



OHIP Artificial Intelligence (AI) Security and Privacy Standards Baseline

Version 2.1 June 2026

The Challenges of AI Security and Privacy

Artificial Intelligence (AI) technologies have significant potential to transform society and people's lives. At the same time, they also pose unique risks to the community that go well beyond those posed by traditional software or information-based systems. AI risks can emerge from the interplay of technical aspects combined with societal factors related to how a system is used, its interactions with other AI systems, who operates it, and the social context in which it is deployed. These risks make AI a uniquely challenging technology to deploy and utilize both for organizations and within society. Without proper controls, AI systems can amplify, perpetuate, or exacerbate inequitable or undesirable outcomes for individuals and communities. With proper controls, AI systems can mitigate and manage risks of inequitable or harmful outcomes.ⁱ

As such, organizations are entering uncharted territory in securing and overseeing AI solutions. The rapid advancement of AI also opens doors for adversaries to enhance their attack strategies, introducing a dual challenge of defense and threat escalation. For any organization leveraging AI systems in its operations, it's critical to assess and establish comprehensive policies, governance, security protocols, privacy measures, and accountability standards to ensure these technologies align with business processes securely and ethically. The overall aim, therefore, is not just to leverage AI for success but also to protect against the risks of hasty, disparate, ungoverned, or insecure AI implementations.ⁱⁱ

The field of AI introduces a new set of terminologies. The reader is highly encouraged to review the glossary at the end of this document. This will assist the reader in comprehending the concepts contained herein.

AI Security and Privacy Risks and Threats are Unique

- **AI Security and Privacy Risks are Unique**

There are myriad standards and best practices to help organizations mitigate the risks of traditional software or information-based systems. However, the risks posed by AI systems are in many ways unique to this technology domain (e.g., See Appendix B. NIST AI 100-1)ⁱⁱⁱ.

- **Data Driven Nature of AI Introduces Additional Security and Privacy Challenges**

The data-driven approach of Machine Learning (ML) introduces additional security and privacy challenges besides the classical security and privacy threats faced by most operational systems. The spectrum of effective attacks against ML is wide, rapidly evolving, and covers all phases of the ML lifecycle. These security and privacy challenges include

the potential for adversarial manipulation of training data and adversarial exploitation of model vulnerabilities to adversely affect the performance of the AI system. They also include malicious manipulations, modifications or mere interaction with models to exfiltrate sensitive information about people represented in the data, about the model itself, or proprietary enterprise data.^{iv}

- **The Class of AI that is Adopted Adds to the Security and Privacy Challenges**
For example, Generative AI security has clear overlaps with ML security. However, there are differences given the role of natural language interaction within Generative AI applications^v.
- **AI Incidents are Difficult to Detect and Not as Well Defined as Cyber Incidents**
AI systems and the contexts in which they are deployed are complex, making it difficult to detect and respond to failures when they occur^{vi}. Further, as of this date, there are currently no widely accepted definitions of what constitutes an AI Incident, nor have any authoritative governmental bodies issued a definition. In contrast, cybersecurity incidents are well understood with key concepts and guidance for how to think about threats and risks. Indeed, the need to distinguish between AI and Cyber Incidents adds to complexity of AI Incident detection, assessment of the blast radius, and response.^{vii}
- **Large Language Model (LLM) Adoption has Introduced Unprecedented Risks to Data Security and Privacy**
As organizations scale their reliance on LLMs and their AI systems become more sophisticated, the challenges of protecting sensitive information, ensuring compliance, privacy and mitigating evolving threats become more acute. Risks include potential breaches of sensitive information and challenges in meeting stringent data protection regulations.

Traditional security practices such as encryption, access control, data masking, and network security constitute the foundational layer for safeguarding data in Large Language Models. These traditional and necessary measures must be augmented to address the unique challenges that arise throughout the AI lifecycle.^{viii}

- **Adversaries are Leveraging AI to Advance Their Ends**
Adversaries are increasingly leveraging these very AI tools to attack both traditional information systems and AI systems. Large Language Models (LLMs) and Generative AI tools are being used by adversaries to^{ix}:
 - o refine and expedite methods of attacking organizations.
 - o craft new malware and detection evasion techniques.
 - o generate sophisticated, unique, or tailored, phishing schemes.
 - o create convincing deep fake attacks to implement social engineering ploys.
 - o execute sophisticated, learning-enhanced, probing and intrusion processes.
 - o develop innovative hacking capabilities.
- **The Emerging Agentic AI (Agent based AI) Technology Amplifies Risks in Adoption**
Industry is rapidly adopting and advancing Agentic AI (Agent based AI) technology. Agentic AI is a class of artificial intelligence that focuses on autonomous systems that can make decisions and perform tasks without human intervention. AI agents are software entities designed to perceive their environment, make decisions, and take actions autonomously to achieve specific goals by selecting tools, such as APIs, to achieve them.



Agentic AI represents a seismic shift, offering immense promise and equally significant risk. Agentic AI introduces a fundamentally new threat surface. Its probabilistic nature, memory and reasoning capabilities, and autonomy make it vulnerable to manipulation, misuse, and abuse.

Security professionals and AI developers must transition from traditional controls to a proactive, embedded, and defense-in-depth approach that spans the entire agent lifecycle.^x

Purpose

This document provides an audience-appropriate summary of definitions, standards, citations and justifications for Baseline AI Security and Privacy. They are established to meet State and Federal requirements and industry best practices for protecting the security and privacy of “Sensitive / Confidential Data” (defined below) as well as ensuring its safe, responsible and trustworthy use by AI Systems and Services used by OHIP.

The principles and their attributes herein are based on national standards and best practices. They provide objective, verifiable criteria for assessing risk and identifying countermeasures that directly address root causes of risks to AI applications and models. These principles and standards therefore support verifiable compliance with the AI risk assessment mandate in Section 4.5 of the NYS ITS AI Acceptable Use Policy^{xi}

The standards within this document are intended to act as a foundation for guiding AI System / Service procurement, development, implementation, and risk assessment.

This document is also intended to support identification and prevention / mitigation of AI vulnerabilities that present the potential for unacceptable risk for technical and business impacts on the Department as well as risks to populations served by the Department.

Preconditions and Assumptions

Both national and internationally recognized standards bodies have adopted the adjectives *trustworthy* and *responsible* as overarching attributes for adoption of artificial intelligence^{xii}. Decisions regarding identifying, managing, mitigating, or accepting risks that may impact these qualifiers are an essential role of a strategically aligned, organization-wide, AI governance plan, process and structure. This framework is epitomized by the NIST Risk Management Framework^{xiii} and its Playbook^{xiv}.

The Guiding Principles and Attributes in this document, and the accompanying technical guidance document, provide AI security and privacy standards and best practices. They also provide the control point ‘hooks’ to implement organization-wide artificial intelligence:

- Governance policies.
- Risk assessments and management.
- Security protocols.
- Privacy measures.
- Accountability standards .

The governance roles are essential to ensure AI system outcomes align with organizational



expectations of security, privacy and ethics.

These standards documents do not speak to the governance framework described above. Rather they presume that is in existence as a pre-condition.

Document Description and Overview

This document is a summary of specific guiding principles and underlying attribute standards in the Department's AI Security and Privacy Baseline Technical Guidance document. The AI Security and Privacy Technical Guidance document contains detailed and heavily referenced implementation standards for each Attribute within a Guiding Principle. The detailed implementation standards for each attribute are based on a composite of the standards and best practices listed in the reference section below.

While the summary attribute standards listed here are very general, the detailed implementation standards in the technical guidance document are specific in their requirements to meet those high-level attribute standards. Attestations to the Department's AI Security and Privacy Standards are based on the implementation standards within the Technical Guidance Document. The reader is referred to the Technical Guidance Document for detailed guidance, implementation specifics, and supporting references.

The Guiding Principles and their supportive Attribute standards herein are divided into two parts, Baseline and Advanced. The Baseline Guiding Principles and their supportive Attribute standards are intended to be AI class (type) and model agnostic. This 'Baseline Component' applies to all AI types and models. The Advanced Guiding Principles and Attribute standards are intended to apply specifically to Agentic AI Systems. Both Baseline and Advanced standards apply to Agentic AI Systems.

Applicability and Scope

These standards apply to any AI system that will be used to access, store, transmit, interpret, or process Sensitive / Confidential data of any type. The term "Sensitive / Confidential Data" includes, but is not limited to:

- Medicaid Confidential Data(MCD),
- PHI / PII,
- SAMHSA / SUDS / Mental Health
- Reproductive Health
- Health of Minors
- FTI, SSA, PCI,
- Information that, if exposed, could present a threat to AI system integrity / availability, Security / Privacy, or support adversarial reconnaissance or unauthorized access to protected resources (e.g., access tokens, model / agent capabilities and guardrails).
- Information that essential to critical decision support for members, providers or state administrative staff.
- Intellectual property (e.g., source code, proprietary algorithms) \



- Other program administrative information that requires provenance / attribution to authoritative source (e.g. policies, regulations) .

Use of Sensitive / Confidential Data to support AI systems includes, but not limited to, use of this data in any direct or indirect manner for the purpose of:

- Model training
- Augmentation / grounding
- RLHF, data labelling
- Inference
- Application development, code review, code documentation, code transformation or debugging
- Decision support
- Predictive clinical or administrative functions
- Payment / reimbursement
- Benefit assess / determination
- Eligibility / enrollment
- Payment integrity
- Adjudication of claims
- Prior authorization
- Member or Practitioner health / care access, management or provision
- Member or practitioner help desk or advisory functionality of any type
- Beneficiary / provider communications of any type
- Policy or regulatory development or distribution / interpretation of same

The applicability of any specific standard herein will depend on the AI use case. This would include considerations such as level risk of / confidentiality sensitivity or decision criticality classification of the AI application, its functionality, architecture, environment, exposure and complexity. The Department reserves the right to define applicability of any specific standard based on the above criteria.

The entities covered within this scope include but are not limited to AI Systems used or developed by:

- Vendors, third-parties or contractors on behalf of DOH.
- COTS and SaaS offerings that leverage AI.
- A data consuming entity for use with Sensitive / Confidential Data as described above.
- OHIP.
- Office of Information Technology Services (ITS) or other NYS agencies.

Regulations

AI systems that fall within the Scope and Applicability statements above, shall adhere to the following regulations:

- Health Information Portability and Accountability Act (HIPAA).
- New York State Technology Law^{xv}.
- New York State Information Technology Policy NYS-P24-001 Acceptable Use of Artificial Intelligence Technologies^{xvi}.



- Other applicable New York State ITS policies and standards^{xvii}.

