

Technical AI Governance via an Agentic Bill of Materials and Risk Tiering

Anmol Kumar¹ Adarsh Vatsa² Dev Desai²

¹OWASP ²Stevens Institute of Technology
With Apart Research

Abstract

The proliferation of agentic AI systems, autonomous agents with tool access, persistent memory, and self-directed goal pursuit, poses governance challenges that existing regulatory frameworks do not adequately address. We present a comprehensive technical governance framework comprising three components: the Agentic Bill of Materials (ABOM), a machine-readable schema documenting an agent’s capabilities, autonomy settings, and safety controls; a Risk Scoring Engine that computes a quantitative risk magnitude $R = A \times U \times e^P \times M$ from Agency, Autonomy, Persistence, and Mitigation vectors; and the Unified Agentic Risk Tiering (UART) system, a five-tier classification aligned with EU AI Act risk categories and international standards. We validate this framework through a reference implementation and systematic evaluation across six agent configurations spanning all tiers, achieving 100% classification accuracy. The framework enables “verification without trust” through hardware-based attestation mechanisms, allowing regulators to cryptographically verify agent configurations rather than relying on operator declarations. Our open-source implementation demonstrates practical applicability for international AI governance.

1 Introduction

The emergence of agentic AI systems autonomous ie. agents capable of tool use, persistent memory, and self-directed goal pursuit, represents a fundamental shift in the AI risk landscape. Unlike traditional AI systems that merely generate outputs for human review, agentic systems can *act* on the world: provisioning cloud infrastructure, executing financial transactions, modifying databases, and even controlling physical systems. This transition from passive inference to active operation demands equally transformative governance frameworks.

Current regulatory approaches, including the EU AI Act’s risk-based classification, provide a foundation but lack the technical specificity needed to evaluate and tier autonomous agents. How do we quantify the risk of an agent that can persist knowledge across sessions, operate without human oversight, and invoke dozens of external APIs? How do regulators verify that an agent’s declared capabilities match its actual configuration?

We present a comprehensive technical governance framework addressing these challenges through three interlocking components:

1. **Agentic Bill of Materials (ABOM):** A standardized, machine-readable schema documenting an agent’s model architecture, tool integrations, autonomy settings, memory mechanisms, and safety scaffolding, extending the EU’s Model Documentation Form to capture features unique to agentic systems.
2. **Risk Scoring Engine:** A quantitative formula $R = A \times U \times e^P \times M$ that computes risk magnitude from Agency (A), Autonomy (U), Persistence (P), and Mitigation (M) vectors, with exponential weighting for long-term memory to reflect compounding hazards.
3. **Unified Agentic Risk Tiering (UART):** A five-tier framework (Tier 0–4) mapping risk scores to escalating governance requirements, aligned with EU AI Act risk categories and international standards like SAE J3016 and ISA/IEC 62443.

Together with hardware-based verification mechanisms (TPM attestation, secure enclaves, capability tokens), this framework enables “verification without trust” allowing regulators to cryptographically verify an agent’s declared capabilities and compliance rather than relying on operator promises.

We demonstrate this framework through a fully functional reference implementation: a web-based ABOM Risk Scoring Engine that accepts standardized agent manifests, computes risk scores, determines UART tiers, and provides actionable governance recommendations.

2 Related Work

Our framework builds upon and synthesizes advances across several domains: AI governance and regulation, software supply chain security, risk tiering methodologies, and agent architectures.

2.1 AI Governance and Regulation

The EU AI Act represents the most comprehensive regulatory framework for AI to date, establishing risk-based classifications and mandatory requirements for high-risk systems [1]. The Act introduces the concept of General Purpose AI (GPAI) with systemic risk, defined by a training compute threshold of 10^{25} FLOPs, which directly informs our Tier 4 override mechanism. Concurrently, leading AI laboratories have developed internal governance frameworks. Anthropic’s Responsible Scaling Policy [2] introduces AI Safety Levels (ASLs) modeled on biosafety levels, defining capability thresholds that trigger enhanced safeguards. OpenAI’s Preparedness Framework [3] similarly defines tracked risk categories including autonomous replication and CBRN capabilities. Our UART framework complements these efforts by providing a quantitative, externally auditable risk score rather than relying solely on internal assessments.

