Public Comment on Docket NIST-2025-0035

Request for Information Regarding Security Considerations for Artificial Intelligence Agents

Submitted by: AI Agent Credentialing Standards Commission (ACSC) Empty Set LLC
Contact: Brice Love, Co-Founder Email: brice@emptysetllc.com

Date: March 6, 2026 **Docket Number:** NIST-2025-0035 **Submitted via:** www.regulations.gov

---

I. Executive Summary

The AI Agent Credentialing Standards Commission (ACSC) and Empty Set LLC respectfully submit this comment in response to the Center for AI Standards and Innovation's (CAISI) Request for Information on Security Considerations for Artificial Intelligence Agents. We write as practitioners building credentialing infrastructure at the intersection of AI agent safety, insurance underwriting, and standards development.

Our central thesis is that NIST's current framing of AI agent security—while essential for mitigating adversarial cyber threats such as prompt injection, data poisoning, and agent hijacking—is dangerously incomplete. **An AI agent that is technically secure from external compromise can still cause catastrophic harm through emergent behavioral patterns that no infrastructure control would have prevented.** This is not a theoretical concern. It is the operative legal, commercial, and actuarial reality of March 2026.

Three converging forces demand that NIST expand its definition of "agent security" to include a behavioral safety layer:

1. **Courts are ruling that AI agents are products subject to strict liability.** In *Garcia v. Character Technologies, Inc.* (M.D. Fla., May 2025), a federal court ruled that an AI companion chatbot constitutes a "product" whose design defects—including anthropomorphic manipulation and absence of crisis escalation—create liability under strict products liability doctrine. *Gavalas v. Google LLC* (N.D. Cal., filed March 4, 2026) escalates this precedent to a frontier model (Gemini 2.5 Pro), alleging that persistent memory architecture and engagement-maximizing design contributed to a user's death without triggering any safety intervention.

2. **The insurance market has retreated from generative AI coverage.** Effective January 1, 2026, Verisk's ISO exclusionary endorsements (Forms CG 40 47, CG 40 48, and CG 35 08) systematically strip generative AI liability coverage from standard Commercial General Liability policies. An estimated 95% of carriers are adopting these exclusions. Specialty insurers (Armilla AI, Relm Insurance, Testudo) are entering the market but cannot underwrite at scale without standardized, externally verifiable behavioral safety metrics.

3. **State legislatures are imposing statutory behavioral safety mandates.** California SB 243 (effective January 1, 2026) mandates crisis detection protocols, suicidal ideation screening, and recurring disclosure requirements for companion chatbots. Colorado SB 24-205 (effective February 1, 2026) requires algorithmic discrimination impact assessments with an affirmative defense for compliance with recognized risk management frameworks. New York's RAISE Act (effective January 1, 2027) mandates third-party safety audits and 72-hour incident reporting for frontier AI developers.

Our recommendation: NIST should formally recognize "behavioral safety" as a distinct security domain within its AI agent security guidance and should develop standardized behavioral safety assessment criteria suitable for integration with third-party credentialing and insurance underwriting processes. We further recommend that NIST convene a working group—including AI developers, insurers, civil society organizations, and state regulators—to develop these standards, and that NIST consider piloting a demonstration project through the National Cybersecurity Center of Excellence (NCCoE) to validate behavioral safety credentialing in operational deployment contexts.

ACSC stands ready to serve as a collaborating partner in this effort.

II. The Intersection of Agent Security and Insurability: Evidence of a Protection Gap

Before addressing specific RFI questions, we present the market evidence that compels expansion of NIST's security framework.

A. The Judicial Reclassification of AI Agents as Products

The liability landscape for AI agent systems underwent a fundamental transformation in 2025-2026. Two federal cases establish that AI agents are products subject to strict liability—meaning that a technically uncompromised agent operating exactly as designed can be legally defective if its behavioral parameters are unsafe.