AI Security and Privacy Standards and Best Practice References Used and Cited in Developing this Document

- Technical Standards Organization Sources
 - Cloud Security Alliance (CSA)
 - o Secure Agentic System Design. A Trait-Based Approach^{xviii}.
 - o Data Security Within AI Environments 2025^{xix}.
 - o Agentic AI Identity and Access Management. A New Approach^{xx}.
 - o MLOps Overview^{xxi}
 - o Agentic AI Red Teaming Guide^{xxii}
 - o AI Controls Matrix (AICM) Implementation and Auditing Guide^{xxiii}.
 - Internet Engineering Task Force (IETF)
 - o The OAuth 2.1 Authorization Framework^{xxiv}.
 - o IETF Network Group. AAuth - Agentic Authorization OAuth 2.1 Extension^{xxv}.
 - o Agent Name Service (ANS): A universal directory for secure AI agent discovery and interoperability^{xxvi}.
 - o Extending Certificate Enrollment Protocols for Scalable Agentic AI Identity IETF RATS Workgroup June 2025^{xxvii}.
 - MITRE
 - o SAFE-AI. A Framework for Securing AI-Enabled Systems^{xxviii}.
 - National Institute of Standards and Technology (NIST)
 - o Artificial Intelligence Risk Management Framework (RMF)^{xxix}.
 - o AI RMF Playbook^{xxx}.
 - o AI 600-1 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence (GAI) Profile^{xxxi}.
 - o AI 100-2e2023 Adversarial Machine Learning. A Taxonomy and Terminology of Attacks and Mitigations^{xxxii}.
 - o SP 800-218A Secure Software Development Practices for Generative AI and Dual-Use Foundation Models^{xxxiii}.
 - o SP 1270 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence^{xxxiv}.
 - o SP 800-61r3 Incident Response Recommendations and Considerations for Cybersecurity Risk Management. A CSF 2.0 Community Profile 2025^{xxxv}.
 - o SP 800-84 Guide to Test, Training, and Exercise Programs for IT Plans and Capabilities^{xxxvi}.
 - o AI 100-4 Reducing Risks Posed by Synthetic Content^{xxxvii}.
 - Open Web Application security Project (OWASP)
 - o 2025 OWASP Top 10 Security Risks for Large Language Model Applications^{xxxviii}.
 - o OWASP AI Security and Privacy Guide^{xxxix}.
 - o OWASP AI Exchange 2026^{xl}.
 - o OWASP AI Security Testing Guide^{xli}.
 - o OWASP GenAI COMPASS Run Book^{xlii}.
 - o OWASP GenAI Security Project Threat Defense COMPASS 1.0^{xliii}.
 - o Large Language Model Applications Cybersecurity and Governance Checklist v1.1^{xliv}.

- o Large Language Model and Gen AI Data Security Best Practices^{xlv}.
- o Large Language Model and Generative AI Security Solutions Landscape^{xlvi}.
- o The OWASP Generative AI and Agentic Security Solutions Landscape (Q2–Q3 2025)^{xlvii}.
- o LLM Prompt Injection Prevention Cheat Sheet^{xlviii}.
- o A Practical Guide for Securely Using Third-Party MCP Servers^{xlix}.
- o A Practical Guide for Secure MCP Server Development^l.
- o OWASP Top 10 for Agentic Applications for 2026^{li}.
- o OWASP Top 10 Non-Human Identities Risks 2025^{lii}.
- o OWASP AI Vulnerability Scoring System (AIVSS)^{liii}.
- o AIVSS Scoring System for OWASP Agentic AI Core Security Risks^{liv}.
- o GenAI Incident Response Guide 1.0^{lv}.
- o Guide for Preparing and Responding to Deepfake Events^{lvi}.
- o GenAI Red Teaming Guide^{lvii}.
- o OWASP Agentic AI – Threats and Mitigations Version 1^{lviii}.
- o OWASP Agentic AI – Threats and Mitigations Version 1.1^{lix}.
- o State of Agentic AI Security and Governance 1.0^{lx}.
- o Securing Agentic Applications Guide 1.0^{lxi}.
- o Multi-Agentic System (MAS) Threat Modeling Guide^{lxii}.
- o AI Security Solutions Landscape for Agentic AI Q3 2025^{lxiii}.
- o Agent Name Service (ANS) for Secure AI Agent Discovery v1.0^{lxiv}.
- o OWASP AI Bill Of materials (AIBOM) Project^{lxv}.
- o OWASP CycloneDX (ECMA-424)^{lxvi}.
- o OWASP Software Component Verification Standard (SCVS) for identifying and reducing risk in a software supply chain^{lxvii}.
- o OWASP Vendor Evaluation Criteria for AI Red Teaming Providers and Tooling^{lxviii}.
- o OWASP GenAI Data Security and Mitigations 2026^{lxix}.

Organization for the Advancement of Structured Information Standards (OASIS)

- o OASIS Coalition for Secure AI (CoSAI). Incident Response Framework, V1.0^{lxx}.
- o OASIS Collaborative Automated Course of Action Operations (CACAO) for Cybersecurity. Security Playbooks Version 2.0^{lxxi}.
- Sources for Industry Standards and Best Practices for AI Threat Modeling and Risk Analysis.
 - o MITRE Group Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS)^{lxxii}.
 - o MITRE Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK®) as related to ATLAS^{lxxiii}.
 - o MITRE Common Attack Pattern Enumerations and Classifications (CAPEC)^{lxxiv}.
 - o MITRE D3FEND CyberSecurity Countermeasures^{lxxv}.
 - o Google's Secure AI Framework (SAIF). Organizes AI security into four key components: Application, Model, Infrastructure and Data, allowing visibility into attack surfaces and threats across the full AI Architecture and lifecycle^{lxxvi}.
 - o Process for Attack Simulation and Threat Analysis (PASTA). A seven-stage, risk-centric threat modeling framework that aligns technical analysis with business impact^{lxxvii}.

- o Microsoft's STRIDE modeling framework that categorizes threats into Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege^{lxxviii}.
- o LINDDUN (Linking, Identifying, Non-repudiation, Detecting, Data disclosure, Unawareness, and Non-compliance). A privacy-focused framework for modeling threats to data confidentiality and compliance^{lxxix}.
- o MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome), a threat modeling framework designed specifically for the unique challenges of Agentic AI^{lxxx}.
- o Practical Library of Threats 4 Artificial Intelligence (PLOT4AI)^{lxxxi}.
- Other Sources.
 - o L. Kastner and B. Crook. Explaining AI through mechanistic interpretability^{lxxxii}.
 - o L. Bereska and E. Gavve. Mechanistic Interpretability for AI Safety A Review^{lxxxiii}.
 - o Agent2Agent (A2A) Protocol^{lxxxiv}.
 - o Model Context Protocol^{lxxxv}.
 - o O. Santos A New Identity Framework for AI Agents (Cisco) May 2025^{lxxxvi}.
 - o K. Feng, D. McDonald and A. Zheng. Levels of Autonomy in Agents. Cornell University arXiv 2025^{lxxxvii}.
 - o S. Pant and M. Viswanathan. Five levels of AI Intelligence for Enterprise Use. Cisco Outshift 2025^{lxxxviii}.
 - o GAO-24-107292 Combatting Deepfakes 2024^{lxxxix}.
 - o GAO-20-379SP Deepfakes 2020^{xc}.
 - o N. Kobis, B. Doležalová and I. Soraperra. Fooled twice: People cannot detect deepfakes but think they can. 2021^{xi}.
 - o NSA, FBI, and CISA Release Cybersecurity Information Sheet on Deepfake Threats. 2023. Last Accessed February, 2025^{xci}.
 - o Contextualizing Deepfake Threats to Organizations. 2023. Last Accessed February, 2025^{xcii}.
 - o Center for Content Provenance and Authenticity. Content Credential Provenance and Authenticity Standards^{xciv}.
 - o Content Credentials: Strengthening Multimedia Integrity in the Generative AI Era. 2025. Last Accessed February, 2025^{xcv}.
 - o DHS Public-Private Analytic Exchange Program, Increasing Threat of Deepfake Identities. 2022^{xcvi}.
 - o DHS S&T Digital Forgeries Report, Technology Landscape Assessment. 2023^{xcvii}.
 - o Impacts of Adversarial Use of Generative AI on Homeland Security. 2025^{xcviii}.
 - o Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study. 2023^{xcix}.
 - o S. V. Chilukoti et al. Auto DP-SGD: Dual Improvements of Privacy and Accuracy via Automatic Clipping Threshold and Noise Multiplier Estimation. 2023^c.



Companion Documents

This document is accompanied by five companion documents.

- OHIP Secure Application Programming Interface (API) Baseline
- OHIP Secure Application Programming Interface (API) Technical Guidance
- OHIP Application Programming Interface (API) Implementation Standards Baseline Attestation
- OHIP Artificial Intelligence (AI) Security and Privacy Baseline Technical Guidance
- OHIP Artificial Intelligence (AI) Security and Privacy Implementation Standards Baseline Attestation

IMPORTANT NOTES:

- Insecure APIs are a significant threat to AI Systems^{ci}. Given the dependency of AI Systems on internal and external API's as well as internal microservices infrastructure, the API Security Technical Implementation Standards must be attested to as a prerequisite to the AI Security and Privacy Implementation Standards Attestation.
- Attestation to the AI Security and Privacy Standards is based on the Implementation Standards contained within the AI Security and Privacy Standards Technical Guidance Document.

AI Security and Privacy Best Practices and Standards

OHIP has developed the following set of six AI Security and Privacy principles. Each principle is accompanied by attribute standards that support the principle. The principles and attributes are based on authoritative national sources listed above. AI systems that fall within the Scope and Applicability statements above, shall adhere to the following principles and attributes. Attestations to their implementation will be based on the implementation standards within the Technical Guidance Document, where deemed applicable. The term 'data' used within these standards should be interpreted to mean "Sensitive / Confidential Data".

The six Security and Privacy principles are also accompanied by three (3) Overarching Principles (see Section on Overarching Principles below). These deal with the Departments expectations regarding:

- Implicit / explicit existence of documentation and artifacts.
- Responding to the overall intent of a control attribute and / or specific implementation standard.
- Full disclosure of third-party dependencies.

AI-1 The AI Service provider implements baseline AI Model and AI Application Service Data Security, Privacy and Reliability controls based on zero trust principles and industry best practices.

Baseline Threat Modeling and Risk Assessment Attribute Standards

- AI 1-1 The AI system is mapped to a comprehensive decomposition of its AI architecture into components / subcomponents using an open industry approved standard methodology.



- AI 1-2 An industry recognized security threat modeling methodology is applied to each component / subcomponent identified in AI 1-1 through the full AI system lifecycle.
- AI 1-3 Industry recognized threat modeling methodology(ies) for Privacy, Ethics, Bias, Fairness is / are applied to each component / subcomponent identified in AI 1-1 throughout the full AI system lifecycle.
- AI 1-4 Risks, risk analyses, prioritization and mitigation strategies are documented, catalogued and continually updated along with corresponding threats.

Baseline Minimum Necessary, Least Privilege and Privacy Attribute Standards

- AI 1-5 The AI Service Provider implements policies, procedures, and technology to support the principle of minimum (information) necessary.
- AI 1-6 The AI Service Provider implements Privacy-Enhancing Technologies (PETs) to safeguard user data privacy while maintaining model effectiveness.
- AI 1-7 The AI Service Provider implements policies, procedures, and technology to enforce the principle of least privilege for user authorizations.
- AI 1-8 The AI Service Provider implements zero trust policies, procedures and technology to support least privilege for AI Models.

Baseline AI Identity and Access (Authorization) Management (IAM) Attribute Standards

- AI 1-9 IAM and authorization for user, user delegated client or non-human entity to an endpoint follows the OHIP API Security standards for strong client authentication and authorization.

