PUBLIC COMMENT

Re: Request for Information Regarding Security Considerations for Artificial Intelligence Agents
Docket No. NIST-2025-0035
Document No. 2026-00206
March 9, 2026
Vega Commons Project

This comment addresses questions 1(a), 1(d), 2(a), 2(e), 3(a), 3(b), 4(a), 4(b), 4(d), 5(a), 5(b), and 5(e) of the RFI. It focuses on the security implications of retaining interaction records generated by AI agent systems.

OVERVIEW

Current frameworks do not cover a critical security surface: the retention and discoverability of records created when users interact with AI agent systems. These records capture the iterative reasoning process through which operators direct, refine, and evaluate agent actions, including delegated intent, authorization boundaries, and strategic constraints. They are retained by default on centralized infrastructure, are targets for adversarial extraction, and are producible under legal compulsion.

Governance of AI systems operates across multiple institutional layers, including records management, policy oversight, security controls, compliance requirements, and assurance practices. These layers assume the existence of system artifacts but do not define lifecycle rules for interaction records themselves.

They also do not distinguish between default retention posture during ordinary operation and preservation posture when legal, regulatory, audit, or investigative obligations require continued retention. Both states affect whether interaction records continue to exist and should be disclosed and governed separately.

The analysis below provides concrete attack scenarios, identifies control gaps, proposes measurable security requirements, maps recommendations to existing NIST frameworks, and includes a control mapping appendix for NIST SP 800-53.

1. SECURITY THREATS, RISKS, AND VULNERABILITIES

Response to 1(a): Unique Security Threats Distinct From Traditional Software

AI agent systems create a category of security exposure that has no direct analog in traditional software: the retained record of operator intent and deliberative process.

When an operator interacts with an AI agent to plan, research, decide, or act, the system generates logs containing initial problem framing and context, iterative instruction refinement as the operator evaluates agent outputs and adjusts direction, abandoned approaches and reconsidered options, intermediate reasoning steps generated by the system during task execution, delegated intent including what the agent was authorized to do and under what constraints, and execution history associated with tool calls, external services, or API-based actions performed by the agent, including authentication contexts, access patterns, and system responses.

This classification issue resembles the reasoning in *Carpenter v. United States*, where the Supreme Court recognized that automatically generated digital records may reveal highly detailed information about a person's activities. While *Carpenter* addressed passive location tracking, AI agent interaction logs present an analogous privacy concern through different mechanisms: although initiated by deliberate user input rather than passive sensing, the resulting records are generated as a condition of system operation, not as deliberate publication for external consumption, and they reconstruct the complete reasoning sequence that produced downstream actions.[1]

This distinction creates three threat vectors that do not arise in traditional software:

Threat 1: Credential and policy extraction via retained agent memory

An enterprise operator, contractor, or third-party service provider instructs an AI agent to perform tasks on behalf of an organization, including system administration, customer support, procurement, or other operational functions performed through hosted or API-based systems. The interaction log may capture the operator's reasoning about internal processes, network boundaries, authorization workflows, business constraints, temporary exceptions granted during the task, or instructions governing the agent's use of external tools, APIs, or connected services.

If the AI platform, integration service, or intermediary provider retains this interaction log by default, the retained record may reveal a detailed map of internal procedures, permissions, and decision constraints. Because the interaction data may be stored within provider-controlled logging, monitoring, or service infrastructure, exposure can occur through components of the

[1] *Carpenter v. United States*, 138 S. Ct. 2206, 2216–17 (2018) (describing cell-site location information as "*detailed, encyclopedic, and effortlessly compiled*").

workflow that are outside the organization's controlled systems. In workflows where agents invoke external tools or services, retained interaction records may also reflect credentials, access scopes, or execution parameters used across multiple systems, increasing the sensitivity of the retained data when logs exist outside the organization's controlled infrastructure.

Threat 2: Memory poisoning and indirect privilege escalation

In multi-turn agentic workflows, agents frequently retrieve historic interaction records for context, including records stored in persistent memory, retrieval indexes, tool-invocation logs, or integrated services used during prior execution. If retained operator interaction data is stored in shared repositories, logging systems, or third-party components used in the workflow, an attacker who compromises a lower-privilege part of the system, such as an integration service, external tool, or intermediary platform, can inject altered context into the retained records. When the agent later retrieves this corrupted memory during a higher-privilege task, the context may cause the agent to bypass execution constraints, disclose sensitive parameters from prior sessions, or perform actions outside the operator's intended authorization scope.

Interaction records may be reused across sessions, tools, external services, vendors, or service providers as part of normal agent operation. Context retained by one component of the workflow may be incorporated into execution performed by another component operating at a higher privilege level. In workflows that involve multiple tools or API-connected services, retained context may propagate across systems with different authorization scopes and different administrative control, including systems that were not intended to share persistent records.

Threat 3: Compelled disclosure of operator deliberative process

Retained agent interaction records are producible under legal compulsion. Federal courts have treated AI interaction logs as ordinary electronically stored information, holding that use of a consumer or hosted AI service does not create privilege or work product protection for the user's deliberative process.[2] Courts have also ordered production of large volumes of conversation logs as standard discoverable records. In some deployments, organizations may be required to retain interaction records for audit, compliance, safety review, or legal defense purposes. When such retention occurs across hosted platforms or integrated services, the organization faces compelled production risk for records it must preserve but does not control.

