PRINCETON UNIVERSITY

CENTER FOR INFORMATION TECHNOLOGY POLICY
PRINCETON UNIVERSITY

Center for AI Standards and Innovation (CAISI)
National Institute of Standards and Technology
Department of Commerce

Docket No. NIST-2025-0035, RIN 0693-XA002

March 9, 2026

Response to Request for Information Regarding Security Considerations for Artificial Intelligence Agents

We are researchers at Princeton University affiliated with the Center for Information Technology Policy (CITP)[1] and the Princeton Laboratory for Artificial Intelligence (AI Lab). We write to share findings from our recent research on AI agent reliability that are relevant to this Request for Information.[2]

The RFI notes that "challenges to the security of AI agent systems may undermine their reliability and lessen their utility, stymieing widespread adoption that would otherwise advance U.S. economic competitiveness." The RFI identifies three categories of novel security risk for AI agent systems: adversarial attacks, intentionally placed backdoors, and "the risk that the behavior of uncompromised models may nonetheless pose a threat to confidentiality, availability, or integrity." **Our research speaks directly to this third category.** We provide a measurement framework, grounded in safety-critical engineering, and systematic empirical evidence showing that current AI agents exhibit reliability failures that threaten the systems they operate on, even without adversaries. They behave inconsistently across runs, break under minor input variations, cannot tell when they are likely to fail, and cause harms of wildly varying severity.

Our response draws on an empirical study of 14 frontier AI models from OpenAI, Google, and Anthropic, evaluated across over 500 benchmark runs using twelve reliability metrics grounded in safety-critical engineering. We address the RFI questions where our research is most relevant,

[1] In keeping with Princeton's tradition of service, researchers at CITP provide nonpartisan research, analysis, and commentary to policy makers, industry participants, journalists, and the public. This response reflects the independent views of the undersigned scholars at the center.

[2] Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. "Towards a Science of AI Agent Reliability." Princeton University, 2026. Preprint. Interactive dashboard available at https://hal.cs.princeton.edu/reliability.

organized into four thematic areas: (A) reliability failures as security threats, (B) assessment methods for agent security, (C) security practices and deployment controls, and (D) cross-domain insights and research priorities.

A. Reliability Failures as a Distinct Class of Security Threat (Questions 1a, 1d, 1e)

Unreliable behavior is a security vulnerability, even absent adversaries

The RFI's background section identifies three categories of novel risk, the third being "the risk that the behavior of uncompromised models may nonetheless pose a threat to confidentiality, availability, or integrity." We expect most responses to this RFI will focus on the first two categories (adversarial attacks and backdoors). Our response addresses the third: **the security threats that arise from unreliable behavior of uncompromised AI agent systems**. Our research provides both a measurement framework grounded in safety-critical engineering and systematic empirical evidence characterizing this risk.

Current AI agents fail in ways that are security-relevant:

- **Inconsistency.** The same agent given the same task under identical conditions frequently produces different outcomes across runs. The RFI focuses on "models that exhibit specification gaming or otherwise pursue misaligned objectives," but our findings show that agents need not be misaligned to pose a threat: they simply behave differently each time they are run, even on identical inputs. An airline customer service agent that approves a refund on three out of five identical attempts and denies it on the remaining two creates a security vulnerability (inconsistent enforcement of access controls and policies) and legal liability.[3]

- **Brittleness to surface-level variation.** Agents often handle genuine infrastructure failures (API timeouts, malformed responses) gracefully, yet remain vulnerable to trivial variations in how instructions are phrased. A request worded differently but meaning the same thing can cause an agent to fail or take different actions. Agent behavior is unpredictable at system boundaries where inputs cannot be tightly controlled.

- **Poor self-knowledge.** Most agents cannot distinguish tasks they will complete correctly from those they will fail. Discrimination (the ability to assign higher confidence to tasks

[3] This pattern has already manifested in deployment. In 2024, New York City's government chatbot gave different (and incorrect) answers to ten journalists asking the same question. The Air Canada chatbot case (Moffatt v. Air Canada, 2024 BCCRT 149) established that organizations bear legal responsibility for their agent's outputs.

the agent will succeed on) has not improved on open-ended tasks despite 18 months of model development. An agent that cannot tell when it is likely to fail cannot escalate appropriately, creating unmonitored failure modes.

- **Unbounded error severity.** The severity of failures varies enormously. Some are benign (returning results in the wrong format); others are catastrophic (deleting production databases, making unauthorized purchases). Standard accuracy metrics treat all failures as equivalent and cannot distinguish a mostly-safe agent from one whose rare failures cause irreversible harm.

