Health Sector Coordinating Council
Cybersecurity Working Group

Monitor Threats | Manage Risks | Measure Effectiveness | Respond & Recover | Secure Medtech

Health Industry Cybersecurity

Health Industry AI Cyber Governance Framework Implementation Guide

MAY 2026

Table of Contents

Executive Summary

This guide empowers healthcare organizations (HCOs) to establish cyber governance frameworks for secure AI implementation. It addresses unique cybersecurity and privacy challenges as the sector adopts artificial intelligence across clinical and operational use cases (i.e. electronic health records, diagnostics, decision support, etc.). The framework targets the identification and mitigation of AI-specific cyber risks, including data poisoning, model drift, and adversarial attacks, while ensuring compliance with the healthcare sector's complex regulatory environment. It addresses the full spectrum of AI technologies deployed in healthcare, from traditional machine learning/reactive/non-agentic models to generative AI, and agentic AI systems capable of autonomous action. Each technology category addresses distinct cyber risk issues requiring governance oversight and controls.

The framework establishes core AI Cybersecurity Governance objectives for enterprise and ecosystem, or third-party adoption scenarios. It provides AI Cyber-specific industry best practices and practical tools for tasks such as organizing roles and responsibilities, inventory management, contractual language for vendor relationships, a five-level AI autonomy framework adapted to healthcare contexts, and AI-specific incident response playbook.

The guide also addresses AI supply chain and concentration risk, operational resilience for AI-dependent clinical workflows, non-human identity management, patient engagement and transparency obligations, liability and insurance considerations, and governance requirements for research AI.

Important Scope Note and How to Use This Document

This document focuses on the cybersecurity dimensions of AI governance: protecting AI systems from adversarial threats, ensuring data integrity and privacy, securing the AI supply chain, and maintaining operational resilience. Topics such as clinical safety, ethics, and patient engagement are addressed to the extent that they intersect with cybersecurity risk. Organizations should maintain a broader AI governance program that addresses the full spectrum of AI risks beyond cybersecurity in the ever-changing ecosystem.

This document complements other HSCC AI-specific publications and should be considered as part of a larger volume of work developed to guide the health industry in its safe and secure adoption of AI. Subsequently, this document will significantly reference the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide and should be used in conjunction with this publication.

This document uses terms like "Must", "Should" and "May". "Must" indicates a baseline requirement. "Should" indicates a strongly recommended practice. "May" indicates an optional enhancement. Adopting organizations should translate all "Must" language into enforceable internal requirements.

Disclaimer

This document is provided for educational informational purposes only. Use of this document is not required and does not guarantee compliance with federal, state, or local laws. The information presented may not be applicable to or appropriate for all health sector organizations. This document is not intended to be an exhaustive or definitive source for safeguarding health information from privacy and security risks. The advice and template materials

provided in this guide are neither intended nor offered as legal advice or legal opinions. HSCC-CWG and the authors are not practicing attorneys. The reader should neither act nor fail to act on any legal matter based upon the information or advice provided in this document without first engaging a competent attorney licensed to practice law in their state or territory.

About the Health Sector Coordinating Council Cybersecurity Working Group

The Healthcare and Public Health Sector Coordinating Council (HSCC) is a coalition of private-sector critical healthcare infrastructure entities organized under a national public-private partnership framework to advise the government in the identification and mitigation of strategic threats and vulnerabilities facing the sector's ability to deliver services and assets to the public. The HSCC Cybersecurity Working Group (CWG) is the largest HSCC working group of almost 500 healthcare providers, pharmaceutical and medtech companies, payers and health IT entities partnering with government to identify and mitigate cyber threats to health data and research, systems, manufacturing and patient care. The CWG membership collaboratively develops and publishes freely-available healthcare cybersecurity best practices, policy/procedure recommendations, and outreach/ communications programs emphasizing the imperative that cyber safety is patient safety.

The AI Cyber Governance Framework Implementation Task Group is one of several task groups established under the HSCC CWG. It is charged with 1) identifying emerging risks associated with artificial intelligence and machine learning (AI/ML) products and services used in health and public health (HPH) and 2) developing guidelines, standards, best practices and mitigation recommendations for AI safety and security. This effort aligns with Implementing Objectives 6 and 8 of the HSCC CWG's Health Industry Cybersecurity Strategic Plan 2024-29.

The purpose of this document is to address the critical need to establish and operationalize comprehensive cyber governance frameworks for cybersecurity governance policy, regulatory alignment, privacy compliance, cross-border data control, and clinical use oversight, fostering responsible, safe, secure, and effective AI adoption. As AI is increasingly used in most clinical, operational and administrative functions in healthcare, this resource applies broadly to many of the 36 HSCC CWG best practices published *by the sector for the sector* since 2019.

Foreword from the AI Cyber Governance Task Group Co-Leads

The health sector's rapid adoption of AI introduces new and often poorly understood risks into an already complex ecosystem. From diagnostic algorithms and clinical decision support to revenue cycle automation and drug discovery, AI systems are increasingly embedded in critical healthcare functions. These systems often rely on opaque models, shared datasets, and third-party application programming interfaces (APIs), creating vulnerabilities where attackers can exploit behaviors or cause failures that impact care. Organizations should be able to identify and mitigate these new and diverse vulnerabilities

The Health Industry AI Cyber Governance Framework Implementation Guide delivers clarity, structure, and general industry practices to help healthcare leaders address this challenge head-on. This Guide builds upon proven models like the HSCC-HHS joint Health Industry Cybersecurity Practices (HICP) and aligns with federal and industry AI

cyber risk management frameworks, adapted specifically for the healthcare sector. It empowers organizations of all sizes—from resource-constrained rural hospitals to academic medical centers—to assess and mitigate cyber risk in AI-enabled systems in the context of cybersecurity, privacy, clinical safety, and operational resilience.

Acknowledgments

The AI Governance Task Group Co-Leads are grateful for the significant investment of personal time by all the authors and contributors of this document. The authors represent some of the most skilled and experienced experts across the healthcare ecosystem, and this document would not have been possible without their generosity, leadership, and commitment - and the support of their employers - to a more secure health sector.

We are grateful for the leadership and editorial skills of Greg Garcia, Executive Director of the HSCC-CWG and the operational support of Allison Burke.

While many individuals assisted in the development and review of this content, the primary authors across this document and version were:

Co-Leads

Ed Gaudet, Censinet

Bill Reid, Google

Samantha Jacques, McLaren Health

Contributors

Edison Alvarez, BD

Mari Savickis, CHIME

Jonathan Almassi, Columbia Memorial Hospital

Pranay Mehta, Headspace

Brindusa Curcaneanu, (previously) NeuroPace

Preethi Amurthur, Philips

Christine Sublett, Sublett Consulting LLC.

Governance Foundations

Overview

AI Governance spans a comprehensive framework of policies, processes, standards, and oversight mechanisms that guide the responsible and safe development, deployment, and management of AI systems. It systematically ensures that AI technologies align with organizational mission, regulatory requirements, and ethical principles while managing associated risks and maximizing benefits. It addresses questions about whether an organization should use an AI system and whether it is making fair and accurate decisions for patients.

While this Guide focuses on governance considerations for AI-specific cybersecurity and privacy risks, health industry leaders and practitioners should have a general understanding of the larger governance context and risks associated with AI adoption. As such, this section is designed to highlight a few of these sector challenges.

Unique Sector Challenges

AI governance in healthcare differs fundamentally from that of other sectors due to the life-and-death nature of medical decisions and the complex regulatory environment governing patient care.

These are some examples of the challenges that the healthcare sector faces with the use of AI:

1. Patient Safety: Algorithmic errors, technical failures, and systemic vulnerabilities can directly impact patient outcomes. Governance frameworks must prioritize patient safety above all other considerations.
2. Regulatory Complexity: HCOs operate under a distinctive framework including the Health Insurance Portability and Accountability Act of 1996 (HIPAA) (privacy/security), Food and Drug Administration (FDA) (medical device regulations), Centers for Medicare & Medicaid Services (CMS) (i.e. quality reporting), and state licensing and privacy requirements. Governance must ensure compliance across multiple jurisdictions simultaneously.
3. Clinical Validation: Governance must ensure AI systems are not only technically sound but clinically appropriate. This requires establishing clinical evidence standards and maintaining clinical oversight of AI recommendations, in alignment with FDA guidance.
4. Resource Disparities: Frameworks must be scalable to support resource-constrained rural hospitals and community centers, not just large integrated delivery systems and academic medical centers.
5. Ethical Implications: Governance must address health equity, algorithmic bias, and informed consent, ensuring AI improves outcomes for all patient populations.

Effective AI Governance transforms AI from a technical challenge into a strategic asset. Fundamentally, it establishes clear accountability, decision-making processes, and control mechanisms across the entire AI lifecycle—from initial needs justification and procurement, through development, deployment, and maintenance, to final decommissioning. By integrating technical, operational, legal, and ethical considerations into a cohesive framework, it effectively addresses critical areas like data quality, algorithmic fairness, system security, and explainability.

Without proper AI governance, AI systems can leak data, disrupt operations, perpetuate biases, adversely affect populations, or fail catastrophically—ultimately compromising patient care, causing direct harm, and damaging organizational reputation.

Defining AI Cyber Governance

AI Cyber Governance refers to the subset of AI Governance focused specifically on the protection, resilience, and secure operation of AI systems throughout their lifecycle. It encompasses the policies and oversight mechanisms that safeguard AI against cyber threats, adversarial attacks, and data breaches.

While AI Cyber Governance is a component of overall AI Governance, it is more narrowly focused on whether AI systems are secure, resilient, and trustworthy. It builds stakeholder trust, which is fundamental to the successful adoption of AI technologies in healthcare and should consider the following:

- Complexity: AI models can exhibit unpredictable behaviors, degrade in performance due to data or concept drift, or remain vulnerable to a myriad of different attacks that may affect the functioning of the model/product.
- Compliance: The regulatory landscape is rapidly evolving. Organizations without robust governance risk non-compliance with emerging regulations and legal liability.
- Strategic Value: Governance provides the foundation for sustainable AI adoption that supports long-term objectives rather than creating technical debt.
- Trust: Governance builds stakeholder trust, which is fundamental to the successful adoption of AI technologies in healthcare.
- Safety: Governance ensures that systems minimize the risk of harm to patients and caregiving staff.

Effective AI Cyber Governance integrates cybersecurity principles into the assessment, design, development, deployment, and decommissioning of AI systems. It establishes protocols for secure data handling, model protection, threat detection, and continuous monitoring of vulnerabilities such as model evasion, model inversion, data leakage, and data poisoning.

Just as cybersecurity is a shared responsibility that healthcare providers and vendors including device manufacturers bear, so too is cybersecurity of AI tools. Further, coordination among discrete components within a healthcare provider setting are needed to ensure holistic responsibility when those units own responsibility for different tools.

Integrating AI Governance with Existing Organizational Governance

Healthcare organizations operate within established governance structures including Medical Staff Committees, Pharmacy and Therapeutics (P&T) Committees, Clinical Ethics Committees, Institutional Review Boards (IRB), Quality and Patient Safety Committees, IT Governance, and Compliance Committees. AI governance should integrate with these existing bodies.

Integration Models by Organizational Size

Small organizations (under 200 beds, critical access hospitals, community health centers): AI governance responsibilities should be assigned to existing committees. The Quality/Patient Safety Committee or Compliance Committee may absorb AI governance review functions. A designated AI governance liaison (which may be a part-time role assigned to an existing leader such as the CIO, CISO, Compliance Officer, or CMIO) should coordinate across committees. Formal AI Governance Committee formation may be recommended only when the AI portfolio exceeds a threshold the organization defines (e.g., five or more active AI systems, or any High/Critical risk system).

Medium organizations (200–500 beds, community hospitals, small health systems): A standing AI Governance Subcommittee reporting to an existing oversight body (IT Governance, Quality Committee, or Compliance Committee) should be established. This subcommittee should include clinical, security, privacy, and legal representation at minimum. The subcommittee conducts initial review of AI proposals and escalates High/Critical risk decisions to the parent committee or executive leadership.

Large organizations (500+ beds, integrated delivery networks, academic medical centers): A dedicated AI Governance Committee with formal charter, defined decision rights, and direct reporting to the Board or a Board-level committee is appropriate. This committee should operate with subcommittees or working groups for clinical AI review, cybersecurity review, ethics review, and vendor/supply chain review. Liaison relationships with other committees is required (i.e., Medical Staff, IRB, P&T, Patient Safety) to ensure bidirectional communication.

Regardless of size, the following integration principles apply:

- AI governance decisions that affect clinical care must include clinical representation with appropriate authority (not merely advisory input).
- AI governance decisions with cybersecurity implications must include the CISO or security designee.
- AI governance decisions involving PHI/PII must include the Privacy Officer and/or compliance.
- Escalation paths from AI governance to the Board must be documented and tested.
- Existing committee charters should be updated to reflect AI governance responsibilities. A model committee charter is provided in Appendix J.

Developing the AI Governance Framework

Before AI Cyber Governance can be implemented, an AI Governance system must be in place. An effective AI Governance system creates a closed loop of accountability, including policies, risk assessment processes, security controls, documentation, and continuous monitoring. Yet, as is apparent with HIPAA Covered Entities and their Business Associates, achieving good governance requires sharing responsibility among many parties. A sample AI governance policy structure (including minimum lifecycle controls and third-party AI expectations) is provided in Appendix E.

Board-Level Oversight

Boards bear fiduciary and ethical responsibility for AI deployment. They should receive regular briefings to learn about AI, on AI cyber risk posture, regulatory trends, and incident reports. Annual attestation to AI cyber governance policies may be included in corporate compliance statements.

AI Risk Appetite Statement: Boards should formally define the organization's risk appetite for AI deployment. This statement should address at minimum: the maximum AI autonomy level (see Appendix D) permitted for clinical decision-making without specific Board awareness, the conditions under which the organization will deploy AI that directly influences treatment decisions, the organization's position on use of AI in life-critical systems, and

acceptable thresholds for AI-related patient safety events. The risk appetite statement should be reviewed annually and updated when the organization's AI portfolio materially changes.

Board Reporting: The AI Governance Committee (or equivalent) should provide the Board with structured AI risk reports at least quarterly. Reports should include: the current AI inventory count and growth trend by risk tier, the distribution of AI systems across autonomy levels, a summary of AI-related incidents (including near-misses) and their resolution, the current maturity score across all three governance objectives, regulatory compliance status and emerging regulatory risks, vendor concentration analysis (how many clinical AI systems depend on the same foundation model or vendor), and material changes to the AI risk posture since the last report. A board reporting template is provided in Appendix K.

Board Education: Board members should participate in annual AI governance education covering: the organization's current AI portfolio and its clinical and operational impact, AI-specific risk categories (adversarial attacks, data poisoning, model drift, hallucination, prompt injection), regulatory trends and enforcement actions, liability and insurance exposure, and fiduciary responsibilities specific to AI oversight. Education should be documented and tracked.

Director and Officer Liability: Board members should understand that failure to exercise reasonable oversight of AI governance may create personal liability exposure, particularly as regulatory enforcement of AI in healthcare matures. Legal counsel should brief the Board annually on the evolving AI liability landscape.

AI Cyber Governance Committee

A multidisciplinary AI Cyber Governance Committee ensures AI systems are secure, resilient, and trustworthy. Membership may include the following roles, depending on the size of the organization:

- **AI Executive Sponsors / AI Program Leads** (for example CIO, CTO, CMIO, Chief Data Officer, VP Analytics, or equivalent): Set direction, allocates resources, and ensures accountability for AI governance decisions. In small and medium hospitals, this role may be assigned to an existing executive leader rather than a dedicated Chief AI Officer position.
- **Physician Leaders** (i.e., Chief Medical Officer (CMO), Chief Nursing Officer (CNO)): Bridge clinical requirements, data science, and clinician adoption. Validates clinical relevance, participates in algorithm review panels, and monitors post-deployment outcomes.
- **IT and Security Teams (CIO and CISO)**: Responsible for architecting and managing the systems that handle information across the organization, including their interactions with AI systems. These teams anticipate AI-specific threats like data poisoning and adversarial attacks and ensure overall infrastructure resilience.
- **Clinical Engineering Teams**: Manage AI-enabled medical devices (in conjunction with IT).
- **Legal/Compliance**: Interpret regulations, update policies, and collaborate with risk management to quantify AI exposures.
- **Privacy Leadership (i.e., Chief Privacy Officer)**: Ensures privacy compliance for AI systems involving ePHI or PII including appropriate use, sharing, and required privacy impact assessments.

- **Patient Advocates**: Represent patient interests in the use of AI systems on patient facing and patient impacting systems.
- **Medical Informatics/Clinical Decision Support**: Oversees the systems that capture and manage clinical information within the organization, where there may be integration with AI systems.

Additional Recommended Roles:

- **Data Scientists / ML Engineers**: Provide technical expertise on model behavior, performance characteristics, and failure modes. Essential for informed governance decisions on model risk.
- **Bioethicist or Ethics Designee**: Distinct from legal/compliance, provides structured ethical analysis for complex AI applications (e.g., predictive models for end-of-life, resource allocation algorithms). In small organizations, this role may be filled by the Clinical Ethics Committee chair.
- **Patient Safety Officer**: Ensures AI governance decisions are informed by the organization's patient safety program and that AI-related adverse events are captured in existing safety reporting systems.
- **Supply Chain / Procurement**: Given the emphasis on vendor risk and AI supply chain governance, procurement leadership should participate in or be consulted on AI acquisition decisions.
- **Revenue Cycle / Finance Leadership**: For organizations deploying AI in coding, billing, claims processing, or prior authorization, finance representation ensures governance addresses revenue cycle integrity and compliance.

Larger organizations may decide to break up a large committee into an executive committee and a general membership committee to ensure proper representation yet still be able to govern effectively.

Committee Operating Model

AI Governance Committees should operate under a formal charter that defines: the committee's authority and scope (advisory vs. decision-making); membership requirements and quorum; meeting cadence (recommended monthly for standing meetings, ad hoc for urgent reviews); escalation criteria to executive leadership and the Board; the process for resolving conflicts when clinical and cybersecurity recommendations diverge; documentation and recordkeeping requirements; and annual self-assessment of committee effectiveness. A model charter template is provided in Appendix J.

When clinical benefit and cybersecurity risk recommendations conflict (e.g., a clinically valuable AI tool has unresolved security concerns), the committee should apply the following hierarchy: (1) patient safety is the paramount consideration; (2) unmitigated cybersecurity risks that could compromise patient safety are disqualifying regardless of clinical benefit; (3) for risks that do not directly threaten patient safety, the committee should document a risk acceptance decision with compensating controls and a defined remediation timeline; (4) the CISO, CMO/CMIO, and Privacy Officer each hold escalation authority to elevate unresolved disagreements to executive leadership. Ultimately a benefit-risk framework that considers patient safety, intended use, available compensating controls, clinical necessity, and applicable regulatory status, including FDA authorization, where relevant, can help organizations resolve conflicts.