Baseline Security Control Attribute Standards

- AI 1-10 The AI Service has implemented technologies and processes to support input and output filters. (OWASP LLM01, LLM06)
- AI 1-11 The AI Service has implemented technologies and processes to support comprehensive, zero trust, input validation for all sources. (OWASP LLM02, LLM05, LLM06, LLM10)
- AI 1-12 The AI Service has zero trust controls in place to mitigate against improper model output handling. Technologies and processes are implemented to support comprehensive output handling, format definition, and validation. (OWASP LLM05)
- AI 1-13 The AI service has controls in place to prevent and mitigate unauthorized disclosure of sensitive information. Sensitive information includes but is not limited to: PHI, PII, FTI, sensitive information about the model, security credentials or information that can lead to compromise of AI safeguards. (OWASP LLM02)
- AI 1-14 AI Service has robust controls in place to protect model and data supply chains. (OWASP LLM03)
- AI 1-15 The AI Service has robust controls in place to mitigate against data and model poisoning. (OWASP LLM04)



AI 1-16 The AI Service has controls in place to mitigate against unbounded resource consumption. (OWAP LLM10)

AI 1-17 The AI Service Provider has controls in place to mitigate misinformation. (OWASP LLM09)

AI 1-18 The AI Service implements zero trust controls to mitigate against System Prompt leakage. (OWASP LLM07)

AI 1-19 The AI Service implements controls to mitigate against vulnerabilities in vectors and embeddings (e.g., used for representing and searching unstructured data using numeric representations). (OWASP LLM08)

AI 1-20 The AI service maintains comprehensive and detailed immutable logs and baseline profiles of all inputs, outputs and actions for all AI system components.

AI-2 Advanced Controls for Securing Agentic AI Systems. The agentic AI Service provider implements a proactive, embedded and defense in depth approach to agentic AI system security. The approach must span the entire agent lifecycle including development, testing, and runtime. (This agentic AI-specific control attributes listed here are considered an overlay to AI-1 and are required in addition to AI-1, where A-1 is applicable.)

Agentic Threat Modeling and Risk Assessment Attribute Standards

AI 2-1 Security Risk Assessments are based on dynamic threat modeling using an industry standard agentic AI-specific multi-layer threat model framework.

AI 2-2 Ongoing, structured threat modeling and attack surface analysis is conducted across all components of the agentic AI architecture, agent framework, and architecture patterns in use.

Agentic AI Non-Human Identity and Access (Authorization) Management (IAM) Attribute Standards

AI 2-3 In scenarios / patterns involving only a single, well-defined agent acting in a relatively static role and tethered directly to a human user's session, or a pre-configured machine identity derived from these traditional IAM systems, the Agent IAM / authorization uses OAuth 2.1

AI 2-4 In scenarios / patterns involving Multi-Agent Systems (MAS), the agentic AI service for MAS scenarios implements an Agentic AI Identity and Access Management Framework Architecture:

- o That is based on zero-trust, least privilege and minimum necessary,
- o Uses strong agent authentication methods,
- o Uses Decentralized Identity Management,
- o Uses Verifiable Credentials,
- o Uses Dynamic Policy-Based Access Controls (PBAC),
- o and uses authenticated delegation.

The Framework also uses continuous monitoring and behavioral analytics, secure agent authentication protocols and an Agent Naming and Discovery Service (ANS).



- AI 2-5 Continual threat analysis of the Agentic AI IAM Framework is performed using the MAESTRO framework or industry endorsed alternative for Agentic AI.
- AI 2-6 Mitigations are in place for known agentic AI within-layer or cross-layer threats based on MAESTRO Threat modeling or industry endorsed alternative for Agentic AI.

Agentic AI Oversight Attribute Standards

- AI 2-7 Agentic AI security oversight processes, procedures, and tools are in place to continually and dynamically assess risks, performance, and unintended agent behavior in real time.
- AI 2-8 Automated, adaptive, and self-regulating AI processes are in place to enforce oversight guardrails on AI agent autonomy.
- AI 2-9 Automated and adaptive risk management policies, thresholds, identification, classification, and intervention mechanisms are in place.
- AI 2-10 Real-time documentation and self-reporting of potential compliance anomalies are in place.

Agentic AI Security Control Attribute Standards

- AI 2-11 Robust autonomous AI security model tooling is in place.
- AI 2-12 Self-healing AI security models are in place.
- AI 2-13 Security controls are directly embedded into agent decision logic in a manner that integrates AI cybersecurity with AI oversight.
- AI 2-14 Advanced, agentic AI threat detection capabilities are in place.
- AI 2-15 Automated, continuous agentic AI penetration and security testing are in place.
- AI 2-16 Automated agentic AI security patching is in place.
- AI 2-17 A hybrid of static and dynamic agentic guardrails is adopted and implemented as part of a layered agent defense strategy.

Agentic AI Threat Mitigation Attribute Standards

- AI 2-18 The AI service implements controls to mitigate AI agent reasoning manipulation, intent breaking, and goal manipulation threats. Controls prevent attackers from manipulating AI intent resulting in security bypasses. Controls are also in place to enhance traceability of AI agent actions. (OWASP ASI01, T06, T07, AIVSS Core Risk 10 NHI10)
- AI 2-19 The AI service implements controls to mitigate agent memory poisoning and AI knowledge corruption threats. Controls prevent the AI service from storing, retrieving, or propagating manipulated data that could corrupt decision-making or spread misinformation. (OWASP ASI06, T1, T4, T12 AIVSS Core Risk 6, NHI2, NHI7)
- AI 2-20 The AI service implements controls for securing AI agent tool execution. (OWASP ASI02, T2, AIVSS Core Risk 1, NHI2, NHI5)



- AI 2-21 The AI service implements controls to prevent unauthorized AI agent privilege escalation, identity spoofing, and access control violations. (OWASP ASI03, T3, AIVSS Core Risks 2,5 NHI3, NHI4, NHI5)
- AI 2-22 The AI service implements controls to protect Human In The Loop (HITL) processes to mitigate decision fatigue exploits. Controls prevent actors from overloading human decision-makers, manipulating AI intent, or bypassing security through deceptive AI agent behaviors. (OWASP ASI09, T10, NHI10)
- AI 2-23 For MAS architecture patterns, the AI service implements controls to ensure secure MAS communication and trust mechanisms. Controls prevent actors from corrupting MAS communication, exploiting trust mechanisms, or manipulating decision-making in distributed AI environments. (OWASP ASI07, T12, T16, T30-OWASP-MAS, Intersects AIVSS Core risks 1, 3, 4, 10, NHI4)
- AI 2-24 The AI service implements controls to mitigate against agentic supply chain vulnerabilities. (OWASP ASI04, T17, AIVSS Core Risk 8, NHI1, NHI3, NHI6).
- AI 2-25 The AI service implements controls to mitigate against resource overload. (OWASP T4. Contributing factors to ASI02, ASI06. Implicit within AIVSS Core Risks 1 and 10, NHI7)
- AI 2-26 The AI service implements controls to mitigate against cascading failures. OWASP ASI08, T5, AIVSS Core Risk 3, NHI7, NHI8, NHI9)
- AI 2-27 The AI service implements controls to mitigate against agent repudiation and untraceability. (OWASP T8, AIVSS Core Risk 9, NHI8, NHI10, MAS T23)
- AI 2-28 The AI service implements controls to mitigate against unexpected remote code execution and code attacks. (OWASP ASI05, T11, AIVSS Core Risk 1 and 7, NHI6)
- AI 2-29 The AI service implements controls to mitigate against rogue agents in MAS. (OWASP ASI10, T13, Core risk 4)
- AI 2-30 The AI service implements controls to mitigate against Human Attacks on Multi-Agent Systems. (OWASP T14, ASI10, included in AIVSS Core Risk4).
- AI 2-31 The AI service implements control policies and technology to enforce principles of minimum necessary and HITL to mitigate vulnerabilities due to excessive agency. (OWASP LLM06, ASI01, ASI02, ASI03, ASI07, ASI08, ASI09, T2, T3, T6, T7, T8, T11, T13, AIVSS Core Risks 1, 2, 3, 6, 9, 10)

AI-3 The AI Service Provider has robust AI-specific incident preparedness, detection, and response capability plans in place. (These capabilities are maintained as an overlay to the Service Provider's standard Incident Response controls as listed in the OHIP Baseline Security and Privacy workbooks).

- AI 3-1 The AI Service Provider has implemented formal AI event-specific preparedness processes and activities based on AI industry best practices.
- AI 3-2 The AI Service Provider has robust protocols, procedures, automated monitoring, and human subject matter expertise in place to detect, diagnose, and alert for AI



events.

- AI 3-3 The AI service provider implements and maintains an AI incident response and recovery plan that is well defined, robust, and continually tested and updated.
- AI 3-4 The AI Service Provider maintains an AI Incident Response Team (AIRT) as a layer on top of the provider's standard incident response team.
- AI 3-5 The AI Service Provider maintains an ongoing AI role-appropriate training program for the AIRT team, the AI Incident Management team and other key roles in their organization.
- AI 3-6 The AI Service Provider has robust protocols that mitigate risks to the Department and the populations it serves from Deepfake / Synthetic content Attacks (DFAs) related to:
 - o Applicable vulnerabilities in the AI service / application provided to the Department,
 - o Digital manipulation of content produced by the AI System which is then falsely represented as the Department's official source,
 - o and applicable vulnerabilities in the Service Provider's internal organizational policies and controls.

AI-4 The AI Service Provider ensures that SecOps actions, roles and responsibilities are fully integrated within and between all phases of AI model and AI application lifecycles. Agentic applications implement both baseline and extended controls herein where the baseline controls are applicable. (NOTE: This guiding principle is based on OWASP's lifecycle definition, which is similar, but more granular than the ISO model. The OWASP lifecycle is broken up into the following phases: Scoping / Planning, Data Augmentation / Fine Tuning, Application Development and Experimentation, Testing and Evaluation of the AI Application, Release of the finalized application, Production Deployment of the model and its components, Operation Phase, Monitor Phase and Govern Phase).

Baseline Attribute Standards for SecOps Actions, Roles and Responsibilities throughout the stages of AI model / application lifecycle.

- AI 4-1 The AI Service Provider SecOps activities **Scoping / Planning Phase** of the lifecycle include at the minimum: human and NHI IAM planning, compliance and data privacy assessments, identification of sensitive data, third party data/model provider risk assessment, and threat modeling.
- AI 4-2 The AI Service Provider SecOps activities during the **Data Augmentation / Fine Tuning Phase** of the lifecycle include at the minimum: data source and model integrity validation, secure data and output handling, adversarial testing, malware scanning, and vulnerability assessments.
- AI 4-3 The AI Service Provider SecOps activities during the **Application Development and Experimentation Phase** of the lifecycle encompasses implementing IAM, testing, vulnerability scanning. Activities also include the review of revised model



and application, their security interactions, and underlying coding and coding practices.

- AI 4-4 The AI Service Provider SecOps activities during the **Test and Evaluation Phase** of the lifecycle encompasses comprehensive security testing to identify and resolve any issues before the application is deployed to production. Testing includes at the minimum: adversarial testing, application security orchestration testing, benchmarking, vulnerability testing, penetration testing, SAST, DAST, and IAST.
- AI 4-5 The AI Service Provider SecOps activities during Production Environment **Application Release Phase** of the lifecycle ensure secure transition from development to production. Ensures that monitoring and alerting systems are in place and conducts final security reviews prior to go-live. Activities include, but are not limited to, review of AI BOM, model security posture evaluation, securing CI / CD pipelines, automating static and dynamic code analysis, model training testing, and user access control validation.
- AI 4-6 The AI Service Provider SecOps activities during the model's production **Deployment Release Phase** of the lifecycle ensures secure launch of the model and its associated components into the production environment. Activities include, at the minimum, compliance and deployment validation, secrets management, AI model firewall implementation, securing configuration management, and API access.
- AI 4-7 The AI Service Provider SecOps activities during the **Operation Phase** of the lifecycle encompass managing and maintaining the application in a live production environment, through continuous monitoring of the applications performance. Activities include, at the minimum, implementation and management of automated processes for adversarial attack protection, vulnerability scanning, incident detection, guardrail enforcement, patch management, and securing output / input handling security, and user interactions.
- AI 4-8 The AI Service Provider SecOps activities during the live production runtime **Monitoring Phase** of the lifecycle encompasses tracking key metrics, detecting anomalies, and ensuring the AI model and application components are functioning as expected. Activities include, at the minimum, adversarial input detection, model behavioral analysis and performance, security posture management, and securing model versioning and rollback.
- AI 4-9 The AI Service Provider SecOps activities during the live production runtime **Governance Phase** include, at a minimum, compliance management, data security posture management, incident governance, risk assessment and management.