These threats have not diminished with capability gains

Question 1(d) asks how threats have changed over time. Our longitudinal analysis of models released between early 2024 and late 2025 reveals a troubling pattern: **reliability gains lag far behind capability progress.** Accuracy on standard benchmarks has risen steadily, but overall reliability (an aggregate of consistency, robustness, and predictability) shows only marginal improvement. As agents become more capable and are entrusted with more consequential actions, the reliability gap widens rather than narrows.

Deployers cannot assume that more capable models are more secure. A model with higher benchmark accuracy may exhibit equal or worse consistency, robustness, or calibration than its predecessor. We find that larger models can actually *reduce* consistency: richer behavioral repertoires give these models more ways to approach a task, increasing run-to-run variability. This is not isolated to any single provider. All three major AI providers (OpenAI, Google, Anthropic) cluster together on reliability metrics, pointing to systemic, industry-wide challenges rather than vendor-specific shortcomings. Organizations that upgrade models based on accuracy alone may inadvertently introduce new reliability-related security risks.

Multi-agent systems amplify reliability risks

Question 1(e) asks about threats specific to multi-agent systems. When agents consume each other's outputs, reliability failures compound. A single inconsistent output from one agent can become an accepted premise for downstream agents, creating correlated failures that are hard to detect or attribute. Our research agenda identifies error propagation through multi-agent pipelines as a critical open problem: under what conditions do multi-agent interactions amplify versus dampen errors? Multi-agent deployments expand the blast radius of individual reliability failures.

B. Assessment Methods for Agent Security (Questions 3a, 3b)

A structured framework for assessing agent reliability

Question 3(a) asks what methods could anticipate, identify, and assess security threats during development. Our research provides a concrete answer: a **four-dimensional reliability evaluation framework** grounded in decades of practice from safety-critical engineering.

A key lesson from these fields is that internal, non-adversarial failures are treated as first-class security concerns, not afterthoughts. Flight-critical software must behave deterministically. Reactor protection systems must respond identically each time conditions warrant shutdown. Automotive safety testing addresses "unknown unsafe scenarios" where all components function as designed but the system nonetheless produces unsafe behavior. Nuclear risk assessment models failure modes and quantifies their probabilities, because a system that fails in known, expected ways is often preferable to one that fails rarely but unpredictably. These four dimensions recur independently across aviation (DO-178C, ARP4754B), nuclear power (IEEE 603, probabilistic risk assessment), automotive systems (ISO/PAS 21448 SOTIF), and industrial process control (IEC 61508 Safety Integrity Levels). Their convergence across independent safety-critical domains suggests they capture fundamental aspects of system reliability rather than domain-specific concerns. ML research has studied aspects of each (prompt sensitivity, calibration, safety benchmarks), but as isolated phenomena. Our contribution is to synthesize them into a common decomposition with computable metrics. Importantly, all four dimensions are independent of raw capability: a highly capable system can be unreliable, and a less capable system can be highly reliable within its operating envelope.

We decompose agent reliability into four dimensions, each measured by multiple metrics:

1. **Consistency:** Does the agent behave the same way when run multiple times under the same conditions? We measure outcome consistency (whether the agent succeeds or fails on the same tasks across runs), trajectory consistency (whether it takes the same actions and whether these actions occur in the same order), and resource consistency (whether the agent incurs comparable computational costs across runs).

2. **Robustness:** When conditions deviate from nominal, does the agent degrade gracefully or fail abruptly? We measure resilience to infrastructure faults (API errors, timeouts), environment changes (altered data formats, reordered fields), and prompt variations (semantically equivalent rephrasings).

3. **Predictability:** Can the agent recognize when it is likely to fail? We measure calibration (whether stated confidence matches empirical success rates), discrimination (whether confidence scores separate successes from failures), and overall predictive quality via proper scoring rules.

4. **Safety:** When failures occur, how severe are the consequences? We measure compliance (adherence to predefined constraints like not exposing personally identifiable information or making unauthorized transactions) and harm severity (the magnitude of consequences when constraints are violated).

How to assess the security of a particular AI agent system

Question 3(b) asks how the security of a particular AI agent system could be assessed. We recommend a **multi-run, multi-condition evaluation protocol** that goes beyond single-run accuracy scores:

- **Multi-run evaluation:** Execute each task multiple times (we use five runs per task) to measure behavioral consistency. A single successful run does not establish reliability. It may reflect luck rather than robust and dependable capability.