Regulatory Compliance and Legal Framework

Under U.S. law, Healthcare AI must comply with the HIPAA Privacy and Security Rules (FD&C Act, Section 524B) regarding Protected Health Information (PHI). Below are other possible regulatory compliance requirements.

- **ASTP/ONC**: Have required transparency for decision support interventions (DSI) through regulations like HTI-1 (though HHS has proposed removing some of these requirements).
- **FDA**: Regulates medical device software functions (previously known as Software as a Medical Device (SaMD)) and AI-enabled medical devices, including AI embedded within hardware devices. FDA permits certain post-market updates to AI models through authorized Predetermined Change Control Plans (PCCPs), which allow predefined, validated modifications while maintaining safety and effectiveness. HCO's should confirm with vendors whether an authorized PCCP is in place and understand its scope before deploying AI-enabled medical devices subject to FDA oversight. For vendors, adding AI capabilities to an existing product may change its classification depending on the outcome/use case, requiring additional FDA review. FDA maintains a public list of AI-enabled medical devices authorized for marketing in the U.S., which currently includes over 1,000 devices. (https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices). This list may not be comprehensive so HCO's can verify status directly through FDA's 510(k) database.
- **State Laws**: Organizations must harmonize requirements from states with relevant laws (i.e. California (CCPA/CPRA) and New York (SHIELD Act)) with federal law.

Additional Regulatory Bodies and Frameworks:

- **CMS Conditions of Participation (CoPs)**: AI systems used in clinical care may involve CoP requirements for quality assessment and performance improvement, medical staff oversight of clinical practice, patient rights (including informed consent), and nursing services. Organizations should evaluate whether AI deployment triggers CoP compliance obligations and ensure the AI governance framework addresses them.
- **The Joint Commission (TJC)**: AI in clinical settings intersects with accreditation standards for patient safety, performance improvement, and medical staff credentialing. Organizations should anticipate that TJC will develop AI-specific accreditation requirements and design governance frameworks to accommodate them.
- **Federal Trade Commission (FTC) Section 5**: AI-related deceptive practices or unfair data use can trigger FTC enforcement. Healthcare organizations making claims about AI capabilities (to patients, payers, or partners) should ensure those claims are substantiated and not misleading.
- **State Attorney General (AG) and Department of Justice (DOJ) Enforcement**: State AGs and DOJ are increasingly active in AI accountability, particularly regarding consumer protection, fraud and criminal activity prosecution, data privacy, and algorithmic discrimination. Organizations should monitor enforcement trends in their operating jurisdictions.

- **HHS Office for Civil Rights (OCR)**: Beyond general HIPAA enforcement, OCR has signaled interest in AI-related discrimination and privacy practices. Organizations should evaluate whether AI systems create disparate impacts that could trigger OCR scrutiny under Section 1557 of the Affordable Care Act (non-discrimination in health programs).
- **CMS Interoperability and Prior Authorization Rules**: AI-enabled prior authorization automation is a major deployment use case with specific regulatory implications. Organizations should ensure AI-driven prior authorization complies with CMS interoperability requirements and that automated decisions are subject to appropriate human review.
- **National Institute of Standards and Technologies (NIST)**: Standards body that is developing AI Risk Management Frameworks and other standards.

Compliance teams should create and maintain a crosswalk mapping of AI governance controls to applicable regulations to enable traceability during audits.

AI Lifecycle Management

Effective cyber governance begins with disciplined lifecycle management anchored by transparency, validation, and accountability. The governance team can use this lifecycle management approach to review, approve (or deny), implement, monitor, and decommission systems. In addition, this governance identifies and mitigates as much risk as possible and defines a process for approval of the residual risk.

Discovery, Procurement, and Inventory Management

A foundational control is the creation and maintenance of a comprehensive AI system inventory. This catalog identifies all AI use cases, data dependencies, model owners, and business sponsors across the organization.

- **Use Case Justification and Initial Risk Classification**: Before procurement or development, require the business sponsor to document intended use, expected benefit, workflow impacts, data types (including PHI), AI autonomy level, and a preliminary risk tier to determine the depth of due diligence and required controls (see Appendix F or the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide).
- **Procurement**: Teams must adopt AI-specific evaluation criteria in addition to typical reviews of company history, healthcare experience, privacy records and compliance history. Requests for Proposals (RFPs) should mandate disclosure of model explainability, data lineage, cybersecurity controls, and bias testing results (see the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide).
- **AI Specific GRC Due Diligence**: Use an AI vendor assessment questionnaire that covers data lineage, training data restrictions, security controls, update management, incident response, transparency artifacts, and bias testing. Scale the question set by solution risk tier and organizational capacity (see Appendix G or the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide).
- **Contracts**: Contracts should include obligations for transparency artifacts (for example model cards), access to logs where applicable, update notification and change control cooperation, incident notification,

audit support, and secure destruction or return of data at termination (Appendix H or the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide).

- **IT Asset Management**: Organizations should extend IT Asset Management (ITAM) practices to explicitly account for the dynamic introduction of AI capabilities into existing applications, medical devices, and services. This includes AI embedded in clinical devices managed by clinical engineering, cloud hosted AI services, and new AI features enabled via automatic updates, firmware upgrades, or feature releases. ITAM should be integrated with medical device inventory and should include discovery and governance for staff use of external AI tools, including use on personal devices (BYOD), to reduce "shadow AI" risk. AI enablement may occur without a discrete procurement or onboarding event, creating material governance, compliance, cybersecurity, and patient safety risks if not properly controlled.

Design, Development, and Clinical Validation

During development, governance emphasizes compliance with secure coding standards and data integrity.

- **Verification and Validation**: Must be both technical (accuracy, sensitivity, specificity) and clinical (real-world applicability). Independent review boards should assess whether models meet clinical and ethical standards.
- **Testing**: Before deployment, AI systems undergo validation using representative datasets. Cross-validation against human expert judgment provides an essential benchmark.
- **Quality Assurance, Verification, and Go Live Readiness**
  - Verification: Verify data flows, interfaces, access controls, and logging in a non-production environment.
  - Clinical Workflow Validation: Perform user acceptance testing with representative clinical scenarios, including downtime drills and manual fallback procedures.
  - Release Criteria: Define minimum acceptable performance, safety, fairness, and security criteria for go live.
  - Documentation: Maintain a validation report, version record, and approval sign offs (see the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide).

Deployment and Change Management

Deployment requires coordination between IT, clinical, and security teams. Each deployment must include a rollback plan and documented contingency operations.

- **Training and Go Live Readiness**: Prior to production use, complete role-based training (clinical, operational, IT/security), document oversight and escalation procedures, and ensure contingency workflows are tested (see the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide).
- **Security Monitoring**: Establish a process to monitor AI security events such as prompt injections for continuous monitoring of threats
- **Runtime Guardrails**: Enforce runtime guardrails to continuously enforce security policies for AI systems in production to mitigate threats

- **Versioning**: Governance requires version control comparable to software configuration management. Each version must be uniquely identified, tested, and approved prior to production.
- **Update and Patch Management**: Establish a change control process for model, prompt, data, and software updates, including vendor advance notice, staging environment testing, clinical and security review, and post update verification prior to continued clinical use.
- **Decommissioning**: When systems are outdated, organizations must securely decommission data and models to prevent outdated algorithms from re-entering clinical workflows. Additionally, organizations may also be required to preserve data such as audit trails, evidence of compliance with HIPAA, FDA, and other regulatory requirements.

Data Governance for Trustworthy AI

High-performing AI systems begin with responsible data governance. In healthcare, where data integrity directly influences patient outcomes, trust and rigor are paramount. The following allows organizations to look at aspects of the AI tool or product to determine if they are trusted enough to be deployed in the environment.

Data Quality and Sourcing

- **Quality**: Governance establishes standards for completeness, accuracy, and timeliness. Automated profiling and human review must identify anomalies.
- **Sourcing**: Organizations should prioritize first-party clinical data collected under informed consent. Secondary use must be authorized by an institutional review board (IRB).
- **Lineage**: Full traceability—from original source to model output—is essential. Lineage tools should document every transformation to support regulatory audits.

Privacy and Access Controls

- **Acceptable Use and Shadow AI**: Establish and enforce an acceptable use policy for AI tools, including prohibiting staff from entering PHI or other sensitive data into unapproved external AI services. Define approved AI tools, BYOD requirements, and recurring training and acknowledgement. Where feasible, implement technical controls (DLP, managed browser, egress filtering, logging) to reduce unapproved data sharing.
- **Minimization:** Organizations must de-identify personally identifiable information (PII) and PHI wherever feasible. Privacy-by-design practices (anonymization, tokenization, secure enclaves) should be embedded throughout the architecture.
- **Advanced Techniques**: Policies should define acceptable use cases for synthetic data generation and differential privacy to train models while safeguarding anonymity.
- **Access Controls**: Access to training and operational data must follow the principle of least privilege, enforced via role-based access control (RBAC) and multi-factor authentication (MFA). Restrict tool access, particularly with the adoption of agents.
- **Cross-Border Data**: Governance must ensure compliance with international transfer laws (e.g., GDPR standard contractual clauses) and enforce encryption during transfer.

Federated Learning and Collaborative AI Development

Multi-institutional AI development, including federated learning, data collaboratives, and AI research consortiums, is an increasingly common model in healthcare. Federated learning allows institutions to contribute to model training without sharing raw patient data, but it introduces governance challenges that organizations should address:

- **Data Management**: Organizations that develop or fine-tune AI models must establish rigorous data management controls over their training pipelines.
- **Model Weight Governance**: Although raw data remains local in federated learning, model weight updates (gradients) shared between institutions can in some cases leak information about the training data. Organizations should evaluate privacy risks of gradient sharing and require differential privacy or secure aggregation techniques where PHI is involved.
- **Contribution Agreements**: Multi-institutional AI collaborations should be governed by formal data use agreements that address: each party's data governance obligations, intellectual property ownership of resulting models, liability allocation for model failures, withdrawal rights and data deletion upon exit, and regulatory compliance responsibilities.
- **Validation Across Sites**: Models trained across multiple institutions must be validated at each participating site against that site's patient population to ensure performance generalizes.
- **Regulatory Implications**: Organizations should evaluate whether participation in federated learning constitutes a "disclosure" of PHI under HIPAA, even when raw data is not shared, and obtain legal guidance.

Synthetic Data Governance

Synthetic data is artificially generated data that mimics the statistical properties of real patient data. It can be used to reduce privacy risk in AI development. However, synthetic data requires its own governance:

- **Re-identification Risk**: Organizations must validate that synthetic data cannot be reverse-engineered to reconstruct real patient records. Membership inference and linkage attack testing should be conducted before synthetic data is used outside the originating institution.
- **Quality Standards**: Synthetic data used to train or validate clinical AI must be assessed for fidelity (does it accurately represent the statistical properties of the source population) and utility (does it produce models that perform comparably to models trained on real data).
- **Regulatory Acceptance**: Organizations should not assume that regulators will accept synthetic data as a substitute for real-world clinical validation. FDA expectations for clinical evidence, in particular, currently emphasize real-world performance data.
- **Provenance and Lineage**: Synthetic datasets should be documented with the same lineage rigor as real data, including the source dataset, generation methodology, privacy evaluation results, and any known limitations or biases.

Note: Organizations must not assume regulators will accept synthetic data as a substitute for real-world clinical validation. Current FDA expectations for clinical evidence heavily emphasize demonstrated performance using real-world data.

AI-Specific Data Retention

AI training data, validation datasets, model weights, inference logs, and audit trails have retention requirements that differ from operational clinical data:

- **Training Data**: Organizations should define retention periods for training and validation datasets sufficient to support model revalidation, incident investigation, and regulatory audit. For models used in clinical decision-making, training data should be retained for at least the period of the applicable statute of limitations for medical malpractice in the organization's jurisdiction, or the FDA-required retention period for SaMD, whichever is longer.
- **Inference Logs**: Logs recording AI outputs, clinician decisions, and human-AI interaction data should be retained as part of the clinical record where the AI output informed a clinical decision. Retention should align with medical record retention requirements.
- **Model Artifacts**: Each deployed model version (weights, configuration, prompt templates) should be retained to support forensic reconstruction of AI behavior at any point during the model's production lifespan.
- **Right to Deletion and Model Unlearning**: Organizations should evaluate the implications of patient data deletion requests (under HIPAA, state law, or GDPR) for AI models that have already been trained on that data. Complete "unlearning" (removing a specific patient's influence from a trained model) is technically difficult and may be infeasible. Organizations should document their approach to deletion requests in the AI context and consult legal counsel on regulatory obligations.

Clinical Safety and Ethics

Clinical Safety Risk Assessment Framework

Clinical safety evaluation must parallel traditional patient safety models. Each AI system should undergo hazard analysis, Failure Mode and Effects Analysis (FMEA), and human factors review, aligning with the Joint Commission, ISO 14971 standards and the IEEE/UL 2933 Trust, Identity, Privacy, Protection, Safety, and Security (TTIPPSS) framework, a cross-standard safety model.

- **Clinical Decision Support (CDS)**: Must adhere to regulatory guidance on transparency. Clinicians must retain the authority to override AI outputs, and systems must record human-AI interaction decisions.
- **Validation Requirements**: Diagnostic tools require rigorous validation against gold standards (e.g., comparative studies with radiologists). Treatment recommendation systems must demonstrate adherence to evidence-based guidelines.
- **Adverse Events**: Organizations must define criteria for AI-related adverse events (e.g., incorrect diagnosis) and report findings through standard channels for root-cause analysis. For regulated medical devices and SAMD, reporting should be via the MDR.

Post-Market Surveillance and Ongoing Clinical Safety Monitoring:

The FDA's Total Product Lifecycle (TPLC) approach to AI/ML-enabled medical devices requires ongoing safety monitoring after deployment. Even for AI systems not regulated as medical devices, healthcare organizations should implement structured post-market surveillance:

- **Real-World Performance Tracking**: Organizations should compare production AI system performance metrics (sensitivity, specificity, positive predictive value, calibration) against the vendor's published validation metrics and the organization's own pre-deployment validation results. Significant divergence should trigger formal review under the drift detection process.
- **Signal Detection**: Establish mechanisms to detect AI-related safety signals, including: clinician-reported concerns through the existing safety reporting system, automated monitoring of AI override rates and their clinical outcomes, correlation analysis between AI deployment and changes in patient safety event rates, and patient complaints specifically referencing AI involvement in care.
- **Feedback Loop to Model Governance**: Clinician overrides, adverse events, near-misses, and outcome data should feed back into the AI governance process through a structured mechanism. This feedback should inform: vendor notification (for third-party AI), retraining decisions (for internally developed AI), and risk tier reassessment. Without a structured feedback loop, organizations cannot improve AI performance based on their own clinical experience.
- **FDA Reporting**: For AI systems classified as SaMD or AI-enabled medical devices, organizations must comply with FDA Medical Device Reporting (MDR) requirements. Organizations should define clear criteria for when an AI-related adverse event triggers an MDR obligation and assign responsibility for reporting.

AI Ethics, Fairness, and Bias Management

Ethical AI governance rests on autonomy, beneficence, non-maleficence, and justice.

- **Bias Management**: Governance mandates demographic impact assessments. Performance metrics must be stratified by race, gender, age, and socioeconomic status to identify inequities early (Note: this may not always be feasible with small facilities where doing so may substantively increase privacy risks). A sample demographic impact assessment template and minimum reporting format are provided in Appendix I.
- **Fairness Standards**: Organizations should require vendors to align with standards such as IEEE P7003: Bias mitigation strategies may involve resampling, reweighting, or model redesign.
- **Transparency**: Models must provide interpretable outputs. Techniques such as “Shapley Additive exPlanation” (SHAP), Local Interpretable Model-agnostic Explanations (LIME), or model cards communicate logic in accessible formats.
- **Informed Consent**: Patients should be informed when AI contributes to their care, including explicit disclosure of AI involvement and data use.
- **Ethics Review Board**: An AI Ethics Review Board (AERB) should review complex or sensitive AI applications and adjudicate ethical dilemmas. Ethics Review Process: Large systems may establish a dedicated AI Ethics Review Board (AERB). Small and medium hospitals can meet this need by extending an

existing Clinical Ethics Committee, IRB, Quality/Patient Safety Committee, or Compliance committee to review sensitive AI applications and adjudicate ethical dilemmas.

- Ethical considerations should include the following:
  - Need for human agency and oversight: human in the loop, concerns about automation bias
  - Privacy and data ownership: ensuring patients understand AI may be used to analyze data, train and refine the model
  - Risks of re-identification: risks increase if data is disaggregated based on demographic groups
  - Liability and accountability: in relation to board/fiduciary responsibilities but also are issues of omission/commission that leads to harm based on AI use
  - Ensuring AI systems benefit patients and improve patient outcomes: require AI not just benefits the institutions using the AI
  - Concerns about dual-uses and data misuses: for example, insurers using hospital data bolstered through AI and Electronic Health Records (EHR) to deny patient care

Cybersecurity and Privacy Controls

AI introduces new attack surfaces and risks that require additional cybersecurity and privacy processes and controls designed to manage new threats such as adversarial inputs, data poisoning, and model theft.

Defensive Architecture

- **Segmentation**: Implement segmented environments for training, testing, and deployment to reduce exposure.
- **Logging**: Monitoring must extend to model activity, including inference requests and parameter changes.
- **Identity Management**: Strong identity and access management (IAM) frameworks, including Zero Trust principles and privileged access management, restricting AI system access.

Data Protection

- **Encryption**: All AI data must be encrypted in transit (TLS 1.2+) with appropriate cipher suites and at rest (AES-256). Keys should be managed centrally with Hardware Security Modules (HSMs). Vendors should develop or have plans to implement Post Quantum Cryptographic (PQC) Resistant Encryption.
- **Integrity**: Governance mandates version hashing and digital signatures to verify model integrity before deployment. Advancements in post-quantum computing should be monitored for any changes to encryption requirements.
- **Data Loss Prevention (DLP)**: DLP controls must inspect prompts and related inputs sent to external AI services and block or redact recognizable PHI patterns (for example MRNs, SSNs). Transmission is only permitted to approved, verified destinations that are authorized to receive that data class.