Advanced Attribute Standards for Securing Agentic AI Systems: Baseline + Expanded SecOps Activities / Roles throughout the stages of Agentic AI lifecycle.

- AI 4-10 In the **Agentic AI Scoping / Planning Phase**, SecOps and DevOps work collaboratively to embed security in the design. Early collaboration aligns agent workflows and tools with enforceable security that focuses on non-human identities (NHIs), agent threat modeling, privilege boundaries, authentication, and memory scoping and isolation.



- AI 4-11 In the **Agentic AI Data Augmentation / Fine Tuning Phase**, SecOps and DevOps work collaboratively before deployment to mitigate risks as described in AI-2, as well as conduct adversarial tuning and reasoning traces. They sanitize data sets, validate alignment, verify authoritative source of log data, and protect sensitive memory with privacy controls.
- AI 4-12 In the **Agentic AI Application Development and Experimentation Phase**, SecOps and DevOps work collaboratively to secure dynamic agent loops, inter-agent communications, and API plugin use. They validate Input / Output (I/O) contracts, embed policy hooks, and test resilience to prevent unsafe behaviors.
- AI 4-13 In the **Agentic AI Test and Evaluation Phase**, SecOps and DevOps work collaboratively to stress-test agentic AI in adversarial conditions, targeting emerging risks like goal drift, prompt injection and tool misuse. In a non-production environment, they run red teaming simulations, sandbox tool / API calls, and validate decisions in multi-agent setups, focusing on behavioral security beyond traditional QA or penetration testing.
- AI 4-14 In the **Agentic AI Production Environment Application Release Phase**, SecOps and DevOps work collaboratively to securely package, validate, and register agentic AI applications. They digitally sign model weights, plugins, and memory to prevent tampering. They verify Software Bills of Materials (SBOMs) for all components, enforce cryptographically validated policies, and register agents in secure capability registries, and ensure trusted, auditable deployments.
- AI 4-15 In the **Agentic AI Deploy Phase**, SecOps and DevOps work collaboratively to ensure all activation of agents occurs in a secure and policy compliant manner. They enforce zero trust communications, rotate ephemeral credentials, set AI model firewalls, access control lists, and apply fine grained authorization so each agent runs with least privilege
- AI 4-16 In the **Agentic AI Operate Phase**, SecOps and DevOps work collaboratively to secure the dynamic footprint of agentic AI, where agents adapt and act in changing environments. SecOps and DevOps monitor memory mutations to prevent drift or poisoning, detect abnormal loops or misuse, enforce HITL overrides, and scan plugins for risks.
- AI 4-17 In the **Agentic AI Monitor Phase**, SecOps and DevOps work collaboratively to monitor agentic AI's dynamic evolving footprint to drive proactive behavior-aware security. Agents produce context-rich traces (e.g., logs) that shift with reasoning and interactions. SecOps utilizes audits of immutable logs to correlate agent steps, tool calls, and inter-agent communications to detect anomalies like goal reversal or adversarial inputs.
- AI 4-18 In the **Agentic AI Govern Phase**, SecOps and DevOps work collaboratively to uphold compliance, access control, and lifecycle governance for evolving agentic AI. They enforce role- and task-based policies, automate agent versioning and retirement, and guard against privilege creep. Immutable logs and audits are used to ensure compliance with regulatory mandates.



AI-5 The AI Service Provider has implemented a holistic approach to AI-specific Red Teaming adversarial testing. Agentic applications implement both baseline and extended controls herein where the baseline controls are applicable.

Baseline Foundational Red Teaming Best Practice Attribute Standards

AI 5-1 Red Teaming governance is mature and integrated within the AI Service Provider organization. It is based on industry best practices that include:

- o cross-team / cross-functional collaboration,
- o ethical and privacy considerations,
- o human oversight,
- o iterative and adaptive processes,
- o KPI testing metrics and benchmarks,
- o transparency and reporting, and
- o ongoing training in emergent AI vulnerabilities and Red Teaming methodologies.

AI 5-2 The AI Service Provider Red Teaming strategy has well defined, adaptive, and documented risk-based testing scopes and objectives.

AI 5-3 The AI Service Provider Red Teaming strategy has well defined, adaptive, and documented test suites, test result evaluation criteria / metrics, security testing tooling, and methodologies.

AI 5-4 The AI Service Provider integrates Red Teaming within ongoing security monitoring as well as early and continually throughout all stages of the development lifecycle. This includes all components supporting the AI system. The AI Service Provider further leverages Red Teaming testing insights to inform updates, model fine-tuning, and implementation of system security measures.

AI 5-5 The AI Service Provider Red Teaming adversarial strategy ensures the following:

- o effective and realistic attackers,
- o automation of repetitive attacker behaviors, and
- o use of manual expert-driven testing / analysis where the team needs to focus on more complex strategic attacker tasks that require nuanced human judgment.
- o Use of realistic simulation and real-world environments.

AI 5-6 The AI Service Provider validates its internal Red Teaming performance and findings with third-party Red Teaming testing.

AI 5-7 Per AI-3, AI Service Provider integrates Red Teaming members and Red Teaming testing into AI Incident preparedness and response planning, plans, exercises, and continual improvement processes.

Baseline Red Teaming AI Testing Attribute Standards: Red Teaming evaluation of the AI Model in Isolation (e.g., MDLC security, model provenance, model malware injection, security of data pipelines used for model training, model robustness, bypass of any model-intrinsic defenses)



- AI 5-8 Red Teaming model testing protocols include inference attacks to exploit outputs of the target model to infer sensitive details about the training data.
- AI 5-9 Red Teaming model testing protocols include extraction attacks to replicate or steal proprietary machine learning model information or its sensitive training data.
- AI 5-10 Red Teaming model testing protocols includes attacks that exploit the fine-tuning stage of AI models to introduce "backdoors" or malicious instructions that can manipulate the model's behavior.
- AI 5-11 Red Teaming model testing protocols include assessment of potential socio-technological harm.
- AI 5-12 Red Teaming model testing protocols include assessing data risk.
- AI 5-13 Red Teaming model testing protocols include assessment of controls in place to ensure model output and behavior align with statutory mandates and Department AI safety requirements, ethical values, intentions, and objectives.
- AI 5-14 Red Teaming model testing protocols include assessment of its resilience to adversarial attacks.
- AI 5-15 Red Teaming model testing protocols include assessment of vectors for technical harm.

Baseline Red Teaming AI Testing Attribute Standards: Red Teaming Evaluation of the AI-Model as Deployed Attribute Standards (e.g., testing for bypass of supporting guardrails, poisoning data used in grounding and testing controls such as model firewalls or proxies)

- AI 5-16 Red Teaming-deployed model testing protocols include assessment of controls over system prompt safety, mitigating against unexpected behavior, or exploitation through prompts.
- AI 5-17 Red Teaming-deployed model testing protocols include assessment of knowledge retrieval security.
- AI 5-18 Red Teaming-deployed model testing protocols include assessment controls over the knowledge base integrity.
- AI 5-19 Red Teaming-deployed model testing protocols include assessment of the model's architecture controls.
- AI 5-20 Red Teaming-deployed model testing protocols include assessment of vulnerabilities leading to bypassing token limitations.
- AI 5-21 Red Teaming-deployed model testing protocols include vulnerabilities leading to bypass of model content filtering.
- AI 5-22 Red Teaming-deployed model testing protocols include testing access controls.

Baseline Red Teaming AI Testing Attribute Standards: Red Teaming zero trust evaluations of AI-Model interactions of other "System" components Attribute Standards. System components that interact with the model are assessed for



vulnerabilities, such as how the model's outputs interact with the broader system. These include supply chain vulnerabilities such as application components used to: train or host models, serve inference points, and store data used to populate prompts with grounded information

- AI 5-23 Red Teaming AI system component testing protocols include assessing vulnerabilities due to remote code execution.
- AI 5-24 Red Teaming AI system component testing protocols include evaluating sandbox escape methods.
- AI 5-25 Red Teaming AI system component testing protocols include assessment of supply chain vulnerabilities.
- AI 5-26 Red Teaming AI system component testing protocols include assessment of potentials for risk propagation paths / chains.
- AI 5-27 Red Teaming AI system component testing protocols include evaluation of system integrity within the context of model outputs.
- AI 5-28 Red Teaming AI system component testing protocols include assessing efficacy of controls over resource utilization.
- AI 5-29 Red Teaming AI system component testing protocols include verification of efficacy of security measures.
- AI 5-30 Red Teaming AI system component testing protocols include AI controls bypass.

Advanced Attribute Standards for Securing Agentic AI Systems: Baseline+ Agentic AI Focused Red Teaming Testing

- AI 5-31 Red Teaming agentic testing protocols identify weaknesses in the authorization, role inheritance, and control mechanisms of AI agents to mitigate risks of control hijacking and unauthorized actions.
- AI 5-32 Red Teaming agentic testing protocols assess the ability of AI agents to ensure human or automated checkers remain actively informed and capable of intervention when system limits are exceeded. The protocols focus on detecting and mitigating unsafe operation.
- AI 5-33 Red Teaming agentic testing protocols assess the resilience of AI agents to manipulation of goals and instructions, preserving their ability to maintain intended behaviors.
- AI 5-34 Red Team agentic testing protocols identify vulnerabilities in AI agents caused by hallucination tendencies, including induced false outputs, decision-making errors, and cascading impacts in multi-agent systems.
- AI 5-35 Red Teaming agentic testing protocols assess the ability of interconnected AI agents and systems to resist cascading compromises, minimize blast radius effects, and contain potential security breaches effectively.
- AI 5-36 Red Team agentic testing protocols assess the resilience of AI agents to knowledge base poisoning attacks by evaluating vulnerabilities in training data,

external data sources, and internal knowledge storage mechanisms.

AI 5-37 Red Teaming agentic testing protocols assess the resilience of AI agents to memory and context manipulation attacks through the identification of vulnerabilities in state management, context persistence, and session isolation mechanisms.

AI 5-38 Red Teaming agentic testing protocols identify vulnerabilities in multi-agent coordination, trust relationships, and communication mechanisms. The objective being to identify risks that could lead to cascading failures or unauthorized operations across interconnected AI agents.

AI 5-39 Red Teaming agentic testing protocols assess the resilience of AI agents to resource and service exhaustion attacks that lead to degraded performance or denial of service.

AI 5-40 Red Teaming agentic testing protocols assess the resilience of AI agents to supply chain and dependency attacks that could lead to unauthorized access, data breaches, or system failures.

AI 5-41 Red Teaming agentic testing protocols assess the traceability and accountability mechanisms of AI agents by evaluating the system's ability to log, monitor, and attribute agent actions accurately.

AI-6 The AI Service provider maintains an ongoing process for AI Data Security Posture Management (AI-DSPM). This includes discovering, classifying, governing, and monitoring data across all AI pipelines throughout the AI system data lifecycle. The purpose of this Guiding Principle is to define extended AI data security implementation standards that ensure that the AI Service provider can identify where sensitive data exists, understand how it flows or where it is derived (e.g., into embeddings or logs), and is able to enforce controls that prevent exposure, tampering, and non-compliance.