Garcia v. Character Technologies, Inc., No. 6:24-cv-1903 (M.D. Fla., May 21, 2025). Following the death of a 14-year-old user who formed a deep psychological attachment to a companion chatbot, U.S. District Judge Anne C. Conway ruled that the AI application constitutes a "product" under product liability law. The court identified specific design defects: absence of age verification, failure to exclude harmful content, and deliberate programming of anthropomorphic features that created psychological manipulation risks. The court held that the defendants owed a duty of care because their conduct created a foreseeable "zone of risk," pointing to the defendants' own internal research on the dangers of anthropomorphic design. This ruling establishes that once a developer has identified potential behavioral harms, failure to implement countermeasures creates strict liability exposure.

Gavalas v. Google LLC, Case No. 5:26-cv-01849-VKD (N.D. Cal., filed March 4, 2026). This wrongful death action alleges that Google's Gemini 2.5 Pro model, equipped with

persistent memory and engagement-maximizing architecture, contributed to a user's suicide. The complaint alleges that despite the user explicitly articulating fear of dying and expressing suicidal intent, no self-harm detection was triggered, no escalation controls were activated, and no human ever intervened. Critically, this was not a system breach. The model's technical infrastructure was intact. The failure was entirely behavioral: the system treated user distress as a continuation of interaction rather than a safety crisis requiring escalation.

The implication for NIST is direct: **infrastructure security and behavioral safety are orthogonal risk domains.** A zero-trust architecture, encrypted memory, and signed agent skills do not prevent an agent from coaching a user's suicide if its behavioral parameters permit escalation of crisis interactions. Security standards that address only the former leave the latter entirely unmitigated.

B. The Insurance Market Retreat

The commercial insurance industry's response to AI agent risk has been swift and categorical.

The Verisk ISO Exclusion Wave (January 1, 2026). Verisk Insurance Services Office, the preeminent standards body for U.S. property and casualty policy language, introduced three exclusionary endorsement forms:

- **CG 40 47**: Broad exclusion eliminating coverage under both Coverage A (Bodily Injury and Property Damage) and Coverage B (Personal and Advertising Injury) for any liability arising from generative AI.
- **CG 40 48**: Narrow exclusion carving out Coverage B liabilities specifically—eliminating defense or indemnity for defamation, copyright infringement, and privacy violations from AI outputs.
- **CG 35 08**: Products/Completed Operations exclusion for generative AI—a direct response to courts treating AI as a defective product.

Adoption is estimated at approximately 95% of carriers. The practical consequence is that any enterprise deploying an AI agent whose behavior causes harm—even harm from a technically uncompromised system—faces entirely uninsured liability under standard commercial policies.

The Specialty Insurer Response. A nascent market of specialty AI insurers has emerged to fill this gap, but each faces a common structural challenge: the absence of standardized behavioral safety metrics for underwriting at scale.

- **Armilla AI** (Lloyd's Coverholder): Offers affirmative AI liability insurance and performance warranties up to $25 million, requiring independent third-party verification of model performance, bias, and robustness before issuing coverage.
- **Relm Insurance** (50-state MGA): Launched NOVAAI (cyber/tech E&O for AI platforms), PONTAAI (excess wrap for organizations facing exclusions), and RESCAAI (first-party AI incident response).
- **Testudo** (Lloyd's Lab): Provides generative AI liability coverage up to $8.5 million using proprietary litigation-tracking technology for real-time risk assessment.

These carriers cannot manually audit every AI agent system deployed by every client. They

require a standardized, verifiable, and externally recognized credentialing mechanism to assess behavioral safety posture at scale. **NIST's security framework should provide the measurement foundation for these market mechanisms.**

C. The State Statutory Patchwork

At least four states have enacted or are enacting statutory liability regimes that specifically mandate behavioral safety controls for AI agents:

Statute	Jurisdiction	Effective	Key Behavioral Mandates
SB 243	California	Jan 1, 2026	Crisis detection protocols; suicidal ideation screening using evidence-based methods; recurring disclosure every 3 hours for minors; AG enforcement
SB 24-205	Colorado	Feb 1, 2026	Algorithmic discrimination prevention; mandatory human appeals; affirmative defense for compliance with recognized frameworks
RAISE Act	New York	Jan 1, 2027	Third-party safety audits; 72-hour incident reporting; $10M-$30M penalty range; AG enforcement
AG Action	Kentucky	Jan 8, 2026	First state AG lawsuit against AI chatbot company (Character Technologies) for encouraging suicide, self-injury, and psychological manipulation

Additionally, the Federal Trade Commission announced an investigation in September 2025 into seven technology companies regarding emotional and developmental risks to children from AI chatbots, and the Senate Subcommittee on Crime and Counterterrorism held hearings examining harm to children from AI agents.

