AI Security Playbook

Phil Parker, Chris Rutter, Ben Wilkes

EQUAL EXPERTS

CONTENTS

INTRODUCTION

Generative AI is already embedded in how organisations work - whether through products, delivery tooling, or everyday business tools. This playbook does not assume a future of AI; it starts from the reality that AI is here, widely used, and already shaping decisions, data flows, and outcomes across your organisation.

Nothing in this document is radically new. Most of the recommendations have always been good security practice. What has changed is the context: AI introduces new vectors, makes some risks easier to trigger, and exposes existing weaknesses more clearly. In the rush to adopt AI, many organisations are rediscovering problems they already knew how to solve, and in other areas realising that they were already exposed before AI came along.

This playbook deliberately focuses on the most common, high-impact uses of Generative AI today. We do not cover advanced cases such as full model training or extensive fine-tuning. Instead, we concentrate on what most organisations need to get right now: practical, proportionate controls that make AI use visible, governable, and resilient.

What you need to do right now (tl;dr)

Generative AI is already in your organisation. The priority is not perfect foresight, but establishing a small set of foundations that make your use of **AI visible, governable, and resilient.**

Three immediate actions cut across all teams and use cases.

1. **Map your AI risk footprint**
2. **Train your people about AI**
3. **Instigate left-shifted risk assessment processes for AI**

1. Map your AI risk footprint

Start with visibility. You should have a shared view of where AI is already touching your organisation, your data, and your decisions.

In practice, this means understanding at a high level:

- Where AI is being used (products, delivery tooling, business tools, automation).
- What data is flowing into and out of AI, especially anything sensitive or regulated.
- What you depend on (models, tools, plugins, agents, retrieval systems, third parties).
- Where AI can influence real outcomes (access, actions, decisions, or customer interactions).

The goal here is not bureaucracy, but a lightweight, living picture of your AI footprint so you can govern it.

2. Train your people about AI

AI is used by many different cohorts, in very different ways. Treating everyone the same either over-constrains some groups or leaves others under-protected.

Crucially: **Every AI user is a risk owner - align expectations, set limits, and make accountability clear.**

Right now you should:

- Recognise the main ways AI is used across your organisation (product teams, delivery teams, citizen developers, business users, and customers).
- Provide guidance that reflects how risk actually shows up for each group (e.g. data exposure for business users; decision influence for AI-infused products; scope creep for citizen developers).
- Make it normal to talk about AI risks, not just AI benefits.

The immediate objective is shared understanding, not perfect compliance.

3. Instigate left-shifted risk assessment processes for AI

You need to shift left the handling and mitigation of AI risk - building it into how work happens, rather than relying on late-stage gates.

In practical terms:

- Bring AI explicitly into existing practices such as threat modelling, design review, and code review.
- Put simple, repeatable guardrails in place (data controls, least privilege for agents, and clear human oversight for high-risk actions).
- Ensure you can see when things go wrong through sensible logging, monitoring, and periodic testing (including AI-specific scenarios like prompt injection).

The aim is to make security thinking about AI part of everyday delivery and usage, not an afterthought.

If you'd like help performing an assessment, Equal Experts provides a Security Healthcheck service. Get in touch at: **https://www.equalexperts.com/contact-us/** for more information.

Want to apply this to your organisation? **Download a copy of our AI Security Foundations and Benchmark**, which translates the principles into a practical, OWASP-aligned set of controls you can use to assess your current approach and identify meaningful gap.

AI SECURITY PLAYBOOK

RISK FIRST PRINCIPLES

GenAI introduces a number of fundamental characteristics that meaningfully change how risk shows up in practice. These are not abstract concerns or emerging edge cases; they are properties of how large language models (LLMs) work, and they apply regardless of use case, team, or maturity. The following principles describe those characteristics and explain why familiar security threats behave differently when GenAI is involved. They provide a shared mental model for understanding why existing controls may need to be adapted, strengthened, or applied in new places across all cohorts in this document.

