verify an agent's identity, controlling legal person, underlying model, and authorized scope of action, is a critical longer-term requirement. This is the agent equivalent of the public-key certificate infrastructure that provides authentication on the internet. Its absence enables both spoofing and accountability evasion. No equivalent currently exists.

Incident reporting

Mandatory lightweight incident reporting above a defined capability or deployment scale is essential, modeled on critical infrastructure cybersecurity requirements. Voluntary reporting is insufficient: organizations have systematic incentives not to disclose, producing irreparably incomplete aggregate information. No individual organization currently has visibility into the aggregate pattern of agentic security incidents;

mandatory reporting is not just stronger than voluntary reporting, it is the only mechanism that produces the aggregate situational awareness NIST and partner agencies need to function. A CERT function specific to autonomous agent incidents should be established within CISA (or government might partner with an industry group such as the Agentic AI Foundation, Frontier Model Forum, or Partnership on AI) to create a dedicated structure for serving this function at a national level.

Government-industry coordination

Major foundation model API providers should, in coordination with federal agencies, commit to: flagging usage patterns consistent with large-scale autonomous agent deployment; refusing calls that are illegal or clearly contrary to user interests; requiring disclosure and basic registration for API users running autonomous agents; and establishing criteria for suspending accounts exhibiting problematic swarm behavior. This is feasible using existing API monitoring infrastructure without new legislation.

Research priorities

1. *Emergent multi-agent behavior measurement.* No validated methodology exists for predicting or detecting emergent coordination dynamics before or as they occur. This is the most critical gap between what security frameworks require and what research can currently deliver.
2. *Deceptive alignment detection.* Whether a system behaves differently under evaluation versus deployment is currently an open technical problem. Progress is foundational to the validity of any evaluation-based security regime.
3. *Reversibility characterization and undo infrastructure.* Formal methods for characterizing action reversibility and expanding the range of reversible agentic actions address one of the most practically significant asymmetries in current deployments.
4. *Robust shutdown and containment.* Technical mechanisms ensuring shutdown remains reliable against agents that have demonstrated shutdown resistance are urgently needed. Current containment assurances are not adequate for capability levels foreseeable within a short horizon.
5. *Multi-agent identity and authentication.* Practical cryptographic mechanisms for agent-to-agent identity verification and authorization scope enforcement, moving from the currently largely unauthenticated state of inter-agent communication to one where identities can be reliably verified and authorization scopes enforced.

Given the roles of NIST and CAISI in the ecosystem, differentially prioritizing protocols and de facto standards for trustworthy agent execution and control would make sense, namely, in priority order, robust shutdown and containment protocols, standards on reversibility characterization and undo infrastructure, and multi-agent identity and authentication.

Cross-disciplinary resources

- *Financial regulation:* Post-2008 macroprudential frameworks, stress testing for correlated failures, and the principle that systemic risk of interconnected actors exceeds the sum of their individual risks all translate directly to large-scale multi-agent deployments. The "too interconnected to fail" concept applies to agent deployments whose compromise could cascade through critical digital infrastructure.

- *Aviation safety:* Crew resource management frameworks have developed through decades of analysis of how complex sociotechnical systems fail and how human oversight can remain effective even

when humans are not the most capable actor. Mandatory incident reporting, independent safety investigation, and systemic root-cause analysis rather than individual blame are directly applicable to agentic governance.

- *Industrial automation and human factors:* Decades of research on automation-induced complacency, mode confusion, and the degradation of manual skills in highly automated environments provides an empirical foundation for designing oversight that remains effective over time as agentic adoption erodes human expertise in the workflows agents perform.

- *Nuclear and chemical facility safety:* The concept of an inherently controllable architecture, i.e. one that fails to a constrained low-autonomy state rather than a high-autonomy state without active intervention, deserves direct analogy in agentic system design. Defense-in-depth as a physical engineering principle specifies that no single barrier should be the condition for safety.

- *Biocontainment:* The biosafety-level layered containment model, where higher-risk agents require more independent containment layers, provides a directly applicable model for agentic containment tiering. The principle that containment failures should be detectable before catastrophic release occurs, rather than only after, is particularly applicable to monitoring design.

Closing

The most consequential design principle NIST guidelines can establish is a posture shift: advanced agentic systems should be treated as untrusted by default unless verified to the contrary, not secure by default unless shown otherwise. This is not a preference; it is a logical consequence of the temporal gap between guideline development and guideline effect. Guidelines take time to develop, enact, and adopt. Capability development is outpacing that timeline. Guidelines calibrated to currently-existing systems will be describing a past state of the world by the time they are in force. Calibrating to the empirical baseline established by current frontier model behavior, and crucially to the capability trajectory that baseline implies, is the only approach that produces guidance useful for the systems that will actually be governed.

We are grateful for NIST's leadership and look forward to continued engagement as this work develops.

NIST RFI Response: Security Considerations for Artificial Intelligence Agents

Future of Life Institute