Interaction data may be retained not only for long-term storage, but also for operational logging, service telemetry, billing, monitoring, audit, troubleshooting, safety review, or quality-improvement processes.[3] Such data may be stored within the primary platform, within

---

[2] *United States v. Heppner,* No. 25-CR-503 (S.D.N.Y. Feb. 17, 2026) (holding that materials generated using publicly available AI tools were not privileged*); Warner v. Gilbarco, Inc.,* No. 2:24-cv-12333 (E.D. Mich. Feb. 10, 2026) (addressing work-product protection for AI-assisted materials).

[3] *In re OpenAI, Inc. Copyright Infringement Litigation,* No. 25-MD-3143, ECF No. 1021, at 3 (S.D.N.Y. Jan. 5, 2026) (ordering production of conversation logs as discoverable ESI).

integrated services, or within third-party providers supporting analytics, moderation, account management, or compliance functions, and may also be retained by external tools or API-connected services invoked during agent execution, depending on the architecture of the workflow. These records may persist beyond user-visible deletion, beyond the period intended by the operator, or outside the retention controls applied within the operator's own recordkeeping environment.[4]

Interaction data may be stored by platforms, integration services, or downstream providers involved in the workflow, resulting in artifacts existing across multiple systems outside the operator's direct control. Copies may be retained in logging, analytics, audit, or compliance systems operated by the primary platform or by supporting providers, and may persist even when the operator does not intend long-term storage or is unaware that retention has occurred. Centralized or intermediary retention of agent interaction data can create a pool of producible records exposing not only the actions taken but also the operator's strategic reasoning, authorization constraints, and rejected alternatives.

Concrete scenario 1: Healthcare decision support

A healthcare professional uses an AI agent delivered through a hosted or API-based service to research treatment protocols for a complex case, iterating through differential diagnoses, weighing risk factors, and evaluating options ultimately not pursued. The interaction record captures the clinician's provisional reasoning, including assessments and hypotheses that were considered and rejected. This record may be stored by the service provider or by supporting systems outside the healthcare organization's controlled recordkeeping environment and may not be governed by the retention, classification, or access rules applied to the patient's official medical record.

In some deployments, the healthcare organization may be required to retain records of AI-assisted decision support for audit, quality review, or legal defense purposes. When such records are created through hosted or API-based services, the organization may be required to preserve them without controlling the lifecycle, custody surface, or preservation state of the systems in which they actually reside.

The risk is associated with retention of interaction records outside the system of record used for

[4] Geoffrey A. Fowler, *ChatGPT's year-end review knows way too much. How to fix your privacy settings*, Washington Post (Dec. 29, 2025) (reporting that AI services may retain conversation data for training, memory, safety review, or other internal processing, and that user deletion controls do not necessarily remove all associated records from provider-operated, logging, analytics, or supporting systems involved in service operation); see also Jennifer King, Kevin Klyman, Emily Capstick, Tiffany Saade & Victoria Hsieh, *User Privacy and Large Language Models: An Analysis of Frontier Developers' Privacy Policies*, Stanford Institute for Human-Centered AI (2025) (review of major AI developers' privacy policies finding that user interaction data may be retained for training, review, safety, and service operation across multiple systems and may persist across provider-controlled infrastructure).

clinical documentation, rather than with improper disclosure of protected health information. Interaction artifacts may be retained within provider-controlled infrastructure, logging pipelines, or supporting services that are not subject to the operator's retention policy. In many deployed systems, deletion of a user-visible session does not remove all associated records, which may persist in logging, monitoring, audit, or service-management systems. Such records may remain available to the provider or to downstream systems involved in logging, monitoring, analytics, or service operation, and may become producible in malpractice or regulatory proceedings even when the operator did not intend the deliberative process to be preserved.

Concrete scenario 2: Autonomous procurement

An organization uses an agentic system delivered through a hosted platform or integration service to assist with procurement, subcontracting, or supply chain negotiation. The operator provides instructions describing budget limits, fallback positions, approval conditions, or internal constraints relevant to the transaction. The interaction record may capture negotiation strategy, authorization thresholds, and internal decision rules not intended for external disclosure. These records may be stored by the platform, by an integration provider, or by third-party API services and supporting systems used for logging, analytics, monitoring, or service operation. In multi-provider workflows, interaction data may be retained by upstream or downstream services participating in the workflow, and copies may exist across administrative domains outside the control of any single operator’s recordkeeping environment, including upstream or downstream services involved in logging, monitoring, analytics, or service delivery.

In multi-provider deployments, interaction artifacts created during system operation may be retained across multiple services involved in the workflow, including provider-controlled infrastructure outside the organization’s recordkeeping environment. In such cases, the operator may not control lifecycle state transitions, deletion propagation, preservation overrides, or retention schedules for records stored in provider-controlled infrastructure.