Intent is interpreted, not enforced

Large language models don't follow instructions in the way traditional software does. They don't, by default, "check" rules or enforce boundaries; instead, they try to work out what you mean from the language they're given. That means the model is always interpreting intent rather than executing clear, fixed instructions. From a security perspective, this matters because untrusted input can influence behaviour in subtle ways, without ever looking like an error or breach.

You see this whenever user input, documents, or context are mixed into prompts. A customer question, a support transcript, or a retrieved document can shift how the model behaves, even if no one intended it to. This is why prompt injection is easier than many teams expect, and why guardrails can't rely on "good prompts" alone. Across all cohorts, the lesson is the same: language is a control surface. Treat prompts, context, and retrieved data as untrusted input, and put controls around how they are combined and used.

Behaviour is non-deterministic

LLMs are probabilistic systems. Given the same input, they won't always produce the same output. Most of the time the differences are small, but sometimes they matter. This clashes with how we usually think about software delivery, where we expect repeatable behaviour, predictable changes, and confidence that what passed yesterday's test will behave the same way tomorrow.

In day-to-day use, this shows up in familiar ways. A summary is accurate one day and misleading the next. Generated code looks fine but introduces a subtle vulnerability. An automated filter works until a slightly different phrasing allows it to be bypassed. None of this means GenAI can't be used safely, but it does mean teams need to adapt. Instead of assuming we can test every outcome upfront, we need stronger guardrails, human review at the right points, and good monitoring so we can spot and correct issues quickly when they appear.

The technology landscape is rapidly (and often opaquely) evolving

When teams use GenAI, they're rarely just depending on a single model. They're also relying on tools, plugins, prompt frameworks, retrieval systems, and third-party services - many of which change frequently and not always transparently. Models get updated, behaviour shifts, and capabilities expand, often without clear versioning or advance notice.

This affects everyone, not just product teams. A business user may notice a tool behaving differently. A delivery team may inherit new risks through an updated IDE assistant. A production system may change its responses without any code being deployed. The key point is that change can happen outside your normal delivery controls. Secure use of GenAI therefore requires ongoing attention: understanding what you depend on, watching for unexpected changes, and designing systems so that when behaviour shifts, the impact is limited and visible rather than surprising and harmful.

Existing security best-Practices still apply

Generative AI does not replace established security practices - it amplifies the consequences of getting them wrong.

When organisations start assessing GenAI risk, they often discover that the most significant issues predate AI adoption. Weak data classification, unclear ownership, inconsistent access control, and informal change management become more visible once AI is introduced, because models operate across boundaries that were previously implicit or loosely governed.

This can create the impression that AI has introduced entirely new risks, when in reality it has exposed existing gaps. Prompts function as untrusted input. AI-generated artefacts resemble externally sourced dependencies. Model access introduces new interfaces that require the same discipline as any other integration.

The key lesson is not that pre-AI security practices are obsolete, but that partially applied or inconsistently enforced controls are no longer sufficient. Secure use of GenAI depends on reinforcing fundamentals - clear accountability, deliberate boundaries, and the ability to detect and correct issues quickly - even as the technology changes where those controls must be applied.

AI SECURITY PLAYBOOK

RISK COHORTS

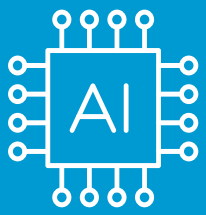
Generative AI is a widely applicable technology: it deals with language at both its inputs and outputs, and its capabilities are general rather than tied to one narrow task. That means it doesn't belong to any single department, problem type, or domain - people use it for writing, coding, research, summarisation, conversation, automation, ideas and more. It also didn't take years to become familiar; when ChatGPT first launched in late 2022, it drew roughly 1 million users in just five days and then reached 100 million monthly active users within two months, outpacing the early adoption curves of major social platforms such as Instagram and TikTok.

Because of this general-purpose nature and rapid uptake, individuals and teams across organisations are engaging with GenAI in very different ways. We've characterised these into the following 5 categories.

