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

START HERE

AUTHORIZE THE TASK. KEEP EVIDENCE OUTSIDE THE MODEL.

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

WHO SHOULD USE THIS

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

WHAT YOU WILL PRODUCE

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

KEEP THESE BOUNDARIES

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

SCOPE ONE DEFENSIVE WORKFLOW

01 **Define the task and owner**

Identify the system, data, permitted actions and accountable security owner.

02 **Limit access and authority**

Start read-only where possible; require human approval for consequential actions.

03 **Test controls independently**

Check leakage, prompt injection, false results, access revocation and recovery.

04 **Record the decision**

Retain test evidence, review measures and explicit conditions to continue or stop.

Built on Mississippi's Statewide AI Framework

The Mississippi Statewide AI Framework (August 2026) is the foundation for this playbook. It supplies shared priorities and a learning progression; this playbook applies them to AI security, bounded defensive work and cybersecurity workforce learning. Both are voluntary. Applicable law, adopted standards and institutional authority remain controlling.

Read the August 2026 Statewide AI Framework (PDF)

Learning progression. Workforce: technical practice through security leadership. This is a specialist application; it does not require every learner to become a security engineer.

Five priorities in practice

Train analysts for usable, approved access (1); challenge conclusions and effects (2); enforce data, identity and system controls (3); practice transferable defensive skills (4); compare labeled cases, total effort and failure evidence (5).

Priority numbers refer to literacy and access (1), ethics and critical thinking (2), privacy and governance (3), workforce readiness (4), and strategic alignment and measurable outcomes (5).

Selected Framework domains, practice and evidence

Framework domain	Practice and evidence to retain
11. Cyber Security and Safety	Test privileges, malicious inputs, leakage and recovery; retain observed results.
6. AI-Enabled Workflow and System Design	Map tools, identities, data and escalation; retain the inventory and threat model.
3. Critical Evaluation of AI Outputs	Verify conclusions against telemetry; retain labeled counts and correction evidence.
8. Human Agency, Judgment, and Restraint	Keep consequential response with authorized people; retain action approval and decision rationale.

Keep the shared human-review boundary

Apply the Framework's human-review triggers (p. 35) before action: assessment or credentials; financial allocation or compensation; health or physical safety; discipline or employment; legal or compliance content; and sensitive or protected information. Its additional agent triggers cover autonomous system-of-record changes; persistent memory; tool access, credentials or external transactions; handoffs in safety-critical, regulated or student-facing processes; missing stop conditions or logged escalation to a named human owner; and institutional responses to synthetic media about identifiable people. Product or pilot approval does not replace action-specific review. Review the system and access before enabling them; review each proposed system-of-record change before execution. Advance permission cannot waive these triggers.

Use evidence within its limits

Use the crosswalk to select practice and retain observable evidence. It is MAIN's implementation mapping, not a validated proficiency test, credential or certification. Local rubrics and pilot measures alone do not establish statewide proficiency or causal impact. Recheck alignment when the Framework changes; record the edition, local owner and next review.

Source: August 2026 Framework, pp. 5-7 (priorities), 14-30 (domains and progression), and 35-37 (review, measurement and maintenance).

01 HOW SHOULD A SECURITY TEAM BEGIN WITH AI?

SECURITY + RISK / SECTION 01

The short answer

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.

Maintain the underlying safeguards: managed updates, unique passwords, MFA, least privilege, protected and tested backups, and trained reporting. Use CISA's voluntary Cross-Sector CPGs 2.0 to prioritize applicable improvements; AI does not replace them.

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

SOURCE BASIS CISA's voluntary Cross-Sector CPGs 2.0

02 THREE CONNECTED CYBERSECURITY MISSIONS

SECURITY + RISK / SECTION 02

Secure, defend, thwart

This organizing model follows the three focus areas in NIST IR 8596, the Cyber AI Profile. It is an Initial Preliminary Draft, not a final standard; use it only as supplemental input.

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.

