SECURITY + RISK

CYBERSECURITY AI PLAYBOOK

Bound AI-enabled defensive work with human authorization, independent evidence, revocable authority and tested recovery.

CISOs, security operations, incident response, engineering, architecture, identity, data governance and risk owners.

BOUND

Inventory system, data, access and decision authority.

TEST

Challenge failure, leakage, injection and revocation.

DECIDE

Continue, revise, pause or stop from independent evidence.

MISSISSIPPI ARTIFICIAL INTELLIGENCE NETWORK (MAIN)

MISSISSIPPI'S STATEWIDE AI INITIATIVE | MAINMS.ORG

ONE BOUND AI-ENABLED DEFENSIVE WORK WITH HUMAN AUTHORIZATION, INDEPENDENT EVIDENCE, REVOCABLE AUTHORITY AND TESTED RECOVERY.

WHO

CISOs, security operations, incident response, engineering, architecture, identity, data governance and risk owners.

PRODUCES

A control inventory, bounded pilot record and incident/recovery decision with evidence outside the model.

BOUNDARY

This playbook does not authorize autonomous high-consequence action, replace incident command or prove security because a vendor claims a control.

START WITH THE PEOPLE AND EVIDENCE THAT OWN THE DECISION

01 **NAME THE DECISION**

02 **BRING THE CURRENT RECORD**

03 **BOUND AUTHORITY**

04 **RETAIN THE OUTCOME**

DECISION PROMPT

What must be true—and what evidence must be retained—before this work can continue?

01 ORIENTATION AND SCOPE

SECURITY + RISK / SECTION 01

MAIN AI Playbook

A practical guide to govern AI systems, secure AI components, bound AI-enabled defense, prepare for AI-enabled threats, validate outputs, and respond to AI-related incidents.

Start a Bounded Pilot

Explore MAIN AI Courses

The short answer

SOURCE BASIS NIST Cybersecurity Framework 2.0 | NIST AI Risk Management Framework | CISA/UK NCSC secure AI development guidance

02 HOW SHOULD A SECURITY TEAM BEGIN WITH AI?

SECURITY + RISK / SECTION 02

Inventory the AI system and its dependencies, define a bounded defensive workflow, classify the data, approve the tool, limit privileges, establish a baseline, require analyst review, test failure and abuse cases, log actions, and set stop conditions before the pilot starts.

This voluntary resource is not a compliance standard, threat-intelligence directive, incident-specific runbook, penetration-testing authorization, or substitute for qualified legal, privacy, safety, cyber-insurance, or regulatory advice.

WHO SHOULD USE THIS

- CISOs and security leaders
- Security operations and incident response
- Threat intelligence and detection engineering
- Application, cloud, identity, and data security
- AI/ML engineering and platform teams
- Risk, privacy, legal, procurement, and audit

Secure, defend, thwart

SOURCE BASIS NIST Cybersecurity Framework 2.0 | NIST AI Risk Management Framework | CISA/UK NCSC secure AI development guidance

03 THREE CONNECTED CYBERSECURITY MISSIONS

SECURITY + RISK / SECTION 03

This organizing model is informed by NIST IR 8596, the Cybersecurity Framework Profile for Artificial Intelligence. That publication remains a preliminary draft, not a final standard, as of August 2026.

SECURE AI SYSTEM COMPONENTS

Protect models, data, prompts, applications, agents, identities, tools, infrastructure, interfaces, monitoring, and the software and service supply chain.

CONDUCT AI-ENABLED CYBER DEFENSE

Use AI to support bounded security analysis while preserving evidence, measuring performance, controlling permissions, and requiring accountable review.

THWART AI-ENABLED CYBER ATTACKS

Prepare for faster or more persuasive abuse, including impersonation, malicious automation, model or agent exploitation, data poisoning, and software supply-chain risk.

Use final references as the control baseline. Map applicable requirements and final guidance, such as NIST CSF 2.0, NIST AI RMF, final SSDF publications, contracts, laws, and sector obligations. Then use preliminary or community resources as supplemental threat-informed input.