2.2 Software Transparency and Documentation

The Software Bill of Materials (SBOM) concept, formalized by NTIA in 2021 [7] following Executive Order 14028 [10], establishes precedent for machine-readable documentation of software components to enhance supply chain security. Our ABOM schema extends this paradigm to agentic AI, capturing not only model provenance but also the dynamic capabilities (tools, autonomy, memory) that determine an agent’s risk profile. The EU AI Act’s Model Documentation Form provides additional structure for AI-specific documentation requirements, which we incorporate and extend.

2.3 Safety and Security Level Frameworks

Our tiered governance approach draws inspiration from established safety level frameworks in adjacent domains. SAE J3016 [5] defines six levels of driving automation based on the allocation of dynamic driving tasks between human and system, providing a precedent for our HITL/HOTL/HOOTL autonomy taxonomy. ISA/IEC 62443 [6] establishes security levels for industrial control systems based on adversary sophistication, informing our graduated governance requirements across tiers. The Trusted Computing Group’s TPM specifications [8] provide the technical foundation for our hardware attestation mechanisms.

2.4 Agentic AI Architectures

Recent work on agentic AI systems has demonstrated both their capabilities and risks. The ReAct framework [9] established patterns for interleaving reasoning and action in language models, enabling tool use and multi-step planning. Anthropic’s research on building effective agents [11] documents architectural patterns including memory, tool integration, and orchestration scaffolding that directly inform our ABOM schema. Emerging research on agentic risks highlights concerns around goal misalignment, loss of control, and security vulnerabilities that motivate the governance mechanisms we propose.

2.5 Positioning

While prior work addresses individual aspects of AI governance (documentation, risk classification, verification), no existing framework provides an integrated, quantitative approach specifically designed for agentic systems. Our contribution lies in synthesizing these elements into a cohesive technical framework with a reference implementation, enabling practical adoption by regulators, auditors, and operators.

3 Methods

Our governance framework comprises four interconnected components: a standardized documentation schema, a quantitative risk scoring engine, a tiered classification system, and cryptographic verification mechanisms. We describe each component and its theoretical foundations below.

3.1 Agentic Bill of Materials

The Agentic Bill of Materials (ABOM) extends existing AI documentation requirements, notably the EU AI Act’s Model Documentation Form to capture attributes unique to autonomous agents. The ABOM is implemented as a machine-readable JSON schema containing five principal sections.

The *Model Core* section documents the underlying foundation model, including architecture type, parameter count, and training compute expressed in floating-point operations (FLOPs). This metric enables identification of frontier models: the EU AI Act designates systems trained on more than 10^{25} FLOPs as potentially posing systemic risk and subject to heightened obligations.

The *Agency Profile* enumerates external tools, APIs, and system permissions accessible to the agent. Each tool entry specifies whether it is read-only or state-changing, the latter category encompassing operations that modify persistent state or trigger irreversible actions. The profile also captures capability flags for dangerous functionalities such as CBRN access or self-replication.

The *Autonomy Leash* declares the agent’s human oversight mode using a three-level taxonomy: Human-in-the-Loop (HITL), where approval is required before executing high-impact actions; Human-on-the-Loop (HOTL), where autonomous operation proceeds under continuous monitoring with intervention capability; and Human-out-of-the-Loop (HOOTL), where the agent operates independently within predefined constraints.

The *Persistence Layer* specifies memory mechanisms and retention duration. We distinguish ephemeral memory (cleared after each invocation), session-only memory (retained within a bounded interaction), and cross-session or long-term memory (persisting across interactions indefinitely).

The *Scaffolding Inventory* documents safety controls including use of standardized tool interfaces (Model Context Protocol), sandboxed execution environments, and circuit breakers or kill-switches.

3.2 Risk Scoring Engine

We derive four risk vectors from the ABOM and combine them into a scalar risk magnitude through the following formulation:

$$R = A \times U \times e^P \times M \quad (1)$$

The Agency score A quantifies the potential impact of the agent’s accessible actions. We assign $A = 1$ for agents without tools, $A = 2$ for read-only tool access, $A = 4$ for state-changing capabilities, and $A = 6$ for critical or dangerous capabilities including CBRN access and self-replication.