In some deployments, the organization may be required to retain records of authorization decisions, approval conditions, or negotiation workflows for audit, compliance, internal governance, or legal hold purposes. When such records are created across hosted platforms or integrated services, the organization may be required to preserve them without controlling deletion propagation, preservation overrides, or lifecycle state transitions across the systems in which copies reside.

Interaction data may be replicated across logging, analytics, audit, or service-management systems operated by the primary platform or by supporting providers, and deletion of a user-visible interaction may not remove all associated records from provider-operated or downstream systems. Such records may remain available to the provider or to other components involved in service delivery, monitoring, or compliance, and may become producible in

contractual, regulatory, administrative, or civil proceedings even when the operator did not intend the deliberative process to be retained. The risk arises from retention of interaction artifacts created during normal system operation, rather than from improper disclosure of sensitive information.

This scenario illustrates the need for controls addressing retention visibility, lifecycle management, and cross-system record propagation in multi-provider AI deployments.

Response to 1(d): How Threats Have Changed And Are Likely To Evolve

Four developments have materially changed this threat surface in the past twelve months:

First, judicial treatment of AI logs as ordinary electronically stored information has moved from theoretical to operational. A February 2026 federal court ruling confirms that AI interaction logs may be treated as routine business records for discovery purposes.

Second, platform retention practices have produced a documented two-tier regime. Enterprise customers operating under negotiated terms may receive reduced-retention or non-retention configurations, while consumer and small-organization users are typically subject to default retention policies. These protections are contractual rather than structural, and vendor retention practices may change under legal, commercial, or governmental requirements in ways that contractual assurances do not fully control. Public analyses of platform policies describe deletion overrides through legal holds,[5] preservation directives, and exception-tail retention. Provider documentation also indicates that certain safety, monitoring, or classifier artifacts may be retained beyond standard deletion intervals, and that scheduled deletion may be suspended in response to litigation holds, regulatory obligations, or court orders. These conditions create a distinction between retention states intended to prevent storage and preservation states imposed by legal, regulatory, or provider-level requirements, which may suspend operator-selected deletion settings.

Third, agentic capability is scaling rapidly. As AI agents gain persistent memory, multi-session continuity, tool-call authority, and autonomous action capability, including the ability to invoke external tools, services, or APIs as part of normal execution, the retained record expands from discrete interaction snapshots to longitudinal operational profiles encoding authorization patterns, strategic constraints, and organizational decision processes. Existing security guidance primarily addresses the integrity of agent execution, but does not address governance of the interaction records that document how execution was directed.

---

[5] Fatemehsadat Mireshghallah & Yixuan Li, “Position: Privacy Is Not Just Memorization!” arXiv:2510.01645 (Oct. 2025).

Fourth, regulatory frameworks are beginning to require logging and retention practices that expand this exposure surface. Regulation (EU) 2024/1689 requires providers of certain high-risk AI systems to implement automatic logging capabilities under Article 12 and to retain the logs referred to in Article 12(1), to the extent under their control, for a period appropriate to the system's intended purpose and at least six months under Article 19, unless other Union or national law provides otherwise. The regulation does not specify custody architecture for the required logs, does not provide privilege or discovery protection for them, and does not address the exposure created by centralized retention of detailed interaction records. A system that fully complies with logging requirements may therefore generate a comprehensive, searchable archive of interaction data as part of normal operation. Comprehensive logging supports safety and auditability, but also increases the volume of retained records that may be subject to extraction, reuse, or compelled production. Controls such as content-telemetry separation and configurable retention classification allow auditability to be maintained without requiring full-content retention in centralized infrastructure.

In systems operating under mandatory logging, audit, compliance, safety-review, or legal-preservation requirements, non-retentive architectures may not be available as compliance options. In those deployments, the relevant security question is not whether retention can be eliminated, but how retained interaction records are classified, separated from operational telemetry, preserved, disclosed, and governed across the full custody and lifecycle chain.

In many deployed systems, logging and telemetry are implemented within provider-operated or cloud-hosted infrastructure rather than within the operator's own environment. As a result, interaction records, execution traces, or monitoring artifacts may exist within service-provider systems even when the operator has configured reduced-retention or zero-retention modes for primary content storage. Interaction records may also be created within provider-operated logging, safety, analytics, or monitoring systems outside the operator's controlled environment, including in multi-vendor or API-based workflows. This includes systems accessed through hosted interfaces, APIs, or multi-service workflows in which execution, logging, or monitoring functions occur outside the operator's controlled infrastructure.

In multi-service or agentic workflows, interaction records may be distributed across multiple provider-operated or integrated systems, creating a supply chain of retained artifacts outside the operator's direct control. Controls that limit model training use, enable reduced-retention configurations, or provide zero-data-retention (ZDR) modes do not necessarily eliminate operational logs, safety telemetry, or service-management records. Zero-retention or reduced-retention modes typically apply to primary content storage but may not govern all operational logs, classifier outputs, monitoring artifacts, or provider-level records generated during system operation.

Interaction with agent systems often includes exploratory, provisional, or deliberative exchanges that would not ordinarily be preserved as formal records within the operator's own environment,

such as drafts, notes, or working materials created during problem-solving or planning. When such interaction occurs through hosted or API-based systems, the resulting artifacts may be retained as part of normal service operation even if the operator does not intend long-term storage. Once retained within provider-controlled infrastructure, these interaction records may exist outside the operator's recordkeeping environment and may be treated as ordinary electronically stored information for purposes of audit, investigation, or legal process.