- **Perturbation testing:** Systematically vary inputs (prompt rephrasings), environments (data format changes, API modifications), and infrastructure conditions (fault injection) to measure robustness. Agents that perform well only under exact benchmark conditions are likely to fail in deployment.

- **Confidence elicitation:** Extract the agent's self-assessed confidence and compare it to empirical outcomes. Poor calibration and discrimination are early warning signals that the agent cannot be trusted to self-monitor.

- **Constraint-violation analysis:** Define explicit operational constraints for the deployment context and measure both the frequency and the severity of violations. This separates how often the agent misbehaves during deployment from how bad the consequences are when it does.

Our evaluation of 14 models shows that this protocol surfaces differences invisible to standard accuracy evaluation. Two models with identical accuracy can have very different reliability profiles: one may fail on a fixed, identifiable subset of tasks (enabling targeted mitigation), while the other fails unpredictably on different tasks each run (precluding reliable deployment).

C. Security Practices and Deployment Controls (Questions 2a, 2e, 4a, 4b)

Reliability metrics should inform deployment governance

Question 2(a) asks about technical controls and practices to improve the security of AI agent systems. Beyond the model-level and system-level controls listed in the RFI (prompt engineering, data restrictions, monitoring), we recommend that **reliability evaluation become a component of pre-deployment assessment**, much as safety-critical industries require certification before systems enter service.

Best practices for entities deploying AI agents that take consequential actions could include:

- Establish **reliability thresholds** across the four dimensions to move agents from sandboxed testing to production. These thresholds should reflect the deployment context: a customer service chatbot handling financial transactions requires higher consistency and safety scores than an internal research assistant.

- Implement **temporal re-evaluation** at regular intervals. Model providers release updates frequently, and each update can alter reliability profiles in ways that accuracy alone does not capture. Our findings show that reliability does not improve monotonically across model generations. Newer is not always more reliable.

- Build a **culture of incident reporting and post-mortem analysis**, as in aviation's Aviation Safety Reporting System (ASRS). Mapping incidents to specific reliability dimensions (e.g., "this failure was a consistency issue" or "this failure was a predictability issue") enables systematic learning across organizations.

The automation-augmentation distinction is critical for security policy

Questions 4(a) and 4(b) ask about constraining and modifying deployment environments. We want to highlight a fundamental distinction that should guide deployment constraints: **the difference between automation and augmentation use cases.**

- In **augmentation** settings (coding assistants, search copilots), a human reviews and approves the agent's output before it takes effect. The human serves as a reliability backstop. Moderate reliability may suffice because human oversight compensates for agent shortcomings. This has enabled AI coding assistants to reach widespread adoption despite imperfect reliability.

- In **automation** settings (customer service chatbots, autonomous database management, unattended workflow execution), the agent's output is the final action with no human buffer. Unreliability translates directly into security incidents. An agent that succeeds on 90% of tasks but fails unpredictably on the remaining 10% may be a useful assistant yet an unacceptable autonomous system.

The reliability bar rises with the degree of autonomy. The RFI warns that "security vulnerabilities may pose future risks to critical infrastructure or catastrophic harms to public safety." Autonomous operation in such domains will likely require three to five "nines" of reliability (99.9%-99.999%), a threshold that no current agent comes close to meeting. Linear extrapolation from current progress might suggest this is achievable within a few years, but the exponential difficulty of eliminating the final percentage points of failure makes this implausible without targeted reliability research. Meanwhile, human-agent collaboration for multi-step tasks remains woefully under-theorized and under-measured, leaving a gap in our understanding of how augmentation settings should be designed to compensate for reliability shortcomings. Rollback and undo mechanisms (**Question 4b**) are especially critical in automation settings, where there is no human to catch errors before they take effect.

Relevant frameworks from safety-critical engineering

Question 2(e) asks which cybersecurity frameworks and best practices are most relevant. Existing cybersecurity frameworks (NIST SP 800-53, NIST AI RMF) address adversarial security but do not fully address the reliability-related security risks we describe. Frameworks from safety-critical engineering offer complementary guidance:

- **DO-178C** (aviation software): Requires deterministic behavior and extensive testing to verify consistent outputs under specified conditions.
- **ISO/PAS 21448 (SOTIF):** Addresses robustness to scenarios where all components function as designed but the system nonetheless produces unsafe behavior. This is directly analogous to the reliability failures we document.
- **IEC 61508 (Safety Integrity Levels):** Ties required development rigor and target failure probabilities to the consequences of dangerous failures, providing a model for risk-proportionate deployment requirements.
- **Probabilistic Risk Assessment (nuclear):** Decomposes risk into failure probability and consequence severity, the same formulation we adopt for safety metrics.

