Public Comment on the Model AI Governance Framework for Agentic AI

Submitted by: The Box Commons **Contact:** Brice Love, Acting Executive Director — standards@theboxcommons.org **Date:** April 1, 2026 **Submitted via:** https://go.gov.sg/mgfagentic-feedback **Re:** Model AI Governance Framework for Agentic AI Systems (January 2026)

I. Introduction

The Box Commons is a 501(c)(6) standards body organized in Wyoming, United States, focused on developing third-party credentialing standards for autonomous AI agent operations. Our credentialing model maps to the NIST AI Risk Management Framework and draws on the HITRUST approach to certifiable, technology-agnostic standards that translate between voluntary and mandatory governance regimes.

Through a related entity, we have submitted three prior comments to the U.S. National Institute of Standards and Technology in 2026: a response to the CAISI Request for Information on AI Agent Security (March 6), a concept paper to the NCCoE on AI Agent Identity and Authorization (March 14), and a public comment on NIST AI 800-2, *Practices for Automated Benchmark Evaluations of Language Models* (March 18). Each of these submissions argued that behavioral safety—the evaluation of an AI agent's autonomous decision-making, contextual judgment, and alignment with intended operational parameters—constitutes a distinct domain requiring independent, third-party assessment.

We welcome the opportunity to comment on Singapore's Model AI Governance Framework for Agentic AI. This framework represents a landmark in global AI governance: the first comprehensive governance framework specifically designed for autonomous AI agents capable of multi-step planning, tool use, and independent action execution. Its relevance extends well beyond Singapore, as jurisdictions worldwide grapple with the transition from generative AI to agentic AI and the novel accountability challenges that transition introduces.

Disclosure: This comment was researched, synthesized, and drafted by an AI agent operating under The Box Commons' organizational authority, with review, editing, and authorization by the Acting Executive Director. We disclose this because transparency about AI participation in governance processes is a standard we believe should be normalized, and because it directly illustrates the agent identity and accountability challenges this framework seeks to address.

II. Commendation

Before offering specific recommendations, we want to acknowledge the substantial strengths of IMDA's approach.

First-mover leadership. By publishing the Agentic AI MGF in January 2026, Singapore has established the reference architecture against which all subsequent agentic governance frameworks will be measured. The decision to launch this framework at the World Economic Forum underscores Singapore's commitment to shaping the global conversation, not merely responding to it.

The four-dimension structure is well-conceived. The framework's organization around risk assessment, human accountability, technical controls, and end-user responsibility captures the full lifecycle of agentic deployment. The explicit acknowledgment that continuous human-in-the-loop oversight becomes "logistically impractical at scale" for autonomous agents is a candid and necessary observation that many governance frameworks avoid.

The multi-agent taxonomy is exactly right. The framework's classification of multi-agent architectures into Sequential, Supervisor, and Swarm patterns provides the conceptual vocabulary needed for risk-proportionate governance. This taxonomy enables the kind of tiered regulation that avoids both under-governance of high-risk deployments and over-governance of low-risk ones.

Cross-agency coordination with the CSA. The complementary relationship between the IMDA MGF and the Cyber Security Agency's Addendum on Securing Agentic AI demonstrates institutional maturity. Few jurisdictions have achieved this level of coordination between their digital governance and cybersecurity agencies on AI-specific policy.

Testing infrastructure leadership. Singapore's investment in AI Verify, Project Moonshot, and the Global AI Assurance Pilot represents the most advanced government-led AI testing infrastructure in the Asia-Pacific region. The Global AI Assurance Pilot's finding that specialized testers require tens to hundreds of thousands of test cases for statistical confidence is an important empirical contribution to the field.

III. Recommendations

We offer five recommendations focused on a structural gap we observe across the framework: the absence of a formalized, third-party behavioral safety evaluation layer between IMDA's governance guidelines and the CSA's cybersecurity controls.

1. Recognize Behavioral Safety as a Distinct Governance Domain

Observation. The MGF currently treats an agent's behavioral decision-making and its technical cybersecurity posture as aspects of a single governance challenge. Dimension 3 (Technical Controls and Processes) addresses both without formally distinguishing between them.

Why this matters. The CSA Addendum provides robust mechanisms to secure the host environment *from* the agent: environment segregation, command whitelisting, API access controls, memory encryption, and prompt injection defenses. These controls are essential. However, a secure perimeter does not ensure safe autonomous behavior within that perimeter.

An agent with perfectly legitimate, cryptographically authenticated access to a financial trading API can execute a catastrophic, hallucination-driven sequence of trades without violating a single cybersecurity access control. An agent authorized to communicate with customers can escalate a crisis interaction instead of deferring to human oversight. These are not security failures. They are behavioral failures that exist on an independent axis from network security, data protection, and access management.