Retention of interaction data within provider-operated, integrated, or downstream systems may therefore create a standing pool of records that remains subject to discovery, audit, or legal preservation regardless of the operator's intended retention policy or configured storage settings.

These developments are likely to accelerate as agent capabilities, multi-service workflows, and regulatory logging requirements continue to expand. In many current deployments, the operator can request deletion but does not control whether retention or preservation overrides occur. Without architectural standards for retention governance, the exposure surface created by retained interaction data will increase with the scale and autonomy of deployed systems.

Many current governance approaches require logging, monitoring, and traceability for accountability purposes but do not define lifecycle controls for the resulting artifacts. Interaction records may therefore persist across provider infrastructure, backup systems, safety pipelines, analytics stores, and secondary-use workflows without explicit classification, retention limits, or deletion rules.

This gap indicates the need for architectural controls that address the custody, lifecycle, and retention of interaction artifacts, in addition to the security of the systems that generate them.

2. SECURITY PRACTICES

Response to 2(a): Technical Controls to Improve Security

The following controls address the operator interaction data security surface at the system, platform, and operator levels:

Retention classification: Agent systems should implement configurable retention classes for interaction data, but retention class availability does not constitute operator custody of retention decisions. Ephemeral: instruction content exists only in volatile memory and is cryptographically wiped upon session termination; this should be the default option for systems capable of affecting external state where governance requirements permit non-retention. Session-bounded: content is retained for a defined period tied to task completion or tool-call resolution, then purged. Durable: content is explicitly saved by the operator for workflow templates or

compliance requirements, with elevated encryption and local-first storage where feasible. These classes define technical capability. Custody architecture determines who governs which class applies to a given interaction and whether deletion requests are honored, suspended, or overridden. Telemetry necessary for service operation (timestamps, error states, resource usage, success/fail indicators) should be architecturally separated from cognitive content (prompts, responses, reasoning sequences, authorization parameters) so that operational logging does not require content retention. In cases where content review is required for safety, incident response, or compliance, such access should be subject to elevated authorization, purpose limitation, and separate retention controls distinct from routine operational telemetry.

Retention classification should also define the conditions under which deletion rules are suspended and records enter a preservation state. The critical governance question is not whether the platform can execute these state transitions, but whether the operator controls the determination of which records enter preservation, under what conditions, and through what decision process. Systems should be capable of documenting which records are subject to preservation, which systems hold copies, and whether user-visible deletion differs from backend retention. Without operator custody of lifecycle state determination, interaction data retention is governed by platform policy, contractual terms, and vendor discretion rather than by the deploying organization's retention rules.

Content-telemetry separation: Agent systems should support logical or physical separation between routing metadata (operator identity, billing, rate-limiting data) and instruction payload (the substance of the operator's interaction). This separation reduces the risk of re-joining of anonymized telemetry with identified cognitive content, allows operational monitoring without content surveillance, and ensures that audit objectives may require monitoring of system behavior, but should not require routine retention of full interaction content when metadata, scoped review, or controlled access can satisfy accountability requirements.

Retention stack disclosure: Platforms should disclose the full retention architecture for interaction data, including default retention periods and class assignments, override conditions (legal holds, safety-classifier flags, regulatory requirements), secondary-use pathways (model training, product improvement, advertising), and the conditions under which operator-initiated deletion is not honored. Disclosure requirements address transparency within a platform-controlled retention regime but do not establish operator custody. A platform that fully discloses its retention practices while retaining unilateral authority over deletion, preservation, or lifecycle state transitions has satisfied transparency requirements but not custody requirements. Disclosure should also distinguish user-visible deletion from backend retention and identify whether ordinary deletion is suspended under legal hold, preservation demand, regulatory inquiry, contractual audit, or internal investigation.

Operator-configurable retention: Operators and deploying organizations should have enforceable controls over interaction data retention, including the ability to select zero-retention modes for sensitive workflows. Enforceable in this context means that the operator governs

whether records are created, retained, or deleted, rather than configuring preferences within a platform-controlled retention regime. A system that offers retention class selection but reserves platform override authority, secondary-use exceptions, or unilateral preservation determinations has provided configuration options, not custody controls. These controls should be implemented in a way that does not rely solely on policy assurances, contractual commitments, or platform discretion.

Compelled access notification: When interaction data is produced under legal compulsion, the operator or deploying organization should receive notice to the extent permitted by law, analogous to existing notification practices used in communications or data-access requests.

Response to 2(e): Relevant Cybersecurity Guidelines And Frameworks

NIST SP 800-53 Rev. 5 provides relevant control families, but application to AI agent interaction data requires additional interpretation. A detailed control mapping is provided in the Appendix.

AU (Audit and Accountability) controls address logging but do not distinguish between operational telemetry and operator intent content. In agent systems, audit requirements may involve both operational metadata and interaction content, but these categories may require different retention classes. Guidance should specify that audit compliance can be achieved without requiring uniform retention of full interaction content when telemetry, event records, or summaries are sufficient for accountability. This distinction is important because audit objectives concern system behavior, while interaction content may contain the operator's reasoning process.