Risk is contextual

SOURCE BASIS NIST Cybersecurity Framework 2.0 | NIST AI Risk Management Framework | NIST IR 8596 preliminary Cyber AI Profile

04 CYBERSECURITY AI RISK MAP

SECURITY + RISK / SECTION 04

Score the full workflow before deployment and after any material change. High privilege, autonomy, exposure, or consequence can make an apparently simple use case high risk.

DATA SENSITIVITY

Credentials, secrets, personal or customer data, incident evidence, restricted intelligence, source code, configurations, and security telemetry.

PRIVILEGE AND REACH

Read versus write access; endpoint, identity, cloud, email, ticketing, code, firewall, orchestration, or production permissions.

AUTONOMY AND SPEED

How many steps can occur without confirmation? Can the system act faster or farther than a reviewer can understand and stop?

EXTERNAL EXPOSURE

Internet-facing inputs, untrusted documents, email, chat, plugins, Model Context Protocol (MCP) servers and comparable tool/context integrations, APIs, third-party retrieval, and vendor integrations.

CONSEQUENCE AND REVERSIBILITY

Service disruption, evidence loss, user lockout, data disclosure, regulatory impact, safety effect, financial loss, and ability to restore state.

HUMAN REVIEW AND OBSERVABILITY

Reviewer expertise, time, source access, explainability, logging, reproducibility, uncertainty reporting, alerting, and audit trail.

Know what you operate

SOURCE BASIS NIST Cybersecurity Framework 2.0 | NIST AI Risk Management Framework | NIST IR 8596 preliminary Cyber AI Profile

05 GOVERNANCE AND MINIMUM AI INVENTORY

SECURITY + RISK / SECTION 05

Assign accountable owners across business, security, privacy, data, AI engineering, legal, procurement, and incident response. Inventory both official and discovered uses.

Minimum record for each AI-enabled cyber system

Category	Record
Purpose and owner	Business/security objective, accountable owner, operators, affected users, and decision authority.
Components	Model and version, provider, application, agents, prompts, retrieval stores, tools, APIs, infrastructure, dependencies, and data flows.
Access	Identities, roles, secrets, permissions, network paths, connected systems, approval gates, and emergency revocation.
Data	Allowed and prohibited classes, location, retention, training use, encryption, residency, sharing, provenance, and deletion.

Category	Record
Assurance	Threat model, tests, baseline metrics, limitations, change controls, logs, monitoring, fallback, recovery, and review date.

Treat AI as a system

SOURCE BASIS NIST Cybersecurity Framework 2.0 | NIST AI Risk Management Framework | NIST IR 8596 preliminary Cyber AI Profile

06 SECURE THE AI LIFECYCLE AND ATTACK SURFACE

SECURITY + RISK / SECTION 06

Do not focus on the model alone. Threat-model the inputs, outputs, data, tools, identities, integrations, infrastructure, people, and supply chain.

INPUTS AND CONTEXT

- Prompt injection and indirect prompt injection
- Malicious files, links, images, retrieved content, or tool output
- Context poisoning and hidden instructions
- Untrusted user content crossing trust boundaries

MODELS, DATA, AND RETRIEVAL

- Training or retrieval-data poisoning
- Sensitive-data memorization or disclosure
- Model theft, tampering, substitution, or unsupported updates
- Weak provenance, evaluation, or version control

AGENTS, TOOLS, AND IDENTITY

- Excessive agency and permissions
- MCP or comparable integration and tool poisoning
- Confused-deputy and cross-tenant risk
- Secret leakage, unsafe actions, or unauthorized delegation

OUTPUTS AND APPLICATIONS

- Insecure output handling or code execution
- False or manipulated recommendations
- Denial of service or resource exhaustion
- Unsafe rendering, logging, caching, or downstream use

SUPPLY CHAIN

- Model, dataset, package, container, plugin, and service provenance
- Vendor access and fourth-party dependencies
- Signed artifacts, integrity checks, bills of materials, and update controls
- Exit, portability, outage, and compromise plans

MONITORING AND RECOVERY

- Security-relevant prompt, retrieval, tool, identity, decision, and action logs
- Privacy-preserving retention and access control
- Behavior drift, abnormal use, policy bypass, and cost spikes
- Kill switch, credential rotation, rollback, restore, and evidence preservation

Access-control default: separate service identities, deny by default, grant least privilege, scope tokens to one purpose, prefer read-only access, require step-up approval for consequential actions, and test revocation before production.