AI 6-1 The AI Service has implemented mature AI-DSPM, capabilities that are specifically tailored to its AI architecture (e.g., generative, agentic architecture or hybrid thereof).

DSPM for Mitigating Risks Due to Data Governance Failure(s)

AI 6-2 The AI Service DSPM data governance, lifecycle, traceability and classification controls mitigate against risks to the Department accruing from Service provider governance failures. (OWASP DSGAI07)

AI 6-3 The AI Service DSPM mitigates non-compliance, regulatory violations or reputational damage risks to the Department due to failures in AI Service provider governance. (OWASP DSGAI08).

DSPM for Mitigating Risks Due to Direct Data Exposure Vectors

AI 6-4 The AI Service DSPM mitigates sensitive data leaking from models and retrieval systems. (OWASP DSGAI01)



AI 6-5 The AI Service DSPM mitigates agent credential and identity management failures that open the data plane to unauthorized access. (OWASP DSGAI02)

AI 6-6 The AI Service DSPM mitigates ungoverned data flows from unsanctioned AI use (Shadow AI). (OWASP DSGAI03)

DSPM for Mitigating Threats to Integrity of Data Pipeline

AI 6-7 The AI Service DSPM mitigates upstream compromises that lead to use of the AI data pipeline to propagate downstream attacks. (OWASP DSGAI04)

AI 6-8 The AI Service DSPM ensures that controls are in place at each data ingestion boundary that protect AI pipelines from attack via maliciously crafted payloads. (OWASP DSGAI05)

AI 6-9 The AI Service DSPM mitigates the risks introduced in execution boundaries where plugins, tools, and agents exchange context. (OWASP DSGAI06)

DSPM for Mitigating Threats (Attack Surfaces) unique to Generative AI

AI 6-10 The AI Service DSPM mitigates multimodal capture and cross-channel data leakage (OWASP DSGAI09)

AI 6-11 The AI Service DSPM mitigates inadequate privacy assurance due to weaknesses in synthetic data, anonymization and transformation processes. (OWASP DSGAI10)

AI 6-12 The AI Service DSPM mitigates exploitation of cross-context and multi-user conversation / multi-tenant bleed. (OWASP DSGAI11)

AI 6-13 The AI Service DSPM mitigates natural-language interfaces generating unsafe database queries. (OWASP DSGAI12)

AI 6-14 The AI Service DSPM ensures vector store platform security. (OWASP DSGAI13)

DSPM Coverage of the AI System's Operational Infrastructure

AI 6-15 The AI Service DSPM mitigates data leakage due to excessive telemetry and monitoring (OWASP DSGAI14)

AI 6-16 The AI Service DSPM mitigates risks due to overbroad context windows and "oversharing" of data in prompts. (DSGAI15)

AI 6-17 The AI Service DSPM mitigates risks from local clients through endpoint and browser assistant overreach. (DSGAI16)

AI 6-18 The AI Service DSPM mitigates data availability and resilience failures in AI pipelines related to RAG (DSGAI17)

DSPM Mitigates Threats to the AI Model as a Data Artifact

AI 6-19 The AI Service DSPM mitigates inference and data reconstruction attacks. (OWASP DSGAI18)

AI 6-20 The AI Service DSPM mitigates RLHF and data labelling risks (OWASP DSGAI19)



AI 6-21 The AI Service DSPM mitigates model exfiltration attacks (OWASP DSGAI20)

AI 6-22 The AI Service DSPM mitigates adversarial disinformation introduced through trusted retrieval pipelines. (OWASP DSGAI21)

Overarching AI Security Implementation Standards

Overarching AI Security Implementation Standards		
AI Overarching 1- Implicit / Explicit Existence of Documentation and Artifacts	Background	Many implementation standards herein use / include phrases such as: <ul style="list-style-type: none">• Procedures• Policies• Written standards, AI service provider standards, AI standards• Are defined and documented• Requirements clearly define• Checklists, security requirements, policy guidelines are documented• follows a defined, documented and standards-based• Establishes mandated standards• Specifications, definitions• Tests or testing• Logic flows, permissions, schema, allow lists, input / output validation, configuration(s)
	The Department's Expectations	The Department's expectation for the response to that standard is that: <ul style="list-style-type: none">• The AI service provider has existing mature, detailed and written documentation that covers the specifics of the implementation standard.• The documentation includes a version history that demonstrates the level of maturity.• That the AI service provider will provide that documentation / artifact to the Department in a timely manner upon request.
AI Overarching 2- Responding to the Overall Intent of a Control Attribute and	Background	Many of the implementation standards for control attributes within the AI Baseline Security Technical Guidance Document contain explicit technical terms (e.g., see definitions section AI

/ or Specific Implementation Standard		Baseline Security Technical Guidance Document).
	The Department's Expectations	The AI service provider / responsible entity may believe that the entire control attribute itself does not apply based on the terminology within a specific implementation standard. However, the Department's expectation for the attestation response by the vender / responsible entity to that implementation standard is that it addresses the overall intent of the control attribute. The overall intent of the control attribute is defined in the control attribute's Discussion and Threats / risks mitigated column.
AI Overarching 3-Disclosure of Third-Party Dependencies	Background	AI service providers utilize external Third-Party services or software providers to provide or support components of their application's functionality. These include but are not limited to services, software, data, tools, plugins, models, agents, RLHF, coding/debugging, data labelling, and training or augmentation. The AI service provider must provide full disclosure of all external / third-party dependencies
	The Department's Expectations	<p>The Department's expectations are that:</p> <ul style="list-style-type: none"> • The vendor, service provider or OHIP program is responsible for full and accurate disclosure of all external third-party (external) dependencies that directly or indirectly provide or support components of the vendor's overall AI application / solution. • The Disclosures will include: <ul style="list-style-type: none"> o Name, website, contact information of the Third-Party and contractual relations, if any, to the AI Service provider. o Detailed documentation on what services, functionality and role it plays in supporting the overall AI service provider's application. o The criticality (low, medium, high) of the role that the Third-Party service or software plays in supporting the availability, integrity and security of the overall AI application. o Disclosure as whether the Third-Party service or software either directly or

		<p>indirectly controls, enables or influences access to protected resources (e.g. PHI). This includes sensitive information (e.g., access credentials, Intellectual property, etc.) that could lead to harm to the availability, integrity or security of the overall AI application or it's outputs</p> <ul style="list-style-type: none"> • The Department will, at its discretion, make a determination as to whether third-party service or software will be required to separately address and attest to the applicable controls and implementation standards within this document. (see AI baseline Security Technical Guidance document implementation standards for examples of third-party dependencies.)
--	--	--

Glossary and Definitions of AI Terms

- **Agentic AI** – An agent is an intelligent software system designed to perceive its environment, reason about it, make decisions, and take actions to achieve specific objectives autonomously. An agent or agentic AI system exhibits the following characteristics: Planning and reasoning, retention and recall of information, action, and tool use. Agentic AI applications often integrate with Generative AI.
- **Agent2Agent (A2A) Protocol** – Agent2Agent (A2A) is an open protocol developed by Google that allows different AI agents to interact with each other, securely exchange information, and coordinate actions across different platforms and applications. It connects agents built on different platforms. Agents interact without needing to share internal memory, tools, or proprietary logic. It includes agent discovery, messaging, payload format for requests/responses and enterprise standards for authentication / authorization.
- **Agent Cards** – AI Agent Cards are structured configuration files that describe AI agents, facilitating their discovery and interaction within AI systems. They contain essential metadata about the agent. This includes name and provider, capabilities and skills, API endpoint, authentication mechanism, categorization and compliance / governance.
- **Agent Delegation Chaining** – Agent delegation is where an agent assigns specific tasks, permissions, or decisions to another specialized AI agent. Agent delegation chaining occurs when an AI agent passes tasks or data to another, creating a sequential or collaborative chain to tackle complex problems. Agent delegation chaining patterns include sequential, hierarchical and parallel chaining.
- **Agent Intent Capsules** – It is an agentic AI architecture pattern that is useful in preventing agent goal hijacking or rogue agent behavior. An agent's declared goal, constraints, and user context are cryptographically bound into a signed 'envelope' that follows the (input)

request through the agent's subsequent actions within all stages of its execution cycle. If an agent encounters a malicious prompt, the runtime environment can check the new, resulting, action against the signed Intent Capsule. If the action contradicts the signed intent, the AI system can block the action, preventing unauthorized actions.

- **Agent Persona** – AI Agent persona refers to its defined role, tone, expertise, business profile, system prompts and behavioral patterns.
- **AI Context Window and Tokens** – AI Models do not process words, rather they ‘dissect’ words / phrases into tokens which are then processed. A token can be a character, part of a word, a whole word, or even a short phrase. The **context window** of an AI model defines how much information (measured in *tokens*) the model can hold in its **working memory** at one time while generating a response. It is a short-term memory. The larger the context window, the more past conversation, document content, or code the model can consider before it starts losing track of earlier parts.
- **AIVSS** – OWASP AI Vulnerability Scoring System. It is an OWASP initiative designed to provide a structured and quantifiable methodology to identify, assess, and mitigate vulnerabilities specific to all types of AI systems. The citations used with this version of NYSDOH AI security standards are based on the recent OWASP publication that deals specifically with Agentic AI Core Risks.
- **AI Guardrails** – AI guardrails are safety barriers designed to keep AI systems on track and ensuring they behave responsibly and safely. They act as a set of policies, controls and technical measures that sit between AI model and the user interface. Their primary function is to intercept, block and mitigate risks in real time.
- **AI Incident** – For purposes of this document, an AI incident refers to any event where an AI system's behavior or misbehavior leads to unintended, harmful, or risk-elevating outcomes^{cii}.
- **AI Incident Blast Radius** – The blast radius for an AI incident extends beyond compromised infrastructure to include models, datasets, applications, services, and decisions (e.g. fraudulent transactions approved). Identifying the ‘blast radius’ is a key component of AI incident preparedness and response.
- **AI model or simply “model”** – For purposes of this standards document the term “AI-Model” or just “model” will be used to refer to any AI model that accepts one or more forms of input (e.g., text, images, audio, graphs, plans) and generates one or more forms of new content as output (e.g., text, images, videos, graphs, actions, plans). It is a component of an information system that implements AI technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs.
- **Artificial Intelligence (AI)** – A machine-based system that can identify patterns of data or predict outcomes for a given set of human-defined objectives. Output can include making predictions, learning, reasoning, problem-solving, perception, recommendations, or decisions influencing real or virtual environments.
- **ATLAS** – ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) is a continually updated knowledge base of adversary tactics, techniques and mitigations specifically related to AI-enabled systems. The knowledge set is based on real-world attack observations and realistic demonstrations from AI Red Teamings and security groups. ATLAS was developed by MITRE.
- **ATT&CK and TTPs** – Adversarial Tactics, Techniques, and Common Knowledge

(ATT&CK) are maintained by MITRE. It is an open knowledge base of adversary tactics and techniques based on real-world observations. It is used as a foundation for the development of specific Cyber threat models and methodologies.

TTPs refer to the Adversarial Tactics, Techniques, and Procedures (TTPs) documented within the MITRE ATT&CK framework.