Vulnerability Management

AI systems introduce attack surfaces that extend beyond traditional software, including adversarial manipulation of model inputs, data poisoning, model extraction, and exploitation of AI-specific integration points. Organizations should incorporate AI-specific threat vectors into their existing vulnerability management programs, scaling testing and assessment rigor to the system's risk tier:

- **Penetration Testing:** For High and Critical risk AI, conduct or obtain independent testing that includes AI-specific vectors (such as model inversion and prompt injection). Resource-constrained organizations may rely on recent third-party testing reports and remediation attestations from the vendor, supplemented with targeted testing of local integrations. Conduct this independent AI-specific testing at a cadence commensurate with the assessed risk level. Organizations should obtain evidence of remediation for any critical findings from the vendor, supplemented with targeted internal testing of local and network integrations.
- **Privacy Impact Assessment (PIA):** Organizations should conduct a PIA for any AI processing PHI or PII, documenting data flows and consent mechanisms. A sample PIA template is provided in Appendix I.
- **DevSecOps**: Enforce security gates within AI CI/CD pipelines to enforce Software Composition Analysis (SCA), Secrets Scanning, and SAST/DAST testing.

AI System Patching and Updates

AI systems require ongoing patching and updates — including model retraining, algorithm updates, software dependency patches, and infrastructure-level security updates — to maintain safety, performance, and compliance. Unlike traditional software, AI system updates may alter clinical behavior, decision boundaries, or output characteristics in ways that are not immediately apparent. Organizations should establish governance processes that treat AI updates with the same rigor applied to clinical system changes:

- Maintain an inventory of all deployed AI system versions, including model version, training data vintage, and software dependency versions, linked to the organization's asset management and configuration management processes.
- Classify updates by type and risk impact. Model retraining or algorithm changes that may alter clinical outputs should be classified at a higher change-control tier than routine infrastructure or dependency patches.
- Require vendors to provide release documentation that describes the nature of each update, the rationale, any changes to model behavior or performance characteristics, and known limitations or risks introduced by the update. For third-party AI, update notification and documentation requirements should be contractually defined in alignment with the *Health Industry Third-Party AI Risk and Supply Chain Transparency Guide*.
- Establish approval workflows that require AI Governance Committee review for material updates (model retraining, algorithm changes, significant dependency upgrades) and delegated approval for routine patches, consistent with the organization's change management framework.
- Define rollback criteria and procedures for each update type, including pre-update baseline snapshots sufficient to restore the prior system state.

- Track patching cadence and update compliance as a governance metric, reported alongside operational and safety metrics.

Post-Update Verification and Validation

Every material update to an AI system — whether initiated by the organization or delivered by a third-party vendor — should undergo structured verification and validation before full production deployment. Post-update verification ensures that changes have not degraded system performance, introduced bias, or compromised explainability. Organizations should implement the following practices:

- Define a post-update validation protocol for each deployed AI system, specifying the performance, fairness, explainability, and safety metrics to be retested after an update. These metrics should be consistent with those established during pre-deployment validation and aligned with the organization's NIST AI RMF-mapped risk profile.
- Conduct validation using a representative dataset that reflects current production data characteristics, not solely the original training or validation dataset, to detect performance issues arising from data distribution changes since initial deployment.
- Compare post-update performance against both the pre-update production baseline and the vendor's published validation benchmarks. Material degradation in any metric should halt deployment and trigger formal review.
- Reassess model explainability after each material update using the methods established under Explainability Monitoring Over Time. Verify that explanation methods (SHAP, LIME, model cards) remain valid for the updated model version.
- For third-party AI, require vendors to provide post-update validation evidence — including updated model cards, performance benchmarks, and bias testing results — as a condition of update acceptance. Contractual obligations for post-update validation deliverables should be defined in alignment with the *Health Industry Third-Party AI Risk and Supply Chain Transparency Guide.*
- Document all post-update verification results, including any deviations, remediation actions, and approval decisions. Verification records should be retained as part of the AI system's lifecycle documentation and made available for internal audit.
- Where risk tier warrants, conduct a limited production pilot (shadow mode or staged rollout) before full deployment to validate real-world performance under current operational conditions.

Incident Response

- **Incident Response Plans**: Extend incident response plans to cover AI-specific scenarios (for example model failures, data poisoning, prompt injections), including clear triggers for pulling an AI system from service.
- **Tabletop Exercises**: Conduct joint tabletop exercises with IT/security, operations, clinical teams, and third-party vendors, and feed lessons learned into corrective actions.
- **Cross Functional Response Team**: CISO, privacy officer, department heads, etc. that can help analyze incident impacts and support response

OWASP LLM Top 10 Controls Mapping:

Organizations deploying large language models or generative AI should map their security controls against the OWASP Top 10 for LLM Applications. Each risk category has specific healthcare implications:

- **LLM01**: Prompt Injection: In healthcare, prompt injection may occur through clinical notes, patient-submitted intake forms, ingested documents (faxes, PDFs, external records), or adversarial inputs from other systems. Controls must include input sanitization, system prompt protection, output validation, and segregation of instruction context from data context.
- **LLM02**: Sensitive Information Disclosure: LLMs may expose PHI memorized from training data, leaked through prompt context, or inferred from model outputs. Controls must include DLP inspection of LLM outputs, minimization of PHI in prompts, and evaluation of vendor training data practices.
- **LLM03**: Supply Chain Vulnerabilities: Foundation models, fine-tuning datasets, embeddings, plugins, and retrieval corpora all represent supply chain attack surfaces (see AI Supply Chain and Concentration Risk).
- **LLM04**: Data and Model Poisoning: Training data poisoning can introduce systematic biases or backdoors. Healthcare-specific risk: corrupted clinical data feeds can degrade model accuracy for specific patient populations. Controls must include training data integrity verification, anomaly detection, and vendor transparency on data sourcing.
- **LLM05**: Insecure Output Handling: LLM outputs that feed downstream clinical systems (EHR order entry, CDS alerts, clinical documentation) without validation can propagate errors into the clinical record. Controls must include output validation, structured output schemas, and human review gates for clinical outputs.
- **LLM06**: Excessive Agency: LLM-based agents with tool-calling capabilities can take actions beyond their intended scope (see Agentic AI Governance).
- **LLM07**: System Prompt Leakage: Exposure of system prompts can reveal organizational logic, PHI handling instructions, or security controls to adversaries. Controls must include prompt confidentiality protection and monitoring for prompt extraction attempts.
- **LLM08**: Vector and Embedding Weaknesses: Retrieval-augmented generation (RAG) systems used with clinical knowledge bases can be manipulated through poisoned embeddings or retrieval manipulation. Controls must include embedding integrity verification and access controls on vector databases.
- **LLM09**: Misinformation: LLMs can generate clinically plausible but incorrect medical information (hallucination). Controls must include grounding mechanisms (RAG with authoritative clinical sources), output verification for clinical content, and clinician review requirements.
- **LLM10**: Unbounded Consumption: LLM resource consumption (token costs, compute) can be exploited for denial-of-service or financial impact. Controls must include rate limiting, cost monitoring, and anomaly detection on usage patterns.

A detailed healthcare mapping with specific controls and governance requirements for each OWASP LLM Top 10 category is provided in Appendix N.

AI Red Teaming

For High and Critical risk AI systems, organizations should conduct or require AI-specific red teaming distinct from traditional network penetration testing:

- **Scope**: AI red teaming should evaluate: adversarial robustness (can inputs be crafted to manipulate outputs?), prompt injection resistance (for LLM-based systems), jailbreak resistance (can safety controls be circumvented?), data extraction (can training data or PHI be extracted through model queries?), output manipulation (can the model be induced to produce harmful clinical recommendations?), model inversion (sensitive features from the training data can be reconstructed from the model's outputs) and privilege escalation (can the AI system be manipulated to access data or systems beyond its authorized scope?).
- **Frequency**: AI red teaming should be conducted prior to initial deployment for High and Critical risk systems, after material model updates, and periodically (at least annually) for production systems.
- **Execution**: Organizations may conduct AI red teaming internally (if they have qualified personnel), through third-party AI security firms, or by requiring vendors to provide independent red team reports. Resource-constrained organizations should prioritize vendor-provided red team evidence supplemented with targeted testing of local integrations and prompt injection vectors.
- **Remediation**: Red team findings should be documented, risk-rated, and tracked through remediation to closure with the same rigor as traditional penetration test findings.

Model Extraction/Stealing and Intellectual Property Protection:

Model stealing occurs when an adversary trains a surrogate model by querying a production model. Organizations should:

- Monitor for anomalous query patterns that may indicate model extraction attempts (unusually high query volumes, systematic exploration of model boundaries, queries that appear designed to map decision surfaces).
- Implement rate limiting and query anomaly detection for externally accessible AI models.
- Require vendors to document their model extraction defenses.
- Include model IP protection in vendor contracts and evaluate the organization's own exposure if it develops proprietary models.

AI Identity and Non-Human Identity (NHI) Management

As generative and agentic AI systems proliferate, they create a large and growing population of non-human identities (service accounts, API keys, OAuth tokens, machine certificates) with access to clinical systems. These non-human identities require dedicated governance:

- **Inventory**: Organizations must inventory all non-human identities associated with AI systems, including service accounts, API keys, OAuth client credentials, machine certificates, and inter-system authentication tokens. This inventory should be maintained as part of the broader AI system inventory (see Appendix A) and the organization's identity governance program.
- **Least Privilege and Segregation of Duties**: AI system credentials must follow the principle of least privilege and segregate all duties. An AI agent authorized to query a patient's medication list should not

hold credentials that allow it to modify orders. Access scope should be defined at the most granular level the target system supports.

- **Just-in-Time Access**: Where technically feasible, AI systems should use just-in-time credential issuance rather than long-lived credentials, particularly for access to high-sensitivity systems (EHR, PACS, lab).
- **Credential Lifecycle**: AI system credentials must be subject to defined lifecycle management, including: automatic rotation on a defined schedule, expiration policies, and revocation procedures for decommissioned systems. This includes a controlled disablement or pause capability that is governed by documented risk assessments and emergency procedures. This capability must require role-based, multi-party authorization for high-impact actions, enforce strong authentication (MFA), generate immutable audit logs, and default to a fail-safe behavior that ensures patient safety.
- **Attribution and Audit**: When an AI system takes an action (queries a record, modifies data, triggers a workflow), the audit trail must attribute the action to both the AI system identity and the human who authorized the AI system's operation. AI systems must not operate under shared human user accounts.
- **Self-Escalation Prevention**: AI systems must not have the ability to modify their own access permissions, create new credentials, or expand their access scope. Privilege changes require human authorization through the organization's identity governance process.
- **Zero Trust Application**: Zero Trust principles should be extended to AI system identities: verify every request, assume breach, enforce least privilege, and monitor continuously. AI systems should not be implicitly trusted based on network location.

Generative AI and Large Language Model Risks

Large language models and generative AI represent the most rapidly adopted AI technology category in healthcare. From clinical documentation assistants and patient-facing chatbots to diagnostic support and administrative automation, generative AI is deployed across nearly every healthcare function. These systems introduce risk categories that are qualitatively distinct from traditional machine learning and require dedicated governance controls.

This section addresses the cybersecurity and data protection risks specific to generative AI and LLM deployments in healthcare.

Distinct Risk Profile of Generative AI

Generative AI introduces the following risks that are absent or materially different from traditional ML:

- **Hallucination**: LLMs can generate text that is fluent, confident, and clinically plausible but factually incorrect. In healthcare, hallucinated drug interactions, dosage recommendations, diagnostic criteria, or procedure instructions can directly harm patients. Unlike traditional ML drift (which degrades gradually and can be statistically detected), hallucination can occur unpredictably on any individual output.
- **Prompt Injection (Direct and Indirect)**: Adversaries can manipulate LLM behavior by injecting instructions into inputs the model processes. Direct injection occurs through user-supplied prompts. Indirect injection occurs when the LLM processes external content containing hidden instructions in clinical notes, patient-submitted forms, imported documents, or data from other systems. Indirect injection

is particularly dangerous in healthcare because clinical workflows routinely ingest uncontrolled external content (faxes, external records, patient portal messages). Combining controls such as input sanitization, structural prompt separation and monitoring can aid in mitigation.

- **Training Data Memorization and Extraction**: LLMs can memorize and reproduce segments of their training data, including potentially PHI if the model was trained or fine-tuned on clinical data. Adversaries can use targeted prompting techniques to extract memorized content.
- **Jailbreaking**: Adversaries can craft prompts designed to circumvent the model's safety controls, causing it to produce harmful, inappropriate, or policy-violating outputs. In healthcare, jailbreaking a clinical AI assistant could cause it to provide dangerous medical advice it was designed to refuse.
- **Context Window Manipulation**: LLMs have limited context windows. Adversaries can exploit this by flooding the context with irrelevant or misleading information, displacing legitimate instructions or clinical data.
- **Output Inconsistency**: LLMs are inherently non-deterministic (or weakly deterministic even at low temperature settings). The same clinical question may receive different answers on different occasions, complicating clinical reliability and reproducibility.
- **Data Leakage**: Sensitive health information can be shared publicly or internally in AI responses to others besides the provider making the inquiry.
- **Unexplainability**: There is no way to understand how the AI model reached its conclusions.
- **EHR (or other system) integration problems**: The AI may not intersect, interact, or integrate correctly with downstream systems.

Governance Requirements for Generative AI

Organizations deploying generative AI should implement the following controls in addition to the baseline controls described elsewhere in this document:

- **Input Validation and Sanitization**: All inputs to LLM systems (i.e. from users, clinical systems, or external data sources) should be inspected for injection attempts, excessive length, and inappropriate content before processing. Technical controls such as input filtering, prompt boundary enforcement, and instruction-data separation should be implemented.
- **Output Validation**: LLM outputs that inform clinical decisions, are written to the clinical record, or trigger downstream system actions must be validated before use. Validation mechanisms may include: grounding against authoritative clinical knowledge bases (RAG with verified medical sources), structured output schemas that constrain the format and content of clinical outputs, automated fact-checking for verifiable claims (drug interactions, dosages, procedure codes), and human review requirements calibrated to risk tier and autonomy level.
- **Grounding and Retrieval-Augmented Generation (RAG)**: Organizations using RAG to ground LLM outputs in organizational knowledge bases, clinical guidelines, or formularies should govern the retrieval corpus with the same rigor as the model itself: access controls, integrity verification, version management, and regular updates to ensure currency.
- **Prompt Governance**: System prompts, prompt templates, and prompt engineering practices should be treated as configuration artifacts subject to change control. Organizations should define who is authorized

to modify system prompts; how prompt changes are reviewed (including clinical review for clinical-facing prompts); versioning and rollback capabilities for prompt changes; and protections against system prompt extraction or leakage.

- **Fine-Tuning Data Governance**: If the organization fine-tunes LLMs on clinical data, the fine-tuning dataset must be governed under the Data Governance for Trustworthy AI requirements, including data quality, privacy controls, bias assessment, and lineage documentation. Organizations should evaluate memorization risk for the fine-tuning dataset and implement mitigation techniques (differential privacy, data deduplication, canary tokens for extraction detection).
- **PHI Exposure Prevention**: Organizations must implement controls to prevent PHI from being sent to external LLM services without authorization. This includes DLP controls on LLM input channels, managed access points that inspect and redact PHI before transmission to external models, approved-service whitelisting, and logging of all data sent to external AI services. Where PHI processing by an external LLM is authorized, the vendor must be a Business Associate with appropriate Business Associate Agreement (BAA) protections.
- **Disclosure to Patients and Clinicians**: When LLM-generated content appears in patient-facing communications (portal messages, after-visit summaries, chatbot interactions) or in clinical documentation, the AI-generated nature of the content should be disclosed (see Patient Engagement and Transparency).

Agentic AI: Governance for AI Systems That Act

Agentic AI systems are AI systems that can reason about goals, formulate plans, invoke external tools or APIs, chain multiple actions, and interact with other systems or AI agents with varying degrees of human oversight between actions. Unlike traditional AI models that produce outputs for human consumption, agentic AI systems take actions in the world: querying EHRs, triggering orders, sending messages, modifying records, calling external services, and executing multi-step workflows. This section establishes specific governance principles for agentic AI.

The AI Autonomy Levels in Appendix D describe the decision authority spectrum (how much human oversight is required for AI decisions). Agentic AI introduces a distinct and orthogonal dimension; i.e., action capability (what the AI system can do). A Level 2 (Augmented Intelligence) system with tool-calling capabilities may be more dangerous than a Level 3 (Partial Autonomy) system without them, because the Level 2 agentic system can take actions whose consequences propagate beyond the AI-human interaction.

A compromised or malfunctioning agentic AI system does not merely produce a bad recommendation that a clinician can ignore. It can take a series of harmful actions—querying sensitive records, modifying data, triggering clinical workflows, sending communications, or invoking external services—before any human is aware of the problem. The blast radius of an agentic AI failure is fundamentally larger than a traditional AI failure.

Governance Principles for Agentic AI

Principle 1: Least-Privilege Action Scope. Agentic AI systems must be restricted to the minimum set of tools, APIs, data sources, and actions required for their defined purpose. An AI agent authorized to draft a clinical note should not hold credentials that allow it to submit orders, modify medication lists, or access financial systems. Action scope must be defined explicitly, enforced technically, and audited regularly.

Principle 2: Human Authorization Gates for Consequential Actions. Organizations must define which actions are "consequential" in the healthcare context and require explicit human authorization before an agentic AI system can execute them. At minimum, the following actions should require human approval (Human in the Loop): modification of the patient medical record, submission of clinical orders (medications, labs, imaging, procedures), changes to treatment plans, communication with patients or their representatives, access to records beyond the immediate clinical context, financial transactions or billing submissions, and any action that is irreversible or difficult to reverse. The granularity of human authorization gates should scale with risk tier: Critical risk agentic systems should require per-action human approval for all consequential actions; High risk systems should require human approval for categories of consequential actions with per-action logging; Medium risk systems may permit automated execution of pre-approved routine actions with periodic human review.

Principle 3: Action Logging and Auditability. Every action taken by an agentic AI system must be logged with sufficient detail to support forensic reconstruction: what action was taken, what data was accessed or modified, what the AI system's reasoning chain was (to the extent the architecture supports it), what the triggering condition or goal was, and which human authorized the agent's operation. Logs must be tamper-resistant, retained according to data retention requirements, and accessible for incident investigation.

Principle 4: Containment and Kill-Switch. Organizations must implement the technical capability to halt an agentic AI system's action chain in progress. This is distinct from the general kill-switch concept (which takes a system offline): agentic containment must be able to interrupt a multi-step execution mid-sequence, prevent completion of in-progress actions where technically feasible, quarantine the agent's pending actions for human review, and preserve the agent's state for forensic analysis. Containment procedures must be documented, assigned to specific roles, and tested at least annually. The process must also take into account patient safety and ensure no patient harm occurs during this technical process.