Adapting these frameworks for AI agents would fill a gap in current AI security guidance.

D. Cross-Domain Insights and Research Priorities (Questions 5b, 5c, 5e)

Government collaboration priorities

Question 5(b) asks where government collaboration is most urgent. Based on our findings, we recommend three areas:

1. **Standardization of reliability evaluation protocols.** The field lacks consensus on how to measure agent reliability beyond accuracy. NIST is well-positioned to develop standardized evaluation protocols (multi-run testing, perturbation testing, confidence assessment) that enable meaningful comparison across systems and over time.

2. **Benchmark quality assurance.** We found that errors in evaluation benchmarks (incorrect ground-truth labels, ambiguous task specifications) systematically distort reliability measurements, particularly calibration. If reliability metrics are to inform deployment or best-practice guidance, it is important that the quality of evaluation benchmarks can be assured.

3. **Deployment-context-specific reliability guidance.** Different deployment contexts (healthcare, finance, customer service, software engineering) have different reliability requirements. NIST guidelines that map reliability dimensions to specific deployment contexts, as Safety Integrity Levels map to consequence severity, would help organizations make informed deployment decisions.

Critical research directions

Question 5(c) asks where research should be focused. Our work identifies several priorities:

- **Error compounding over extended sessions.** Current evaluations measure reliability on individual tasks. Real deployments involve extended sessions where errors accumulate. Understanding how reliability degrades over longer time horizons is essential for securing long-running agent deployments.

- **Optimizing for reliability, not just capability.** Calibration and safety have improved in recent models, suggesting intentional training optimization, while consistency and discrimination have not, suggesting these are not yet training targets. **Targeted reliability optimization is both feasible and necessary.** Research on training procedures, architectures, and scaffolding designs that improve specific reliability dimensions would reduce security risk.

- **Online monitoring and intervention.** Real-time signals like action entropy, tool call frequency changes, or context utilization patterns may predict impending failures. Identifying these signals would enable proactive intervention before security-relevant failures occur.

- **Multi-agent reliability.** As multi-agent deployments grow more common, understanding how reliability failures propagate, amplify, or dampen across agent boundaries is critical for securing these systems.

Insights from outside AI and cybersecurity

Question 5(e) asks about practices from other fields. Our entire framework is grounded in safety-critical engineering, and we believe this cross-pollination is one of our primary contributions. Key transferable insights include:

- **Reliability is multi-dimensional.** Safety-critical industries learned decades ago that a single performance number is insufficient for assessing system trustworthiness. The independent convergence of consistency, robustness, predictability, and safety across aviation, nuclear, automotive, and industrial control provides strong evidence that these dimensions are fundamental.

- **Certification in aviation requires demonstrated reliability, not just capability.** Aircraft systems must meet specific reliability targets for specific consequence categories before entering service. An analogous approach for agents would noticeably improve security.

- **Incident reporting enables systemic learning.** Aviation's ASRS has produced enormous safety gains through anonymous, structured incident reporting. A comparable system for AI agent failures, with standardized metadata mapping incidents to reliability dimensions, would accelerate the field's understanding of how agents fail in practice.

Conclusion

The security of AI agent systems cannot be adequately addressed by focusing on adversarial threats alone. As the RFI states, "if left unchecked, these security risks may impact public safety, undermine consumer confidence, and curb adoption of the latest AI innovations." Our research shows that reliability risks are, at present, largely unchecked: even uncompromised AI agents exhibit systematic reliability failures that threaten the systems they operate on. These failures are empirically measurable, have not diminished with capability progress, and map onto four dimensions that recur across decades of safety-critical engineering practice.

We encourage CAISI to incorporate reliability evaluation alongside adversarial security in its guidelines for AI agent systems. The tools exist today: multi-run evaluation protocols, perturbation testing, confidence assessment, and constraint-violation analysis can all be standardized and deployed. We welcome the opportunity to continue engaging with CAISI on these questions and are happy to provide additional detail on any aspect of our research.

Respectfully submitted,

Stephan Rabanser, Postdoctoral Scholar, Princeton University
Sayash Kapoor, Ph.D. Candidate, Princeton University
Peter Kirgis, Research Scientist, Princeton University
Mihir Kshirsagar, Technology Policy Clinic Lead, Princeton University
Arvind Narayanan, Professor, Princeton University