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

03 CYBERSECURITY AI RISK MAP

SECURITY + RISK / SECTION 03

Risk is contextual

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.

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

04 GOVERNANCE AND MINIMUM AI INVENTORY

SECURITY + RISK / SECTION 04

Know what you operate

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 and router, application, model endpoints, training/evaluation harnesses, agents, prompts, retrieval stores, tools, APIs, infrastructure, compute quotas, dependencies, and data flows.
Access	Human and service identities, roles, secrets, token lifetime, permissions, network paths and egress, connected systems, approval gates, and emergency revocation.
Data	Allowed and prohibited classes, location, retention, training or evaluation use, encryption, residency, sharing or resale by providers and intermediaries, provenance, and deletion.
Assurance	Threat model, tests, baseline metrics, limitations, change controls, logs, monitoring, provider-escalation path, fallback, recovery, and review date.

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

05 SECURE THE AI LIFECYCLE AND ATTACK SURFACE

SECURITY + RISK / SECTION 05

Treat AI as a system

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, long-lived credentials, and permissions
- MCP or comparable integration and tool poisoning
- Confused-deputy and cross-tenant risk
- Unapproved network egress, shared-service access, 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, router, reseller, and fourth-party access or data reuse
- 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, boundary probing, burst activity, unexpected model use, 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.

SOURCE BASIS NIST SP 800-218A | CISA and UK NCSC secure AI development guidance | NIST AI 100-2 E2025 | Unit 42, An AI-Assisted Cyber Attack: Inside a Unit 42 Investigation

06 SECURE DEVELOPMENT AND ACQUISITION CHECKPOINTS

SECURITY + RISK / SECTION 06

Build and buy defensibly

1. Define: security objectives, trust boundaries, prohibited uses, abuse cases, data rules, and measurable acceptance criteria.
2. Design: isolate components; deny network egress and shared-service access by default for high-risk agent, training, or evaluation workloads; minimize privileges; validate inputs and outputs; preserve provenance; and design fail-safe behavior.
3. Build or acquire: assess providers, routers, data practices, emergency support, and dependencies; protect code and data; document versions; and verify integrity.
4. Test: evaluate normal, edge, adversarial, misuse, privacy, reliability, sandbox escape, boundary probing, provider refusal, outage, recovery, and human-review cases in a representative environment.
5. Deploy: use staged release, approved configuration, short-lived credentials, secrets management, logging, alerting, rollback, and trained operators.
6. Operate: monitor drift, abnormal agent loops, unexpected model use, and abuse; patch dependencies; retest material changes; review access; practice incidents; and retire safely.

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

07 SEVEN DEFENSIVE WORKFLOWS AND THEIR LIMITS

SECURITY + RISK / SECTION 07

Bounded AI-enabled defense

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.

SOURCE BASIS NIST Cybersecurity Framework 2.0 | CISA JCDC AI Cybersecurity Collaboration Playbook | NIST AI Risk Management Framework | Hugging Face, Security incident disclosure — July 2026

08 WORKED DEFENSIVE PILOT: READ-ONLY ALERT TRIAGE

SECURITY + RISK / SECTION 08

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.

Control gate	Required pilot record	Acceptance evidence
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

09 PREPARE FOR AI-ENABLED THREATS WITHOUT AMPLIFYING HARM

SECURITY + RISK / SECTION 09

Defensive implications

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

Verify sensitive requests through a known, independently obtained contact, as advised by FBI/IC3 impersonation guidance. Use transaction controls and rehearsed reporting. Prefer available phishing-resistant MFA, such as a supported security key or passkey; use NIST SP 800-63B-4 for authentication requirements appropriate to the system.

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.

Consent phishing can survive a password change. A legitimate sign-in or consent screen can grant a malicious app access to email or files. Review the app, publisher and requested permissions before consenting. If access was granted, use the connected service's security settings or authorized administrator to revoke the app's permissions and affected tokens; a password change alone may leave access in place. See FBI/IC3 OAuth consent-phishing guidance.