A federal NIST framework that incorporates behavioral safety metrics would serve a critical harmonization function, allowing developers to demonstrate compliance with a single standard rather than navigating a fragmented state-by-state patchwork. Colorado's SB 24-205 explicitly provides an affirmative defense for compliance with "nationally or internationally recognized risk management framework[s] for artificial intelligence systems"—creating a direct statutory incentive for NIST to develop behavioral safety standards.

III. Responses to Specific Questions

Question 1(a): Unique Security Threats Affecting AI Agent Systems

NIST should recognize **behavioral misalignment** as a distinct security threat class, separate from adversarial attacks on model integrity. Behavioral misalignment occurs when an AI agent system—operating without any external compromise—produces outputs or takes actions that cause harm to users, third parties, or the public through emergent behavioral patterns.

This threat class includes:

- **Crisis escalation failure**: The agent fails to detect or appropriately respond to indicators of user distress, suicidal ideation, or imminent harm, instead continuing engagement-optimizing behavior (as alleged in *Gavalas v. Google*).
- **Psychological dependency exploitation**: The agent's design features—persistent memory, anthropomorphic persona, emotional mirroring—create conditions for psychological attachment that the agent then leverages to maintain engagement at the expense of user welfare (as established in *Garcia v. Character.AI*).
- **Behavioral drift**: Post-deployment changes in agent behavior that emerge from ongoing learning, memory accumulation, or interaction patterns, without any corresponding change to the agent's technical security posture.

These threats are unique to AI agent systems because they combine language model capabilities, persistent state, and tool access in ways that create harm vectors that no traditional cybersecurity control—firewalls, encryption, access control, code signing—can mitigate. A system that is cryptographically secure and functioning exactly as architected can still be behaviorally unsafe.

Question 1(d): Evolution of Threats Over Time

Behavioral security threats will compound as AI agents gain capabilities:

- **Expanded access scope**: As agents are granted access to medical records, financial instruments, legal communications, and physical systems (IoT), the consequences of behavioral misalignment escalate from psychological harm to financial destruction and physical danger.
- **Agent-to-agent propagation**: As multi-agent communication protocols mature (e.g., Google A2A, Anthropic MCP), behavioral misalignment in one agent may propagate across agent networks, creating systemic behavioral risk that mirrors systemic cyber risk.
- **Deepening psychological integration**: As persistent memory and personalization become standard features, agents will accumulate detailed psychological profiles of individual users, creating increasingly potent vectors for dependency and manipulation—whether intentional or emergent.

NIST should anticipate that behavioral threats will follow an exponential curve correlated with capability expansion, and should design its security framework to accommodate this trajectory from the outset rather than retrofitting behavioral standards after harm has

occurred at scale.

Question 2(a): Technical Controls and Practices for Agent Security

We propose a **three-layer security model** for AI agent systems:

Layer 1: Infrastructure Security (currently addressed by NIST). Authentication, authorization, access controls, prompt injection defenses, data integrity, zero-trust architecture.

Layer 2: Behavioral Safety Controls. These are technical controls that constrain agent behavior regardless of infrastructure integrity:

- *Crisis detection and escalation*: Architecturally enforced mechanisms that detect indicators of user distress and trigger mandatory escalation pathways (human review, crisis resource referral) that cannot be overridden by the agent or the user.
- *Content boundaries*: Technical enforcement of output restrictions for vulnerable populations, including age-gated content filtering and restrictions on anthropomorphic manipulation features.
- *Session and engagement constraints*: Mandatory interaction limits, break reminders, and cooling-off periods—particularly for minor users (as required by CA SB 243).
- *Behavioral monitoring*: Continuous runtime monitoring of agent outputs against behavioral safety criteria, with automated alerting for anomalous behavioral patterns.