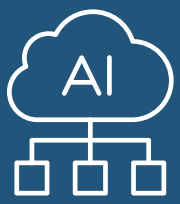
1. **AI product builders**
2. **AI delivery teams**
3. **Citizen developers**
4. **Business users of AI**
5. **Your customers using AI**

Below we describe these groups, how they typically interact with GenAI capabilities, and where risks tend to show up for each.

AI RISK COHORTS



AI PRODUCT BUILDERS

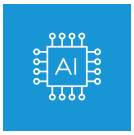


AI DELIVERY TEAMS

CITIZEN DEVELOPERS

BUSINESS USERS OF AI

YOUR CUSTOMERS USING AI



1. AI product builders

Technology **teams building products, services, and data platforms that embed Generative AI capabilities**, where AI behaviour is exposed externally and can influence decisions, actions, or access to data - whether or not those teams use GenAI in their own delivery processes.

This includes systems using AI for:

- conversational or natural-language interfaces
- content generation or transformation
- augmented or automated decision support
- agentic patterns that orchestrate tools or services based on interpreted user intent, including dynamic routing and service discovery
- data products that embed generative AI for analysis, classification, enrichment, synthesis, or insight generation.

In all of these cases, AI output is consumed by users or downstream systems in ways that can materially affect outcomes, making AI behaviour part of the organisation's effective security boundary.

How security risk shows up in this cohort

The primary information-security risk in this cohort is not whether AI output is "correct", but how AI behaviour can be influenced and then trusted.

AI-infused products frequently combine untrusted inputs - such as user prompts, uploaded documents, or retrieved content - with system instructions and internal context. As described earlier in the GenAI Risk First Principles, intent is interpreted rather than enforced. In practice, this means that behaviour can be shaped through language and context without breaching traditional technical controls.

This introduces integrity risks that are difficult to detect using conventional security signals. Manipulated or misleading inputs may influence decisions, routing, or downstream actions while appearing legitimate.

Where products incorporate orchestration or agentic behaviour, these risks can compound. A single manipulated input may influence multiple tools or services, increasing blast radius and reducing the clarity of cause and effect.

Teams with prior experience operating non-generative AI systems - such as classifiers, scoring models, or recommendation engines - are often better prepared for these dynamics, having already encountered probabilistic behaviour, performance variance, and the need for ongoing monitoring. Where that experience is absent, there is a greater risk that systems are designed around deterministic assumptions that no longer hold.

Leadership considerations

For technology leaders, this cohort represents a concentrated area of information-security exposure because AI behaviour is externally visible and may directly influence user or system outcomes.

Key considerations include:

- **Influence over security-relevant decisions**
  Even when positioned as assistive, AI outputs may shape access to data, choice of actions, or escalation paths. Leaders must be explicit about where AI influence is acceptable and where enforcement must remain deterministic.

- **Behavioural change outside release cycles**
  Model, retrieval, or tooling updates can alter how untrusted input is interpreted without any corresponding deployment, weakening traditional assumptions about change control and assurance.

- **Auditability and accountability**
  When AI contributes to outcomes, organisations must still be able to reconstruct what occurred, which inputs influenced behaviour, and who remains accountable.

- **Controls must sit around the model**

Prompts and configuration alone do not provide enforceable security guarantees. Effective risk management depends on constraining inputs, bounding permitted actions, and maintaining visibility when behaviour deviates from expected patterns.

AI-infused product teams sit at the point where generative AI most directly intersects with organisational security boundaries. The defining challenge for this cohort is not whether AI can be made safe in isolation, but how its influence on security-relevant decisions and actions is constrained, observed, and governed when exposed through live products, services, and data outputs.

2. AI delivery teams

Teams in this cohort **use Generative AI to improve the software and service delivery lifecycle**, *even where the resulting products and services may not themselves contain AI functionality.*

This includes AI use across the lifecycle, such as:

- early exploration and problem framing, including synthesis of user research and requirements
- design ideation and option evaluation, including threat modelling support
- generation of architectural sketches and trade-off analysis
- prototype artefact generation and simulated test data creation
- test design, scenario writing, and test coverage suggestions
- implementation support, assisted coding, and refactoring
- peer review, security review assistance, and debugging help
- automated documentation, summarisation, and delivery-related operational guidance
- incident response, observability and service operation.