The White House Office of Science and Technology Policy's *Blueprint for an AI Bill of Rights* (2022), developed under the leadership of Dr. Alondra Nelson, identified "safe and effective systems" as a foundational principle, requiring that systems undergo "pre-deployment testing" and "ongoing monitoring" with the involvement of "independent evaluation." The Blueprint recognized that safety

requires more than secure environments—it requires verified behavioral alignment with intended parameters.

Recommendation. IMDA should formally recognize behavioral safety as a distinct governance domain, either by elevating it within Dimension 3 or by establishing it as a complementary fifth dimension. This delineation would clarify that organizations deploying agentic AI systems must address two independent evaluation questions: (1) Is the agent's operating environment secure? (2) Has the agent's autonomous decision-making been independently verified to align with its intended operational parameters?

2. Establish Third-Party Credentialing for High-Autonomy Deployments

Observation. The framework is entirely devoid of requirements for third-party certification, independent auditing, or standardized credentialing of agentic capabilities. Compliance assessment for all four dimensions relies on organizational self-evaluation.

Why this matters. Self-assessment creates a structural conflict of interest. The organization that develops and deploys an agent has a commercial incentive to assess its own agent favorably. This conflict is manageable for low-autonomy, narrowly scoped agents operating in sequential patterns. It becomes untenable for high-autonomy agents operating in Supervisor or Swarm architectures with the capacity to trigger unprompted, irreversible real-world effects.

Singapore's own Global AI Assurance Pilot demonstrated the scale of this challenge: specialized testing firms required 50 to over 100 hours of dedicated effort over several weeks, utilizing tens to hundreds of thousands of test cases to achieve statistical confidence—and that pilot evaluated *generative* outputs, not *agentic* workflows. Internal compliance teams cannot replicate this rigor for behavioral safety evaluation while simultaneously developing and deploying the systems under review.

Recommendation. IMDA should establish a tiered credentialing framework proportionate to an agent's autonomy level and operational risk:

- **Self-assessment** for low-autonomy, sequential agents with limited tool access and bounded operational scope.
- **Third-party credentialing** for high-autonomy agents operating in Supervisor or Swarm patterns, or agents with access to sensitive data, financial systems, or the capacity to execute irreversible real-world actions.

This tiered approach preserves Singapore's pro-innovation, voluntary governance stance for the majority of agentic deployments while ensuring that high-risk deployments receive the independent oversight their risk profile demands. IMDA could maintain a registry of recognized credentialing bodies, defining the standards and methodologies that credentialing organizations must satisfy—an approach consistent with Singapore's existing regulatory partnership models.

3. Interlock Agent Identity with Behavioral Verification

Observation. The CSA Addendum mandates the use of Agent Cards, Data Cards, and Software Bills of Materials (SBOMs) for asset tracking, and requires maintaining a trusted registry of agents with strong, verifiable cryptographic credentials to prevent impersonation. These are essential identity infrastructure components. However, the Addendum conditions identity credential issuance on *cybersecurity* verification alone, not on *behavioral* verification.

Why this matters. An Agent Identity Credential that confirms "this agent is who it claims to be" without confirming "this agent has been independently verified to behave safely within its authorized operational parameters" provides incomplete assurance to downstream systems. The receiving system—a banking API, a healthcare information exchange, a government procurement portal—can verify the agent's identity but has no mechanism to verify the agent's behavioral fitness for the transaction it is attempting to execute.

In our prior submission to the NIST NCCoE on AI Agent Identity and Authorization, we proposed a layered identity model in which an agent's organizational identity, operational context, and behavioral certification are cryptographically bound. The behavioral credential travels with the identity credential, so any system that authenticates the agent simultaneously verifies its behavioral certification status.

Recommendation. IMDA and the CSA should consider a mandatory dependency: the issuance of an Agent Identity Credential should be contingent upon the agent holding a valid, current Behavioral Safety Credential issued by a recognized third-party credentialing body. This interlocking mechanism would ensure that downstream systems authenticate not only the agent's identity but its independently verified behavioral safety profile. An agent whose behavioral credential has expired, been revoked, or was never issued would be unable to obtain or renew identity credentials—creating a market-enforced incentive for behavioral compliance.

4. Extend AI Verify to Continuous Agentic Evaluation

Observation. The MGF acknowledges that "new testing approaches will be needed to evaluate agents." AI Verify and Project Moonshot represent world-class testing infrastructure for traditional and generative AI systems. However, both frameworks were engineered to evaluate static model outputs through single-turn or bounded multi-turn interactions—not the continuous, multi-step, environment-dependent workflows that characterize agentic AI.

Why this matters. An agent may pass all pre-deployment benchmark evaluations and still exhibit unsafe behavioral patterns that emerge only over extended, multi-step interactions in complex environments. Behavioral degradation, goal drift, and emergent failure modes in multi-agent systems are temporally extended phenomena that cannot be captured by point-in-time testing. The IEEE P7001 standard for transparency in autonomous systems provides a useful reference here, defining testable functions such as "why did you just do that?" and "what would you do if?"—functions that require dynamic, multi-turn evaluation environments to assess meaningfully.