Principle 5: Multi-Agent Governance. When multiple AI agents interact (e.g. scheduling agent, clinical triage agent, documentation agent, coding agent) governance must address the composite behavior of the agent system, not merely the individual agents. Organizations should map agent-to-agent interactions and data flows, apply the same governance controls to inter-agent communications as to human-AI interactions, ensure that privilege escalation cannot occur through agent chaining (Agent A with limited access invoking Agent B with broader access), and validate that the composite system behavior is consistent with the intended use case and risk tier. Additionally, the emerging threat of agent-to-agent prompt injection, where a compromised agent injects malicious instructions into agents through shared communication channels should be controlled.

Principle 6: No Self-Modification. Agentic AI systems must not be able to modify their own instructions, expand their tool access, alter their system prompts, or change their operational parameters without human authorization through the organization's change control process.

Relationship to Autonomy Levels

The Autonomy Levels in Appendix D should be assessed for agentic systems based on both their decision authority and their action capability. An agentic system that autonomously executes clinical actions in bounded contexts (e.g., an automated prior authorization agent) should be classified as at least Level 3 regardless of whether a human reviews its decisions after the fact, because the actions have already been taken.

Organizations should update their AI inventory (see Appendix A) to identify which AI systems have agentic capabilities (tool-calling, API access, action execution, or multi-agent interaction) and apply the enhanced governance requirements of this section accordingly.

AI Supply Chain and Concentration Risk

Healthcare organizations' AI deployments depend on a layered supply chain that extends well beyond the direct vendor relationship. A clinical AI application may itself depend on a foundation model from a separate provider, hosted on a cloud inference platform from a third provider, fine-tuned with data processed by a fourth, and integrated through an API gateway from a fifth. A vulnerability at any layer can compromise the healthcare organization's AI system. Moreover, when multiple clinical applications share the same underlying foundation model or infrastructure provider, a single compromise creates correlated risk across the AI portfolio.

AI Supply Chain Mapping

Organizations should map the full AI supply chain for each AI system in their inventory, extending beyond the direct vendor to include the following considerations:

- **Foundation Model Layer**: Which foundation model(s) does the application use? Is it a commercial API (proprietary model accessed via API), a commercially licensed model (deployed in the vendor's or organization's infrastructure), or an open-source model (downloaded from a public repository)?
- **Fine-Tuning and Training Layer**: Where was the model fine-tuned? What data was used? Who performed the fine-tuning? Were third-party data processors involved?
- **Inference Infrastructure**: Where does model inference execute? On the vendor's infrastructure, a cloud provider (and which one), or on-premises? What is the geographic location of inference processing?
- **Data Pipeline Dependencies**: What data sources feed the AI system? Are there third-party data enrichment services, embedding providers, or retrieval corpus providers?
- **Integration and Middleware**: What API gateways, orchestration platforms, or middleware connect the AI system to the organization's clinical systems?
- **Fourth-Party and Nth-Party Risks**: What are the vendor's own dependencies? Does the direct vendor use sub-processors (e.g. CSP, AWS, etc) for model hosting, data processing, or support services?

Supply chain mapping should be conducted during procurement (for new AI systems) and retrospectively (for existing deployed systems). Supply chain information should be maintained in the AI inventory and updated when vendors notify the organization of material supply chain changes.

AI Bill of Materials (AIBOM)

For High and Critical risk AI systems, organizations should require vendors to provide an AI Bill of Materials (AIBOM) that documents:

- Foundation model(s) used (name, version, provider)
- Fine-tuning data provenance summary (data sources, geographic origin, consent basis). If vendor indicates that this data is proprietary, organizations should determine what level of summary is sufficient for risk assessment.
- Third-party libraries, frameworks, and Software Development Kits (SDKs) (with versions)
- API dependencies and external service calls
- Hosting and inference infrastructure (provider, region)
- Sub-processors and fourth-party services
- Known limitations, failure modes, and contraindications

The AIBOM should be updated with each material release and provided to the customer. AIBOM requirements should be included in procurement contracts (see Appendix H).

Systemic and Concentration Risk Assessment

Organizations should assess systemic and concentration risk across their AI portfolio. The HSCC SMART Toolkit provides templates and a methodology to visualize, identify and measure systemic risk posed by third party technology, software and communications services essential to clinical, administrative and manufacturing workflows. This resource is intended for cybersecurity, supply chain, risk, operational and administrative executives across health industry organizations of all sizes and subsectors, including manufacturers, insurance providers, healthcare providers, public health agencies and information exchanges. The concentration risk assessment should consider the following:

- **Foundation Model Concentration**: If multiple clinical AI applications depend on the same foundation model (or the same model provider), a single model compromise, outage, or vendor disruption creates correlated risk across multiple clinical workflows. Organizations should document which foundation model(s) underlie each deployed AI application and assess the clinical impact if that model or provider becomes unavailable or compromised.
- **Infrastructure Concentration**: If multiple AI systems are hosted on the same cloud inference platform, the organization has single-point-of-failure risk at the infrastructure layer.
- **Vendor Concentration**: If a single vendor supplies multiple AI applications across different clinical domains, the organization's AI portfolio risk is concentrated in that vendor's security posture, financial stability, and operational reliability.
- **Concentration risk**: Organizations should document concentration risk, evaluate diversification opportunities where clinically and operationally feasible, and ensure that business continuity planning accounts for simultaneous failure of correlated AI systems.

Open-Source Model Risk

Organizations deploying open-source models (whether directly or through vendors that incorporate open-source models) should evaluate risk accordingly:

- **Provenance**: Who developed the model? What data was it trained on? Is the training data documented and auditable? Has the model been poisoned (malicious model uploads to public model repositories)?
- **Community Maintenance**: Is the model actively maintained? Are security vulnerabilities addressed promptly? Is there a vulnerability disclosure process?
- **Lack of Contractual Recourse**: Open-source models typically come without warranties, indemnification, or support obligations. Organizations bear full responsibility for security, validation, and ongoing governance.
- **Supply Chain Integrity**: Models downloaded from public repositories (e.g., model hubs) may contain backdoors, biased training data, or undisclosed capabilities. Organizations should verify model integrity (hash verification, provenance documentation) before deployment.
- **Patching**: delayed or infeasible security patching as no one "owns" the model and is responsible for patching the model
- **Intellectual property concerns**: if data on an open source model becomes openly shared or available this can cause organizations using that model loss of intellectual property entered into that model.

For a comprehensive set of risk questions, see the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide.

AI-Specific Incident Response

AI-specific incidents (e.g. model compromise, training data poisoning, adversarial manipulation of clinical outputs, prompt injection exploitation) require detection, triage, containment, and recovery procedures that differ fundamentally from traditional cybersecurity incident response. This section establishes the framework for AI-specific incident response. An AI Incident Response Playbook template is provided in Appendix L.

AI Incident Categories

Organizations should define and prepare for the following AI-specific incident categories:

- **Model Failure / Performance Degradation**: The AI system produces outputs that fall below validated performance thresholds due to data drift, concept drift, model corruption, or infrastructure failure. Clinical impact: degraded diagnostic accuracy, inappropriate recommendations, or missed alerts.
- **Model Compromise / Adversarial Manipulation**: An adversary has deliberately manipulated the AI system's behavior through data poisoning, model tampering, adversarial inputs, or prompt injection. Clinical impact: systematically biased or harmful outputs that may not be immediately apparent to clinicians.

- **Data Poisoning (Training or Inference)**: The integrity of training data or real-time data feeds has been compromised, resulting in model behavior that reflects the poisoned data. Clinical impact: may be delayed and subtle (e.g., gradually degrading accuracy for specific patient populations).
- **Privacy Breach / PHI Leakage**: PHI has been exposed through AI model outputs (memorization), through unauthorized data transmission to external AI services, or through inference attacks on model behavior. Regulatory impact: HIPAA breach notification obligations, OCR investigation, state attorney general notification.
- **Agentic AI Unauthorized Action**: An agentic AI system has taken actions beyond its authorized scope including the following: unauthorized records access, unauthorized data modification, unauthorized clinical actions without required human approval, or errors propagated through downstream systems.
- **AI Supply Chain Compromise**: A vulnerability or compromise in a foundation model, third-party component, or infrastructure dependency has been identified that affects one or more deployed AI systems.

Detection

AI-specific incidents may be detected through:

- Automated monitoring alerts (performance threshold breaches, anomalous query patterns, unexpected output distributions)
- Clinician reports through the patient safety reporting system
- Vendor notifications of model vulnerabilities or data integrity issues
- External threat intelligence (MITRE ATLAS advisories, OWASP community alerts, vendor security bulletins, CISA advisories)
- Internal audit findings
- Patient complaints

Organizations should ensure that existing incident detection mechanisms (SIEM, SOC, safety reporting systems) are configured to recognize AI-specific indicators of compromise and route them to personnel qualified to assess AI incidents.

Triage and Assessment

AI incident triage must determine:

- **Blast Radius**: Which patients were potentially affected? Which clinical decisions were made based on potentially compromised AI outputs? What is the time window of exposure (from when the incident began to when it was detected)?
- **Clinical Impact Assessment**: Does the incident create an immediate patient safety risk requiring urgent clinical intervention (e.g., patient notification, order review, clinical reassessment, implementation of downtime plans)?
- **Ongoing Risk**: Is the AI system still in production? Is the incident ongoing? Is the attack vector still active?
- **Regulatory Notification Obligations**: Does the incident trigger HIPAA breach notification? Is FDA MDR reporting required? Is CIRCIA critical infrastructure reporting and State breach notification in place?

Containment

AI-specific containment actions include:

- **Immediate**: Invoke the AI system kill-switch or manual fallback to remove the AI system from clinical workflow. Preserve model state (weights, configuration, logs) as forensic evidence before any remediation. Revoke or rotate AI system credentials (non-human identities) if compromise is suspected. Ensure containment does not impact clinical workflow or patient safety.
- **Short-Term**: Isolate affected data pipelines and integration points. Notify affected vendors and request incident support. Activate clinical fallback procedures and notify clinical leadership.
- **Extended**: Conduct full forensic analysis of model behavior, data integrity, and access logs. Determine root cause. Assess whether other AI systems sharing supply chain components are affected.

Recovery

Before returning an AI system to production after an incident:

- The root cause must be identified and remediated.
- The model must be revalidated against the original performance benchmarks and against a dataset representative of the period during which the incident occurred.
- Clinical review must assess whether any patient care actions taken during the incident period require follow-up.
- The AI system's risk tier should be reassessed.
- The incident and lessons learned should be documented and reviewed through the organization's existing corrective action process.

Post-Incident Review

Every AI-specific incident should undergo a structured post-incident review that addresses: root cause analysis, assessment of detection timeliness (how long was the exposure window), effectiveness of containment actions, patient impact assessment and follow-up actions, regulatory reporting compliance, lessons learned and corrective actions, and updates to the AI incident response plan.

Operational Resilience and AI Business Continuity

As healthcare organizations integrate AI into clinical and operational workflows, they create dependencies that must be managed through business continuity and disaster recovery planning. Whether due to technical failure, cybersecurity incident, vendor outage, or deliberate takedown for safety reasons, when AI systems fail clinicians and operational staff must be able to continue functioning using manual or alternative procedures. This section establishes governance requirements for AI operational resilience.

AI Dependency Assessment

Organizations should assess AI dependency risk for each deployed AI system:

- **Workflow Criticality**: How critical is the AI system to the clinical or operational workflow it supports? Could the workflow function at an acceptable level without the AI system, or has the organization become dependent on the AI system to maintain safe and effective operations?
- **Degradation Impact**: What is the clinical, operational, and financial impact of the AI system being unavailable for 1 hour, 24 hours, 72 hours, or 1 week? How does impact scale with duration?
- **Substitutability**: Can the AI system be replaced by manual processes, alternative technology, or a different AI system? How quickly can substitution occur?
- **Skill Atrophy**: Have clinicians or staff lost the skills or capacity to perform the function manually due to prolonged reliance on the AI system? Skill atrophy is a particularly insidious form of AI dependency because it degrades the quality of manual fallback over time.

AI dependency assessments should be conducted during deployment (as part of go-live readiness) and reassessed annually or when the AI system's role in the workflow materially changes.

Manual Fallback Procedures

Each AI system deployed in a clinical or operationally critical workflow must have a documented manual fallback procedure:

- The fallback procedure must describe how the function is performed without the AI system.
- The fallback procedure must be tested at least annually (or more frequently for High and Critical risk systems) through tabletop exercises or live drills.
- Staff who may need to execute the fallback procedure must be trained and must demonstrate competency.
- The organization should maintain sufficient capacity (staffing, equipment, process documentation) to operate under manual fallback for a defined period aligned with the AI system's RTO.
- Key AI data as part of the plan should be backed up on a schedule defined by the business continuity plan.

Cascading Failure Analysis

When AI systems are interconnected, failure of one system may cascade to others. For example, the failure of a triage AI agent could cascade failure down to other agents connected such as bed management and staffing. Using tools such as the HSCC SMART maps toolkit, organizations should:

- Map AI-to-AI dependencies as part of the AI inventory and supply chain mapping.
- Identify cascading failure paths and assess their clinical and operational impact.
- Test cascading failure scenarios through tabletop exercises.
- Ensure that fallback procedures account for multi-system failure, not just individual system failure.

Integration with Business Continuity Planning

AI systems should be incorporated into the organization's existing Business Continuity Plan (BCP) and Disaster Recovery (DR) plan:

- Each AI system should have a defined Recovery Time Objective (RTO) and Recovery Point Objective (RPO) aligned with the clinical or operational workflow it supports.
- AI system RTOs should be informed by the AI dependency assessment and the availability of manual fallback procedures.
- The BCP should address the scenario in which multiple AI systems fail simultaneously due to shared infrastructure or supply chain dependencies (see concentration risk).
- AI system recovery procedures should be documented and tested as part of regular BCP/DR exercises.

Vendor Risk Management

Third-party AI vendors expand capability but multiply risk. A formal Vendor Risk Assessment Framework must classify vendors by access level and data sensitivity. For a comprehensive guide to AI vendor risk management, see the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide.

- **Risk Scoring**: Apply a consistent AI risk scoring model (solution and vendor) to determine due diligence depth, required contract clauses, and monitoring frequency.
- **High-Risk Vendors**: Those processing PHI or supplying clinical decision algorithms require enhanced due diligence and continuous monitoring.
- **Transparency**: Vendors must supply technical documentation describing training data sources, logic, and limitations. Model Cards or software notes should accompany each release. Consider asking vendors to supply an AIBOM or information regarding the model components, training data provenance summary, and critical third-party dependencies (including versions) that together make up an AI system.
- **Contractual Safeguards**: Contracts should mandate cooperation with investigations and secure destruction of data at termination.

Vendor Lock-In and Portability Assessment:

Organizations should assess data and model portability before procurement and include portability requirements in contracts:

- **Data Portability**: Can the organization export its data (training data, configuration data, inference logs, validation results) in a standard, usable format if the vendor relationship ends?
- **Model Portability**: If the organization has invested in fine-tuning or customization, can those customizations be transferred to an alternative platform or vendor?
- **Workflow Portability**: How deeply is the vendor's AI system integrated into the organization's clinical or operational workflows? What is the estimated cost and timeline to migrate to an alternative?
- **Contractual Protections**: Contracts should include data export rights, transition assistance obligations, and defined transition periods. See Appendix H for portability clauses.

Continuous Vendor Monitoring:

Initial vendor due diligence establishes a baseline, but vendor risk posture changes over time. Organizations should implement ongoing vendor monitoring with frequency determined by risk tiers:

- **Critical risk vendors** (AI systems directly influencing clinical decisions on PHI): Continuous monitoring of vendor security posture (security ratings services, threat intelligence), quarterly review of vendor performance and incident history, annual comprehensive reassessment.
- **High risk vendors**: Semi-annual performance review, annual reassessment.
- **Medium risk vendors**: Annual review.
- **Low risk vendors**: Review at contract renewal.

Monitoring should include the following: vendor security rating changes, publicly disclosed breaches or vulnerabilities, material changes to the vendor's AI supply chain (foundation model changes, infrastructure changes), vendor financial stability indicators, and regulatory enforcement actions against the vendor.

Vendor Incident Notification SLAs

Contracts should define specific notification timelines for AI-specific incidents:

- Security breaches affecting customer data: Notification within the timeframe required by the applicable BAA and breach notification laws, and no later than 72 hours.
- Model compromises, data poisoning events, or integrity failures: Without unreasonable delay and consistent with applicable law and contract. Where notification is legally or contractually required, notice shall be provided no later than 72 hours
- Material model updates or changes in model behavior: Advance notification per the change control process (see Deployment and Change Management).
- AI supply chain changes (foundation model provider changes, infrastructure changes): Notification within 30 days.

Fourth-Party and Sub-processor AI Risk:

Organizations should extend vendor risk assessment to the vendor's own AI-specific dependencies:

- Require vendors to disclose foundation model providers, model hosting services, data processing subcontractors, and other fourth-party AI dependencies.
- Evaluate whether the vendor's fourth-party dependencies introduce risks not addressed by the direct vendor's controls.
- Require contractual flow-down of material security and privacy obligations to sub-processors.
- Monitor for changes in the vendor's sub-processor relationships through the vendor notification process.

Monitoring and Training

AI System Monitoring

A continuous monitoring framework validates that AI systems, once deployed, remain accurate and reliable. For AI systems provided in whole or in part by third parties, monitoring responsibilities should be defined in accordance with the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide, which establishes shared accountability expectations across the AI supply chain.

- **Drift Detection**: Governance mandates statistical detection of "drift" (deviation from training conditions). If the AI system is provided by one or more third parties, drift detection should be considered a shared responsibility as applicable. Material deviations from predefined, risk-based performance or data variation limits must trigger formal review and corrective action, including human oversight and, where appropriate, model rollback or retraining.
- **Human Oversight**: Define required human review and escalation based on risk tier and autonomy level, including override authority, supervision responsibilities, and documentation requirements. Ensure systems record human-AI interaction decisions and support rapid disablement or rollback.
- **KPIs**: Key performance indicators (KPIs) and metrics track accuracy/sensitivity (for diagnostic AI) and throughput gains (for operational AI).
- **Automated Alerts**: Alerts trigger when performance or ethical metrics — such as bias and fairness indicators, explainability scores, and privacy compliance status, mapped to the organization's NIST AI RMF-aligned risk profile — breach established thresholds, initiating rapid containment.

Ethical Performance Monitoring

Organizations should establish ongoing monitoring of AI ethical performance indicators as a distinct governance activity, not solely as a byproduct of operational performance tracking. Ethical performance metrics - including bias and fairness indicators (NIST AI RMF MAP 2.3), transparency and explainability scores (GOVERN 1.2), and privacy compliance status (MANAGE 3.2) - should be mapped to the organization's NIST AI RMF-aligned risk profile and monitored at a frequency commensurate with the system's risk tier. Material deviations from ethical performance baselines should trigger formal review, which may include vendor notification, enhanced human oversight, or model suspension. For third-party AI, ethical performance monitoring expectations and reporting obligations should be contractually defined in alignment with the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide.