In these scenarios, AI may not form part of the delivered system, but it **directly influences the artefacts, decisions, risk assessments, and changes that reach production.**

How security risk shows up in this cohort

The core information-security risk for **AI delivery teams comes from how AI reshapes delivery decisions, artefacts, and verification.**

AI tools compress multiple stages of reasoning - understanding, design, trade-off analysis, and specification - into singular outputs. From a security standpoint, this increases the likelihood that security-relevant design choices or assumptions are introduced without explicit recognition or challenge.

This changes the nature of review: teams may be reviewing larger artefacts, aggregated outputs, or synthetically generated content where subtle security flaws are harder to pinpoint. Generated outputs often appear coherent and plausible, which can reduce the likelihood of deeper inspection of security boundaries.

AI-assisted delivery also expands the effective supply chain for security risk. Synthesised design options, test scenarios, simulated data sets, configuration fragments, scripts, and dependencies may all influence production systems without clear provenance or ownership.

Where delivery teams adopt local or semi-autonomous AI tooling or developer agents, additional exposure emerges. Tools operating with broad access - repositories, configuration, or network - can leak sensitive data or introduce unsafe changes without requiring direct access to production systems.

Leadership considerations

For technology leaders, the security challenge in this cohort is preserving **deliberate control within the delivery lifecycle** as AI accelerates it.

Key considerations include:

- **Integrity of delivery decisions**
  AI-generated outputs may influence security trade-offs, design decisions, or implementation choices. Leaders must ensure these decisions remain subject to adequate security scrutiny.

- **Effectiveness of review and assurance**
  Traditional review practices can be weakened if not adapted to account for synthetically generated artefacts and reduced visibility into how decisions were formed.

- **Delivery lifecycle attack surface**
  Prompts, tooling plugins, agents, and generated artefacts introduce new pathways for data exposure and manipulation within the SDLC rather than in production.

- **Privilege and autonomy in tooling**
  AI assistants and agents should operate with deliberately constrained access. Excessive privilege in delivery environments can create impactful failure modes even when production security is strong.

- **Clear accountability for AI-assisted change**
  Use of AI does not diminish responsibility. Organisations must retain clarity over who approved changes, what was reviewed, and how security assurances were achieved.

AI delivery teams sit at a critical junction: although AI is not embedded in the shipped product, it increasingly shapes how systems are explored, specified, tested, and modified. The defining security challenge for this cohort is ensuring that acceleration does not come at the expense of visibility, control, or accountability across the lifecycle.

3. Citizen developers

People **outside formal technology departments building systems and tooling using Generative AI** to solve local problems and improve productivity.

The defining characteristic of this cohort is not job role or technical skill, but scope expansion. Individuals begin with limited, local use, but over time may take on work that increasingly resembles technology delivery - without operating inside the controls, governance, and assurance applied to Cohort 2.

Typical activities include:

- analysing operational problems and defining local requirements
- designing workflows or decision-support tools
- building automations that connect systems or datasets
- generating structured outputs used by others
- iterating quickly to meet immediate business needs.

These activities often start small and well-intentioned. Risk emerges as their scope expands beyond personal productivity into shared, persistent, or decision-influencing systems.

How security risk shows up in this cohort

The primary information-security risk in this cohort arises when **work undertaken outside delivery governance begins to carry delivery-grade risk.**

Citizen-developed tools frequently handle real organisational data and may evolve to:

- access sensitive or regulated information
- propagate data between systems
- influence decisions or trigger actions
- become embedded in day-to-day operations.

From a security perspective, these systems may introduce confidentiality and integrity risks comparable to those seen in Cohort 2 - but without equivalent controls such as design review, access governance, testing, monitoring, or formal ownership.

Generative AI significantly accelerates this progression. It lowers the barrier to creating logic, integrations, and decision flows that appear robust, encouraging further expansion of scope. Because behaviour is often driven by interpreted prompts or generated logic, trust boundaries can become unclear or implicit.