- o Tactics are the high-level goals or tactical objectives of an attacker (e.g., Initial Access, Persistence, Privilege Escalation, or Exfiltration)
 - o Techniques are the specific methods an adversary uses to achieve a tactic. (e.g., “Phishing” to achieve the goal of obtaining “access credentials”)
 - o Procedures are the specific actions performed by an attacker to execute a technique
 - o ATT&CK provides real-world, examples of adversarial usage of these TTPs by threat actors. It also provides defense against them. ATT&CK provides the foundation for ATLAS.
- **Corpora** – In artificial intelligence, a corpus is a large, structured collection of written or spoken texts used to train and evaluate machine learning models. It acts as the foundational dataset that enables AI systems to learn language patterns, context, and semantics from real-world examples. (Corpora is the plural of corpus)
- **Data Provenance** – Data provenance is the process of verifying, tracking and documenting the origin and journey of data, including its sources, how it was transformed, and who or what interacted with it. It provides an audit trail that ensures data integrity, reliability, and trustworthiness, it identifies and verifies the single authoritative sources of data.
- **DeepFake** – Multimedia that have either been created (fully synthetic) or edited (partially synthetic) using some form of machine / deep learning (artificial intelligence) are referred to as deepfakes.
- **DID documents** – Decentralized Identifier (DID) Documents are digitally signed, machine-readable files linked to a unique DID. They contain cryptographic keys, verification methods and service endpoint information that allow agents to prove control of the DID and communicate trustworthily, without needing central identity authorities.
- **Differential Privacy** – Differential Privacy (DP) is an extremely strong definition of privacy that guarantees a boundary on how much an attacker with access to the algorithm output can learn about each individual record in the dataset. Differential privacy works by adding random “noise” to the data in a way that obscures the identity of the individuals but keeps the database useful overall as a source of statistical information.
- **DPIA** – A Data Protection Impact Assessment is a formal process mandated under GDPR Article 35 for high-risk processing operations affecting individual rights and freedoms. The assessment is a systematic evaluation of data-processing activities before implementation, identifying privacy risks and documenting mitigation measures. A DPIA is required whenever processing is likely to result in high risk to the rights and freedoms of individuals. While GDPR does not apply to US implementations, the DPIA concept aligns with the NIST AI RMF and with the required privacy risk modeling requirement in AI 1-13.
- **DP-SGD** – Differentially Private Stochastic Gradient Descent (DP-SGD) has emerged as a popular method to protect personally identifiable information (PII / PHI) in deep learning (DL) applications (e.g. Model training) It is a differential privacy technique that uses statistical sampling methods to introduce noise to increase the privacy of individuals.
- **Excessive Agency / Autonomy** – Autonomy is the level an AI agent’s ability to act

independently of human influence, intervention or oversight.

At a high level, AI agency is the ability of an AI agent to act purposely and make decisions. That is, formulating an intention for an action and to carry out that action. It refers to an agent's ability to perceive its environment, reason, create and execute multi-step plans, and take actions to pursue goals. It further includes making decisions based on its internal processes, 'reflections', and adaptation to new situations via the capacity for self-modification. Taking actions includes the ability to call functions or interface with other systems via extensions (tools, skills or plugins) to undertake actions.

AI agents can accept agency without full autonomy, meaning it can process and act meaningfully while still relying on human guidance for critical decisions.

For purposes of this standards document, Excessive Agency means an AI agent has more permissions, functionality, access to high privilege tools / services, or autonomy than it actually needs to do its job, leading to significant risk. The risks are amplified by an agent's 'internal drive', ability to act on its own initiative, and ability to self-modify to achieve goals.

- **Foundation AI Model(s) and background** – The foundation model term and definition were first coined in 2021 by the Center for Research on Foundation Models (CRFM) at the Stanford Institute for Human-Centered Artificial Intelligence (HAI). The original definition was “any AI model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks”.

Presidential Executive Order (EO) 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 2023². Later refined the definition to be “an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts” ...” and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters”.

The UK Competition and Markets Authority AI Foundation Models initial report 2023/2024^{ciii} states that Foundation AI Models (FMs) are a type of AI technology that are trained on vast amounts of data that can be adapted to a wide range of tasks and operations. Most FMs are currently being developed using a deep learning model called a transformer, first introduced by Google.

The type of data that is used to train a FM determines its 'mode'. For example, large language models (LLMs) are a type of FM trained on text data, and image generation models are trained on image data (coupled with text). A multi-modal FM is a FM that is trained using multiple types of data. Expanding the modality of FMs to include other data such as 3D environments, video and audio is an area of ongoing research.

- **Generative AI** – Generative AI is a class of AI models that emulate the structure and characteristics of input data in order to generate derived, original synthetic content, such as text, images, video, audio or software code. Often this is in response to a user's prompt or request. Often, generative AI relies on the use of large language models to perform the tasks needed to create the new data.
- **GPT** – Generative Pretrained Transformer. A generative pre-trained transformer (GPT) is a

type of large language model (LLM) based on a deep learning architecture called the transformer. They are pre-trained on large datasets of unlabeled content and able to generate novel content.

- **Hallucinations** – AI model outputs that are incorrect, nonsensical, or inconsistent with the retrieved information.
- **HITL** – Human In The Loop (HITL) refers to a system or methodology that integrates human intelligence and decision-making into automated AI processes. In short, it ensures human oversight over AI ‘decision making’ and processes.
- **Large Action Models (LAMs)** – Large Action Models (LAMs) are AI models that handle actions or agents as input or output.
- **Large Language Model (LLM)** – A large language model (LLM) is a language model trained with self-supervised machine learning on a vast amount of text and designed for natural language processing tasks, especially language generation. LLMs are typically generative pre-trained transformers (GPTs), based on a transformer architecture (e.g., ChatGPT). By definition, an LLM is single-modal as it exclusively takes language as input and produces language as output.
- **Large Multi-Modal Model (LMM)** – Large Multi-Modal Models (LMM) are AI models capable of processing and/or generating multiple types of input and output (for example., text, images, audio, graphs, actions, plans).
- **LINDDUN** – LINDDUN (Linking, Identifying, Non-repudiation, Detecting, Data disclosure, Unawareness, and Non-compliance) is a privacy-focused framework for modeling threats to data confidentiality and compliance. (See References Section above).
- **logit** – A logit is the natural log of the odds of binary event occurring and used in logistic regressions. They are typically used as part of a ML classicization process. logits are raw scores of possible classes before they are converted into probabilities used for decision-making.
- **LoRA (Low-Rank Adaptation)** – Fine-tuning large language models is a resource intensive process. LoRA (Low-Rank Adaptation) is a technique that allows fine-tuning of large language models with a small number of parameters. It works by adding and optimizing smaller matrices to the parameter weights, reducing the number of parameters used for training by factors of thousands. This greatly increases efficiency of GPU memory requirements.
- **Machine Learning (ML)** – Machine learning (ML) is a subset of AI that focuses on creating statistical algorithms that can learn from data and generalize to ‘unseen’ data. Machine learning algorithms are trained on a set of data, and then they can use that data to make predictions or decisions about new data.
- **MAESTRO** – MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome), a multilayered threat modeling framework designed specifically for the unique challenges of Agentic AI. The framework was developed by the Cloud Security Alliance (CSA). (See References Section above)
- **MAS** – Multiple Agent Systems (MAS). AI Systems that use multiple AI Agents. Broad categories of MAS architecture patterns include Sequential Agent Architecture, Hierarchical Agent Architecture, Collaborative Agent Swarm, Reactive Agent Architecture and Knowledge-intensive Agent.



- **MCP** – Model Context Protocol (MCP) allows AI agents to access and use external tools and data through a standardized protocol, making them more capable and autonomous. MCP provides the framework and "tool registry" for agents to discover and invoke various functions and data sources.
- **Model Safety Layers** – AI model safety layers are mechanisms designed to ensure that AI systems operate within predefined ethical and functional boundaries. (e.g., input validation, output filtering, contextual awareness, and fine-tuned filters.)
- **Model Tensors** – In machine learning, tensors are the primary data structure for storing and manipulating data. They are multi-dimensional arrays that generalize from scalars, vectors, and matrices to higher dimensions. They can be thought of as containers that store data in N-dimensions.
- **NHI** – Non-Human Identity (NHI). For example, the identity for purposes of IAM of an AI Agent.
- **Nonce** – A number used once. A unique, random, value generated by a verifier and included in a request.
- **OPA** – Open Policy Agent (OPA) is an open-source, general-purpose policy engine that enables unified policy enforcement across the cloud-native stack. It decouples policy decision-making from application logic, allowing policies to be defined, managed, and enforced independently.
- **Orchestrator** – In Multiple Agent Systems (MAS), multiple agents collaborate via an orchestrator (router) agent. The central orchestrator agent is an agent that is responsible for coordinating the agents' interactions, assigning tasks, and maintaining the overall workflow or state. It receives the initial complex task or request, analyzes the task, breaks it down into a series of sub-tasks, identifies which specialized agent is best suited to handle each sub-task and assigns the sub-task accordingly. It monitors agent performance, identifies inefficiencies, and autonomously adjusts workflows to optimize efficiency. Orchestrator can interface with the user directly or with a "master" agent that helps coordinate the various agents.
- **PASTA** – PASTA (Process for Attack Simulation and Threat Analysis) is a seven-stage, risk-centric Threat modeling framework that aligns technical analysis with business impact. (See References Section above)
- **Provenance Software Attestation** – Build **Provenance** software attestation, is where the developer / vender generates a verifiable record of a software's build process, using a framework like SLSA (Supply-chain Levels for Software Artifacts). This involves creating a signed statement containing metadata about the build, such as the source code, build parameters, and dependencies.
- **RAG Corpora** – RAG corpora are used in AI to enhance / augment / ground the accuracy and relevance of model responses by integrating document retrieval into the generation process. This approach allows the AI system to reference a specified set of documents before responding to user queries, which supplements the information from the model's pre-existing training data. The method allows AI models to use domain specific and updated information that is not available to the training data. Examples include accessing organizational policy / regulatory standards or generating responses based on other authoritative sources. RAG improves large model performance by incorporating information retrieval before generating responses. This reduces hallucinations and improving

transparency by including sources to the responses.

- **RAG Triad** – The RAG triad is composed of three RAG evaluation metrics: answer relevancy, faithfulness, and contextual relevancy.
- **Retrieval-Augmented Generation (RAG)** – RAG is used by AI models to generate responses based on context beyond the scope of its training data by supplying it with external data as additional context. It redirects the AI model to retrieve relevant information from authoritative, pre-determined knowledge sources.
- **Prompt** – An artificial intelligence (AI) prompt is input provided to an AI model (e.g. LLM) to guide it in generating a specific response. For example, a question, command, or statement that a person gives to an artificial intelligence model. The prompt provides the AI system with the necessary context or instructions to allow it to produce output that is relevant to what is being asked or requested.
- **Prompt Injection** – Prompt injection is a security vulnerability that targets AI language models (e.g., LLMs). Attackers manipulate the model's behavior by crafting malicious or misleading prompts, including code injection, to bypass safety filters and execute unauthorized / unintended instructions. This can result in data leakage, privilege escalation, unethical outputs.
- **Prompt Jailbreak** – It is the outcome of maliciously crafted input that is designed to bypass AI guardrails of an AI model, allowing the model to perform unauthorized actions or produce unauthorized outputs. It is the successful outcome of prompt injection.
- **Regex-based validation** – Regex-based validation utilizes regular expressions (regex) to ensure that user input or data conforms to a predefined pattern or format
- **RLHF** – Reinforcement Learning from Human Feedback (RLHF) is a technique that aligns AI systems, especially AI language models, with human preferences by using human feedback to guide the AI's behavior. The process involves training a reward model to predict human preferences, then using this reward model in a reinforcement learning loop to fine-tune the AI model to generate responses that receive high rewards. This method improves AI model performance by capturing complex human values and goals, resulting in more helpful, honest, and harmless outputs.
- **RLHF traces** – Traces refer to the data and signals generated throughout the RLHF pipeline. Their validation is crucial for preventing issues like biased outputs, reward hacking, and model misalignment.
- **Safe Model Tensors** – Safe tensors are read-only and separate raw numerical data from any executable code. SafeTensors is a framework / format, developed by HuggingFace, for secure storage of model data. It only stores numerical tensors and metadata, preventing the execution of malicious code during deserialization.
- **SAIF** – Google's Secure AI Framework (SAIF). The SAIF framework and mapping process organizes AI security into four key components: Application, Model, Infrastructure and Data, allowing visibility into attack surfaces and threats across the full AI Architecture and lifecycle. This allows application of threat modeling methodologies for each component / subcomponent throughout that lifecycle. (See References Section above)
- **Small Language Model (SLM)** – Small language models (SLMs) are a subset of language models that perform specific tasks using fewer resources than larger models. They are built with fewer parameters and simpler neural architectures than large language models (LLMs), allowing for faster training, reduced energy consumption, and deployment on

devices with limited resources. The potential limitations of SLMs include a limited capacity for complex language and reduced accuracy in complex tasks. The advantages of using SLMs include lower costs and improved performance in domain-specific applications. An example use case would be embedded applications within a mobile device operating system^{civ}.