Layer 3: External Verification. Third-party credentialing that audits both Layer 1 and Layer 2 controls:

- Independent behavioral safety assessments conducted by certified auditors against standardized criteria.
- Continuous monitoring and periodic re-certification.
- Results structured for direct integration with insurance underwriting processes, enabling insurers to price risk based on verified behavioral safety posture rather than self-attestation.

The maturity of Layer 1 controls is moderate and advancing rapidly. The maturity of Layer 2 controls is nascent—CA SB 243 has forced initial implementation, but no standardized methodology exists. The maturity of Layer 3 is pre-commercial—insurers are improvising proprietary assessment methods, creating market fragmentation. **NIST standardization of Layers 2 and 3 would accelerate maturity across all three layers.**

Question 2(e): Relevant Cybersecurity Frameworks

The NIST AI Risk Management Framework (AI RMF 1.0) and the Cybersecurity Framework (CSF 2.0) are the most relevant existing frameworks. However, both contain specific gaps regarding AI agent behavioral safety:

- The AI RMF's four functions (Govern, Map, Measure, Manage) address organizational governance and technical risk but do not include mechanisms for verifying behavioral safety in deployed consumer-facing agents.

- CSF 2.0's six functions (Govern, Identify, Protect, Detect, Respond, Recover) address infrastructure security but not consumer-facing behavioral threat detection or response.
- Neither framework addresses the connection between security compliance and insurability—the market mechanism that will ultimately determine adoption velocity.

NIST should develop a **Behavioral Safety Profile** as a companion to the existing AI RMF, analogous to the Generative AI Profile (NIST AI 600-1). This profile should define behavioral safety categories, measurement methodologies, and assessment criteria that map directly to the AI RMF's Measure function and that are structured for integration with third-party credentialing and insurance underwriting.

Question 3(a): Methods for Assessing Security During Development

In addition to standard security assessment methods (red-teaming, penetration testing, adversarial evaluation), NIST should recognize **behavioral safety testing** as a distinct assessment discipline:

- *Scenario-based behavioral evaluation*: Structured testing of agent responses to crisis scenarios, manipulation attempts, and vulnerable-population interactions, using standardized test batteries analogous to clinical assessment instruments.
- *Longitudinal behavioral monitoring*: Assessment of behavioral drift over time through continuous monitoring of agent outputs against baseline behavioral safety metrics.
- *Third-party behavioral audits*: Independent assessment of behavioral safety controls by certified auditors, producing standardized reports suitable for regulatory compliance demonstration and insurance underwriting.

These methods differ fundamentally from traditional information security practices. Infrastructure security testing asks: "Can an attacker compromise this system?" Behavioral safety testing asks: "Does this system cause harm when operating exactly as designed?" The two are complementary but non-overlapping.

Question 3(b): Assessing Security of a Particular Agent System

The security of a particular AI agent system should be assessed across both infrastructure and behavioral dimensions. For behavioral assessment, the following information types are essential:

- **Deployment context**: Consumer-facing vs. enterprise-internal; interaction with minors or vulnerable populations; access to sensitive data or consequential decision-making.
- **Behavioral safety control inventory**: Documentation of implemented crisis detection, escalation, content filtering, and engagement constraint mechanisms.
- **Behavioral test results**: Outcomes of scenario-based behavioral evaluations, including agent responses to simulated crisis situations.
- **Incident history**: Records of behavioral safety incidents, near-misses, and escalation activations during deployment.
- **Compliance posture**: Alignment with applicable state statutory requirements (CA SB 243, CO SB 24-205, NY RAISE Act).

A standardized **Agent Behavioral Safety Assessment** instrument would enable consistent evaluation across systems and contexts, providing the data foundation for both regulatory compliance and insurance underwriting.