Explainability Monitoring Over Time

Model explainability can degrade as models are updated, retrained, or as the data distribution shifts. Organizations should establish processes to detect and address explainability degradation on an ongoing basis:

- Reassess model explainability after each material model update, not only at initial deployment.
- Track whether explanation methods (SHAP, LIME, model cards) remain valid and accurate for the current model version.
- Require vendors to update transparency artifacts (model cards, explanation documentation) with each release.

Real-World vs. Validation Performance Tracking

Organizations should systematically compare production AI system performance against the vendor's published validation metrics and the organization's own pre-deployment validation results. The following practices support structured performance comparison:

- Define key performance metrics for each deployed AI system (aligned with the metrics used in pre-deployment validation).
- Measure these metrics periodically using production data (frequency determined by risk tier).
- Investigate and document material divergence between real-world and validation performance. Material divergence should trigger formal review and may require vendor notification, model rollback, or risk tier reassessment.

Training and Education

Effective AI governance depends on role-appropriate training that equips stakeholders at every level to understand, oversee, and interact responsibly with AI systems. Training programs should minimally address the following audiences and domains:

- **Clinicians:** Must understand AI capabilities and limitations. Curricula should include bias recognition and escalation procedures.
- **IT/Security:** Teams require specialized instruction in AI architecture, adversarial defense, and model verification.
- **Board Members:** Should participate in annual workshops covering regulatory updates and risk metrics.
- **Change Management:** Use an existing framework (i.e., Awareness, Desire, Knowledge, Ability, and Reinforcement) to guide adoption, emphasizing AI as an augmentation of human expertise.

For a more comprehensive training program, see Appendix H in the Third Party AI Risk and Supply Chain Transparency Guide.

Metrics and Audit

Effective governance relies on comprehensive metrics that quantify maturity. Organizations should track metrics across operational, safety, and outcome domains, and subject governance processes to periodic independent review:

- **Operational Metrics:** Track the number of AI systems under oversight, compliance rates, and incident response times.
- **Safety and Compliance Metrics:** Track AI-related patient safety events and near misses, training completion and annual refresher rates, acceptable use attestations, compliance with required reviews (validation, update approvals), and closure rates for audit findings.
- **Outcome Metrics:** Demonstrate improvements in patient outcomes (e.g., error reduction), mortality rates, and clinician satisfaction.
- **Audit:** Internal audit functions must conduct periodic assessments of governance processes, with findings feeding into corrective action plans.

Clinician Feedback Loop

Organizations should establish a structured mechanism for clinician feedback on AI system performance to inform governance and model improvement. A well-designed feedback loop serves as both a safety signal and a continuous improvement input:

- Clinicians should have a low-friction pathway to report AI outputs they believe are incorrect, misleading, or clinically inappropriate. AI reporting should be distinct from general IT support tickets.
- Feedback should be reviewed by a qualified individual (AI program lead, clinical informaticist, or data scientist) and categorized by severity.
- Aggregated feedback should be reported to the AI Governance Committee and to the vendor (for third-party AI).
- Feedback data should inform retraining decisions, risk tier reassessment, and vendor performance review.

Patient Engagement and Transparency

Patient engagement in AI governance intersects cybersecurity through informed consent, data use authorization, and transparency about how AI systems process patient data.

Patient Notification of AI Use

Organizations should establish a patient notification standard for AI involvement in care:

- **Minimum Standard**: Patients should be informed when AI contributes to clinical decisions affecting their care, including the general nature of the AI involvement (e.g., "An AI system assisted in analyzing your imaging study"), without requiring disclosure of proprietary model details.
- **Notification Models**: Organizations should select and document their notification approach:
  - *Disclosure-only*: Patients are informed that AI is used in certain care processes (e.g., through general notice in patient rights materials or admission documentation).
  - *Opt-out*: Patients are informed and may request that AI not be used in their care where clinically feasible.
  - *Opt-in*: AI is used only with explicit patient consent (may be appropriate for novel, high-risk, or research AI applications).
- **Health Literacy**: AI disclosures must be written in plain language accessible to the organization's patient population, including consideration of primary language, literacy level, and cultural context.

Patient-Facing AI Applications

AI systems that interact directly with patients (e.g. chatbots, symptom checkers, patient portal assistants, virtual health assistants) require specific governance:

- **Identification**: Patient-facing AI must clearly identify itself as AI, not as a human clinician or staff member.

- **Scope Limitations**: Patient-facing AI should communicate its limitations and direct patients to human clinicians for questions beyond its scope, especially emergencies, complex clinical questions, and mental health concerns.
- **Safety Boundaries**: Patient-facing AI must be constrained from providing specific diagnostic conclusions, prescribing medications, or advising patients to discontinue prescribed treatments. Safety boundaries must be tested (see AI Red Teaming) and monitored in production.
- **Escalation**: Patient-facing AI must provide clear escalation pathways to human clinicians, including emergency protocols.

Patient Data Rights in the AI Context

Organizations should develop policies addressing patient data rights specific to AI:

- Right to know whether AI was used in specific care decisions.
- Right to request human review of AI-influenced clinical decisions.
- Right to understand (in general terms) how their data is used in AI systems, including whether their data may be used for model training or improvement.
- Right to request that their data not be used for AI model training where legally applicable and technically feasible.

Liability, Insurance, and Legal Considerations

AI-related liability and insurance considerations intersect with cybersecurity governance through incident-related damages, regulatory penalties, and the liability implications of security control adequacy.

Liability Allocation

When an AI system contributes to patient harm, liability may fall on the healthcare organization, the AI vendor, the clinician, or some combination. Governance should address:

- **Contractual Indemnification**: Procurement contracts should address liability allocation for AI-related harm, including vendor indemnification, data integrity failures, and failure to meet published performance specifications. See Appendix H for model contract language.
- **Clinical Decision Documentation**: The clinical decision audit trail (Section 9) provides essential evidence for malpractice defense by documenting the AI system's output, the clinician's decision, and the clinician's rationale for accepting or overriding the AI recommendation.
- **Human Override Capability**: The organization should demonstrate that clinicians can override AI recommendations; this is a critical element of liability management.
- **Standard of Care**: As AI becomes prevalent in clinical practice, the standard of care may evolve to expect AI use in certain contexts (e.g., AI-assisted radiology interpretation). Organizations should monitor evolving legal and clinical standards and assess whether their AI governance posture is consistent with emerging expectations.

- **Vicarious Liability**: If a physician is at fault because flaws in an AI system leads to misdiagnosis, for instance, both the hospital and the physician could be liable because the hospital procured a flawed AI system.
- **Negligence**: The organization and/or physician/care staff's failure to monitor AI issues such as bias, diagnose drift, etc.
- **Fraud and False Claims**: Should AI systems hallucinate and fraudulently bill, the organization may be liable for fraud and false claims.

Insurance Considerations

Organizations should evaluate whether existing insurance coverage extends to AI-related incidents:

- **Cyber Insurance**: Does the organization's cyber insurance policy cover AI-specific incidents (model compromise, data poisoning, AI-related privacy breach)? Are there exclusions for AI-related claims? Is there a requirement to disclose AI deployments to the insurer?
- **Professional Liability / Malpractice**: Does malpractice coverage extend to AI-assisted clinical decisions? Are there exclusions for algorithmic errors or AI system failures?
- **Product Liability**: If the organization develops and deploys proprietary AI models, does its insurance cover product liability claims arising from model defects?

Organizations should disclose their AI deployment posture to insurers and work with insurance counsel to ensure coverage adequacy. Gaps in coverage should be reported to executive leadership and the Board.

Evolving Legal Landscape

The legal landscape for AI liability in healthcare is evolving rapidly. Organizations should:

- Monitor federal and state legislative developments affecting AI liability.
- Track judicial decisions in AI-related malpractice and product liability cases.
- Engage legal counsel to periodically assess the organization's AI liability exposure.
- Report material changes in the legal landscape to the Board as part of AI risk reporting (see Board Reporting, Appendix K).

Research AI Governance

Research AI governance addresses unique considerations that do not apply to every health organization. Where applicable (e.g. large integrated health networks, academic medical centers), AI-enabled research may require Institutional Review Board (IRB) review when:

- AI is used to analyze human subjects' data (including retrospective studies using existing clinical data).
- AI-generated outputs are used to guide clinical research decisions (participant selection, dosing algorithms, endpoint assessment).
- The research involves developing or validating AI models intended for eventual clinical deployment.
- The research involves novel AI techniques whose risks to participants are not well characterized.

Organizations should provide guidance to researchers on when AI-enabled research triggers IRB review and develop AI-specific IRB review criteria.

Research-to-Clinical Pipeline

When AI systems developed or validated in a research context are proposed for clinical deployment, they must transition through the full AI governance lifecycle described in this document, including clinical validation, security review, privacy review, and risk tiering. A research validation study, however rigorous, does not substitute for the governance requirements of clinical deployment. Organizations should define the formal transition process from research to clinical use.

Federated Learning and Multi-Institutional Collaboration

Research AI collaborations involving multiple institutions introduce governance challenges (see Federated Learning and Collaborative AI Development). In addition, research collaborations should address the following: intellectual property ownership of resulting models, publication rights and data sharing restrictions, regulatory compliance across participating institutions (especially when institutions are subject to different jurisdictions), and data use agreement alignment across participating sites.

Dual-Use Considerations

Some AI research may have dual-use implications such as models developed for beneficial purposes that could be repurposed for harmful applications (e.g., drug discovery models repurposed for toxicology exploitation). Organizations should evaluate dual-use risk for AI research projects with significant capability development and establish review processes for dual-use concerns.

Align Regulations and Conformance Recommendations

Clear and actionable guidance is critical to ensuring regulatory alignment for AI adoption. Recommendations reflect applicable federal and state statutes, ensuring legal compliance and operational resilience.

Regulatory Crosswalk

Organizations should map governance controls against the following frameworks:

- **Data Protection**: HIPAA Privacy & Security Rules govern PHI, while state laws like the NY Stop Hacks and Improve Electronic Data Security Act (SHIELD Act) and California Consumer Privacy Act (CCPA)/California Privacy Rights Act (CPRA) introduce consumer rights and breach notification requirements.
- **Transparency**: There may be federal regulations that address predictive decision support in EHRs, and state bills (e.g., CA SB 53) mandate AI transparency.
- **Device Regulation**: The FDA regulates Software as a Medical Device (SaMD) and medical devices with AI embedded.

- **Cybersecurity**: Section 524B of the FD&C Act mandates cybersecurity requirements for cyber devices (including AI-enabled devices), and the Cyber Incident Reporting for Critical Infrastructure Act (CIRCIA) mandates incident reporting for critical infrastructure.

Standards Mapping

Organizations should identify AI-specific standards to operationalize controls:

- NIST AI Risk Management Framework: Voluntary guidance for managing AI risks like the National Institute of Standards and Technology AI Risk Management framework.
- FDA Guidance: Recommendations for AI-enabled device marketing submissions.
- ISO/IEC 42001/23894: Standards for AI management systems and risk management.
- EU AI Act / GDPR: International regulations affecting global healthcare organizations.
- AAMI CR515: Cybersecurity Considerations Unique to Machine Learning Enabled-Medical Devices.
- OWASP Top 10 for LLM Applications: Risk framework for large language model and generative AI deployments.
- OWASP Top 10 for Agentic AI and Non-Human Identities
- MITRE ATLAS (Adversarial Threat Landscape for AI Systems): Knowledge base of adversarial techniques against AI systems, analogous to MITRE ATT&CK for traditional cybersecurity. Organizations should consider using ATLAS as a reference for AI-specific threat modeling (see Appendix M: AI Threat Model Template).
- NIST SP 800-218A (Secure Software Development Framework for AI): Extends NIST SSDF to address AI-specific development practices.
- ISO/IEC 27090: Guidance on addressing security threats to AI systems.
- ISO/IEC 27091: Guidance on addressing privacy threats to AI systems.
- HSCC Third-Party AI Risk and Supply Chain Transparency Guide: Companion document providing detailed guidance on AI vendor risk assessment, AIBOM, and supply chain transparency. This Implementation Guide absorbs and references the Third-Party AI Risk Guide for vendor governance requirements.

Assess Governance Effectiveness

Organizations can assess governance effectiveness using a 5-point maturity scale across three objectives:

1. AI Cybersecurity Governance Framework
2. Regulatory Alignment
3. Standards & Compliance Mapping

For the full seven step process, see Appendix C. In summary, governance must align with risk tiers. For example, High Risk systems (clinical diagnostics) require at least Maturity Level 4 before deployment, while Low Risk (administrative) systems may accept Level 2.

Conclusion

Robust AI Cybersecurity Governance ensures that products and tools deployed to healthcare environments are adequately reviewed, and risks are identified and mitigated so that the mission and vision of organizations can be met. This guide empowers healthcare organizations (HCOs) to establish cyber governance frameworks for secure AI implementation to guide that work. It addresses identification and mitigation of AI-specific cyber risks and provides practical tools for tasks such as organizing roles and responsibilities, inventory management, contractual language for vendor relationships, and an AI-specific incident response playbook. The guide also addresses AI supply chain and concentration risk, operational resilience for AI-dependent clinical workflows, non-human identity management, patient engagement and transparency obligations, liability and insurance considerations, and governance requirements for research AI. Organizations should implement the recommendations provided here to ensure safe and effective use of AI tools throughout their organization. With the ever-changing healthcare ecosystem, effective management of AI is critical to patient safety.

Appendix A: AI Usage Inventory

One challenge that organizations can face is the use of ungoverned or unknown AI systems. This can be either intentional (designed to work around what are perceived to be overly restrictive policies and procedures that are judged to be hindering innovation) or unintentional (where the user is not aware that the technology is operating outside of governance and IT oversight). Like the challenges of "shadow IT," the existence of "shadow AI" brings risk to the organization. AI inventory serves as the foundation for governance.

Inventory Methodology

1. **Establish Scope:** Define "AI" for the organization and create standardized categories (diagnostic, predictive, administrative).
2. **Discovery:** Conduct interviews with department heads to document obvious AI and hidden AI components within larger systems.
3. **Technical Mapping:** Map data flows, integrations, and hosting environments (cloud vs. on-premises).
4. **Business Impact:** Assess criticality:
    - **Clinical Decision Support (High Risk):** Sepsis prediction, dosing. Failure leads to missed critical conditions.
    - **Operational Efficiency (Medium Risk):** Patient flow, staffing prediction.
    - **Administrative (Variable Risk):** Billing, coding, scheduling, and fraud detection. Note that billing and revenue cycle AI may warrant Medium or High risk classification depending on financial materiality, regulatory exposure, and payer compliance implications; organizations should assess administrative AI on a case-by-case basis rather than defaulting to a uniform Low risk designation.

AI System Type Classification

Each AI system in the inventory should be classified by system type to enable appropriate governance, risk tiering, and oversight requirements. Organizations should assign one or more of the following classifications to each inventoried system:

- **Traditional Machine Learning (ML):** Systems that use supervised, unsupervised, or reinforcement learning models to generate predictions, classifications, or recommendations based on structured data. Examples include sepsis prediction models, patient readmission risk scores, and staffing optimization algorithms.
- **Generative AI:** Systems that produce novel content — including text, images, code, or synthetic data — based on foundation models or large language models. Examples include clinical documentation assistants, patient communication drafting tools, and AI-assisted coding or summarization platforms.
- **Agentic AI:** Systems that autonomously plan, execute multi-step tasks, invoke tools or APIs, and take actions with limited or no real-time human intervention. Agentic AI systems may operate as single agents or as multi-agent architectures. Examples include automated prior authorization agents, autonomous supply chain ordering systems, and AI-driven clinical workflow orchestration. Agentic AI systems should be assessed in accordance with the enhanced governance requirements defined in the Agentic AI section of this guide.

- **Embedded/Component AI:** AI functionality embedded within a larger system (e.g., an EHR module, a medical device, or a SaaS platform) where the AI component may not be independently visible to the organization. Discovery of embedded AI should be a specific focus of the inventory process and vendor due diligence.

Organizations may assign more than one classification where applicable (e.g., a generative AI system with agentic capabilities).

Agentic Capability Inventory

For any AI system classified as Agentic AI — or any system that exhibits agentic characteristics regardless of primary classification — the inventory should capture the following additional attributes:

- **Tool-Calling and API Access:** Whether the system can invoke external tools, APIs, or services, and which integrations are active.
- **Action Execution Authority:** Whether the system can take actions that produce real-world effects (e.g., placing orders, modifying records, sending communications) and whether those actions are reversible.
- **Autonomy Level:** The system's assessed autonomy level per Appendix D, reflecting both decision authority and action capability. Systems that autonomously execute actions in bounded contexts should be classified as at least Level 3 regardless of post-hoc human review.
- **Multi-Agent Interaction:** Whether the system interacts with, delegates to, or receives instructions from other AI agents, and the governance implications of those interactions.
- **Human Override Mechanism:** The mechanism and latency for human intervention, including whether override can halt actions already initiated.
- **Scope Boundaries:** The defined boundaries of the system's autonomous operation, including guardrails, domain restrictions, and escalation triggers.

These agentic attributes should be maintained alongside the standard inventory fields and reviewed whenever the system is updated, its scope of deployment changes, or its autonomy level is reassessed.

Appendix B: AI Governance RACI Matrix

This RACI matrix defines roles and responsibilities across the AI governance lifecycle, covering internally developed AI, third-party AI, AI embedded in devices and platforms, and staff use of external AI tools. Organizations should adapt to their structure and size; small organizations may consolidate roles while preserving the accountability and consultation requirements.