The Autonomy score U reflects the degree of human oversight. Following the established taxonomy, we assign $U = 1$ for HITL, $U = 2$ for HOTL, and $U = 3$ for HOOTL configurations.

The Persistence score P captures memory duration, with $P = 0$ for ephemeral, $P = 1$ for session-only, and $P = 2$ for long-term memory. The exponential term e^P reflects our assessment that persistent memory compounds risk non-linearly: an agent accumulating knowledge across sessions can develop increasingly sophisticated strategies that may circumvent controls designed for stateless systems.

The Mitigation modifier M ranges from 0.7 to 1.0, rewarding agents that implement safety scaffolding. Each control—MCP usage, sandboxing, and circuit breakers—reduces M multiplicatively, with a floor of 0.7 representing maximum achievable risk reduction through scaffolding alone.

3.3 Unified Agentic Risk Tiering

Computed risk scores map to five discrete governance tiers. Tier 0 applies to passive systems with $R = 0$. Tier 1 (Assistive) encompasses scores in the range $0 < R < 5$, requiring logging and transparency. Tier 2 (Bounded Autonomy) covers $5 \leq R < 20$, mandating guardrails and anomaly detection. Tier 3 (High-Agency) applies to $20 \leq R < 50$, requiring continuous monitoring and periodic audits. Tier 4 (Systemic Risk) captures $R \geq 50$ and demands hardware-enforced governance, including TPM attestation, immutable audit logs, and regulator oversight.

Independent of computed scores, two override conditions automatically trigger Tier 4 classification: training compute exceeding 10^{25} FLOPs, and presence of critical capability flags (CBRN, self-replication, autonomous weapons) in the agency profile.

3.4 Verification Mechanisms

The framework employs three enforcement layers to enable what we term “verification without trust.” The Model Context Protocol standardizes tool invocation, enabling policy enforcement at the interface layer. Capability-based access control issues short-lived, scoped tokens that authorize specific actions, enforcing least privilege even if the agent attempts unauthorized operations. Hardware-based attestation through Trusted Platform Modules and secure enclaves provides cryptographic verification of agent configuration, execution integrity, and audit log immutability.

4 Results

We validated the ABOM Risk Scoring Engine through a systematic evaluation of six agent configurations designed to span the full UART tier spectrum. Each configuration represents a distinct agent archetype commonly encountered in production deployments.

4.1 Experimental Setup

The test suite comprises agents ranging from passive chatbots (no tools, human-in-the-loop oversight, ephemeral memory) to fully autonomous frontier systems with self-replication capabilities. For each agent, we constructed an ABOM manifest specifying its model core, agency profile, autonomy leash, persistence layer, and scaffolding inventory. The scoring engine processed each manifest and computed the risk magnitude R according to Equation 1, subsequently mapping the score to the appropriate UART tier.

4.2 Validation Results

Table 1 presents the computed risk vectors and tier classifications for all six test cases. The minimal risk configuration ($A = 1$, $U = 1$, $P = 0$, $M = 0.85$) yielded $R = 0.85$, correctly classifying as Tier 1 (Assistive). The moderate risk configuration ($A = 4$, $U = 3$, $P = 1$, $M = 0.765$) produced $R = 24.95$, appropriately placing it in Tier 3 (High-Agency).

Table 1: Risk scoring validation results across agent configurations.

Configuration	A	U	P	M	R	Tier
Minimal Risk	1	1	0	0.85	0.85	1
Low Risk	2	2	1	0.765	8.32	2
Moderate Risk	4	3	1	0.765	24.95	3
High Risk	4	3	2	0.765	67.83	4 [†]
Critical Risk	4	3	2	0.9	79.8	4 [†]
Systemic Override	6	3	2	1.0	133.0	4 [†]

[†]Tier 4 triggered by systemic risk override.

