# Document Metadata

author: jfraze@mycomp.org

title: Risk Treatment and Controls for Generative AI

nist\_function: Manage

priority\_phase: Must

last\_reviewed: 2025-06-24

status: draft

## **Purpose**

To define structured controls and treatment strategies for the risks posed by generative AI technologies, including hallucination, prompt injection, PHI leakage, model drift, and supply chain exposure.

## **Scope**

Applies to all GenAI systems developed, deployed, or integrated into hospital IT environments, covering clinical, operational, and administrative use cases.

## **Policy Statement**

All generative AI deployments must include clearly defined risk treatment strategies that align with the organization’s risk tolerance, mapped threats, and lifecycle management practices. Controls must address both technical and organizational safeguards and reflect secure AI engineering principles.

## **Roles and Responsibilities**

• Risk Officers: Approve mitigation strategies for high/critical GenAI risk categories.

• AI Product Owners: Implement controls for hallucination management, access restriction, and use case boundaries.

• DevOps / Security: Ensure execution of prompt controls, token budget limits, and anomaly detection during runtime.

• Compliance & Privacy: Validate that controls align with HIPAA, consent directives, and safe data practices.

## **Implementation Phases**

### **Must Do**

• Apply prompt sanitization, token limit thresholds, and retry behavior capping (SP 800-218A §4.3).

• Implement hallucination mitigation via rule-based post-processing, ensemble filtering, or confidence gating (AI 600-1 §2.3.1).

• Enforce default opt-out of PHI sharing in model input/output interfaces.

• Treat model drift and misuse as notifiable risk triggers (NIST AI RMF MANAGE-2.1).

### **Should Do**

• Integrate misuse detection classifiers or human-in-the-loop review workflows.

• Use embedded telemetry for hallucination/error rate tracking and escalation.

• Store structured logs for replayable incident triage.

• Apply system-level “kill switches” or query throttling for suspect behaviors.

### **Recommended**

• Create model-specific fallback strategies: defaulting to human escalation on ambiguous queries.

• Use watermarking or fingerprinting to detect unauthorized reuse or re-identification (AI 600-1 §2.4.1).

• Maintain a model-specific Risk Control Register (RCR) linked to drift and threat reports.

## **References**

• NIST AI RMF: MANAGE-1, MANAGE-2.3, MANAGE-3.1

• NIST AI 600-1: §2.3.1, §2.4.1, §2.5

• NIST SP 800-218A §4.1–5.2

• Top Vulnerabilities: #11 Hallucination, #1 Prompt Injection, #6 Automation Bias, #16 Alignment Drift

## **Review Cycle**

• Bi-annually or upon critical incident

• At any change to model deployment scope, access policy, or underlying architecture