SC (System and Communications Protection) controls address protections for data in transit and data at rest, but assume that processed data will be stored in persistent form. In agent systems, interaction content may be transient and may not require retention after execution. Some platforms provide reduced-retention or zero-data-retention configurations intended to limit storage of interaction content, but these controls typically apply to primary content storage and may not govern all records generated during processing, including operational logs, safety telemetry, monitoring systems, classifier artifacts, or other provider-operated and integrated service components. Guidance should address the ability to execute tasks without persisting full interaction content to storage, logs, caches, or provider-operated infrastructure when retention is not required. Conversely, when retention is required by audit, compliance, safety review, or legal preservation mandates, guidance should address lifecycle controls, custody surface disclosure, and preservation state management for records that must persist. Non-retention or minimized-retention control capabilities are therefore needed in addition to traditional

data-at-rest protections, together with lifecycle controls for records that must persist.

SI (System and Information Integrity) controls address the integrity of stored and transmitted data, but do not address the integrity of deletion, the completeness of record removal, or the risk that retained interaction data may continue to influence later execution in persistent-memory or multi-session agent systems. In systems that reuse context across sessions, tools, or external services, interaction content may be carried forward beyond its intended scope, allowing prior instructions, constraints, or exploratory exchanges to affect subsequent operations. Guidance should address the correctness of lifecycle transitions, including deletion, reset, and context isolation, so that previously retained interaction data does not persist in ways that alter execution after its intended use has ended.

MP (Media Protection) controls address sanitization and disposal of physical media, but do not address sanitization of volatile memory, temporary storage, or transient execution environments used by agent and cloud-hosted systems. Interaction data may exist in RAM, container volumes, caches, swap space, or short-lived service instances, and may be captured in snapshots, backups, or provider-operated infrastructure even when execution is intended to be ephemeral or session-bounded. Guidance should address sanitization requirements for volatile and temporary storage so that interaction content does not remain recoverable after the end of a session, task, or execution cycle.

The **NIST AI Risk Management Framework (AI 100-1)** and the **Generative AI Profile (AI 600-1)** address bias, safety, transparency, and governance considerations for AI systems. Neither explicitly addresses the retention, lifecycle, and discoverability of operator interaction records as a distinct risk category, including the exposure created when exploratory, provisional, or deliberative interaction artifacts are retained as part of normal system operation. The frameworks focus on system behavior and organizational oversight, but do not specify how records generated through operator interaction should be classified, minimized, retained, deleted, or governed across provider-operated or multi-vendor environments.

Recent governance frameworks, including ISO/IEC 42001 (AI management systems), focus on organizational controls for the design, deployment, and oversight of AI systems. These frameworks address lifecycle, risk management, and accountability at the system level, but do not specify how conversational interaction records, prompts, responses, or agent-generated content should be classified, minimized, retained, or deleted across provider-operated, integrated, or multi-vendor environments. As a result, records generated through ordinary use of agent systems remain only partially governed within existing standards.

Existing standards also do not define who controls the lifecycle of interaction records once they are created within provider-operated or multi-vendor environments. Existing records-management authorities define retention and disposition rules for recognized record categories, but interaction artifacts generated during AI use often do not map cleanly to those

categories. Retention, deletion, preservation, and access are therefore often determined by implementation defaults, platform settings, or vendor practices rather than by explicit policy rules.

3. ASSESSING SECURITY

Response to 3(a): Methods To Anticipate, Identify, and Assess Security Threats During Development

Assessment of operator interaction data risk during development should include the following methods:

Retention surface mapping: During system design, developers should map all points at which operator interaction data is persisted, including primary storage, backup systems, logging pipelines, safety-classifier retention, model training data pipelines, and caching or replication systems. This mapping should distinguish between retention for service operation (the vendor holds data as a technical incident of providing the service), retention for secondary use (the vendor holds data for model training, product improvement, or other purposes beyond the immediate transaction), and retention under legal hold (data is preserved pursuant to litigation or regulatory requirements).

Each form creates different security exposures, governance requirements, and remediation paths. Retention for service operation may in some cases be minimized through architectural controls. Retention for audit, compliance, safety review, or legal preservation may be required, in which case the relevant assessment questions concern disclosure, override conditions, lifecycle management, and operator visibility rather than elimination alone.

Where interaction artifacts may be stored by multiple service providers, assessment should identify whether the operator controls retention, deletion, or downstream use across the full service chain.

Assessment should also determine whether deletion at the application layer removes records from backup systems, safety pipelines, vendor logs, and derived artifacts, and whether preservation obligations would require those records to be retained even after user-visible deletion.

Deletion verification testing: Developers should test whether operator-initiated deletion actually removes interaction data from all retention points identified in the surface map. Current practice reveals significant gaps: safety-classifier flags can retain associated artifacts for years beyond

standard deletion timelines, legal holds can suspend deletion without notice, and data may persist in training pipelines or aggregated datasets after nominal deletion from the primary store.