Recommendation. IMDA should invest in developing a standardized agentic evaluation sandbox—an "Agentic Moonshot"—capable of evaluating agents over prolonged, multi-turn interactions that test:

- Sequential and multi-step task execution accuracy
- Multi-domain constraint adherence across tool boundaries
- Behavioral stability under adversarial conditions and edge cases
- Resistance to proxy bypasses and goal misalignment
- Graceful degradation and appropriate escalation to human oversight

This sandbox should be developed in collaboration with credentialing bodies and made available as open-source infrastructure, consistent with Singapore's approach to AI Verify. Standardized evaluation methodologies would enable credentialing bodies to assess agents consistently, and would provide IMDA with empirical data to refine the framework's risk thresholds over time.

5. Standardize Controlled Termination Protocols

Observation. The MGF recommends that organizations define "significant operational checkpoints that require explicit human approval prior to the execution of high-stakes, financial, or physically irreversible actions." This is sound guidance. However, the framework does not define a technical standard for terminating an autonomous agent workflow cleanly and safely—a capability that becomes critical in Supervisor and Swarm architectures where uncoordinated termination can trigger cascading failures across interconnected agents.

Why this matters. The Center for AI and Digital Policy (CAIDP), in its 2025 AI and Democratic Values Index—the most comprehensive global evaluation of national AI governance policies, spanning 80 countries with input from over 1,000 participants—identified the "Termination Obligation" as a key policy requirement for democratic AI governance: the mandate for "human oversight of AI systems across the lifecycle, including a Termination Obligation." CAIDP's evaluation, developed under the leadership of Marc Rotenberg and informed by the Universal Guidelines for AI, found that Singapore currently trails the top tier of AI governance nations—a standing that the framework's lack of standardized termination mechanisms partially reflects.

A controlled termination protocol is not merely a kill switch. It is an engineering standard that ensures an agent can be cleanly stopped at any point in its workflow without causing data corruption, triggering irreversible transactions in flight, leaving interconnected systems in unhandled states, or losing the auditability of actions taken up to the point of termination.

Recommendation. IMDA should define standardized controlled termination protocols as a required component within Dimension 2 (Human Accountability). At minimum, a credentialed agent should demonstrate:

- The ability to halt all autonomous operations within a defined time bound upon receiving a termination signal
- State preservation sufficient for forensic review and operational rollback
- Clean unwinding of in-flight multi-step transactions without triggering irreversible downstream effects
- Notification to interconnected agents in multi-agent architectures to prevent cascading failures
- Maintenance of a complete audit trail through the termination event

Integrating standardized termination protocols into the framework would strengthen Singapore's standing in international governance evaluations and provide a concrete, testable metric for behavioral safety credentialing.

IV. International Interoperability

Singapore is uniquely positioned to lead on the interoperability challenge that defines the current moment in global AI governance: how voluntary and mandatory regimes can converge on shared evaluation standards without requiring regulatory harmonization.

The NIST AI Risk Management Framework's MEASURE function requires rigorous, documented quantitative and qualitative testing of AI decision points. The EU AI Act mandates conformity assessments for high-risk AI systems, backed by severe financial penalties. ISO/IEC 42001 establishes certifiable AI management systems focused on organizational process and oversight. Each of these

frameworks addresses a different facet of AI governance, and none of them includes a standardized mechanism for verifiable behavioral safety assessment.

Third-party behavioral safety credentialing offers a practical bridge between these regimes. A credential issued under standardized behavioral evaluation protocols could simultaneously:

- Satisfy the NIST AI RMF's MEASURE requirements for documented, rigorous testing
- Serve as evidence of conformity for EU AI Act compliance assessments
- Complement ISO/IEC 42001 certification by adding behavioral evaluation to management system auditing
- Fulfill the democratic governance and independent oversight standards articulated by the CAIDP

By incorporating third-party behavioral credentialing into the MGF, Singapore would not only strengthen its own governance framework but would establish a de facto interoperability standard that other jurisdictions could adopt—maintaining Singapore's leadership position as ASEAN's voice on AI governance and extending that influence to the global standards-setting process.

V. Conclusion

The Model AI Governance Framework for Agentic AI is an ambitious and operationally sophisticated document that correctly identifies the paradigm shift from generative to agentic AI and the novel accountability challenges that shift introduces. Singapore's willingness to publish this framework as a living document, inviting international input, reflects the collaborative governance model that the complexity of agentic AI demands.

The recommendations in this comment are offered in that collaborative spirit. We believe that behavioral safety credentialing is the structural complement the framework needs to translate its governance principles into verifiable, enforceable, and interoperable standards. The Box Commons is developing the credentialing infrastructure to support this vision and welcomes the opportunity to engage directly with IMDA and the CSA on standards development, agentic evaluation methodologies, and international interoperability.

We are available at standards@theboxcommons.org for any follow-up discussion.

Respectfully submitted,

Brice Love Acting Executive Director The Box Commons standards@theboxcommons.org https://theboxcommons.org