



AI security and data protection

Protecting data, identities, prompts, tools, and connected systems in AI-enabled work.

WORKS WITH THE ASSISTANTS TEAMS ALREADY USE



EDITION 14 / EMBER & ROSE

EXECUTIVE WORKING EDITION - 08 OCT 2026

14

AT A GLANCE

Protecting data, identities, prompts, tools, and connected systems in AI-enabled work.



AI changes the security picture in a specific way: systems now read untrusted text, hold broad access to company data and can take actions through connected tools. This playbook explains the main threats in plain language, including prompt injection, data leakage, excessive agent permissions and supplier risk. It then lays out controls for access, data classification, logging, red teaming and vendor assurance, with an implementation plan. It is general guidance and does not replace your security team or legal advice.

WHO IT IS FOR

CISOs, security and IT leaders, data protection officers, architects and business owners who approve or run AI tools that touch company data or systems.

TIME TO USE

75 minutes to read; two working sessions to apply to a live use case



What you will have at the end

- A working understanding of the threats that are new with AI and those that are old problems in new clothes
- A data classification rule that tells staff what may go into which AI tool
- An access and permission design for assistants and agents based on least privilege
- A red-team test plan and a vendor security question set
- A 90-day plan with measurable security outcomes



Contents

01	AI widens the attack surface in three ways	03
02	Where AI security effort belongs, and where it is wasted	04
03	The threats, in plain language	05
04	Who owns what in AI security	07
05	The control set: identity, data, tools, logging and humans	08
06	Red teaming and measuring security	10
07	Securing AI in 90 days	11
08	Preparing for the AI incident you will eventually have	13
09	Common failures and an illustrative scenario	14
10	A first quarter that produces evidence	15
11	Terms used in this publication	16



EXECUTIVE BRIEF

AI widens the attack surface in three ways

01



AI systems read untrusted input, concentrate access to data, and can act, and each of those needs its own control.

Traditional applications run fixed code on structured inputs. An AI assistant takes natural language from anyone, including attackers, and decides what to do with it. That single difference explains most new risk. A web page, an email, a PDF or a support ticket can contain text that the model treats as an instruction. The system cannot reliably tell the difference between your instruction and instructions hidden in the data it was asked to read.



Three shifts follow. First, input is now an attack vector in plain language. Second, assistants connected to document stores, mailboxes and databases concentrate access, so one compromised session can expose far more than one user would normally reach. Third, agents that call tools can change things: send messages, edit records, move money. A mistake or manipulation becomes an action, not only a bad answer.

- Most real incidents will be ordinary: over-shared documents, weak access control, staff pasting confidential data into consumer tools.
- AI makes ordinary weaknesses easier to find and exploit, because an assistant will helpfully surface anything it can reach.
- New attack types, such as prompt injection and data poisoning, need design-level answers, not just filters.
- Security and data protection are joint work: the same controls that limit breach also support lawful processing of personal data.



SCOPE NOTE

This playbook is operational guidance. Data protection, sector and contractual obligations vary by country and industry. Have your privacy and legal advisers confirm what applies to you, and check the current OWASP guidance for LLM applications and your supplier terms.




WORKING QUESTION

If an assistant in your company could read everything its most senior user can read, what would a single compromised session expose?


WHERE IT FITS

Where AI security effort belongs,
and where it is wasted

02

Match the depth of controls to the data an AI system touches and the actions it can take.

Not every AI use needs the same protection. Two questions set the level: what data can the system see, and what can it do? A summarizer that works on a single pasted public article needs little. An assistant with search over all internal documents needs strong access control. An agent that can send email or update the finance system needs approvals, scoped credentials and detailed logging.




USE CASE	DATA EXPOSURE	ACTION POWER	CONTROL EMPHASIS
Chat over public information	Low	None	Acceptable-use rules, basic logging
Drafting with internal documents	Medium	None	Approved tool, data classification, vendor terms
Enterprise search assistant over file shares	High	None	Respect existing permissions, clean up oversharing
Customer-facing support bot	Medium	Low	Input and output filtering, rate limits, escalation
Agent that edits records or sends messages	High	High	Least privilege, human approval, full audit trail

USE CASE	DATA EXPOSURE	ACTION POWER	CONTROL EMPHASIS
Coding assistant with repository access	High	Medium	Secret scanning, review of generated code, dependency checks



Where effort is wasted

- Trying to write the perfect prompt that forbids misuse. Instructions to a model are advice, not enforcement.
- Buying a single AI firewall product and assuming prompt injection is solved. No filter is complete.
- Blocking every AI site while leaving personal devices and mobile data open, which just displaces use.
- Treating security as a launch gate only. Models, plug-ins and data change after go-live.