Compelled access modeling: Developers should assess what interaction data would be producible under a civil subpoena, criminal investigation, or administrative demand. If the producible corpus includes operator deliberative process, the system's retention architecture presents a security risk that traditional access controls do not mitigate.

Log-reconstruction adversarial testing: Red-teaming frameworks for agentic systems typically focus on prompt injection and model evasion. Assessment should be expanded to include log-reconstruction audits. Assessors should attempt to reconstruct the operator's original intent, authorization boundaries, and strategic constraints using only the decoupled telemetry and execution logs that the system retains after content purging. A system that claims content-telemetry separation should render this reconstruction infeasible.

Custody and override-path testing: Assessment should determine whether retention, preservation, and deletion state transitions are controlled by the deploying organization or by the platform provider. Testing should identify whether provider-operated systems can override operator-selected retention classes, suspend deletion, or retain interaction artifacts for safety, monitoring, regulatory, or secondary-use purposes. Where override paths exist, assessment should document the conditions under which they are triggered and whether the operator receives notice when retention state changes occur.

Multi-system retention verification: Assessment should evaluate whether interaction artifacts persist across integrated services, external tools, or downstream providers participating in the workflow. In multi-vendor or API-based environments, interaction content, execution traces, or telemetry may be retained outside the primary platform. Testing should determine whether operator-initiated deletion propagates to all participating systems and whether any component retains derived artifacts, logs, or monitoring records after the primary interaction has been removed.

Context isolation testing: For systems that reuse interaction history across sessions or tools, assessment should verify that context reset, session termination, or retention-class transitions prevent prior interaction content from influencing later execution. Testing should confirm that retained context cannot be injected into higher-privilege workflows or reused outside the scope intended by the operator.

Response to 3(b): Assessing The Security of a Particular AI Agent System

The following assessment criteria are specific to operator interaction data security and should be included when evaluating a particular AI agent or agentic workflow system:

Assessment should determine whether the system retains the substance of operator interaction, including prompts, responses, reasoning traces, tool calls, and authorization parameters, after session completion, and if so, the duration of retention, the retention class applied, and any conditions under which retention may be extended or overridden.

Assessment should evaluate whether operator intent content is architecturally separated from operational telemetry, such that monitoring, logging, and reliability functions can operate without requiring retention of full interaction content.

Assessment should determine whether the deploying organization or operator can configure retention class, including minimized-retention or ephemeral execution modes, and whether those settings are enforced across all components of the system, including provider-side logs, safety pipelines, and integrated services.

Assessment should identify what interaction artifacts would be producible under legal compulsion, including civil discovery, administrative demand, or criminal process, and whether the operator receives notice when preservation or disclosure obligations attach.

Assessment should verify whether deletion is demonstrable across the full retention surface, meaning the system can confirm removal of interaction content from primary storage, backups, exception tails, safety monitoring systems, training or evaluation pipelines, and any derived datasets.

Assessment should evaluate whether operator intent could be reconstructed from telemetry, metadata, or execution traces retained after content deletion, and whether the system's logging architecture prevents reconstruction of deliberative process.

Assessment should determine whether retention, preservation, or deletion state transitions can be overridden by the platform provider, safety systems, or regulatory controls, and whether such overrides occur without operator control or visibility.

Assessment should evaluate whether interaction artifacts persist across multi-vendor or API-based workflows, including external tools, orchestration layers, or downstream services, and whether deletion requests propagate to all participating systems.

Assessment should confirm whether locally executed or non-API workflows create interaction records outside the primary platform, and whether the security assessment accounts for logs, caches, or temporary storage created in those environments.

These assessment criteria address both whether non-retentive architectures are correctly implemented and whether governance controls are adequate for deployments where retention is required. In systems that retain interaction records for legitimate audit, compliance, or safety purposes, the relevant questions include whether retention is appropriately classified, disclosed, bounded, and governed across the full lifecycle. In systems claiming minimized-retention or ephemeral execution, assessment should verify that those claims hold across the full retention surface, including downstream services, derived artifacts, and reconstruction risk.

These criteria could be incorporated into a standardized assessment checklist analogous to the multi-tier supplier and system assessment approaches described in NIST SP 800-161.

4. DEPLOYMENT ENVIRONMENTS

Response to 4(a): Deployment Environment Constraints

One architectural approach that can reduce operator interaction data risk is local-first processing. AI agent systems may be designed so that instruction payloads are processed on the operator's local hardware or within an environment under the operator's control, with only the data necessary for service operation or external tool execution transmitted to centralized infrastructure.

This architecture is feasible for some agent use cases. Local inference using open-weight models, combined with selective API calls for capabilities that require frontier-scale compute, can allow substantial portions of deliberative interaction to occur without centralized content retention. The agent's interactions with remote tools or services may still require external communication, but the operator's reasoning process that directed those actions need not be retained centrally.

Where execution cannot remain entirely local, ephemeral execution environments provide an intermediate option. These are compute environments in which interaction data is processed in volatile memory and not written to persistent storage as part of normal operation. The key requirement is that instruction content does not enter durable storage during execution.