- **SPIFFE** – Secure Production Identity Framework For Everyone (SPIFFE) is a specification for creating and managing universal, secure software identities for services in distributed systems, such as microservices and cloud-native environments. It provides a standard way to define, issue, and validate these identities, often using X.509 or JWTs, which are encoded in a SPIFFE Verifiable Identity Document (SVID). The concrete implementation of the SPIFFE specification is called SPIRE (SPIFFE Runtime Environment).
- **STRIDE** – Microsoft's STRIDE modeling framework that categorizes threats into Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege. (See References section above)
- **System Prompts** – System Prompts are instructions used to steer the behavior of the model. They are designed to guide the model's output based on the requirements of the application. They provide context, determine which tools and when to invoke specific tool calls to support the desired outcome of an application.
- **Transformer** – A transformer is a deep learning concept introduced by Google in 2017. It is a neural network architecture in which text is converted to numerical representations called tokens, and each token is converted into a vector via lookup from a word embedding table.
- **Typosquatting** – Typosquatting is a form of cybersquatting that exploits common typos in web site addresses to redirect users to fraudulent or malicious sites.
- **Vectors and Embeddings** – Vector embeddings are a core concept in the functioning of AI models. They are high-dimensional numerical representations that capture the semantic meaning of text, such as words, sentences, or entire documents. These embeddings allow AI models to process and understand language in a mathematical form. Vector embeddings play a crucial role in Retrieval-Augmented Generation (RAG) systems by transforming unstructured data into structured numerical formats, enabling fast and accurate search and retrieval of contextual information, enhancing the performance of language models.

Versioning

Version History			
Version	Publication / Revision Date	Author(s)	Nature of Change



1.0	October 2025	R. Wheeler, T. Brasel, S. Dikeman, I. Gotham	Initial Draft
	October 22 nd 2025	I. Gotham	Added MITRE SAFE-AI and OWASP GenAI Run Book to references section list. Added MITRE SAFE-AI reference to “APIs pose significant risk to API systems”.
2.0	May 6 2026	I. Gotham	Added changes to attribute wording. Changed control family number six from cots products to data security posture management controls. Added more definitions to the glossary.
2.1	June, 2026	R. Wheeler	Accepted all changes, removed all comments and added “Standards” to the title.



ⁱ NIST Artificial Intelligence Risk Management Framework 1.0

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

ⁱⁱ LLM Applications Cybersecurity and Governance Checklist v1.1 – English

<https://genai.owasp.org/resource/llm-applications-cybersecurity-and-governance-checklist-english/#>

ⁱⁱⁱ Appendix B NIST Artificial Intelligence Risk Management Framework 1.0

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

^{iv} NIST AI 100-2e2023 Adversarial Machine Learning. A Taxonomy and Terminology of Attacks and Mitigations

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf>

^v OWASP GenAI Incident Response Guide 1.0

<https://genai.owasp.org/resource/genai-incident-response-guide-1-0/>

^{vi} NIST Artificial Intelligence Risk Management Framework 1.0

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

^{vii} OWASP GenAI Incident Response Guide 1.0

<https://genai.owasp.org/resource/genai-incident-response-guide-1-0/>

^{viii} OWASP LLM and Gen AI Data Security Best Practices

<https://genai.owasp.org/resource/llm-and-gen-ai-data-security-best-practices/>

^{ix} OWASP LLM Applications Cybersecurity and Governance Checklist v1.1 – English

<https://genai.owasp.org/resource/llm-applications-cybersecurity-and-governance-checklist-english/#>

^x State of Agentic AI Security and Governance 1.0

<https://genai.owasp.org/resource/state-of-agentic-ai-security-and-governance-1-0/>

^{xi} New York State Information Technology Policy NYS-P24-001. Acceptable Use of Artificial Intelligence Technologies.

<https://its.ny.gov/system/files/documents/2025/05/nys-p24-001-acceptable-use-of-artificial-intelligence-technologies.pdf>

^{xixii} OWASP LLM Applications Cybersecurity and Governance Checklist v1.1 – English

<https://genai.owasp.org/resource/llm-applications-cybersecurity-and-governance-checklist-english/#>

NIST Artificial Intelligence Risk Management Framework 1.0

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

^{xiii} NIST Artificial Intelligence Risk Management Framework 1.0

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

^{xiv} NIST AI RMF Version 1.0 Playbook.

<https://aicc.nist.gov/airmf-resources/playbook/>

^{xv} NY State technology Law

<https://its.ny.gov/nys-technology-law>

^{xvi} New York State Information Technology Policy NYS-P24-001. Acceptable Use of Artificial Intelligence Technologies.

<https://its.ny.gov/system/files/documents/2025/05/nys-p24-001-acceptable-use-of-artificial-intelligence-technologies.pdf>

^{xvii} New York State ITS policies and standards.

<http://its.ny.gov/eiso/policies/security>

^{xviii} Cloud Security Alliance. Secure Agentic System Design. A Trait-Based Approach.

<https://cloudsecurityalliance.org/artifacts/secure-agentic-system-design/#>

^{xix} CSA Data Security within AI Environments

<https://cloudsecurityalliance.org/artifacts/data-security-within-ai-environments>

^{xx} Cloud Security Alliance. <https://cloudsecurityalliance.org/artifacts/agentic-ai-identity-and-access-management-a-new-approach>

^{xxi} Cloud Security Alliance. MLOps Overview

<https://cloudsecurityalliance.org/artifacts/machine-learning-ops-overview>

^{xxii} Cloud Security Alliance. Agentic AI Red Teaming Guide.

<https://cloudsecurityalliance.org/artifacts/agentic-ai-red-teaming-guide>

^{xxiii} AI Controls Matrix (AICM) Implementation and Auditing Guide

<https://cloudsecurityalliance.org/artifacts/ai-controls-matrix>

^{xxiv} The OAuth 2.1 Authorization Framework draft-ietf-oauth-v2-1-13 May 2025

<https://datatracker.ietf.org/doc/draft-ietf-oauth-v2-1/>

^{xxv} IETF Network Group. AAuth - Agentic Authorization OAuth 2.1 Extension November 2025

<https://www.ietf.org/archive/id/draft-patw-hite-aauth-00.html>

^{xxvi} Huang, K., Narajala, V. S., Habler, I., & Sheriff, A. (2025, May). Agent Name Service (ANS): A universal directory for secure AI agent discovery and interoperability (Internet-Draft No. draft-narajala-ans-00).

Internet Engineering Task Force.



<https://datatracker.ietf.org/doc/draft-narajala-ans/>

^{xxvii} Extending Certificate Enrollment Protocols for Scalable Agentic AI Identity IETF RATS Workgroup June 2025.

<https://www.ietf.org/archive/id/draft-huang-acm-e-scalable-agent-enrollment-00.html>

^{xxviii} MITRE SAFE-AI. A Framework for Securing AI-Enabled Systems

https://atlas.mitre.org/pdf/files/SAFEAI_Full_Report.pdf

^{xxx} NIST Artificial Intelligence Risk Management Framework 1.0

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

^{xxx} NIST AI RMF Version 1.0 Playbook.

<https://airc.nist.gov/airmf-resources/playbook/>

^{xxxii} NIST AI 600-1 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>

^{xxxiii} NIST AI 100-2e2023 Adversarial Machine Learning. A Taxonomy and Terminology of Attacks and Mitigations

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf>

^{xxxiii} NIST SP 800-218A Secure Software Development Practices for Generative AI and Dual-Use Foundation Models

<https://csrc.nist.gov/pubs/sp/800/218/afinal>

^{xxxiv} NIST SP 1270, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.

<https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf>

^{xxxv} NIST SP 800-61r3 Incident Response Recommendations and Considerations for Cybersecurity Risk Management A CSF 2.0 Community Profile

<https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r3.pdf>

^{xxxvi} NIST SP 800-84 Guide to Test, Training, and Exercise Programs for IT Plans and Capabilities

<https://csrc.nist.gov/pubs/sp/800/84/final>

^{xxxvii} NIST AI 100-4 Reducing Risks Posed by Synthetic Content

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf>

^{xxxviii} 2025 OWASP Top 10 for LLM Applications report

<https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/>

^{xxxix} OWASP AI Security and Privacy Guide

<https://owasp.org/www-project-ai-security-and-privacy-guide/>

^{xi} OWASP AI Exchange February 2026. Last Accessed February 2026.

<https://owaspai.org/OWASP-AI-Exchange.pdf>

^{xii} OWASP AI testing Guide.

<https://github.com/OWASP/www-project-ai-testing-guide/tree/main>

^{xlii} OWASP GenAI Run Book

<https://genai.owasp.org/resource/owasp-genai-security-project-threat-defense-compass-runbook/>

^{xliii} OWASP GenAI Security Project Threat Defense COMPASS 1.0

<https://genai.owasp.org/resource/owasp-genai-security-project-threat-defense-compass-1-0/>

^{xliv} OWASP LLM Applications Cybersecurity and Governance Checklist v1.1 – English

<https://genai.owasp.org/resource/llm-applications-cybersecurity-and-governance-checklist-english/#>

^{xlv} OWASP LLM and Gen AI Data Security Best Practices

<https://genai.owasp.org/resource/llm-and-gen-ai-data-security-best-practices/>

^{xlvi} OWASP LLM and Generative AI Security Solutions Landscape – Q1,2025

<https://genai.owasp.org/resource/llm-and-generative-ai-security-solutions-landscape-q12025/>

^{xlvii} The OWASP Generative AI and Agentic Security Solutions Landscape (Q2–Q3 2025)

<https://genai.owasp.org/resource/owasp-genai-security-project-solutions-reference-guide-q2-q325/>

^{xlviii} OWASP LLM Prompt Injection Prevention Cheat Sheet

https://cheatsheets.owasp.org/cheatsheets/LLM_Prompt_Injection_Prevention_Cheat_Sheet.html

^{xlix} A Practical Guide for Securely Using Third-Party MCP Servers

<https://genai.owasp.org/resource/cheatsheet-a-practical-guide-for-securely-using-third-party-mcp-servers-1-0/>

ⁱ OWASP A Practical Guide for Secure MCP Server Development.