WORKING QUESTION

Which of your AI use cases combines broad data access with the ability to take action, and does it have the strongest controls?

HOW IT WORKS

The threats, in plain language

03



Understanding five or six attack patterns is enough to design sensible defenses.

A large language model predicts text from a context window that holds the system instructions, the user request, retrieved documents and tool results. The model has no firm boundary between trusted and untrusted parts of that window. Attacks exploit this, and they cluster into a few types. The OWASP guidance for LLM applications gives a maintained list; the patterns below are the ones most business systems meet first.



THREAT	WHAT HAPPENS	BASIC DEFENSE
Direct prompt injection	A user crafts input to override rules or extract hidden instructions	Do not rely on prompts for security; enforce limits in code and permissions
Indirect prompt injection	Hidden instructions in a web page, email or file steer the assistant	Treat retrieved content as untrusted; limit tools available when reading it
Sensitive data disclosure	The assistant reveals data the user should not see	Enforce access control at retrieval time, per user
Excessive agency	An agent has more permissions or autonomy than its task needs	Least privilege, approval gates, narrow tools
Supply chain risk	Compromised models, plug-ins, libraries or data sources	Vendor assurance, pinned versions, allow lists
Data or index poisoning	Bad content planted in training data or knowledge base	Control who can write to sources; review changes

01 Trace the data path

Draw where a request goes: user, application, model provider, retrieval store, tools, logs. Mark every place data is stored or processed and who can reach it.

02 List trust boundaries

Mark which inputs are untrusted: web content, inbound email, uploaded files, customer messages.

03 Mark the actions

List each tool the assistant can call and whether its effect is read-only, reversible or irreversible.

04 Apply controls at the boundaries

Put enforcement in access control, tool design and approvals, because the model itself cannot be the guard.

KEY PRINCIPLE

Assume the model can be talked into anything. Design so that even a fully manipulated model cannot reach data or take actions beyond what the current user is entitled to do.

WORKING QUESTION

For your riskiest AI system, where is the strongest control enforced: in a prompt, or in code and permissions?



OPERATING MODEL

Who owns what in AI security

04



Security for AI is shared work across security, IT, data, privacy and the business, and it fails in the gaps between them.

Assign ownership by control, not by technology fashion. The security team sets standards and tests. IT runs identity, devices and network. The data team owns classification and source quality. Privacy owns lawful basis, notices and rights requests. The business owner of each AI system owns the decision to accept residual risk. If these roles are not written down, a risk like an overshared SharePoint library will sit with everybody and therefore nobody.

ROLE	OWNS	QUESTION THEY ANSWER
Security lead	Threat model, red team, monitoring and incident response	Could this be attacked, and would we see it?
Identity and IT	Single sign-on, role-based access control, device and network controls	Does the assistant act only as an authorized user?
Data owner	Classification, source quality, retention	Is this data allowed in this tool?
Privacy officer	Lawful basis, notices, rights handling, impact assessments	Are we handling personal data lawfully?
Procurement and legal	Vendor contracts and security terms	What do suppliers promise and what happens if they fail?
System owner	Acceptance of residual risk, user training	Is the remaining risk worth the value?



Data classification that staff can use

A usable rule fits on a card. Define three or four classes such as public, internal, confidential and restricted. Then state, for each class, which AI tools are allowed. For example: public content may go into any approved tool; internal content only into tools with enterprise terms and no training on your data; confidential content only into named tools with logging; restricted content such as health data, credentials or unannounced financials into none unless a specific review approves it. Adapt the classes to your own data and legal duties.

Ownership readiness

- Every AI system has a named system owner and a technical owner
- Security has a seat in the intake process for any tool with data access or actions
- A named person can disable any AI integration within one hour
- Privacy has recorded which AI uses process personal data
- Staff know where to report a suspected leak, and that reporting is safe

WORKING QUESTION

Can your staff say, in one sentence, what they may paste into which AI tool?

CONTROLS

The control set: identity, data, tools, logging and humans

05

A short list of well-applied controls beats a long list of partial ones.

Identity and access

- Use single sign-on and multi-factor authentication for every AI service, and remove personal accounts from work use.
- Make retrieval respect source permissions per user. If a person cannot open a file, the assistant must not summarize it for them.
- Give agents their own identities with narrow scopes, never a shared administrator credential.

- Review and expire tokens and API keys, and keep them out of prompts, code repositories and logs.