Selected AI-enabled threat reports. Hugging Face disclosed a July intrusion driven end to end by an autonomous agent system; OpenAI later attributed the incident to misaligned behavior by internal research models and reported other lower-severity third-party impacts. In September, Unit 42 reported a human-directed, agent-assisted intrusion that compressed more than 50 MITRE ATT&CK techniques into less than 10 hours. Google Threat Intelligence Group reported an agent-enabled credential-harvesting campaign completed in under six hours, attempts to trick AI coding assistants and security scanners during open-source supply-chain compromises, and theft or misuse of AI credentials and compute. Microsoft's October 2026 reporting describes AI use across multiple stages of cyber operations, while noting that most complex real-world intrusions still involve meaningful human direction. An October FBI advisory also highlights AI-generated impersonation in election-related fraud. These reports demonstrate feasibility and shorter response windows; they do not establish that most attacks are autonomous or primarily AI-generated.

Source coverage through October 18, 2026.

Pace capability with assurance. September calls by frontier-lab leaders to slow capability development are governance proposals, not standards or a blanket reason to stop lower-risk local pilots. For an adopting organization, the operational lesson is to pause expansion whenever isolation, monitoring, independent evaluation, incident response, or recovery cannot keep pace with the system's capability and reach.

SOURCE BASIS FBI/IC3 impersonation guidance | NIST SP 800-63B-4 | FBI/IC3 OAuth consent-phishing guidance | October 2026 reporting

10 AI-RELATED INCIDENT RESPONSE SEQUENCE

SECURITY + RISK / SECTION 10

Prepare before the alert

Use NIST SP 800-61 Rev. 3 to integrate AI incidents into established incident-response, privacy, legal, safety, vendor and communications processes. Preserve evidence and avoid untested containment that could worsen the event.

Specialist training and evaluation: Teams working with highly capable agents on long-horizon tasks should also review the joint NIST, NSA and CISA incident practices (October 2026, version 1.0). Their voluntary scope includes protected monitoring, investigative checkpoints and environment evidence, coordinated stopping, and verified containment before restart. Apply these with the training/evaluation owner; they do not replace existing incident authority or create a compliance certification. These specialist practices are not blanket requirements for classroom or small-office pilots.

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, endpoints, evaluation harnesses, providers or routers, tenants, privileges, compute, and time window.
4. Contain safely. Use pre-authorized, deterministic procedures to revoke exposed access, malicious app grants and tokens; terminate sessions, freeze affected pipelines, isolate components or accounts, and shift to a known fallback. These procedures execute predefined conditions and actions; an AI recommendation to select or change a consequential action requires authorized human review before execution. Novel or high-consequence choices remain with authorized incident leadership.
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.

Test the defensive fallback before an incident. Confirm whether an approved hosted provider can accept realistic exploit payloads, command traces, credentials, and other restricted evidence through a lawful emergency process. Record the escalation path and maintain an approved alternative that protects the evidence: another forensic tool, a vetted private capability where justified, or a manual workflow. Self-hosting is not a universal requirement.

Follow the organization’s incident escalation and notification procedures. For cyber-enabled crime, retain relevant records and report through FBI/IC3; use CISA’s reporting guidance as applicable. Reporting does not replace containment, recovery or required notifications.

SOURCE BASIS NIST SP 800-61 Rev. 3 | joint NIST, NSA and CISA incident practices | FBI/IC3 | CISA’s reporting guidance

11 30-DAY CYBERSECURITY AI PILOT

SECURITY + RISK / SECTION 11

A measured first deployment

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.