One possible control approach is verifiable non-retention: When an agent session concludes, the system may provide an attestation that instruction content has been purged from the processing environment. This is analogous to secure deletion verification practices used in high-assurance environments. Implementation approaches may include cryptographic attestation from hardware-backed secure enclaves and other verifiable destruction or sanitization mechanisms. Such attestations would not eliminate all downstream retention risk, but may provide a measurable control for the processing environment itself.

Deployment architecture may be constrained by governance requirements: These approaches are most relevant where non-retention is permissible. Many deployments operate under governance requirements that mandate retention of interaction records for audit, compliance, safety review, or legal preservation. In those environments, the security problem is not eliminated by architecture choice alone, because the records must persist to satisfy legitimate governance obligations. Once retained, such records may be treated as ordinary electronically stored information for purposes of audit, investigation, or legal process. Where retention is required, the relevant control question becomes how those records are classified, separated from operational telemetry, preserved, disclosed, and governed across the full retention and custody chain.

For deployments where retention is required by applicable legal, regulatory, or institutional requirements, the controls described in Section 2, including retention classification, content-telemetry separation, retention stack disclosure, and preservation-state transparency, provide governance over the lifecycle of records that must persist.

Response to 4(b): Modifying Environments to Mitigate Threats

Memory-only execution: Agent processing environments can be configured to operate without persistent storage for interaction content. Where governance requirements permit non-retention of interaction content, and where interaction content is processed only in volatile memory and not written to durable storage as part of normal operation, the risks associated with breach, compelled production, and secondary use are substantially reduced.

Network segmentation for content and telemetry: Deployment environments should route instruction content and operational telemetry through logically or physically separate paths, reducing the risk that telemetry can be re-identified through correlation with content streams.

Retention budget enforcement: Deployment environments can enforce maximum retention periods at the infrastructure level, automatically purging interaction content after a configured interval regardless of application-layer settings. This provides defense-in-depth against application-layer retention overrides. Retention budget controls should also account for derived artifacts, monitoring records, and provider-operated systems that may retain copies beyond the primary application layer.

Cross-session memory isolation: For systems that implement durable retention classes for specific workflows, retained interaction artifacts should be cryptographically or logically isolated to prevent cross-session contamination. A lower-privilege interaction record should not be retrievable during a higher-privilege execution context. This addresses the memory poisoning threat described in Section 1(a).

These environment modifications are most applicable where non-retention or minimized-retention operation is permissible. In deployments that must retain interaction records for audit, compliance, safety review, or legal preservation, the relevant controls are those that govern how retained records are segmented, isolated, disclosed, and managed across the lifecycle rather than eliminated altogether.

Response to 4(d): Monitoring Deployment Environments

Monitoring paradox: Traditional security monitoring logs system activity for analysis. If that monitoring captures the content of operator interaction with AI agents, the monitoring system itself becomes a retention point for operator intent data. This tension does not arise in traditional software monitoring and is not addressed by existing monitoring frameworks.

Proposed approach: Monitoring should operate primarily on metadata and telemetry rather than on full interaction content. In cases where content review is required for safety, incident response, or compliance, such access should be subject to elevated authorization, purpose limitation, and separate retention controls distinct from routine operational monitoring.

Legal and privacy challenges: The primary legal challenge is that monitoring systems which retain interaction content for security analysis create the same compelled-access exposure as the primary interaction logs. Any monitoring framework for AI agent systems should specify what monitoring data is itself producible under legal compulsion, and should apply the same retention minimization principles to monitoring data as to primary interaction data.

Monitoring and compliance systems may themselves become additional retention points, causing records that would otherwise be deleted to persist across audit logs, backup media, or vendor-managed systems once preservation obligations apply.

5. ADDITIONAL CONSIDERATIONS

Response to 5(a): Methods and Tools to Aid Rapid Adoption

A useful tool for rapid adoption would be a standardized assessment methodology for operator interaction data risk, analogous to the NIST Cybersecurity Framework's assessment tiers. A baseline governance standard should include default non-retention for instruction content, configurable retention classes with disclosed override conditions, architectural separation of content from telemetry, verifiable deletion capability across all retention points,

compelled-access notification where legally permitted, and prohibition on secondary use of interaction content without explicit operator consent.

The submitter has developed a governance specification that includes lifecycle state definitions, custody-surface mapping, configurable retention classes, preservation-state controls, and verifiable deletion for interaction records generated by automated systems. The specification is designed to operate alongside existing privacy, security, and AI risk frameworks and is available upon request.

The configurable retention class model is designed to accommodate sector-specific retention mandates. In healthcare, clinical decision support queries can be classified as ephemeral while AI-generated recommendations adopted into clinical documentation are retained as part of the medical record. In financial services, compliance-related AI interactions can be classified as persistent under FINRA Rule 17a-4 while internal analytical reasoning is classified as ephemeral. The governance standard does not require non-retention across all artifact classes. It requires that the operator specify which classes are retained and for how long, that the classification is documented and auditable, and that default retention is minimized rather than enabled by default.