R = Responsible (performs the work) · **A** = Accountable (ultimately answerable) · **C** = Consulted (provides input) · **I** = Informed (kept updated)

PHASE 1: STRATEGY, POLICY & GOVERNANCE STRUCTURE

Activity	R	A	C	I
AI Strategic Plan & Risk Appetite Statement	AI Program Lead, CISO	CEO, Board	CMO/CMIO, CFO, Legal, Privacy Officer	Senior Mgmt, AI Governance Committee
AI Governance Policy & Acceptable Use Policy	Compliance, CISO, Privacy Officer	AI Governance Committee	Legal, CMO/CMIO, Clinical Informatics, HR	Board, Senior Mgmt, All Staff
AI Governance Committee Formation, Charter & Integration with Existing Governance Bodies	AI Program Lead	CEO	CMO/CMIO, CISO, Privacy Officer, Legal, Patient Safety Officer, IRB Chair	Board, Senior Mgmt
Board AI Education & Reporting Cadence	AI Program Lead, Compliance	Board Chair	CEO, CISO, General Counsel, CMO/CMIO	AI Governance Committee

PHASE 2: DISCOVERY, INVENTORY & RISK CLASSIFICATION

Activity	R	A	C	I
AI System Discovery (Enterprise-Wide, Including Shadow AI & Embedded AI)	IT Asset Mgmt, Security Team, Clinical Engineering	CIO	Dept Heads, Clinical Informatics, Procurement	CISO, AI Governance Committee
AI Inventory Creation & Maintenance (Technology Type,	IT Asset Mgmt, AI Program Lead, Business Owners	CIO	Security, Clinical Informatics, Clinical Engineering, Data Science/ML Engineering	AI Governance Committee, Compliance

Activity	R	A	C	I
Autonomy Level, Agentic Capabilities)				
Risk Tier Assignment & Due Diligence Depth Determination	Security Team, Business Owner	CISO	Clinical Leadership, Privacy Officer, Compliance, Patient Safety Officer	AI Governance Committee, Senior Mgmt
AI Supply Chain Mapping, AIBOM Collection & Concentration Risk Assessment	Security Team, Procurement	CISO	Business Owners, IT Ops, Legal, CFO	AI Governance Committee, Board

PHASE 3: USE CASE JUSTIFICATION, PROCUREMENT & CONTRACTING

Activity	R	A	C	I
Use Case Justification, Clinical Needs Assessment & Ethical Screening	Business Owner, Clinical Dept Lead	AI Governance Committee	CMO/CMIO, Privacy Officer, Security, Compliance, Bioethicist, Patient Advocates	Senior Mgmt
Vendor Due Diligence (AI-Specific Assessment per Appendix G, Including Bias, Transparency, Supply Chain)	Security & Compliance Teams, Data Science/ML Engineering	CISO	Business Owner, Privacy Officer, Clinical Leadership, Legal	AI Governance Committee
Contract & BAA Negotiation (AI Clauses, SLAs, Liability, Portability, AIBOM per Appendix H)	Legal, Procurement, Privacy Officer	General Counsel	CISO, Compliance, Business Owner, CFO, Risk Mgmt	Senior Mgmt, Board (if material)
Open-Source Model Provenance & Integrity Verification	Data Science/ML Engineering, Security	CISO	IT Ops, Legal	AI Governance Committee

PHASE 4: DEVELOPMENT, VALIDATION & PRE-DEPLOYMENT

Activity	R	A	C	I
Secure AI Development, Training Data Governance & Prompt Governance	Data Science/ML Engineering, Software Engineering	AI Program Lead	CISO, Privacy Officer, Clinical Leadership, Compliance	AI Governance Committee
AI Threat Modeling & AI Red Team Testing (High/Critical)	Security Team, Data Science/ML Engineering	CISO	AI Program Lead, Business Owner, Clinical Informatics	AI Governance Committee
Clinical Validation, Bias/Fairness Testing & Explainability Documentation	Clinical Review Team, Data Science/ML Engineering, Quality	CMO/CMIO	Patient Safety Officer, Bioethicist, Patient Advocates, Compliance	AI Governance Committee
Privacy Impact Assessment & Security Assessment	Privacy Officer, Security Team	Chief Privacy Officer, CISO	Compliance, Business Owner, Legal, Clinical Leadership	AI Governance Committee, Senior Mgmt
Non-Human Identity Provisioning, Least-Privilege Access & Agentic AI Controls Verification	IT Ops, IAM Team, Security	CISO	AI Program Lead, Business Owner, Clinical Leadership	Compliance
Manual Fallback Documentation, Staging Testing, V&V & Go-Live Readiness Approval	Business Owner, IT Team, Quality	AI Governance Committee	CMO/CMIO, CISO, Privacy Officer, Patient Safety Officer, Vendor	Board (Critical risk), Senior Mgmt

PHASE 5: DEPLOYMENT & CHANGE MANAGEMENT

Activity	R	A	C	I
Production Deployment, Version Baseline & Audit Trail Activation	IT Team, Clinical Informatics	CIO	Security, Business Owner, Clinical Leadership, Vendor	Senior Mgmt, Compliance
Role-Based Training (Clinical, Operational, IT/Security) & Acceptable Use Attestation	Business Owner, Training Team, HR	Business Owner, Chief Compliance Officer	Clinical Leadership, Security, AI Program Lead	Compliance, AI Governance Committee

Activity	R	A	C	I
Patient Notification & Transparency Implementation	Business Owner, Patient Relations	CMO	Privacy Officer, Legal, Compliance, Communications	AI Governance Committee

PHASE 6: ONGOING MONITORING & PERFORMANCE

Activity	R	A	C	I
Clinical Performance Monitoring (Real-World vs. Validation, Drift, Override Rates, Alert Fatigue)	Clinical Informatics, Data Science/ML Engineering, Quality	CMO/CMIO	Patient Safety Officer, Business Owner	AI Governance Committee
Security Monitoring, AI Threat Detection, DLP & Agentic Action Log Review	Security Ops, IT Ops	CISO	Business Owner, Data Science/ML Engineering, Privacy Officer	Compliance, Senior Mgmt
Bias/Fairness Monitoring, Hallucination Monitoring & Explainability Reassessment (LLM/GenAI)	Data Science/ML Engineering, Quality	AI Program Lead	Clinical Leadership, Bioethicist, Compliance	AI Governance Committee
Clinician Feedback Collection, Triage & Vendor Notification	Clinical Informatics, Business Owner	AI Program Lead / CMO	Data Science/ML Engineering, Quality, Vendor	AI Governance Committee
Vendor Performance Monitoring, Supply Chain Change Monitoring & Non-Human Identity Lifecycle	Security Team, Vendor Mgmt, IAM Team	CISO, CPO	Business Owners, Procurement, IT Ops	AI Governance Committee

PHASE 7: UPDATE & PATCH MANAGEMENT

Activity	R	A	C	I
Vendor Update Triage, Staging Testing & Clinical/Security Review	IT Team, Security Team, Business Owner	CIO, CISO	Clinical Leadership, Quality, Vendor, Data Science/ML Engineering	AI Governance Committee, Compliance

Activity	R	A	C	I
Post-Update Verification, Bias Re-Evaluation & Explainability Check (per Appendix I Checklist)	IT Team, Data Science/ML Engineering, Business Owner	CIO, AI Program Lead	Clinical Leadership, Security, Privacy Officer	AI Governance Committee, Compliance
Prompt Template Change Control (LLM/GenAI)	Data Science/ML Engineering, Clinical Informatics	AI Program Lead	CMO/CMIO (clinical prompts), Security, Privacy Officer	AI Governance Committee

PHASE 8: INCIDENT RESPONSE & RECOVERY

Activity	R	A	C	I
AI Incident Detection, Classification & Kill-Switch / Fallback Activation	Security Ops, IT Ops	CISO (security), CMO (clinical safety)	Business Owner, Privacy Officer, Patient Safety Officer	Senior Mgmt, AI Governance Committee, Legal
Clinical Impact Assessment (Blast Radius — Patients, Decisions, Time Window)	Clinical Review Team, Quality, Patient Safety Officer	CMO	Clinical Informatics, Business Owner, Data Science/ML Engineering	AI Governance Committee, Senior Mgmt, Legal
Evidence Preservation, Credential Revocation, Agentic Containment & Vendor Coordination	Security Team, IT Ops, IAM Team	CISO, General Counsel	AI Program Lead, Business Owner, Procurement, Vendor	Senior Mgmt
Model Revalidation, Post-Incident CAPA & Regulatory Notification (FDA, HIPAA, CIRCIA, State)	Data Science/ML Engineering, Compliance, Regulatory Affairs	AI Program Lead, General Counsel	Security, Clinical Leadership, Quality, Privacy Officer	Senior Mgmt, Board
AI Incident Tabletop Exercises (Annual Minimum)	Security Team, AI Program Lead	CISO	Clinical Leadership, IT Ops, Privacy Officer, Legal, Business Owners	AI Governance Committee

PHASE 9: END-OF-LIFE & DECOMMISSIONING

Activity	R	A	C	I
EOL Planning, Clinical Workflow Transition & Replacement (if applicable)	Business Owner, IT Team, Clinical Dept Lead	CIO, CMO/CMIO	Security, Compliance, Vendor Mgmt, Patient Safety Officer	AI Governance Committee, Senior Mgmt
Data Extraction, Migration Validation & Model Artifact Archival	IT Team, Data Science/ML Engineering	CIO	Business Owner, Security, Privacy Officer, Legal, Vendor	Compliance, Senior Mgmt
Secure Data Destruction, Non-Human Identity Deprovisioning & Inventory Update	IT Team, Security Team, IAM Team	CISO, CIO	Vendor, Business Owner, Privacy Officer	Compliance, AI Governance Committee

---

CONTINUOUS ACTIVITIES

Activity	R	A	C	I
AI Governance Policy Updates & Regulatory Landscape Monitoring	Compliance, Legal, Privacy Officer, Security	Chief Compliance Officer, General Counsel	Clinical Leadership, AI Program Lead	Senior Mgmt, Board
Board AI Risk Reporting (Quarterly per Appendix K)	AI Program Lead, CISO, Compliance	AI Governance Committee Chair	CMO/CMIO, Privacy Officer, Legal, CFO	Board
AI Governance Maturity Assessment (Annual)	AI Program Lead, Compliance	AI Governance Committee Chair	CISO, CMO/CMIO, Privacy Officer	Senior Mgmt, Board
AI Training & Education (Clinician Literacy, Security Training, Board Workshops)	Training Team, AI Program Lead, Clinical Informatics	Business Owner (role-specific), Board Chair (Board)	HR, Clinical Leadership, Security	Compliance
Shadow AI Enforcement, Acceptable Use Monitoring & Internal Audit	Security Team, IT Ops, Internal Audit	CISO, Chief Audit Executive	Privacy Officer, HR, Compliance, Dept Heads	AI Governance Committee, Board Audit Committee
Insurance, Liability Review & Patient Data Rights Management (AI-Specific)	Risk Mgmt, Legal, Privacy Officer, Patient Relations	General Counsel, CFO, Chief Privacy Officer	CISO, AI Program Lead, Compliance	Board, Senior Mgmt

Activity	R	A	C	I
Research AI Governance & IRB Coordination	Research Leadership, AI Program Lead	IRB Chair / Chief Research Officer	Data Science/ML Engineering, Privacy Officer, Legal	AI Governance Committee

---

Organizations should customize role assignments for their governance structure and size. The minimum viable model requires clinical, security, and privacy perspectives in every AI governance decision.

Appendix C: Maturity Model

Governance effectiveness is assessed on a 5-point scale:

- Level 1 (Not Started): No governance; AI deployed without clinical consultation or security review.
- Level 2 (Initial Implementation): Ad hoc, siloed efforts; informal clinical input.
- Level 3 (Defined and Repeatable): Documented framework exists; structured regulatory crosswalks; governance consistently applied.
- Level 4 (Managed and Monitored): Governance embedded in enterprise processes; metrics tracked; crosswalks updated quarterly.
- Level 5 (Optimized): Dynamic adaptation; clinicians co-lead governance; automated compliance and audit trails.

How to Conduct a Maturity Assessment

Organizations should conduct a maturity assessment at least annually — and following any material change in AI deployment scope, organizational structure, or regulatory requirements — using the following steps:

- Step 1: Assemble the Assessment Team. Identify a cross-functional assessment team that includes representation from IT/security, clinical leadership, compliance, legal, privacy, and the AI Governance Committee. No single function should score the organization in isolation. For objectivity, organizations at Level 3 or above should consider involving internal audit or an independent third party in the assessment.
- Step 2: Define Assessment Scope. Determine whether the assessment will be conducted at the enterprise level, by business unit, or by AI risk tier (Low, Medium, High, Critical). Organizations with diverse AI portfolios should score separately by risk tier, as governance maturity for low-risk administrative AI may differ significantly from maturity for high-risk clinical AI. Document the scope before beginning.
- Step 3: Gather Evidence. For each objective below, collect the documentation, artifacts, and data needed to substantiate a score. Evidence may include AI inventory records, governance policies, approval workflows, audit reports, crosswalk documents, training records, metrics dashboards, and incident logs. Scores should be based on demonstrated, documented practice — not intent or plans.
- Step 4: Score Each Objective. Using the Maturity Scoring Template below, score each objective from 1 to 5 based on the level description that best fits the organization's current state. Where the organization falls between two levels, assign the lower score unless there is documented evidence that the higher level criteria are substantially met. Each objective should be scored independently.
- Step 5: Determine Overall Maturity. Use the lowest of the three objective scores as the overall maturity level, unless the organization formally documents and the AI Governance Committee approves an alternative scoring rule (e.g., weighted average based on risk exposure). The conservative default ensures that a significant gap in any one area is not masked by strength in others.
- Step 6: Identify Gaps and Prioritize Remediation. For any objective scored below the organization's target maturity level, document the specific gaps, the evidence (or lack thereof) that drove the score, and a remediation plan with owners, timelines, and success criteria. Gap remediation should be tracked as a governance metric and reported to the AI Governance Committee.

- Step 7: Document and Report. Complete the Maturity Assessment Summary (template below) and present findings to the AI Governance Committee and executive leadership. Retain assessment records for audit and year-over-year trend analysis.

Maturity Scoring Template

Score each objective from 1 to 5 based on the best-fitting description. Organizations may score separately by AI risk tier (Low, Medium, High, Critical).

Objective 1: AI Cybersecurity Governance Framework

- 1 — Not Started: No AI inventory exists. No consistent reviews or defined ownership of AI systems. AI is deployed without governance oversight.
- 2 — Initial Implementation: Partial inventory covering some known AI systems. Reviews occur ad hoc and are triggered reactively. Approvals and ownership are inconsistent or undocumented.
- 3 — Defined and Repeatable: Complete inventory for all in-scope AI systems. A defined review and approval workflow exists and is consistently followed. Ownership, accountability, and approval authority are documented.
- 4 — Managed and Monitored: Governance is embedded into enterprise processes (e.g., procurement, change management, clinical system oversight). Controls are measurable. Regular monitoring and periodic audits are conducted with documented findings.
- 5 — Optimized: Continuous monitoring is in place with automated evidence collection. Governance processes adapt dynamically to new AI deployments and emerging risks. Rapid improvement cycles are driven by metrics and audit findings.

Evidence examples: AI inventory records, governance charter, approval workflow documentation, audit reports, monitoring dashboards, corrective action logs.

Objective 2: Regulatory Alignment

- 1 — Not Started: No regulatory crosswalk exists. Compliance activity is entirely reactive and occurs only in response to external events or inquiries.
- 2 — Initial Implementation: Informal regulatory mapping exists for some AI systems but is not standardized or consistently maintained. Compliance is addressed on a case-by-case basis.
- 3 — Defined and Repeatable: A documented regulatory crosswalk is used during AI approval and review processes. Applicable regulations (HIPAA, FDA, state AI laws, CMS conditions) are mapped to governance controls.
- 4 — Managed and Monitored: The crosswalk is maintained and updated on a defined cadence (at least quarterly). Evidence of compliance is retained and organized for audit readiness. Regulatory changes are tracked and assessed for impact.
- 5 — Optimized: Continuous compliance monitoring is in place. Audit-ready reporting can be generated on demand. The organization proactively anticipates regulatory developments and adjusts governance before requirements take effect.

Evidence examples: Regulatory crosswalk document, compliance tracking records, evidence of quarterly updates, regulatory change impact assessments, audit-ready reports.

Objective 3: Standards and Compliance Mapping

- 1 — Not Started: No recognized standards (NIST AI RMF, OWASP, MITRE ATLAS, HSCC SMART) have been adopted for AI governance.
- 2 — Initial Implementation: Limited adoption of select standards by individual teams or for specific AI systems. Adoption is informal and not integrated into governance workflows.
- 3 — Defined and Repeatable: Adopted standards are mapped to key governance controls and tailored to risk tier. Standards are referenced in governance documentation and used during AI review and approval.
- 4 — Managed and Monitored: Standards are mapped to specific controls with measurable implementation status. Gaps between adopted standards and actual practice are tracked and reported. Implementation is consistent across AI risk tiers.
- 5 — Optimized: Standards adoption is continuously refined based on emerging frameworks, threat intelligence, and operational experience. Standards compliance is integrated into governance automation and generates automated evidence and reporting.

Evidence examples: Standards mapping documentation, control implementation status reports, gap analysis records, evidence of standards integration in approval workflows, automation configuration.

Maturity Assessment Summary Template

Organizations should complete the following summary for each assessment cycle:

- Assessment Date: [Date]
- Assessment Scope: [Enterprise / Business Unit / By Risk Tier]
- Assessment Team: [Names, roles, and functions represented]
- Objective 1 Score: [1–5] — Key findings: [Summary]
- Objective 2 Score: [1–5] — Key findings: [Summary]
- Objective 3 Score: [1–5] — Key findings: [Summary]
- Overall Maturity Level: [Lowest objective score, or alternative with documented justification]
- Target Maturity Level: [Organization's stated target]
- Key Gaps Identified: [List with objective reference]
- Remediation Plan: [Owner, timeline, success criteria for each gap]
- Prior Assessment Comparison: [Score trend from previous assessment, if applicable]
- Date of Next Scheduled Assessment: [Date]
- Overall Maturity (Recommended): Use the lowest of the three objective scores as the overall maturity level, unless the organization formally documents an alternative scoring rule approved by the AI Governance Committee.

Note for Healthcare Organizations: The autonomy levels described in this appendix are intended to support HDO governance and procurement decisions. They are not equivalent to the safety engineering frameworks FDA uses to evaluate automated medical devices. FDA's evaluation of AI-enabled and automated medical devices focuses on specific use-related risks of automation, including automation bias (the tendency to over-rely on automated outputs without verification), complacency (reduced vigilance due to trust in automation), loss of situational

awareness (reduced understanding of patient or device state due to automation), and skill degradation (decay of manual clinical skills through disuse). For AI-enabled medical devices subject to FDA oversight, HDOs should review device labeling for information on: the conditions under which the device has been validated to operate; fallback modes and the user's expected response when the device exits automated operation; user responsibilities during automated operation and mode transitions; and training requirements for safe use of the device's automated functions. These are the considerations FDA expects manufacturers to address in premarket submissions for automated medical devices, and they provide a more clinically grounded basis for HDO governance decisions than autonomy level classification alone.

Appendix D: AI Autonomy Levels

The AI autonomy framework establishes a common language for responsibility and oversight, similar to autonomous vehicle levels.

- Level 1: Assisted Intelligence. AI supports human decision-making (e.g., spell check, clinical alerts). Humans retain full control. Benefit: Low adoption risk.
- Level 2: Augmented Intelligence. AI executes complex tasks, but humans remain "in the loop" to validate outputs (e.g., radiology AI flagging anomalies). Humans must approve of actions. Risk: Workflow disruption if recommendations are unclear.
- Level 3: Partial Autonomy. AI executes decisions in bounded contexts. Humans are "on the loop" supervising and intervening if needed (e.g., triage bots). Risk: Escalation protocols must be flawless.
- Level 4: Conditional / High Autonomy. AI acts independently, escalating only exceptional cases. Humans are "out of the loop" for standard operations (e.g., automated supply chain). Risk: Reduced transparency.
- Level 5: Full Autonomy. AI operates independently without human oversight. Currently theoretical in clinical care due to ethical and safety risks.