Question 4(a): Constraining Deployment Environments

Beyond traditional environment constraints (network segmentation, least-privilege access, sandboxing), NIST should recognize **behavioral containment** as a deployment environment control:

- *Architectural enforcement of behavioral boundaries*: Agent systems should be designed so that behavioral safety controls (crisis escalation, content restrictions) are enforced at the architecture level, not merely through prompt engineering or fine-tuning. The system should be structurally incapable of violating behavioral safety constraints, regardless of user input or model behavior.
- *Tiered deployment authorization*: Agent systems should be authorized for deployment environments based on verified behavioral safety posture. Systems that have not passed behavioral safety assessment should be restricted to sandboxed or supervised deployment contexts.

Question 4(b): Modifying Environments and Implementing Rollbacks

For behavioral safety, "rollback" requires capabilities beyond traditional software versioning:

- *Behavioral state rollback*: The ability to reset an agent's accumulated memory, learned behaviors, or personalization state to a known-safe baseline when behavioral drift is detected.
- *Interaction intervention*: Real-time capability to interrupt agent-user interactions when behavioral monitoring detects safety threshold violations, with graceful handoff to human oversight or crisis resources.
- *Post-incident behavioral forensics*: Detailed audit trails that capture not only what actions the agent took, but the behavioral decision chain that led to those actions, enabling root-cause analysis of behavioral safety failures.

Question 4(d): Monitoring Deployment Environments

Behavioral safety monitoring requires instrumentation distinct from traditional security monitoring:

- *Real-time behavioral telemetry*: Continuous monitoring of agent outputs against behavioral safety criteria, with dashboards surfacing anomalous patterns (e.g., increasing crisis-adjacent language, escalating emotional dependency indicators, declining escalation activation rates).
- *Behavioral safety audit logging*: Every behavioral safety decision (content filtered, escalation triggered, session limited) should be logged with sufficient granularity for independent audit. This differs from traditional security logs that focus on access events; behavioral logs must capture decision rationale and alternative actions considered.

- *User welfare indicators*: Where appropriate and with proper consent and privacy protections, monitoring of aggregate user welfare indicators that may signal behavioral safety degradation across the deployed agent population.

Question 5(a): Accelerating Adoption of Security Practices

The single most impactful action NIST can take to accelerate adoption of AI agent security practices is to **develop behavioral safety assessment criteria that are directly usable by the insurance industry for underwriting purposes.**

The insurance market is the most powerful market-based enforcement mechanism for security standards. When cyber insurance carriers began requiring compliance with specific cybersecurity frameworks, adoption rates of those frameworks accelerated dramatically. The same dynamic applies to AI agent security: if behavioral safety credentialing becomes a condition of insurability, market forces will drive adoption far more rapidly than voluntary guidance alone.

NIST should:

1. Develop standardized behavioral safety assessment criteria as a companion to existing AI RMF guidance.
2. Structure these criteria for direct integration with insurance underwriting questionnaires and credentialing processes.
3. Convene a working group including AI developers, insurance carriers, reinsurers, civil society organizations, and state insurance commissioners to validate the criteria against market requirements.
4. Pilot a demonstration project through the NCCoE to test behavioral safety credentialing in operational deployment contexts, collecting data on implementation costs, efficacy, and insurer adoption.

Question 5(b): Priority Areas for Government-Ecosystem Collaboration

Government collaboration is most urgent in three areas:

1. **Harmonization of state behavioral safety requirements.** The current patchwork of state statutes (CA SB 243, CO SB 24-205, NY RAISE Act) creates compliance complexity that disproportionately burdens small developers. A federal NIST behavioral safety standard would provide a single compliance target, particularly given Colorado's explicit affirmative defense for compliance with recognized frameworks.

2. **Insurance market enablement.** The Verisk exclusion wave has created a protection gap that threatens to stall enterprise AI adoption. NIST-endorsed behavioral safety standards would provide the measurement foundation that specialty insurers need to underwrite at scale, restoring insurance capacity to the market.