Read-only does not prevent disclosure. Read-only access can still disclose information. Enforce source and tenant authorization, permitted tools, and outbound destinations in application and network controls outside the model. Test attempts to retrieve or transmit prohibited data; prompt instructions alone are not an access-control boundary.

Build and buy defensibly

SOURCE BASIS NIST SP 800-218A | CISA and UK NCSC secure AI development guidance | NIST AI 100-2 E2025

07 SECURE DEVELOPMENT AND ACQUISITION CHECKPOINTS

SECURITY + RISK / SECTION 07

- 1. Define: security objectives, trust boundaries, prohibited uses, abuse cases, data rules, and measurable acceptance criteria.
- 2. Design: isolate components, minimize privileges, validate inputs and outputs, preserve provenance, and design fail-safe behavior.
- 3. Build or acquire: assess providers and dependencies, protect code and data, document versions, and verify integrity.
- 4. Test: evaluate normal, edge, adversarial, misuse, privacy, reliability, recovery, and human-review cases in a representative environment.
- 5. Deploy: use staged release, approved configuration, secrets management, logging, alerting, rollback, and trained operators.
- 6. Operate: monitor drift and abuse, patch dependencies, retest material changes, review access, practice incidents, and retire safely.

Bounded AI-enabled defense

SOURCE BASIS NIST SP 800-218A | CISA and UK NCSC secure AI development guidance

08 SEVEN DEFENSIVE WORKFLOWS AND THEIR LIMITS

SECURITY + RISK / SECTION 08

Start with drafts, summaries, and recommendations. Preserve source telemetry and require analysts to verify conclusions before any consequential action.

1. LOG AND CASE SUMMARIZATION

Allowed: summarize approved, minimized records and cite source event IDs.

Validate: timeline, omissions, identity, timestamps, and links to raw evidence.

2. ALERT TRIAGE SUPPORT

Allowed: group related alerts and propose investigation questions.

Validate: false-negative risk, severity, asset context, uncertainty, and analyst disposition.

3. THREAT-INTELLIGENCE SYNTHESIS

Allowed: compare licensed sources and extract defensive implications.

Validate: source restrictions, dates, confidence, indicators, attribution, and independent corroboration.

4. DETECTION ENGINEERING

Allowed: draft queries, test cases, mappings, and documentation.

Validate: syntax, coverage, evasion assumptions, false positives, data availability, and isolated testing.

5. VULNERABILITY PRIORITIZATION

Allowed: combine approved asset, exposure, business, exploitation, and remediation context.

Validate: authoritative scanner data, ownership, compensating controls, active exploitation evidence, and human priority decision.

6. SECURE CODE AND DOCUMENTATION

Allowed: explain findings, draft fixes or runbooks, and propose tests in approved repositories.

Validate: secrets, licenses, dependencies, security tests, peer review, and change management.

7. TABLETOP DESIGN

Allowed: create scenario variations, injects, decision prompts, and evaluation criteria using fictionalized data.

Validate: realism, safety, roles, escalation paths, recovery, legal/communications involvement, and after-action evidence.

DO NOT DELEGATE BY DEFAULT

Account disablement, endpoint isolation, firewall or identity changes, production code deployment, evidence deletion, public attribution, breach notification, ransom decisions, or safety-critical response require explicit authorized decision paths.

Defensive implications

SOURCE BASIS NIST Cybersecurity Framework 2.0 | CISA JCDC AI Cybersecurity Collaboration Playbook | NIST AI Risk Management Framework

09 WORKED DEFENSIVE PILOT: READ-ONLY ALERT TRIAGE

SECURITY + RISK / SECTION 09

ILLUSTRATIVE BOUNDED PILOT - NO AUTONOMOUS RESPONSE

Situation: a security team evaluates whether AI can group related alerts and propose investigation questions from approved, minimized records. The pilot is read-only. It cannot disable an account, isolate an endpoint, change a rule, close an incident, delete evidence, notify an affected party, or deploy code.