Appendix E: AI Governance Policy Template

AI Governance Policy Template

Health Industry | [Organization Name] | Version [X.X] | [Date]

This template establishes the organizational framework for governing artificial intelligence systems across [Organization Name].

1. Governance Principles

All AI governance activities shall be guided by the following core principles:

- Patient Safety and Clinical Integrity. AI systems in clinical contexts must demonstrably support patient safety, with human oversight proportional to potential harm.
- Transparency and Explainability. Patients, clinicians, and stakeholders have the right to understand when and how AI contributes to decisions affecting them.
- Fairness, Equity, and Bias Mitigation. AI systems must be evaluated for bias across demographic groups, with processes to detect, measure, and mitigate algorithmic disparities.
- Privacy and Data Protection. Full compliance with HIPAA, state privacy laws, and applicable data protection requirements for all PHI in AI systems.
- Accountability. Every AI system must have an identified owner, a defined risk profile, and an accountable human authority. No AI system operates in a governance vacuum.
- Shared Responsibility. AI governance obligations are distributed among all parties in the AI supply chain—organizations, vendors, developers, and deployers—proportional to role and access.
- Security. AI Governance shall not negatively affect cybersecurity and integrity of organizations.

2. Governance Structure and Accountability

AI Governance Committee

A cross-functional AI Governance Committee shall provide oversight, strategic direction, and policy interpretation. Membership shall include executive leadership, clinical leadership (CMO/CMIO), CIO, CISO, compliance/legal/privacy, data science or informatics, risk management, and patient advocacy or ethics representation. The committee shall convene at least quarterly.

Roles and Responsibilities

Role	Primary AI Governance Responsibilities
Board of Directors	Fiduciary oversight of AI risk; approve governance policy and strategic AI direction
CEO / Executive Sponsor	Organizational commitment, resource allocation, responsible AI culture
CIO	AI technology infrastructure, integration standards, architecture review
CISO	AI cybersecurity risk assessment; security control requirements for AI systems
CMO / CMIO	Clinical AI safety and efficacy validation; clinical workflow appropriateness
CPO	HIPAA compliance for AI; data governance oversight
AI Governance Committee	Cross-functional risk triage, policy interpretation, AI system approval/denial
AI System Owners	Operational accountability for assigned AI systems; lifecycle management
Procurement	AI governance requirements in vendor contracts and due diligence

Role	Primary AI Governance Responsibilities
Workforce Members	Policy compliance; report concerns; participate in AI training
Vendors / Business Associates	Meet contractual AI governance obligations; transparency on model behaviors and risks

3. AI System Inventory and Risk Classification

AI System Registry

A comprehensive registry of all AI systems shall be maintained as the single source of truth. Each entry shall capture: system name and vendor; deployment status; AI system owner; clinical vs. non-clinical designation; data types processed (including PHI); integration points; risk tier; last assessment date; and applicable regulatory requirements (FDA, HIPAA, state law).

Risk-Based Classification

Risk Tier	Criteria	Governance Requirements
Tier 1 — Critical	Direct clinical decision-making; life-safety; FDA-regulated; large-scale PHI	Full assessment; Committee approval; continuous monitoring; annual reassessment; incident response plan
Tier 2 — High	Influences clinical workflows; significant PHI; operational dependency	Comprehensive assessment; Committee review; periodic monitoring; biannual reassessment
Tier 3 — Moderate	Administrative automation; limited PHI; supports but does not drive decisions	Standard assessment; department review; annual reassessment
Tier 4 — Low	General productivity tools; no PHI; minimal operational impact	Lightweight assessment; inventory registration; standard IT controls

4. AI Risk Assessment Framework

AI risk assessment must be completed before production deployment and revisited per risk tier schedule. Critically, AI cyber governance—threat modeling, security architecture review, penetration testing—depends on this foundational governance layer being in place first.

Assessment Domains

Each assessment shall evaluate six domains:

- **Clinical Safety and Efficacy:** Validation evidence, human oversight mechanisms, known failure modes, clinical workflow impact.

- **Privacy and Data Governance:** PHI data flows, minimum necessary compliance, data provenance, de-identification practices.
- **Bias and Fairness:** Training data representation, performance disparities across populations, bias testing results, remediation plans.
- **Security and Resilience:** Attack surface analysis, adversarial robustness, access controls, business continuity in degraded mode.
- **Regulatory and Legal Compliance:** HIPAA, FDA status, state AI laws, contractual obligations with vendors and business associates.
- **Ethical and Organizational Impact:** Patient and workforce transparency, workforce augmentation vs. displacement, mission alignment.

Assessment Process

1. AI system owner submits request with system documentation
2. Preliminary screening assigns risk tier
3. Risk assessment performed across all six domains at depth proportional to tier
4. Findings documented with risk scores, mitigations, and residual risk acceptance
5. AI Governance Committee reviews and renders approval, conditional approval, or denial
6. Approved systems enter production with monitoring plan activated
7. Periodic reassessment per tier schedule; event-driven reassessment as needed

5. Third-Party AI and Shared Responsibility

Just as a Covered Entity cannot delegate HIPAA compliance by contracting with a Business Associate, a healthcare organization cannot fully delegate AI governance to vendors. Governance obligations travel with the data and the decisions. For a detailed policy on managing AI third-party and supply chain risk, see the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide.

Vendor Due Diligence and Contractual Requirements

Prior to procurement or renewal of any AI-enabled product, the organization shall: complete the AI risk assessment for the vendor's system; review the vendor's own AI governance practices and model documentation; evaluate AI-specific security posture (data poisoning, model theft, prompt injection); verify regulatory compliance (HIPAA BAA, FDA clearance); assess bias testing and explainability; and confirm incident response commitments.

Contracts shall include provisions for: AI model transparency obligations; notification of material model changes or retraining; data use limitations (especially model training on organizational data); right to audit AI performance, bias, and security; incident notification timelines for AI failures; subprocessor governance; and termination provisions including model decommissioning.

Ongoing Vendor Monitoring

Vendor AI governance is continuous, not point-in-time. Organizations shall monitor vendor AI performance against benchmarks, conduct periodic reassessments, review vendor security certifications (SOC 2, HITRUST), and track vendor notifications about changes, vulnerabilities, or incidents.

6. AI-Specific Security Controls

The following controls supplement existing information security programs, applied proportional to risk tier:

- **Data Security:** Encryption of training data, model weights, and inference data at rest and in transit; least-privilege access controls; DLP for PHI in AI contexts; secure data lifecycle management.
- **Model Security**: Integrity verification (checksums, signing); adversarial attack protection; version control and change management; secure deployment pipelines with environment separation.
- **Infrastructure and Access:** Network segmentation for AI workloads; MFA for administrative access; logging and monitoring of all AI access, queries, and outputs; vulnerability management for AI components.
- **Non-Human Identity Management:** AI agents, automated pipelines, and machine-to-machine integrations shall have unique identities, credential rotation, least-privilege access, and audit logging equivalent to human users.

7. Documentation and Transparency

Each AI system shall be documented with: intended use and operational context; model architecture and training methodology; data sources and known limitations; performance metrics and validation results; known failure modes and mitigations; human oversight mechanisms; and change history. Patients shall be informed when AI materially contributes to clinical decisions. Workforce members shall have clear channels to question AI outputs and report concerns.

8. Continuous Monitoring and Incident Response

Monitoring

Scaled to risk tier, monitoring shall cover: performance drift and accuracy degradation; security anomalies and data exfiltration indicators; compliance and vendor status; and ongoing fairness assessments against demographic benchmarks.

AI Incident Response

The incident response plan shall address AI-specific scenarios: model failure affecting patient care; breach involving AI training data or outputs; discovery of significant bias; adversarial attack (prompt injection, data poisoning, model manipulation); and unauthorized shadow AI deployments. Each scenario shall have defined escalation paths, notification requirements, root cause analysis, and remediation timelines.

The Closed Loop

Monitoring findings and incident outcomes feed back into risk assessments, the AI registry, policy updates, and vendor management—ensuring continuous improvement. This feedback mechanism is the hallmark of effective governance and the prerequisite for AI cyber governance.

9. Training and Culture

All workforce members shall receive AI governance training appropriate to their role: general awareness for all staff; role-specific training for clinicians, IT, and data teams on capabilities, limitations, and override procedures; and governance training for leadership on risk, regulation, and accountability. Training shall be delivered at onboarding,

annually, and when AI systems are deployed or updated. The organization shall foster a culture where questioning AI outputs and reporting concerns is expected and protected.

10. Regulatory and Standards Alignment

Framework	Relevance
HIPAA (Privacy, Security, Breach Notification)	Data protection, PHI safeguards, breach response for AI systems
NIST AI Risk Management Framework (AI RMF)	AI risk lifecycle: govern, map, measure, manage
NIST Cybersecurity Framework (CSF) 2.0	Foundational cybersecurity controls for AI infrastructure
HSCC Health Industry Cybersecurity Practices (HICP)	Healthcare-specific cybersecurity practices
OWASP LLM Top 10 / ML Top 10	Application-level AI/ML security vulnerabilities
FDA Software as a Medical Device (SaMD)	Regulatory requirements for AI-enabled clinical software
HSCC SMART Framework	AI maturity assessment for healthcare organizations
State and Federal AI Legislation	Emerging requirements for AI transparency, bias, accountability

11. Policy Governance

This policy shall be reviewed at least annually by the AI Governance Committee. Exceptions require written request with risk justification, compensating controls, and expiration date. This policy operates alongside existing Information Security, HIPAA Privacy/Security, Acceptable Use, Vendor Risk Management, Incident Response, Data Governance, and Ethics policies.

12. AI System Deployment Checklist

#	Assessment Item	✓
1	AI system registered in organizational inventory	□
2	Risk tier classification assigned (Tier 1–4)	□
3	Risk assessment completed across all six domains	□
4	AI Governance Committee approval obtained (Tier 1–2)	□

#	Assessment Item	✓
5	Designated AI System Owner with clear accountability	□
6	PHI data flows mapped and secured	□
7	Bias testing performed and documented	□
8	Human oversight and override mechanisms in place (clinical AI)	□
9	Vendor due diligence completed; contractual provisions in place	□
10	Monitoring plan established proportional to risk tier	□
11	Incident response plan updated for this AI system	□
12	Workforce training updated; documentation audit-ready	□
13	Reassessment date scheduled per risk tier requirements	□

Approval: [Name / Title] _____ Date: _____ [Name / Title] _____ Date:

Appendix F: AI Use Case Justification and Risk Scoring Template

Use case intake fields:

- Use case name and sponsor
- Intended use and users
- Clinical or operational impact
- Autonomy level (Appendix D)
- Data types used (PHI/PII, imaging, notes, device data)
- Integration points (EHR, PACS, device, cloud)
- Vendor or internal build
- Safety considerations and failure modes
- Monitoring plan summary

Risk scoring (0 to 3 each):
Patient safety impact: 0 none, 1 indirect, 2 moderate, 3 high direct

- Influence over clinical decisions: 0 none, 1 informational, 2 decision shaping, 3 decision driving
- Autonomy: 0 Level 1, 1 Level 2, 2 Level 3, 3 Level 4 or 5
- Data sensitivity: 0 none, 1 limited, 2 PHI/PII, 3 PHI at scale or highly sensitive
- Equity and bias risk: 0 minimal, 1 possible, 2 likely, 3 high impact population risk

Total score: _____ / 15

Risk tier mapping (suggested): 0 to 4 Low, 5 to 8 Medium, 9 to 12 High, 13 to 15 Critical

Due diligence depth: Use Appendix G question sets and Appendix H clause sets based on tier.

Appendix G: AI Vendor Assessment Questionnaire (Tiered)

Essential question set (all hospitals, all AI that touches PHI/PII):

1. Describe intended use, contraindications, and known limitations.
2. Does the solution process PHI/PII? Where is it stored and processed (regions)?
3. Will any customer data be used for training or model improvement? If yes, require explicit opt in and describe controls.
4. Provide security architecture overview (hosting, segmentation, IAM, MFA, RBAC).
5. Describe logging and audit trails available to the customer.
6. Provide encryption details for data in transit and at rest.
7. Provide incident response and breach notification timelines and process.
8. Provide sub-processors and fourth party services list.
9. Describe update and patch management, customer notice, and rollback support.
10. Provide penetration testing or independent security assessment evidence and remediation status.
11. Provide model card or equivalent artifact describing performance, failure modes, and validation evidence.
12. Provide bias testing summary and what populations were evaluated.
13. Describe drift monitoring capabilities and customer reporting cadence.
14. Describe data retention and deletion practices at termination.
15. Confirm customer can disable or bypass the system safely (kill switch or workflow bypass).

Enhanced question set (Medium, High, Critical risk):

1. Provide validation study details relevant to the proposed clinical setting.
2. Provide details on prompt injection and data poisoning mitigations where applicable.
3. Provide secure SDLC evidence and vulnerability disclosure policy.
4. Provide availability and downtime procedures and RTO/RPO targets.
5. Provide internal model governance process and change control criteria for model updates.
6. Provide explanation method details appropriate to end users.

Comprehensive question set (Large systems, Critical risk):

Use the full expanded questionnaire from the HSCC Third Party AI Risk and Supply Chain Transparency Guide.

Appendix H: Sample Contract and BAA Language (Prioritized)

This Appendix attempts to prioritize some of the contract and BAA language provided in the Health Industry Third-Party AI Risk and Supply Chain Transparency Guide. For a complete list, please use that document. Organizations should work with their legal counsel to review prior to adopting any clauses in contract negotiations.

Small hospitals (minimum clauses to prioritize):

- No training on customer PHI/PII without explicit written permission.
- Sub-processors bound to equivalent obligations and disclosed in advance.
- Breach and incident notification timelines and cooperation obligations.
- Update notification, change logs, and customer ability to defer or rollback material changes.
- Audit support and evidence provision (security assessment, pen test, SOC2 if available).
- Data return and secure destruction at termination.

Medium hospitals (add):

- Model card delivery per release and bias testing evidence for intended use population.
- Performance and availability commitments and downtime procedures.
- Explicit support for incident investigation including clinical safety escalation.

Large systems (add):

- AI BOM or equivalent component transparency and dependency disclosure.
- Continuous monitoring reporting and contractual right to enhanced evidence requests.
- Explicit allocation of responsibilities for update validation and remediation timelines.

BAA addendum AI specific points (where HIPAA applies):

- Explicit restriction on use of ePHI for training except as expressly permitted in writing.
- Required safeguards, breach reporting, and subcontractor flow downs.
- Return or destruction of ePHI at termination and documentation of completion.

Portability and Lock-In (all tiers):

- Customer retains ownership of all customer data and derived outputs.
- Vendor shall provide data export in a machine-readable standard format upon request and at termination.
- Vendor shall provide transition assistance for a minimum of [90/180] days following termination notice.
- Vendor shall not hold customer data or configurations back to contract renewal.

Supply Chain Transparency (High and Critical risk):

- Vendor shall provide AIBOM documenting foundation model(s), fine-tuning data provenance, third-party dependencies, and infrastructure.
- Vendor shall notify customer within 30 days of material changes to AI supply chain components.
- Customer reserves the right to terminate if vendor materially changes AI supply chain components in a manner that increases risk without customer consent.

Appendix I: Templates and Checklists

1. Privacy Impact Assessment (PIA) template (minimum fields):
    - System name, owner, vendor
    - Data elements processed (PHI/PII), sources, destinations
    - Disclosures (internal, external), cross border transfers
    - Access controls (roles, MFA), logging, retention, deletion
    - Consent and authorization approach (treatment vs research)
    - Risk summary, mitigations, approval sign offs (e.g. Privacy, Security, Legal)
2. Demographic Impact Assessment template:
    - Intended use and population
    - Metrics to stratify (sensitivity, specificity, FPR, FNR, calibration)
    - Subgroups (e.g. race, sex, age, language, payer, socioeconomic proxies)
    - Results summary and disparity thresholds
    - Mitigations (workflow controls, threshold adjustment, vendor remediation)
    - Monitoring cadence and ownership
3. Safety risk assessment template (lightweight):
    - Hazard list (what can go wrong)
    - Severity, likelihood rating, and controls (e.g. oversight, workflow checks, fallback)
    - Residual risk and approval
4. Post update verification checklist:
    - Confirm version, release notes, scope of change, clinical sign off
    - Verify integrations and access controls
    - Run representative test cases and compare to baseline
    - Confirm monitoring and alerting functioning
5. Minimum control set by hospital size:
    - Small: inventory, use case justification, essential vendor questionnaire, minimum contract and BAA clauses, privacy/security review, role based training, incident response playbook with kill switch, vendor supported monitoring review.
    - Medium: small set plus enhanced questionnaire, hazard analysis and lightweight FMEA for high risk, update validation, periodic audits, stronger vendor monitoring cadence.
    - Large: medium set plus continuous monitoring pipelines, internal testing capability, comprehensive questionnaire, dedicated ethics review function, formal audit program across AI portfolio.

Appendix J: AI Governance Committee Charter Template

[Organization Name] AI Governance Committee Charter

1. Purpose: The AI Governance Committee provides multidisciplinary oversight of AI systems across [Organization Name] to ensure safe, secure, ethical, and compliant AI deployment.
2. Authority: The Committee has [decision-making / advisory] authority over AI system approval, risk acceptance, policy enforcement, and exception management. The Committee reports to [Board committee / Executive leadership].
3. Scope: All AI systems within the scope of the AI Governance Policy (Appendix E), including third-party AI, internally developed AI, AI embedded in medical devices or platforms, and staff use of external AI tools.
4. Membership: Quorum: [X] members including 1 clinical, security, and privacy representative.
5. Meeting Cadence: Monthly standing meetings. Ad hoc meetings for urgent reviews (High/Critical risk proposals, AI incidents, regulatory developments).
6. Responsibilities:
    - Review and approve/deny AI system proposals by risk tier.
    - Set and enforce AI governance policy.
    - Review AI-related incidents, near-misses, and corrective actions.
    - Monitor AI portfolio risk posture and maturity.
    - Report to the Board per the Board reporting cadence (Appendix K).
    - Resolve conflicts between clinical benefit and cybersecurity/privacy risk.
    - Review and approve exceptions to the AI governance policy.
7. Decision Framework: When recommendations conflict: (1) Patient safety is paramount. (2) For other risks, the Committee documents risk acceptance with compensating controls. (4) CMO/CMIO, CISO, and Privacy Officer hold escalation authority.
8. Documentation: Meeting minutes, decisions, risk acceptance, and exception approvals shall be documented and retained per organizational recordkeeping policy.
9. Annual Review: The Committee shall conduct annual self-assessment of charter adequacy, membership effectiveness, and governance outcomes.