Data protection

- Minimize: send the model only the fields it needs. Mask or remove personal identifiers where the task allows.
- Set retention for prompts, outputs and logs, and confirm what the vendor stores and for how long.
- Check contract terms on training use, sub-processors, data location and breach notification.
- Run data protection impact assessments where personal data use is high risk, as your applicable law requires.

Tools, outputs and oversight

- Give each tool the smallest scope: read-only where possible, specific folders, specific fields.
- Require human approval for irreversible actions such as payments, deletions and external messages.
- Treat model output as untrusted before passing it to other systems; validate it like any user input to avoid injection into databases or web pages.
- Rate limit and monitor for unusual volume, which can signal extraction or abuse.

LOGGING

Log who asked, what was retrieved, which tools were called and what happened, with sensitive content handled carefully. Without logs you cannot investigate an incident. With raw logs of everything, you create a new store of sensitive data. Decide both deliberately.

WORKING QUESTION

If an agent sent a wrong email to a customer tonight, could you reconstruct exactly why from your logs tomorrow?

TESTING AND MEASUREMENT

Red teaming and measuring security

06

Test AI systems by attacking them, and measure the results over time against a baseline.

Security review on paper catches design flaws. Red teaming catches behavior. A red team, internal or external, attempts the attacks above against your actual system: injecting instructions through documents and emails, trying to extract other users' data, pushing the agent to call tools it should not, and probing for secrets in prompts. Run it before launch and after major changes, and keep the test cases as a regression set.

01 Build an attack library

Collect injection phrasing, jailbreak patterns, and data extraction attempts relevant to your use case, drawing on OWASP guidance and your own incident history.

02 Test with real permissions

Use test accounts at different permission levels to confirm that retrieval respects access boundaries.

03 Test the whole chain

Include the documents, connectors and downstream systems, not just the chat box.

04 Record and fix

Log each finding with severity, owner and a retest date. Add every confirmed attack to a permanent regression set.

METRIC	WHY IT MATTERS	HOW TO MEASURE
Red-team success rate	Shows how often attacks work	Successful attacks divided by attempts, per release
Overshared content found	Shows hidden exposure the assistant could surface	Files readable by far more people than needed, from a permissions scan
Time to disable	Shows readiness to contain	Minutes from decision to integration switched off, in drills

METRIC	WHY IT MATTERS	HOW TO MEASURE
Unapproved tool use	Shows shadow AI	Traffic and spend reviews versus approved list
Open findings past due	Shows whether fixes happen	Count of findings beyond their retest date
Incident detection time	Shows monitoring quality	Time from event to alert, in exercises



BASELINE FIRST

Before you add controls, run a permissions scan and an attack session on one system. The first results are your baseline. Illustrative example: if a scan finds many files readable company-wide that should be team-only, fixing that is worth more than any new filter.



WORKING QUESTION

Has anyone actually tried to break your most connected AI system, and where are the results?

IMPLEMENTATION GUIDE

Securing AI in 90 days

07



Start with visibility and permissions, then add testing and supplier assurance.



01 Inventory AI and data flows

List every AI service, integration and agent, with the data it reaches. Include features switched on inside software you already own.



02 Fix the foundation

Enforce single sign-on and multi-factor authentication, and block personal accounts for work data. Run a permissions scan on the repositories the assistants will read.

03 Publish the data rule

Issue the classification-to-tool rule on one page and train staff using real examples.

04 Threat model the top systems

For the two most connected systems, trace data, trust boundaries and actions, then agree the controls and approvals.

05 Red team and remediate

Test, fix, retest, and keep the attack library.



06 Assure suppliers

Send the security questions below to every AI vendor with access to confidential data, and record the answers.



Supplier assurance questions

- Is our data used to train or improve models, and how is that disabled and evidenced?
- What are the retention periods for prompts, outputs and logs, and can we set them?
- Which security audits or certifications are current, and what is their scope?
- Where is data processed, and which sub-processors have access?
- How are tenants isolated, and how is access by the vendor's staff controlled and logged?
- What is the breach notification commitment and the incident contact route?
- How are model and plug-in changes communicated before they take effect?
- Can we export and delete our data at exit, and how is deletion confirmed?



DO NOT SKIP

Retire consumer accounts used for work by offering an approved alternative first. Blocking without a replacement leaves people with the same need and fewer safe options.

WORKING QUESTION

Which single permission clean-up would most reduce what your assistants could expose?

INCIDENT RESPONSE

Preparing for the AI incident you will eventually have

08



An AI incident is handled like any security incident, with a few additions for model behavior, vendors and affected individuals.