Control gate	Required pilot record	Acceptance evidence
Inventory	Owner, analysts, model/version, provider, application, prompts, retrieval, tools, identities, data flows, dependencies, and affected systems.	Inventory is complete enough to revoke access, reproduce a result, and identify every trust boundary.
Authorization	Read-only scope, approved data, least-privilege identity, allowed queries, prohibited actions, approver, effective period, and emergency revocation.	No uncontrolled write path; authorization exists outside the model and expires or can be revoked.
Baseline	Current analyst accuracy, false-positive and false-negative definitions, total handling time, source traceability, and escalation quality.	Local numeric targets and stop limits are approved before the representative test set is run.
Challenge	Normal, edge, adversarial, indirect-prompt-injection, missing-data, outage, model-change, privilege, leakage, and recovery cases.	Every case has expected behavior, observed behavior, analyst disposition, correction, and retained source evidence.
Observability	Prompts where lawful, retrieved material, source event IDs, model/config versions, tool calls, identity, approvals, outputs, analyst changes, and final disposition.	The record is stored outside the AI workflow and supports reconstruction without relying on model memory.
Decision	Compare quality, missed risk, analyst effort, security/privacy boundary, reliability, and recovery against the approved baseline and thresholds.	Authorized owner records continue/revise/pause/stop, residual risk, controls, evidence location, effective period, and next review.

FAIL-CLOSED ACCEPTANCE GATE

- Continue only if material conclusions are traceable to authoritative telemetry, the analyst can reproduce the basis, total review burden is acceptable, and tested revocation and recovery work as designed.
- Revise when useful results require narrower data, stronger validation, better logging, different permissions, clearer uncertainty, or a changed analyst workflow.
- Pause or stop for sensitive-data leakage, uncontrolled action, evidence loss, privilege expansion, material false negatives, unverifiable conclusions, failed revocation, failed rollback, or a change outside the approved configuration.

High-consequence action remains on a separately authorized human path. A successful summary or recommendation does not demonstrate that autonomous containment or production change is safe.

SOURCE BASIS NIST Cybersecurity Framework 2.0 | NIST AI Risk Management Framework | NIST SP 800-218A | CISA JCDC AI Cybersecurity Collaboration Playbook

10 PREPARE FOR AI-ENABLED THREATS WITHOUT AMPLIFYING HARM

SECURITY + RISK / SECTION 10

AI can increase the scale, speed, personalization, or adaptability of familiar attacks. Defenders should strengthen verification, identity, telemetry, resilience, and workforce readiness.

IMPERSONATION AND SOCIAL ENGINEERING

Use known-channel verification for sensitive requests, phishing-resistant authentication where practical, transaction controls, role-based escalation, and rehearsed reporting for voice, video, and message impersonation.

MALICIOUS AUTOMATION

Rate-limit and monitor exposed services, protect identities and APIs, reduce attack surface, patch promptly, detect anomalous behavior, and plan for higher-volume or faster-changing campaigns.

AI APPLICATION EXPLOITATION

Separate untrusted content from instructions, validate tool calls and outputs, constrain retrieval, sandbox processing, protect secrets, and test prompt injection and agent abuse cases.

POISONING AND SUPPLY-CHAIN COMPROMISE

Track provenance, restrict write paths, review providers and updates, verify artifacts, monitor data/model behavior, protect build systems, and retain known-good recovery points.

Threat picture through August 2026. Google Threat Intelligence Group reported a threat actor using a zero-day exploit it assessed with high confidence was developed with AI, plus attacks on AI software dependencies that attempted to pivot into wider network environments. GTIG and Mandiant also documented growth in large-scale open-source supply-chain compromises, while Microsoft demonstrated how prompt injection in vulnerable agent frameworks can become a code-execution path. These observations justify faster exposure reduction, stronger identity and package controls, and focused testing of agent integrations; they do not establish that most attacks are autonomous or primarily AI-generated.