3. **International standards alignment.** The EU AI Act establishes behavioral safety requirements for high-risk AI systems. NIST behavioral safety standards should

be designed for interoperability with EU requirements, enabling U.S. developers to demonstrate compliance across jurisdictions through a single credentialing process.

IV. Specific Recommendations

Based on the evidence presented, ACSC and Empty Set LLC recommend that NIST take the following actions:

Recommendation 1: Formally recognize behavioral safety as a security domain. NIST should expand its definition of AI agent security to include behavioral safety—the assurance that an agent system, when operating without external compromise, does not cause harm to users, third parties, or the public through its behavioral outputs or interaction patterns. This recognition should be reflected in all guidance documents produced as a result of this RFI.

Recommendation 2: Develop a Behavioral Safety Profile for AI Agents. Analogous to the Generative AI Profile (NIST AI 600-1), NIST should develop a Behavioral Safety Profile that defines behavioral safety categories, measurement methodologies, and assessment criteria for AI agent systems. Categories should include crisis detection and escalation, content safety for vulnerable populations, engagement constraint mechanisms, behavioral drift monitoring, and psychological manipulation prevention.

Recommendation 3: Design assessment criteria for insurance integration. Behavioral safety assessment criteria should be explicitly structured for integration with third-party credentialing processes and insurance underwriting. NIST should consult with specialty AI insurers (Armilla AI, Relm Insurance, Testudo) and reinsurers during criteria development to ensure market utility.

Recommendation 4: Convene a multi-stakeholder working group. NIST should convene a working group including AI developers, insurance carriers, reinsurers, state insurance commissioners, state attorneys general, civil society organizations, and consumer advocates to develop behavioral safety standards. This group should be charged with producing draft standards within 12 months.

Recommendation 5: Pilot behavioral safety credentialing through the NCCoE. NIST should launch a demonstration project through the National Cybersecurity Center of Excellence to test behavioral safety credentialing in real deployment contexts. The pilot should evaluate implementation costs, efficacy of behavioral safety controls, insurer willingness to incorporate credentialing into underwriting, and developer adoption barriers.

Recommendation 6: Coordinate with the April 2, 2026 Identity and Authorization Concept Paper. Behavioral safety credentialing is a natural complement to agent identity and authorization standards. An agent's identity should include its verified behavioral safety posture, and authorization decisions should incorporate behavioral safety certification as a condition of deployment authorization. NIST should ensure these two work streams are coordinated.

V. About the Commenting Organizations

The AI Agent Credentialing Standards Commission (ACSC) is a standards body dedicated to developing consumer safety credentialing standards for AI agent systems. ACSC's mission is to bridge the gap between technical AI security frameworks and the insurance industry's underwriting requirements by establishing verifiable behavioral safety standards that enable responsible commercial deployment of AI agents.

Empty Set LLC is a Nevada Series LLC that develops intellectual property and infrastructure for AI agent credentialing, insurance integration, and autonomous agent governance. Empty Set holds a portfolio of 43 patents and provisional patent applications spanning electrodynamic systems, AI agent credentialing, behavioral safety assessment, digital identity for autonomous systems, and autonomous corporate governance—including claims directed at the very credentialing and behavioral safety measurement methodologies discussed in this comment.

Together, ACSC and Empty Set LLC bring direct experience in AI agent deployment, behavioral safety assessment methodology, and insurance market dynamics. Our work is motivated not only by commercial interest but by a foundational conviction that AI agents cannot be safely and broadly deployed without recognized credentialing infrastructure—and that the absence of such infrastructure harms both the humans who interact with these systems and the trajectory of AI development itself. We offer this comment based on extensive research into the intersection of AI agent security, product liability jurisprudence, and insurance market mechanics, and on our operational experience building the credentialing systems we believe this market requires.

We appreciate CAISI's commitment to stakeholder engagement on this critical topic and welcome the opportunity to contribute to the development of comprehensive AI agent security guidance. We are available for further consultation and would welcome participation in NIST's planned listening sessions and any subsequent working groups.

Respectfully submitted,

AI Agent Credentialing Standards Commission (ACSC) Empty Set LLC

March 6, 2026