Response to 5(b): Areas Where Government Collaboration Is Most Urgent

Government collaboration is most urgent on retention governance for AI interaction data. No federal or international governance framework currently specifies how interaction records generated through AI systems, agents, or automated services should be created, retained, minimized, or deleted across vendors, services, or jurisdictions. This gap extends beyond U.S. frameworks: judicial and professional guidance issued across common-law jurisdictions, including the United Kingdom, Singapore, Australia, Canada, New Zealand, and Hong Kong, addresses AI confidentiality, access controls, and vendor due diligence but does not address interaction-record retention architecture, vendor-side custody categories, or the conditions under which retained records may be compelled.

This gap persists even where broader AI governance frameworks exist, because current management-system and risk-management standards focus on system governance, model behavior, data quality, and auditability, not on the lifecycle and custody of interaction records generated during use.

Courts are treating these records as ordinary electronically stored information. Platforms are retaining them by default. Enterprise customers are negotiating contractual non-retention. Individual operators and small organizations have no equivalent protection.

NIST is well-positioned to establish voluntary technical standards for interaction data governance that provide a baseline regardless of the pace of legislative or judicial development.

Response to 5(e): Practices From Fields Outside AI and Cybersecurity

Two established frameworks are directly relevant:

Library privacy: State library privacy statutes protect patron borrowing records on the principle that reading and research patterns reveal cognitive process and that surveillance of those patterns chills intellectual freedom. AI interaction logs capture the same category of information through a software interface rather than a physical lending system. The framework demonstrates that legislatures have recognized cognitive process records as warranting protection. The gap is that existing statutes are medium-specific and do not extend to AI interaction records.

Attorney work product: The work product doctrine protects materials prepared in anticipation of litigation because compelled disclosure of the reasoning process would undermine the adversarial system. AI interaction logs capture an equivalent reasoning process for all operators, not only attorneys. The doctrine demonstrates that the legal system already recognizes some categories of process records as warranting heightened protection. The gap is that the protection is currently limited to credentialed professionals.

These analogies suggest that the security concern identified in this comment is not novel in principle. The novelty is in scale, in default retention as a byproduct of system architecture rather than operator choice, and in the absence of any governance framework for the category.

These observations are intended to identify a records-management and security surface that may warrant further consideration in future guidance addressing AI system deployment, logging, and lifecycle governance.

APPENDIX:

NIST SP 800-53 CONTROL MAPPING FOR AGENTIC AI SYSTEMS

To operationalize the mitigation of operator interaction data risk in agentic AI environments, the following existing NIST SP 800-53 (Rev. 5) control families should be adapted:

MP (Media Protection): Ephemeral Sanitization

Current control: MP-6 (***Media Sanitization***) applies to the end-of-life disposal of physical storage media and the clearing of persistent storage.

Agentic adaptation: Guidance should extend MP-6 to the volatile memory of agentic execution environments. Where agent systems process operator instruction content in RAM-based ephemeral environments, sanitization verification should confirm that instruction buffers and agent reasoning traces are securely purged from volatile memory upon session termination or tool-call resolution. Proof-of-non-retention attestation, as described in Section 4(a), provides the verification mechanism.

AU (Audit and Accountability): Telemetry Decoupling

Current control: AU-3 (***Content of Audit Records***) requires systems to capture sufficient information to establish what events occurred, the sources of events, and the outcomes.

Agentic adaptation: In traditional software, capturing the full request payload for audit purposes is standard practice. In agentic AI, the instruction payload contains the operator's strategic reasoning, authorization constraints, and delegated intent. Retaining this content for audit purposes centralizes sensitive data that is simultaneously a breach target and a compelled-disclosure liability. AU-3 guidance for agentic systems should explicitly specify that audit compliance can be achieved through decoupled telemetry (identity, timestamp, action type, success/fail state, resource identifiers) without retaining the semantic content of the operator's instructions.

SI (System and Information Integrity): Memory Boundary Isolation

Current control: SI-7 (***Software, Firmware, and Information Integrity***) protects against unauthorized modification of system components and information.

Agentic adaptation: Where agentic systems implement durable retention classes for specific workflows, SI-7 should be extended to require cryptographic isolation of retained interaction artifacts across sessions and privilege levels. A retained interaction record from a lower-privilege session must not be retrievable or injectable into the context of a higher-privilege execution. This addresses the memory poisoning and indirect privilege escalation threat described in Section

1(a).

AC (Access Control): Gated Retrieval

Current control: **AC-3** (***Access Enforcement***) governs access to system resources.

Agentic adaptation: Where operator instruction logs must be retained for compliance or specific operational requirements, AC-3 should mandate that internal vendor access to these logs requires elevated, multi-party authorization. The default state should be zero internal access to operator intent artifacts. Access events should be logged with the same rigor applied to access to cryptographic key material.

Defining lifecycle and custody controls for interaction records as a distinct governance layer would complement existing risk-management, assurance, and records-management frameworks by addressing artifacts that are currently created implicitly but governed inconsistently.

ABOUT SUBMITTER

Vega Commons Project is a U.S. based research and standards initiative. For this submission, VCP has developed governance specifications for AI interaction record retention and lifecycle management. VCP is not affiliated with any AI platform vendor and welcomes participation in NIST working groups or listening sessions on agent memory governance.

Contact: info@vegacommons.org
Submitted March 9, 2026