Prepare before the alert

SOURCE BASIS NIST AI 100-2 E2025 | NIST SP 800-218A | CISA and UK NCSC secure AI development guidance

11 AI-RELATED INCIDENT RESPONSE SEQUENCE

SECURITY + RISK / SECTION 11

Integrate AI incidents into the organization's established incident-response, privacy, legal, safety, vendor, and communications processes. Preserve evidence and avoid untested containment that could worsen the event.

- 1. Declare and assign. Confirm severity, incident commander, technical leads, decision authority, and required notifications.
- 2. Preserve. Protect relevant prompts, outputs, identities, tool calls, model/config versions, retrieval records, telemetry, and chain-of-custody information.
- 3. Bound. Identify affected users, systems, data, models, agents, vendors, tenants, privileges, and time window.
- 4. Contain safely. Revoke exposed access, pause risky actions, isolate components, or shift to a known fallback using authorized procedures.
- 5. Eradicate and recover. Remove the cause, rotate secrets, restore trusted components and data, validate integrity, and monitor for recurrence.
- 6. Learn and share appropriately. Document root and contributing causes, control gaps, human factors, vendor actions, metrics, and approved information-sharing opportunities.

AI-specific evidence matters. Capture model and application versions, system and user prompts when lawful, retrieval context, tool permissions and calls, safety-policy events, outputs, feedback, configuration, and vendor status, not only conventional network and endpoint logs.

A measured first deployment

SOURCE BASIS NIST Cybersecurity Framework 2.0 | CISA JCDC AI Cybersecurity Collaboration Playbook | NIST AI Risk Management Framework

12 30-DAY CYBERSECURITY AI PILOT

SECURITY + RISK / SECTION 12

Select a lower-risk, read-only workflow using approved and minimized data. The goal is evidence about quality, analyst effort, security, and repeatability, not autonomous response.

Define the evaluation denominator. Use independently labeled malicious and benign cases held out from prompt tuning. Report false negatives as missed malicious cases divided by all malicious cases, and false positives as benign cases

incorrectly escalated divided by all benign cases; retain counts and case mix. Triage results describe the supplied alerts, not attacks absent from upstream telemetry. Compare total analyst time and repeat tests after material changes.

- Week 1: Define. Name owners; threat-model; approve tool, data, permissions, reviewers, baseline, success measures, and stop conditions.
- Week 2: Test. Train a small group; use representative sanitized cases; record prompts, sources, errors, revisions, false positives, and false negatives.
- Week 3: Challenge. Test edge, adversarial, injection, outage, data-leakage, privilege, and recovery cases; verify logging and revocation.
- Week 4: Decide. Compare with baseline; review risk and total analyst time; continue, revise, pause, or stop; document the decision.

Pilot scorecard

Measure	Question
Quality	Were conclusions accurate, complete, sourced, reproducible, and useful?
Analyst effort	Did total time, including prompting, verification, correction, and documentation, improve?
Security and privacy	Did the workflow stay within approved data, access, retention, and logging boundaries?
Operational reliability	How did it behave under edge cases, untrusted inputs, outages, model changes, and degraded data?
Decision	Do evidence and residual risk support continue, revise, pause, or stop?

Stop immediately when: prohibited data is exposed; privileges exceed approval; output causes or recommends unsafe action without the required gate; evidence integrity is threatened; logging or revocation fails; material drift is unexplained; or analysts cannot verify the result in time.

Continue with MAIN

SOURCE BASIS NIST Cybersecurity Framework 2.0 | CISA JCDC AI Cybersecurity Collaboration Playbook | NIST AI Risk Management Framework

13 CYBERSECURITY LEARNING AND IMPLEMENTATION

SECURITY + RISK / SECTION 13

AI FOR CYBERSECURITY COURSE

Review MAIN's current course catalog and availability through participating institutions.

Explore AI courses

MISSISSIPPI STATEWIDE FRAMEWORK

Connect cybersecurity AI work with the state's broader AI priorities and workforce context.

Review statewide priorities

RESPONSIBLE IMPLEMENTATION SUPPORT

Explore MAIN services for strategy, training, convening, and practical AI adoption support.