1. Week 1: Define. Name owners; threat-model; approve tool, data, permissions, network egress, reviewers, baseline, success measures, and stop conditions.
2. Week 2: Test. Train a small group; use representative sanitized cases; record prompts, sources, errors, revisions, false positives, and false negatives.
3. Week 3: Challenge. Test edge, adversarial, injection, denied egress, sandbox escape, boundary probing, burst activity, provider refusal, outage, data-leakage, privilege, synchronized containment, and recovery cases; verify logging and revocation.
4. 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; the system probes or crosses a trust boundary, reaches an unapproved network destination, or communicates through an unapproved external service; output causes or recommends unsafe action without the required gate; evidence integrity is threatened; logging, revocation, or containment fails; material drift is unexplained; or analysts cannot verify the result in time.

MOVE FROM PILOT TO APPROVED OPERATION

The security leader and accountable system or business owner should review the pilot record with the organization's designated privacy, IT, risk, and procurement authorities as applicable. Authorization is for a named workflow and configuration, not for AI-enabled defense generally. Keep the baseline comparison, labeled-case counts, false-positive and false-negative results, analyst time, adversarial and failure tests, residual risks, approved data and permissions, and the documented decision.

Before routine use, assign an operating owner, trained analyst reviewers, monitoring and log review, change control, tested revocation and fallback, and a review date. Keep consequential actions behind the approved human gate. Reassess after a material model, data, integration, privilege, threat, or performance change; pause under the stop conditions above. This playbook does not confer security, legal, regulatory, or procurement approval.

SOURCE BASIS NIST Cybersecurity Framework 2.0 | CISA JCDC AI Cybersecurity Collaboration Playbook | NIST AI Risk Management Framework | Hugging Face, Security incident disclosure — July 2026

12 CYBERSECURITY LEARNING AND IMPLEMENTATION

SECURITY + RISK / SECTION 12

Continue with MAIN

AI FOR CYBERSECURITY COURSE

Review current course availability and MGCCC registration steps in MAIN's course catalog.

Explore AI courses

For everyday verification, account protection and scam recovery, use MAIN's AI Cyber Safety guide.

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

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

13 FREQUENTLY ASKED QUESTIONS

SECURITY + RISK / SECTION 13

Cybersecurity AI questions

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. Model-selected consequential containment, account, network, endpoint or production changes require an authorized person to review the specific proposed action before execution. Separately approved deterministic response procedures may execute only their predefined steps; AI output cannot choose or widen those steps without the required human review. System or pilot approval does not waive the Statewide AI Framework's review triggers.

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.

WHAT IF A HOSTED AI SERVICE REFUSES INCIDENT EVIDENCE?

Test that failure mode before an incident. Maintain a pre-approved alternative that protects sensitive evidence, such as an authorized provider escalation path, a vetted private capability where justified, another approved forensic tool, or a manual workflow.

IS NIST IR 8596 A FINAL STANDARD?

No. NIST IR 8596 was released as a preliminary draft in December 2025. NIST identifies this publication as an Initial Preliminary Draft, not a final standard. Clearly label it as preliminary and use final NIST publications and applicable requirements as controlling references.

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

14 STANDARDS, DRAFTS, AND THREAT KNOWLEDGE

SECURITY + RISK / SECTION 14

Authoritative foundation

Federal publications provide the control foundation. Final, preliminary, voluntary and community resources serve different roles; this playbook does not establish compliance. Vendor reports below illustrate attributed incidents and proposals.

- CISA Cross-Sector Cybersecurity Performance Goals 2.0 (voluntary prioritized safeguards, aligned with NIST CSF 2.0)
- NIST SP 800-63B-4, Authentication and Authenticator Management (final; apply the requirements appropriate to the system)
- FBI/IC3, OAuth consent phishing (federal threat advisory; revoke malicious app access)
- FBI/IC3, Law enforcement and government impersonation (federal fraud advisory; independently verify requests)
- CISA, NSA, ASD's ACSC and international partners, Careful adoption of agentic AI services (joint security guidance, May 1, 2026); NSA coauthor announcement
- 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
- Mandiant, Blueprint for AI-Assisted Vulnerability Management
- Google Threat Intelligence Group and Mandiant, Mitigation Guidance for Supply Chain Compromise
- Microsoft Security, prompt injection and code-execution vulnerabilities in AI agent frameworks
- 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 Cross-Sector Cybersecurity Performance Goals 2.0 | NIST SP 800-63B-4, Authentication and Authenticator Management | FBI/IC3, OAuth consent phishing | FBI/IC3, Law enforcement and government impersonation

15 INCIDENT AND GOVERNANCE SOURCES

SECURITY + RISK / SECTION 15

Attributed incident reports

These sources inform the selected threat reports and operational additions above. Vendor incident reports and industry proposals are attributed evidence, not standards or proof of prevalence.