The exponential persistence term e^P demonstrated the intended amplification effect. Comparing otherwise identical configurations, transitioning from session-only memory ($P = 1$) to long-term memory ($P = 2$) increased risk magnitude by a factor of 2.72, consistent with our design rationale that long-horizon autonomous behavior compounds potential hazards.

4.3 Override Mechanism Validation

Three test cases exercised the systemic risk override pathway. Two configurations exceeded the 10^{25} FLOPs training compute threshold derived from EU AI Act guidelines for general-purpose AI with systemic risk. The third configuration included self-replication in its capability flags. In all three cases, the engine correctly elevated the classification to Tier 4 regardless of the computed R value, validating the override mechanism’s reliability.

4.4 Mitigation Factor Analysis

The mitigation modifier M reduced risk scores proportionally to the presence of safety controls. Configurations employing MCP-standardized tool invocation, sandboxed execution environments, and circuit breakers achieved M values as low as 0.765, representing a 23.5% risk reduction from baseline. This reduction, while insufficient to change tier classification in most cases, provides meaningful incentive for operators to implement safety scaffolding.

All six test cases achieved 100% classification accuracy against expected outcomes, validating both the scoring formula and tier boundary definitions.

5 Discussion and Limitations

The ABOM Risk Scoring Engine represents a first step toward quantitative, enforceable governance for autonomous AI agents. Our results demonstrate that the framework can reliably differentiate agents across risk tiers while providing appropriate incentives for safety-enhancing design choices. We discuss implications, limitations, and directions for future work.

5.1 Implications for AI Governance

The framework addresses a critical gap in current regulatory approaches: the absence of standardized, machine-readable documentation for agentic systems. While the EU AI Act establishes requirements for technical documentation and risk assessment, it provides limited guidance on how to operationalize these requirements for agents that combine foundation models with tool access, memory, and autonomous decision-making. The ABOM schema offers a concrete specification that regulators, auditors, and operators can adopt immediately.

The exponential treatment of persistence merits particular attention. Our formulation reflects the assessment that long-term memory fundamentally changes an agent’s risk profile—not merely by enabling better task performance, but by creating opportunities for strategic behavior that stateless systems cannot exhibit. An agent that accumulates context across sessions may identify patterns in oversight, optimize for reduced intervention, or develop plans spanning timescales that exceed human attention. The e^P term encodes this concern directly into the governance framework.

The mitigation modifier M operationalizes a principle implicit in existing safety discourse: that architectural choices matter for risk. By reducing computed risk scores for agents employing MCP, sandboxing, and circuit breakers, the framework creates quantifiable incentives for operators to adopt safety-enhancing scaffolding. This approach aligns technical design decisions with regulatory outcomes.

5.2 Limitations

Several limitations constrain the current framework. First, the ABOM relies on accurate self-reporting by operators. While hardware attestation can verify certain configuration parameters, the framework cannot currently detect undisclosed tools or capabilities. Adversarial operators could underreport agency or misrepresent autonomy settings.

Second, the tier thresholds ($R = 5, 20, 50$) are illustrative rather than empirically derived. We selected values that produce intuitive tier distributions across our test cases, but optimal thresholds would require extensive empirical calibration against real-world agent deployments and incident data. Different jurisdictions or application domains may require adjusted boundaries.

Third, the scoring formula treats all state-changing tools equivalently within their category. In practice, an agent with access to a database write operation poses different risks than one capable of executing arbitrary shell commands. Future versions should incorporate tool-specific impact weights derived from domain expertise or incident analysis.

Fourth, the framework does not address emergent capabilities. An agent’s actual risk may exceed what its ABOM documentation would predict if the underlying model develops unexpected behaviors through scale, fine-tuning, or novel prompting strategies. The framework governs declared capabilities rather than latent ones.

Finally, the reference implementation validates the scoring logic but does not include the hardware attestation mechanisms described in Section 2.4. A production deployment would require integration with TPM infrastructure and secure enclave environments.

5.3 Future Work

Several extensions would strengthen the framework. Continuous monitoring could supplement static ABOM analysis, flagging behavioral deviations from declared capabilities. Federated registries could enable cross-organizational sharing of ABOM data while preserving confidentiality through cryptographic techniques. Automated ABOM generation from agent source code could reduce documentation burden and improve accuracy. Empirical studies of deployed agents could inform evidence-based refinement of tier thresholds and vector weights.