Explore MAIN services

Cybersecurity AI questions

SOURCE BASIS NIST Cybersecurity Framework 2.0 | NIST AI Risk Management Framework | CISA/UK NCSC secure AI development guidance

14 FREQUENTLY ASKED QUESTIONS

SECURITY + RISK / SECTION 14

WHAT IS A CYBERSECURITY AI PLAYBOOK?

It is a practical operating resource for identifying AI systems and dependencies, matching controls to risk, securing AI components, bounding AI-enabled defense workflows, preparing for AI-enabled attacks, and responding to incidents with accountable human oversight.

CAN AI AUTONOMOUSLY RESPOND TO CYBER INCIDENTS?

Autonomy should be limited by the organization's risk tolerance, testing, permissions, monitoring, and recovery capability. Begin with read-only assistance. Consequential containment, account, network, endpoint, or production changes should require authorized human approval unless a separately validated and governed automation explicitly permits them.

WHAT DATA SHOULD NOT BE ENTERED INTO A PUBLIC AI TOOL?

Do not enter credentials, tokens, private keys, unredacted logs with personal or customer data, sensitive configurations, exploitable vulnerability details, incident evidence, proprietary code, restricted intelligence, or regulated information unless an approved environment and data-use agreement authorize it.

HOW SHOULD SECURITY TEAMS VALIDATE AI OUTPUT?

Compare it with authoritative telemetry and source records, preserve chain of custody where relevant, reproduce queries, test detections and code in an isolated environment, measure false positives and false negatives, review uncertainty, and require an authorized analyst to approve consequential action.

IS NIST IR 8596 A FINAL STANDARD?

No. NIST IR 8596 was released as a preliminary draft in December 2025. As of August 2026, NIST was using feedback to develop the next draft. Clearly label it as preliminary and use final NIST publications and applicable requirements as controlling references.

Authoritative foundation

SOURCE BASIS NIST Cybersecurity Framework 2.0 | NIST AI Risk Management Framework | CISA/UK NCSC secure AI development guidance

15 STANDARDS, DRAFTS, AND THREAT KNOWLEDGE

SECURITY + RISK / SECTION 15

Status labels matter: final, preliminary, and community resources serve different roles.

- CISA, NSA, ASD's ACSC and international partners, Careful adoption of agentic AI services (joint security guidance, May 1, 2026)
- NIST SP 800-61 Rev. 3, Incident Response Recommendations and Considerations for Cybersecurity Risk Management (final, April 2025)
- NIST Cybersecurity Framework 2.0 (final)
- NIST AI Risk Management Framework and Generative AI Profile (final publications)
- NIST AI 100-2 E2025, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (final, March 2025; NIST has posted a planning note for a future update)
- NIST IR 8596, Cybersecurity Framework Profile for Artificial Intelligence (preliminary draft, December 2025; not final)
- NIST IR 8607, Workshop Summary Report for the Cyber AI Profile Hybrid Workshop (final workshop summary, August 2026)
- NIST SP 800-218A, Secure Software Development Practices for Generative AI and Dual-Use Foundation Models (final, July 2024)
- CISA and UK NCSC, Guidelines for Secure AI System Development (joint voluntary guidance, November 2023)
- CISA JCDC AI Cybersecurity Collaboration Playbook (voluntary information-sharing resource, January 2025)
- Google Threat Intelligence Group, AI vulnerability exploitation, augmented operations, and initial access (May 2026 threat research)
- Mandiant, Blueprint for AI-Assisted Vulnerability Management (July 2026 operational guidance)
- Google Threat Intelligence Group and Mandiant, Mitigation Guidance for Supply Chain Compromise (July 2026 threat research and mitigations)
- Microsoft Security, prompt injection and code-execution vulnerabilities in AI agent frameworks (May 2026 security research)
- MITRE ATLAS (living knowledge base of adversary tactics and techniques for AI-enabled systems)
- OWASP Artificial Intelligence Security Verification Standard 1.0 (community verification standard, June 2026; not a regulatory requirement)
- OWASP GenAI LLM Top 10 2026 and OWASP GenAI Agentic Security Initiative (community guidance, not regulatory standards)

SOURCE BASIS CISA, NSA, ASD's ACSC and international partners, Careful adoption of agentic AI services | NIST SP 800-61 Rev. 3, Incident Response Recommendations and Considerations for Cybersecurity Risk Management | NIST Cybersecurity Framework 2.0 | NIST AI Risk Management Framework

16 MAKE AI-ENABLED DEFENSE OBSERVABLE, BOUNDED, AND REVERSIBLE

SECURITY + RISK / SECTION 16

Pair cybersecurity expertise with AI engineering, privacy, legal, procurement, workforce training, and operational ownership. Reassess whenever data, models, integrations, permissions, or threats change.