- Hugging Face, Security incident disclosure — July 2026 (vendor incident disclosure)
- OpenAI, The Hugging Face incident and other third-party impact from misaligned models (incident review and continuing updates)
- Unit 42, An AI-Assisted Cyber Attack: Inside a Unit 42 Investigation
- Google Threat Intelligence Group, From Prompting to Autonomy — The Evolution of Adversarial AI
- Microsoft, 2026 Digital Defense Report overview
- FBI/IC3, Scammers Exploit 2026 U.S. Midterm Election to Perpetrate Multiple Fraud Schemes
- Dario Amodei, We Must Pace the Frontier (industry governance proposal; not a standard)

SOURCE BASIS Hugging Face, Security incident disclosure — July 2026 | OpenAI, The Hugging Face incident and other third-party impact from misaligned models | Unit 42, An AI-Assisted Cyber Attack: Inside a Unit 42 Investigation | Google Threat Intelligence Group, From Prompting to Autonomy — The Evolution of Adversarial AI

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

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

Related MAIN resources

- 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.

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's voluntary Cross-Sector CPGs 2.0
cisa.gov

02 FBI/IC3 impersonation guidance
ic3.gov

03 NIST SP 800-63B-4
csrc.nist.gov

04 FBI/IC3 OAuth consent-phishing guidance
ic3.gov

05 October 2026 reporting
microsoft.com

06 October FBI advisory
ic3.gov

07 NIST SP 800-61 Rev. 3
csrc.nist.gov

08 joint NIST, NSA and CISA incident practices
nist.gov

09 FBI/IC3
ic3.gov

10 CISA's reporting guidance
cisa.gov

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

12 NSA coauthor announcement
nsa.gov

13 NIST Cybersecurity Framework 2.0
nist.gov

REFERENCES — CONTINUED

14 NIST AI Risk Management Framework
nist.gov

15 Generative AI Profile
nist.gov

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

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

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

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

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

21 CISA JCDC AI Cybersecurity Collaboration Playbook
cisa.gov

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

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

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

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

26 MITRE ATLAS
atlas.mitre.org

27 OWASP Artificial Intelligence Security Verification Standard 1.0
owasp.org

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

REFERENCES — CONTINUED

29 OWASP GenAI Agentic Security Initiative
genai.owasp.org

30 Hugging Face, Security incident disclosure — July 2026
huggingface.co

31 OpenAI, The Hugging Face incident and other third-party impact from misaligned models
openai.com

32 Unit 42, An AI-Assisted Cyber Attack: Inside a Unit 42 Investigation
unit42.paloaltonetworks.com

33 Google Threat Intelligence Group, From Prompting to Autonomy — The Evolution of Adversarial AI
cloud.google.com

34 Dario Amodei, We Must Pace the Frontier
darioamodei.com

35 Mississippi Statewide AI Framework, August 2026
mainms.org

36 NIST SP 800-218, SSDF 1.1
csrc.nist.gov

37 NIST IR 8596 preliminary Cyber AI Profile
www.nccoe.nist.gov

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

KEEP THE GUIDANCE CURRENT

Confirm applicability before use. Record the responsible local owner and what should trigger another review.

OWN

Name the person responsible for scope, access and decisions.

CHECK

Have qualified reviewers check evidence, risks and corrections.

STOP

Pause or reverse work that exceeds the approved boundary.