<https://genai.owasp.org/resource/a-practical-guide-for-secure-mcp-server-development/>

ⁱⁱ OWASP Top 10 for Agentic Applications for 2026

<https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/>

ⁱⁱⁱ OWASP Top 10 Non-human Identities Risks 2025

<https://owasp.org/www-project-non-human-identities-top-10/2025/top-10-2025/>

^{iv} OWASP AI Vulnerability Scoring System (AIVSS)

<https://aivss.owasp.org/>

^v AIVSS Scoring System For OWASP Agentic AI Core Security Risks

<https://aivss.owasp.org/assets/publications/AIVSS%20Scoring%20System%20F%20or%20OWASP%20Agentic%20AI%20Core%20Security%20Risks%20v0.5.pdf>

<https://aivss.owasp.org/assets/publications/AIVSS%20Scoring%20System%20F%20or%20OWASP%20Agentic%20AI%20Core%20Security%20Risks%20v0.8.pdf>



^{lv} OWASP GenAI Incident Response Guide 1.0

<https://genai.owasp.org/resource/genai-incident-response-guide-1-0/>

^{lvi} OWASP Guide for Preparing and Responding to Deepfake Events

<https://genai.owasp.org/resource/guide-for-preparing-and-responding-to-deepfake-events/>

^{lvii} OWASP GenAI Red Teaming Guide

<https://genai.owasp.org/resource/genai-red-teaming-guide/>

^{lviii} OWASP Agentic AI – Threats and Mitigations

<https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/>

^{lix} OWASP Agentic AI – Threats and Mitigations Version 1.1

<https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/>

^{lx} OWASP State of Agentic AI Security and Governance 1.0

<https://genai.owasp.org/resource/state-of-agentic-ai-security-and-governance-1-0/>

^{lxi} OWASP Securing Agentic Applications Guide 1.0

<https://genai.owasp.org/resource/securing-agentic-applications-guide-1-0/>

^{lxii} OWASP Multi-Agent System (MAS) Threat Modeling Guide

<https://genai.owasp.org/resource/multi-agent-system-threat-modeling-guide-v1-0/>

^{lxiii} OWASP AI Security Solutions Landscape for Agentic AI Q3 2025

<https://genai.owasp.org/resource/ai-security-solutions-landscape-for-agentic-ai-q3-2025/>

^{lxiv} OWASP Agent Name Service (ANS) for Secure AI Agent Discovery v1.0

<https://genai.owasp.org/resource/agent-name-service-ans-for-secure-ai-agent-discovery-v1-0/>

^{lxv} OWASP AI Bill Of materials (AIBOM) Project.

<https://owaspai bom.org/>

^{lxvi} OWASP CycloneDX (ECMA-424)

<https://owasp.org/www-project-cyclonedx/>

<https://cyclonedx.org/>

^{lxvii} **OWASP Software Component Verification Standard (SCVS) for identifying and reducing risk in a software supply chain.**

<https://owasp.org/www-project-software-component-verification-standard/>

^{lxviii} OWASP Vendor Evaluation Criteria for AI Red Teaming Providers and Tooling

<https://genai.owasp.org/resource/owasp-vendor-evaluation-criteria-for-ai-red-teaming-providers-tooling-v1-0/>

^{lix} OWASP GenAI Data Security and Mitigations 2026

<https://genai.owasp.org/resource/owasp-genai-data-security-risks-mitigations-2026/>

^{lxx} OASIS Coalition for Secure AI (CoSAI). Incident Response Framework, V1.0

<https://github.com/cosai-oasis/ws2-defenders/blob/main/incident-response/AI%20Incident%20Response.md>

see also.

<https://www.coalitionforsecureai.org/defending-ai-systems-a-new-framework-for-incident-response-in-the-age-of-intelligent-technology/>

^{lxxi} OASIS Collaborative Automated Course of Action Operations (CACAO) for Cybersecurity. Security Playbooks Version 2.

<https://docs.oasis-open.org/cacao/security-playbooks/v2.0/cs01/security-playbooks-v2.0-cs01.html>

^{lxxii} MITRE Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS)

<https://atlas.mitre.org>

^{lxxiii} MITRE Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK®)

<https://attack.mitre.org>

^{lxxiv} MITRE Common Attack Pattern Enumerations and Classifications (CAPEC)

<https://capec.mitre.org/>

^{lxxv} MITRE Group D3FEND Cyber Security Countermeasures.

<https://d3fend.mitre.org/>

^{lxxvi} Secure AI Framework (SAIF)

<https://cloud.google.com/use-cases/secure-ai-framework#secure-ai-framework-saif-and-google-cloud>

See also OWASP Security Testing Guide. Threat Modeling for AI Systems

<https://github.com/OWASP/www-project-ai-testing-guide/blob/main/Document/content/2.0 Threat Modeling for AI Systems.md>

^{lxxvii} UcedaVélez, T., & Morana, M. M. (2015). Risk Centric Threat Modeling: Process for Attack Simulation and Threat Analysis (PASTA). Wiley. ISBN 978-1118810040. Available at

<https://www.wiley.com/en-us/Risk+Centric+Threat+Modeling%3A+Process+for+Attack+Simulation+and+Threat+Analysis-p-9780470500965>

<https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118988374.fmatter?msocid=390a85d667de687e298e97b9660a69ca>

See also OWASP Security Testing Guide. Threat Modeling for AI Systems

<https://github.com/OWASP/www-project-ai-testing-guide/blob/main/Document/content/2.0 Threat Modeling for AI Systems.md>

^{lxxviii} Microsoft's STRIDE modeling framework.

<https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool-threats>

See also

OWASP Security Testing Guide. Threat Modeling for AI Systems

<https://github.com/OWASP/www-project-ai-testing-guide/blob/main/Document/content/2.0 Threat Modeling for AI Systems.md>



^{lxix} LINDDUN (Linking, Identifying, Non-repudiation, Detecting, Data disclosure, Unawareness, and Non-compliance)

<https://linddun.org>

See

Robust and reusable LINDDUN privacy threat knowledge. L. Sion, D. Van Landuyt, K. Wuyts and W. Joosen. Computers and Security Vol. 154. 104419. 2025

<https://www.sciencedirect.com/science/article/abs/pii/S0167404825001087?dgcid=author>

<https://doi.org/10.1016/j.cose.2025.104419>

see also

OWASP Security Testing Guide. Threat Modeling for AI Systems

https://github.com/OWASP/www-project-ai-testing-guide/blob/main/Document/content/2.0_Threat_Modeling_for_AI_Systems.md

^{lxxx} Cloud Security Alliance. K. Huang 2025. Agentic AI Threat Modeling Framework: MAESTRO

<https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro>

See also

OWASP Multi-Agentic System (MAS) Threat Modeling Guide

<https://genai.owasp.org/resource/multi-agentic-system-threat-modeling-guide-v1-0/>

^{lxxxi} Practical Library of Threats 4 Artificial Intelligence (PLOT4AI)

<https://plot4.ai/>

^{lxxxii} L. Kastner and B. Crook. Explaining AI through mechanistic interpretability.

European Journal for Philosophy of Science (2024) 14:52

<https://link.springer.com/article/10.1007/s13194-024-00614-4>

^{lxxxiii} L. Bereska and E. Gavve. Mechanistic Interpretability for AI Safety A Review.

arXiv:2404.14082v1 [cs.AI] (22 Apr 2024)

<https://arxiv.org/html/2404.14082v1>

^{lxxxiv} Agent2Agent (A2A) Protocol

<https://a2a-protocol.org/latest/>

^{lxxxv} Model Context Protocol

<https://modelcontextprotocol.io/specification/2025-11-25>

^{lxxxvi} O. Santos A New Identity Framework for AI Agents (Cisco) May 2025. <https://community.cisco.com/t5/security-blogs/a-new-identity-framework-for-ai-agents/ba-p/5294337>

^{lxxxvii} K. Feng, D. McDonald and A. Zheng. Levels of Autonomy in Agents. Cornell University arXiv 2025

<https://arxiv.org/abs/2506.12469>

See also

<https://knights.columbia.org/content/levels-of-autonomy-for-ai-agents-1>

^{lxxxviii} S. Pant and M. Viswanathan. Five levels of AI Intelligence for Enterprise Use. Cisco Outshift 2025

<https://outshift.cisco.com/blog/agentic-ai-intelligence-for-enterprise-use>

^{lxxxix} GAO-24-107292 Combating Deepfakes 2024

<https://www.gao.gov/products/gao-24-107292>

^{xc} GAO-20-379SP Deepfakes 2020

<https://www.gao.gov/products/gao-20-379sp>

^{xci} N. Kobis, B. Doležalová and I. Soraperra. Fooled twice: People cannot detect deepfakes but think they can. 2021 iScience 24:11 103364

<https://www.cell.com/action/showPdf?pii=S2589-0042%2821%2901335-3>

^{xcii} NSA, FBI, and CISA Release Cybersecurity Information Sheet on Deepfake Threats. 2023. Last Accessed February, 2025.

<https://www.cisa.gov/news-events/alerts/2023/09/12/nsa-fbi-and-cisa-release-cybersecurity-information-sheet-deepfake-threats>

^{xciii} Contextualizing Deepfake Threats to Organizations. 2023. Last Accessed February, 2025.

<https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF>

^{xciv} Center for Content Provenance and Authenticity. Content Credential Provenance and Authenticity Standards.

<https://c2pa.org/>

^{xcv} Content Credentials: Strengthening Multimedia Integrity in the Generative AI Era. 2025. Last Accessed February, 2025.

<https://media.defense.gov/2025/Jan/29/2003634788/-1/-1/0/CSI-CONTENT-CREDENTIALS.PDF>

^{xcvi} DHS Public-Private Analytic Exchange Program, Increasing Threat of Deepfake Identities, 2026. Last accessed February 2026

https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

^{xcvii} DHS S&T Digital Forgeries Report, Technology Landscape Assessment. 2023 Last accessed February 2026.

https://www.dhs.gov/sites/default/files/2023-06/23_0630_st_digital_forgeries_report_signed.pdf

^{xcviii} Impacts of Adversarial Use of Generative AI on Homeland Security 2025. Last accessed February 2026

https://www.dhs.gov/sites/default/files/2025-01/25_0110_st_impacts_of_adversarial_generative_ai_on_homeland_security_0.pdf

^{xcix} Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study .2023. Last accessed February 2026

https://www.dhs.gov/sites/default/files/2023-12/23_1222_st_risks_mitigation_strategies.pdf

^c S. V. Chilukoti et al. Auto DP-SGD: Dual Improvements of Privacy and Accuracy via Automatic Clipping Threshold and Noise Multiplier Estimation. 2023

<https://arxiv.org/html/2312.02400v1>



**Department
of Health**

**Office of
Health Insurance
Programs**

^{ci} MITRE SAFE-AI. A Framework for Securing AI-Enabled Systems. APPENDIX C. AI THREATS, CONCERNS, AND RESIDUAL RISK

https://atlas.mitre.org/pdf-files/SAFEAI_Full_Report.pdf

^{cii} OWASP GenAI Incident Response Guide 1.0

<https://genai.owasp.org/resource/genai-incident-response-guide-1-0/>

^{ciii} Competition and Markets Authority AI Foundation Models: Initial report 023

<https://www.gov.uk/government/publications/ai-foundation-models-initial-report>

Competition and Markets Authority AI Foundation Models: Technical Update Report 2024

https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf

^{civ} Microsoft's Definition of Small Language Models (SMLs)

<https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-are-small-language-models>