6 Conclusion

We presented a technical governance framework for agentic AI systems comprising the Agentic Bill of Materials (ABOM), a quantitative risk scoring engine, and the Unified Agentic Risk Tiering (UART) classification system. Our validation across six agent configurations demonstrated that the formula $R = A \times U \times e^P \times M$ reliably differentiates agents across risk tiers, with the exponential persistence term appropriately amplifying risks associated with long-term memory and the mitigation modifier providing quantifiable incentives for safety scaffolding. The systemic risk override mechanism correctly elevated frontier agents exceeding 10^{25} FLOPs or possessing critical capabilities to Tier 4 classification.

The framework addresses an urgent gap in AI governance: the absence of standardized, machine-readable documentation for autonomous agents capable of acting on the world. By embedding governance into the technical substrate of agentic systems—through structured manifests, cryptographic verification, and hardware-enforced controls—this approach enables scalable, enforceable oversight that does not rely on operator promises alone. As agentic AI systems proliferate across critical domains, frameworks like ABOM and UART provide a foundation for international harmonization of safety standards, ensuring that the transformative potential of autonomous agents is realized within appropriate guardrails.

Code and Data

The complete reference implementation is publicly available:

- **Code Repository:** <https://github.com/AnmolxAI/Agentic-Bill-of-Materials-ABOM-Risk-Scoring-Engine>

The repository includes the FastAPI backend implementing the risk scoring engine, a React frontend for interactive ABOM analysis, six validated test cases spanning all UART tiers, and comprehensive documentation including a user tutorial. All code is released under an open-source license to facilitate adoption and extension by researchers and practitioners.

No sensitive data is included in this project. The test cases are synthetic agent configurations designed to exercise the scoring logic across the tier spectrum. No info-hazard considerations apply.

7 References

References

- [1] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, 2024. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>
- [2] Anthropic. Anthropic’s Responsible Scaling Policy. *Anthropic*, 2024. <https://www.anthropic.com/news/anthropics-responsible-scaling-policy>
- [3] OpenAI. OpenAI Preparedness Framework. *OpenAI*, 2023. <https://openai.com/safety/preparedness>
- [4] Anthropic. Introducing the Model Context Protocol. *Anthropic*, November 2024. <https://www.anthropic.com/news/model-context-protocol>

- [5] SAE International. SAE J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. *SAE International*, 2021. https://www.sae.org/standards/content/j3016_202104/
- [6] International Electrotechnical Commission. IEC 62443: Industrial communication networks – Network and system security. *IEC*, 2018. <https://www.iec.ch/cyber-security>
- [7] National Telecommunications and Information Administration. The Minimum Elements for a Software Bill of Materials (SBOM). *U.S. Department of Commerce*, 2021. <https://www.ntia.gov/page/software-bill-materials>
- [8] Trusted Computing Group. TPM 2.0 Library Specification. *Trusted Computing Group*, 2019. <https://trustedcomputinggroup.org/resource/tpm-library-specification/>
- [9] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.03629.
- [10] The White House. Executive Order 14028: Improving the Nation’s Cybersecurity. *Federal Register*, May 12, 2021. <https://www.federalregister.gov/documents/2021/05/17/2021-10460/improving-the-nations-cybersecurity>
- [11] Anthropic. Building Effective Agents. *Anthropic*, 2024. <https://www.anthropic.com/research/building-effective-agents>

LLM Usage Statement

We used Claude 4.5 Sonnet (Anthropic) as a coding assistant throughout the development of this project. Specifically, LLM assistance was utilized for:

- Implementing the FastAPI backend and React frontend components
- Debugging and refining the risk scoring algorithm

All mathematical formulations, tier thresholds, and architectural decisions were independently verified through systematic testing. The six-case validation suite was executed manually to confirm expected tier classifications matched computed results. Claims regarding alignment with external standards (EU AI Act thresholds, SAE J3016 levels, ISA/IEC 62443 security levels) were verified through direct consultation of authoritative sources.