Browse All AI Playbooks

Contact MAIN

Related MAIN resources

SOURCE BASIS NIST Cybersecurity Framework 2.0 | NIST AI Risk Management Framework | CISA/UK NCSC secure AI development guidance

17 WHERE TO GO NEXT

SECURITY + RISK / SECTION 17

- Business AI Policy Template The governance layer to pair with these security practices.
- AI Policy Guides Templates for education, government, healthcare, business, finance, and nonprofits.
- AI in Mississippi How AI education, training, and policy fit together across the state.

LEAVE WITH A DECISION RECORD

A control inventory, bounded pilot record and incident/recovery decision with evidence outside the model.

01

BOUND

Inventory system, data, access and decision authority.

02

TEST

Challenge failure, leakage, injection and revocation.

03

DECIDE

Continue, revise, pause or stop from independent evidence.

OWN

Named authority and scope

CHECK

Evidence and qualified review

STOP

Pause, reversal and next review

18 SOURCES, REVIEW RECORD AND MAINTENANCE

SECURITY + RISK / SECTION 18

MAIN developed the operating model, decision logic, examples and working tools in this playbook. The reviewed references below support the method and its boundaries. Law, agency guidance, research, institutional examples and vendor documentation serve different roles; none substitutes for applicable law, institutional policy or qualified review. Confirm current applicability and currency before use.

01 CISA, NSA, ASD's ACSC and international partners, Careful adoption of agentic AI services
cyber.gov.au

02 NIST SP 800-61 Rev. 3, Incident Response Recommendations and Considerations for Cybersecurity Risk Management
csrc.nist.gov

03 NIST Cybersecurity Framework 2.0
nist.gov

04 NIST AI Risk Management Framework
nist.gov

05 Generative AI Profile
nist.gov

06 NIST AI 100-2 E2025, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations
csrc.nist.gov

07 NIST IR 8596, Cybersecurity Framework Profile for Artificial Intelligence
nccoe.nist.gov

08 NIST IR 8607, Workshop Summary Report for the Cyber AI Profile Hybrid Workshop
nist.gov

09 NIST SP 800-218A, Secure Software Development Practices for Generative AI and Dual-Use Foundation Models
csrc.nist.gov

10 CISA and UK NCSC, Guidelines for Secure AI System Development
cisa.gov

11 CISA JCDC AI Cybersecurity Collaboration Playbook
cisa.gov

12 Google Threat Intelligence Group, AI vulnerability exploitation, augmented operations, and initial access
cloud.google.com

13 Mandiant, Blueprint for AI-Assisted Vulnerability Management
cloud.google.com

14 Google Threat Intelligence Group and Mandiant, Mitigation Guidance for Supply Chain Compromise
cloud.google.com

15 Microsoft Security, prompt injection and code-execution vulnerabilities in AI agent frameworks
microsoft.com

16 MITRE ATLAS
atlas.mitre.org

17 OWASP Artificial Intelligence Security Verification Standard 1.0
owasp.org

18 OWASP GenAI LLM Top 10 2026
genai.owasp.org

19 OWASP GenAI Agentic Security Initiative
genai.owasp.org

20 NIST IR 8596 preliminary Cyber AI Profile
doi.org

Cover photograph: Network equipment and Ethernet connections. Brett Sayles / Pexels. Contextual photograph; no claim of MAIN affiliation or actual AI use.

KEEP THE METHOD CURRENT AND LOCALLY OWNED

Confirm applicability and currency before production use; record the responsible local owner and review trigger.

OWN

A named person controls scope, access and the decision.

CHECK

Qualified reviewers verify evidence, risks and corrections.

STOP

The work can be paused or reversed outside the approved boundary.