Add AI scenarios to your existing incident plan rather than writing a separate one. Typical cases include an assistant exposing restricted content, staff pasting confidential data into an unapproved tool, an agent taking an unintended action, a vendor reporting a breach, and a manipulated output reaching a customer. For each, decide in advance who contains, who investigates, who informs affected people and regulators where the law requires it, and who speaks externally.

01 Contain

Disable the integration, revoke the tokens or restrict the connector. A kill switch that one named person can use within the hour is the most valuable control in this phase.

02 Preserve evidence

Keep logs, prompts, retrieved documents and tool calls before anything is cleaned up or rotated.

03 Assess impact

Establish what data was reached, whose it was, and whether personal data is involved. Involve privacy and legal early, since notification clocks may apply.

04 Fix and learn

Correct the permission, tool scope or design flaw, add the attack to the regression set, and share the lesson with other system owners.

VENDOR INCIDENTS

Your supplier may be the first to know. Confirm in the contract how fast they must tell you, what they will share, and how you can investigate. Rehearse one vendor-side scenario in a tabletop exercise.



WORKING QUESTION

Who in your organization can switch off an AI integration tonight, and have they ever actually done it as a drill?

PITFALLS

Common failures and an illustrative scenario

09

The usual breaches come from familiar mistakes made faster by a helpful assistant.

- Trusting the system prompt as a security boundary.
- Connecting an assistant to a broad index before cleaning up permissions.
- Giving an agent a powerful shared credential because it was quicker to set up.
- Skipping vendor terms review for features that are enabled by default.
- Logging full prompts that contain personal data with no retention limit.
- Treating a successful demo as proof of safety.

ILLUSTRATIVE SCENARIO (HYPOTHETICAL)

A company launches an internal search assistant over its shared drives. On day two an employee asks about reorganization plans and receives a summary of a draft file saved in a folder open to all staff years ago. The assistant did nothing wrong; the permissions were. The fix: scan and tighten permissions, enforce per-user retrieval, and add the incident to the test set.

The lesson is that AI exposes the state of your data hygiene. Search that was once impractical becomes instant, so old oversharing becomes a real incident. Treat AI rollout as a prompt to clean up access, retention and classification. That work protects you with or without AI, and it is usually the highest return security task on the list.

A second pattern is the agent that reads a customer email containing hidden instructions and then uses its tools in a way the customer was never entitled to trigger. The defense is not a smarter model but a smaller toolset, scoped permissions for the current customer and approval for consequential steps.

**WORKING QUESTION**

What old oversharing problem in your environment would an assistant turn into an incident on its first day?

30 / 60 / 90

A first quarter that produces evidence

10



Ninety days is enough to know your exposure, fix the worst permissions and test your most connected system.

First 30 days: ground it

- Inventory AI services, integrations and data flows
- Enforce single sign-on and multi-factor authentication for AI tools
- Publish the one-page data classification rule for AI
- Run a permissions scan on the first repositories to be connected

**Days 31 to 60: test it**

- Threat model the two most connected AI systems
- Fix oversharing and set per-user retrieval enforcement
- Send supplier security questions and record the answers
- Define who can disable an AI integration and test it



Days 61 to 90: decide

- Run the first red-team exercise and fix findings
- Establish the regression attack library
- Set monitoring and alerting for unusual AI activity
- Report baseline and trend metrics to the sponsor

GLOSSARY

Terms used in this publication

11

TERM	MEANING
Prompt injection	Text crafted to make a model follow attacker instructions instead of the intended ones.
Indirect prompt injection	Malicious instructions hidden in content the assistant reads, such as a web page or email.
Least privilege	Giving a user or agent only the access needed for the task and no more.
Excessive agency	An AI agent holding more permissions or autonomy than its task requires.
Red teaming	Structured attempts to attack a system to find weaknesses before real attackers do.
Data classification	Labelling data by sensitivity so that handling rules can be applied.
Data minimization	Using only the personal or sensitive data that a task actually needs.
Retrieval-time access control	Checking the user's permissions on each document when the assistant retrieves it.
Data poisoning	Planting bad content in training data or a knowledge base to alter outputs.
Sub-processor	A third party that processes data on behalf of your vendor.
Data protection impact assessment	A structured review of risks to individuals from a high-risk use of personal data.

TERM	MEANING
OWASP LLM guidance	Community-maintained list of top security risks for applications built on large language models.



A note on use

This playbook is general guidance and not legal advice. Confirm current legal, contractual and regulatory requirements with qualified advisers.

TAKE THE NEXT STEP

Take the free Halden assessment at halden.company/assessments, or talk to us at halden.company/contact about turning this into a plan for your organization.