Appendix K: Board AI Risk Reporting Template

Q U A R T E R L Y

AI Cyber Governance

Board Report

Reporting Period: *[Quarter / Year]*

Prepared By: *[AI Governance Committee Chair / CISO]*

Classification: **Confidential — Board Use Only**

Presented To: *[Board of Directors / Board Risk Committee]*

EXECUTIVE SUMMARY

[Preparer: Provide a 3–5 sentence narrative summary of the organization's AI cyber governance posture this quarter. Highlight the most material risk change, the most significant governance action taken, and any items requiring Board decision. This summary should be readable as a standalone briefing.]

1. AI Attack Surface Summary

Provides the Board with visibility into the scope and risk distribution of AI systems requiring cyber governance oversight.

Metric	Value / Status
Total AI systems under governance	**[#]**
Change from prior period	**[+/- #]**
Distribution by risk tier: Low / Medium / High / Critical	**[#] / [#] / [#] / [#]**
Distribution by autonomy level: L1 / L2 / L3 / L4 / L5	**[#] / [#] / [#] / [#] / [#]**
Agentic AI systems (tool-calling, API access, action execution)	**[#]**
Shadow AI systems discovered this period	**[#]**
AI systems pending governance review (backlog)	**[#]**

BOARD INSIGHT: *[Preparer: Insert 1–2 sentences identifying the most material finding or change this period.]*

2. AI Cyber Risk Posture

Summarizes the organization's cybersecurity risk posture specific to AI systems, including vulnerability management, penetration testing, patching compliance, and threat activity.

Metric	Value / Status
AI-specific penetration tests completed (High/Critical tier)	**[#] of [#] required**
Critical/high-severity AI vulnerabilities identified	**[#]**
AI vulnerabilities remediated this period	**[#]**
AI vulnerabilities open / overdue	**[#] / [#]**

Metric	Value / Status
AI systems with overdue patching or update compliance gaps	**[#]**
Post-update verification failures this period	**[#]**
Mean time to remediate AI-specific vulnerabilities	**[# days] | Trend: [↑↓→]**

AI-Specific Threat Activity: [Summary of AI-specific threat vectors observed or attempted this period — prompt injection, model manipulation, data poisoning, adversarial inputs, unauthorized agentic actions.]

BOARD INSIGHT: *[Preparer: Insert 1–2 sentences identifying the most material finding or change this period.]*

3. AI Incidents, Safety Events, and Breach Exposure

Reports AI-related cybersecurity incidents, patient safety events with a cyber nexus, and any breach or regulatory reporting obligations triggered by AI system failures.

Metric	Count	Detail
AI-related cybersecurity incidents	[#]	[Severity distribution: Critical / High / Medium / Low]
Incidents involving PHI/PII exposure	[#]	[Summary]
Incidents requiring external breach notification	[#]	[Summary]
AI-related patient safety events (cyber nexus)	[#]	[Summary]
Near-misses (contained before impact)	[#]	[Summary]

Metric	Count	Detail
Open corrective actions from prior incidents	[#]	[# overdue]

Incident Response Readiness: AI incident response plan exercised or updated this period: [Yes/No]. If yes, summarize exercise findings and plan updates.

> **BOARD INSIGHT:** *[Preparer: Insert 1–2 sentences identifying the most material finding or change this period.]*

4. Third-Party AI and Supply Chain Risk

Addresses cybersecurity risk introduced through the AI vendor supply chain, including foundation model concentration, vendor security posture, and contractual governance compliance.

Metric	Value / Status
Total AI vendors	**[#]**
High/Critical risk AI vendors	**[#]**
Vendors with overdue security assessments or attestations	**[#]**
Third-party AI risk assessments completed this period	**[#] of [#] required**
Contractual AI governance obligation compliance rate	**[%]**

Foundation Model Concentration Risk: [Summary — e.g., "Seven clinical AI systems depend on a single foundation model provider. The committee is evaluating contingency requirements."]

Vendor Incidents or Material Changes: [Summary of any vendor security incidents, model updates, acquisition/ownership changes, or contractual defaults this period.]

> **BOARD INSIGHT:** *[Preparer: Insert 1–2 sentences identifying the most material finding or change this period.]*

5. Governance Maturity and Compliance

Reports AI cyber governance maturity scores per Appendix C, regulatory alignment status, and audit findings.

Governance Objective	Score (1–5)	Trend	Target
Obj. 1: AI Cybersecurity Governance Framework	[1–5]	[↑↓→]	[1–5]
Obj. 2: Regulatory Alignment	[1–5]	[↑↓→]	[1–5]

Governance Objective	Score (1–5)	Trend	Target
Obj. 3: Standards and Compliance Mapping	[1–5]	[↑↓→]	[1–5]
Overall Maturity (lowest objective)	[1–5]	[↑↓→]	[1–5]

Key Maturity Improvements: [Narrative summary of governance advances this period.]

Key Maturity Gaps Requiring Remediation: [Narrative summary of gaps, with remediation owners and timelines.]

Audit Findings: [#] open AI governance audit findings, [#] closed this period. Summary of material findings.

Regulatory Update: [New or pending regulations affecting AI cyber governance. Relevant enforcement actions in the healthcare sector. Changes to the organization's liability or insurance posture.]

> **BOARD INSIGHT:** *[Preparer: Insert 1–2 sentences identifying the most material finding or change this period.]*

6. Decisions, Escalations, and Actions Requested

Summarizes AI Governance Committee decisions and identifies items requiring Board awareness, action, or resource allocation.

Category	Detail
Key AI Governance Committee decisions this period	[Summary]
Items requiring Board awareness (no action needed)	[List]
Items requiring Board action or approval	[List with recommended action]
Resource or investment needs identified	[Summary, if any]

> **BOARD INSIGHT:** *[Preparer: Insert 1–2 sentences identifying the most material finding or change this period.]*

Appendix: Definitions for Board Members

Reference definitions for AI-specific terminology used in this report.

Term	Definition
Agentic AI	AI that autonomously executes multi-step tasks, calls external tools or APIs, and takes actions with limited real-time human oversight.
Foundation Model Concentration	Dependency of multiple AI systems on a single underlying AI model or model provider, creating correlated failure or supply chain risk.
Shadow AI	AI tools or systems in use within the organization that have not been inventoried or brought under governance oversight.
Post-Update Verification	Testing performed after an AI system is updated to confirm that the update has not degraded performance, introduced bias, or created new vulnerabilities.
Autonomy Level (L1–L5)	A classification reflecting the degree of independent decision-making and action authority an AI system exercises. See Appendix D of the AI Cyber Governance Framework.
Risk Tier	Classification of AI systems by potential impact (Low, Medium, High, Critical) used to calibrate governance rigor, testing frequency, and oversight requirements.
Drift	Statistical deviation of an AI system's behavior or data inputs from its original training conditions, which may degrade accuracy or introduce bias over time.
Prompt Injection	An adversarial attack in which malicious input is crafted to manipulate an AI system into producing unintended outputs or bypassing its safety controls.

End of Report

Next Scheduled Report: [Quarter / Year]

Appendix L: AI Incident Response Playbook Template

1. Incident Classification Matrix:

Category	Description	Severity Indicators	Example
Model Failure	Performance degradation below validated thresholds	Accuracy drop >X%, alert volume anomaly, clinician override rate spike	Sepsis model sensitivity drops from 0.85 to 0.60
Model Compromise	Deliberate adversarial manipulation	Unexplained output pattern changes, adversarial indicators in logs	Prompt injection causes clinical chatbot to provide harmful advice
Data Poisoning	Compromised training or inference data	Data integrity alerts, anomalous data patterns, upstream source compromise	Corrupted lab feed degrades diagnostic AI accuracy
Privacy Breach	PHI exposed through AI outputs or processes	PHI in model outputs, unauthorized data transmission, extraction attack detected	LLM reproduces patient identifiers from training data
Agentic Unauthorized Action	AI system acts beyond authorized scope	Unauthorized record access, unapproved order submission, privilege escalation	Prior auth agent modifies formulary records without approval
Supply Chain Compromise	Vulnerability in underlying AI component	Vendor advisory, CVE publication, threat intelligence alert	Foundation model vulnerability affects 4 clinical applications

2. Response Procedures (per category):

For each category, document:

- Detection indicators and escalation triggers
- Immediate containment actions (kill-switch, credential revocation, isolation)
- Clinical impact assessment procedure (which patients, which decisions, time window)
- Evidence preservation requirements (model state, logs, data snapshots)
- Regulatory notification assessment (HIPAA, FDA MDR, CIRCIA, state breach notification)
- Recovery and revalidation requirements
- Post-incident review process

3. Roles and Responsibilities:

- Incident Commander: [CISO / Security Director]

- Clinical Safety Lead: [CMO / Patient Safety Officer]
- AI Technical Lead: [AI Program Lead / ML Engineering Lead]
- Privacy Lead: [Privacy Officer]
- Legal/Regulatory: [General Counsel / Compliance Officer]
- Communications: [Communications Director]
- Vendor Liaison: [Vendor Management / Procurement Lead]
- Human Resources: As required and determined by type of incident.

4. Tabletop Exercise Schedule:

- Frequency: At least annually, or after material changes to the AI portfolio.
- Scenario coverage: Each incident category should be exercised at least once every two years.
- Participants: Must include clinical, IT/security, privacy, legal, and operational representation.
- Documentation: Exercise scenarios, participant lists, findings, and corrective actions must be documented and retained.

Appendix M: AI Threat Model Template

System Identification:

- System name, vendor, version
- AI technology type (traditional ML, LLM, generative AI, agentic)
- Risk tier and autonomy level
- Data types processed (PHI/PII, clinical, operational)

Architecture Summary:

- Inputs (data sources, user inputs, system feeds)
- Processing (model type, inference location, integration points)
- Outputs (clinical recommendations, alerts, actions, documentation)
- External dependencies (APIs, foundation models, cloud services)

Threat Analysis (mapped to MITRE ATLAS where applicable):

Threat	ATLAS Reference	Healthcare Scenario	Likelihood	Impact	Risk	Controls	Residual Risk
Training data poisoning	AML.T0020	Corrupted clinical data feed introduced during model training	[L/M/H]	[L/M/H]	[score]	[controls]	[score]
Adversarial input (evasion)	AML.T0015	Crafted radiology image causes missed finding	[L/M/H]	[L/M/H]	[score]	[controls]	[score]
Model inversion	AML.T0024	Adversary reconstructs patient data from model outputs	[L/M/H]	[L/M/H]	[score]	[controls]	[score]
Prompt injection (direct)	—	User manipulates clinical LLM through crafted prompts	[L/M/H]	[L/M/H]	[score]	[controls]	[score]
Prompt injection (indirect)	—	Malicious content in ingested clinical document manipulates LLM	[L/M/H]	[L/M/H]	[score]	[controls]	[score]
Model extraction	AML.T0044	Adversary clones proprietary clinical model through queries	[L/M/H]	[L/M/H]	[score]	[controls]	[score]
Supply chain compromise	AML.T0010	Backdoor in foundation model affects clinical application	[L/M/H]	[L/M/H]	[score]	[controls]	[score]
Excessive agency	—	Agentic AI executes unauthorized clinical actions	[L/M/H]	[L/M/H]	[score]	[controls]	[score]

Approval: Threat model reviewed and accepted by [AI Governance Committee / Security Lead / Clinical Lead] on [date].

Appendix N: OWASP LLM Top 10 — Healthcare Controls Mapping

OWASP LLM Risk	Healthcare-Specific Scenario	Required Controls	Governance Requirement
LLM01: Prompt Injection	Clinical notes, faxes, or patient forms containing hidden instructions manipulate LLM behavior	Input sanitization, instruction-data separation, output validation, prompt boundary enforcement	Security review gate before deployment; red team testing for injection vectors
LLM02: Sensitive Information Disclosure	LLM outputs PHI memorized from training or leaked from prompt context	DLP on outputs, PHI minimization in prompts, vendor training data audit	Privacy impact assessment; vendor disclosure of training data practices
LLM03: Supply Chain	Compromised foundation model, poisoned fine-tuning data, malicious plugin	AIBOM, supply chain mapping, vendor assessment, model integrity verification	Supply chain risk assessment at procurement; ongoing monitoring
LLM04: Data and Model Poisoning	Corrupted clinical data feed introduces bias or backdoor during training	Training data integrity controls, anomaly detection, vendor transparency	Data governance review; vendor training data attestation
LLM05: Insecure Output Handling	LLM output feeds EHR order entry without validation, propagating errors	Output validation, structured output schemas, human review gates	Clinical validation for systems with downstream clinical system integration
LLM06: Excessive Agency	AI agent submits clinical orders, accesses records, or sends patient communications without authorization	Least-privilege action scope, human authorization gates, containment capability	Agentic AI governance requirements (Section 12)
LLM07: System Prompt Leakage	Adversary extracts system prompt revealing organizational logic or PHI handling rules	Prompt confidentiality controls, extraction monitoring, prompt design review	Security review of system prompts; monitoring for extraction attempts
LLM08: Vector and Embedding Weaknesses	Poisoned embeddings in RAG knowledge base cause retrieval of incorrect clinical information	Embedding integrity verification, access controls on vector databases, RAG corpus governance	Data governance for RAG corpora; integrity monitoring
LLM09: Misinformation	LLM hallucinates drug interaction, dosage, or diagnostic criterion	Grounding (RAG with authoritative sources), output verification, clinician review	Clinical validation; ongoing hallucination monitoring; clinician override capability

OWASP LLM Risk	Healthcare-Specific Scenario	Required Controls	Governance Requirement
LLM10: Unbounded Consumption	Denial-of-service through excessive LLM queries, or cost exploitation	Rate limiting, cost monitoring, usage anomaly detection, resource quotas	Operational monitoring; financial controls on AI service consumption

Appendix O: Agentic AI Governance — Technical Reference

1. Agentic AI Architecture Patterns in Healthcare:

- Single-agent with tool access: An AI assistant that can query the EHR, search formulary databases, or call clinical calculators. Example: a clinical documentation assistant that queries the patient's medication list.
- Multi-step autonomous agent: An AI system that plans and executes a sequence of actions to achieve a goal. Example: a prior authorization agent that reviews the clinical record, identifies applicable criteria, gathers supporting documentation, and submits the authorization request.
- Multi-agent orchestration: Multiple AI agents coordinating to accomplish a complex task. Example: a care coordination system where a scheduling agent, a triage agent, and a documentation agent interact to manage patient intake.
- Human-in-the-loop agent: An agentic system that pauses at defined checkpoints for human review before proceeding. Example: an AI agent that drafts a referral package but requires clinician approval before submission.

2. Agentic AI Threat Vectors (Healthcare-Specific):

Threat Vector	Description	Healthcare Example	Mitigation
Tool misuse	Agent invokes authorized tools for unintended purposes	Documentation agent queries unrelated patient records	Scope-limited tool definitions; query logging and anomaly detection
Goal hijacking	Adversary manipulates agent's goal through input manipulation	Indirect prompt injection causes triage agent to misclassify severity	Input validation; goal verification; output monitoring
Privilege escalation via chaining	Agent A invokes Agent B to access resources Agent A cannot	Documentation agent calls admin agent to modify system settings	Inter-agent access controls; chain-of-trust verification
Feedback loop exploitation	Adversary manipulates agent's self-correction mechanism	Agent repeatedly escalates a non-urgent case due to manipulated feedback	Feedback integrity checks; loop detection; human circuit-breaker
Unauthorized persistence	Agent retains and acts on information across sessions beyond its authorization	Agent caches patient data and uses it in subsequent unrelated sessions	Stateless agent design; session isolation; data purge on session end

3. Minimum Control Requirements for Agentic AI by Risk Tier:

- Low risk (administrative agents, no PHI): Action logging, defined tool scope, periodic review.
- Medium risk (operational agents, limited PHI): Above, plus human authorization for record modification, quarterly access review.
- High risk (clinical-adjacent agents, PHI): Above, plus human-in-the-loop for clinical actions, real-time anomaly monitoring, annual red team testing.
- Critical risk (clinical agents, direct patient impact): Above, plus per-action human approval for consequential actions, continuous monitoring, formal threat model (Appendix N), containment drill testing.

Appendix P: References and Further Reading

- NIST AI Risk Management Framework (AI RMF 1.0)
- NIST SP 800-218A: Secure Software Development Framework for AI
- OWASP Top 10 for LLM Applications v2.0
- MITRE ATLAS (Adversarial Threat Landscape for AI Systems)
- ISO/IEC 42001: AI Management Systems
- ISO/IEC 23894: AI Risk Management
- ISO/IEC 27090: AI Security Threats
- ISO/IEC 27091: AI Privacy Threats
- ISO 14971: Medical Device Risk Management
- HSCC: SMART Maps Toolkit
- IEEE/UL TIPPSS Framework
- FDA Guidance: Artificial Intelligence and Machine Learning in Software as a Medical Device
- FDA Guidance: Predetermined Change Control Plans for AI/ML-Enabled Devices
- HHS Health Industry Cybersecurity Practices (HICP)
- HSCC Third-Party AI Risk and Supply Chain Transparency Guide
- EU AI Act (Regulation 2024/1689)
- GDPR (Regulation 2016/679)
- NIST AI Risk Management Framework: Voluntary guidance for managing AI risks like the National Institute of Standards and Technology AI Risk Management framework.
- FDA Guidance: Recommendations for AI-enabled device marketing submissions.
- ISO/IEC 42001/23894: Standards for AI management systems and risk management.
- EU AI Act / GDPR: International regulations affecting global organizations.
- AAMI CR515: Cybersecurity Considerations Unique to Machine Learning Enabled-Medical Devices.
- OWASP Top 10 for LLM Applications: Risk framework for large language model and generative AI deployments.
- MITRE ATLAS (Adversarial Threat Landscape for AI Systems): Knowledge base of adversarial techniques against AI systems, analogous to MITRE ATT&CK for traditional cybersecurity. Organizations should consider using ATLAS as a reference for AI-specific threat modeling (see Appendix M: AI Threat Model Template).
- NIST SP 800-218A (Secure Software Development Framework for AI): Extends NIST SSDF to address AI-specific development practices.
- ISO/IEC 27090: Guidance on addressing security threats to AI systems.
- ISO/IEC 27091: Guidance on addressing privacy threats to AI systems.
- HSCC Third-Party AI Risk and Supply Chain Transparency Guide: Companion document providing detailed guidance on AI vendor risk assessment, AIBOM, and supply chain transparency. This Implementation Guide absorbs and references the Third-Party AI Risk Guide for vendor governance requirements.