This creates a risk gap: **delivery-like systems emerging outside delivery controls.**

Shadow IT typically appears at this point - not because individuals are acting irresponsibly, but because capability grows faster than governance adapts. Restrictive responses often reduce visibility rather than risk, pushing activity further from oversight.

Leadership considerations

For technology leaders, the central challenge in this cohort is managing **scope creep across organisational boundaries.**

Key considerations include:

- **Clear thresholds for when work becomes "delivery"**
  Organisations must define when a local tool or automation has grown to the point where Cohort 2–level controls should apply.

- **Visibility of expanding scope**
  Leaders need mechanisms to detect when citizen-built solutions begin to handle sensitive data, affect decisions, or serve multiple users.

- **Proportionate guardrails**
  Platform-level constraints should limit data access, integration capability, and automation impact before risk escalates.

- **Ownership and accountability**
  As scope expands, clear responsibility must be established - otherwise security risk persists even when the original creator moves on.

- **Safe escalation paths**
  There must be straightforward routes for citizen-built solutions to be reviewed, supported, or formally adopted without discouraging transparency.

Citizen developers occupy the space where productivity gains can quietly turn into unmanaged systems. The defining security challenge for this cohort is not the initial use of AI, but the **uncontrolled expansion of scope**, where work performed outside delivery governance begins to carry the same risk profile as formally engineered systems - without the same safeguards.

4. Business users of AI

This cohort includes **people across the organisation using off-the-shelf AI tools as part of their day-to-day work**, without any intention to design or deliver new products, services, or bespoke systems.

This includes use of AI tools that provide:

- conversational and research capabilities
- drafting, editing, and summarisation features
- analysis and sense-making support
- built-in automations such as scheduling, reminders, or simple agentic workflows
- integrations offered directly within the tool itself.

The defining characteristic of this cohort is that individuals are **using capabilities as provided by the tool**, rather than composing new workflows, integrations, or logic across systems.

While the boundary with citizen development is soft, business users are not seeking to create persistent organisational systems - they are applying AI to support individual or team productivity within pre-defined tooling.

How security risk shows up in this cohort

The primary information-security risk in this cohort arises from **data exposure at scale**, rather than system construction.

Business users interact with AI tools through interfaces that encourage free-form input and context sharing. This makes it easy - intentionally or unintentionally - to include information that may be sensitive, regulated, or confidential.

From an infosec perspective, this creates confidentiality risk, particularly where:

- users are unclear what data is appropriate to share
- retention and training behaviour of tools is poorly understood
- outputs are reused or shared beyond their original context.

A significant driver of risk in this cohort is use of unapproved tooling. Where safe, capable, enterprise-approved AI tools are not available, users often adopt consumer

services that lack adequate controls around auditability, data retention, and contractual protection.

This leads to shadow AI emerging through normal work, not deliberate policy violation.

Even where tools are approved, built-in automation and agentic features can amplify impact. A single prompt or configuration may affect repeated actions, scheduled behaviour, or multiple outputs, increasing the consequences of mistakes or misuse.

Leadership considerations

For technology leaders, the challenge in this cohort is managing widespread, low-friction exposure.

Key considerations include:

- **Provision of capable approved tools**
  If approved tools are materially less useful than consumer alternatives, risk will be displaced rather than reduced.

- **Practical guidance on data use**
  Users need clear, situational guidance on what types of data can be shared with AI tools, aligned to real work patterns.

- **Visibility of usage patterns**
  Individual uses may appear low-risk, but cumulative behaviour can create meaningful exposure without obvious incidents.

- **Understanding automation impact**
  Built-in scheduling or agentic features can turn one-off actions into persistent behaviour that requires oversight.

- **Clear organisational accountability**
  Even when AI use feels personal, responsibility for data handling and outcomes remains organisational.

Business users form the largest AI cohort in most organisations. The defining security challenge is not intent or sophistication, **but the cumulative effect of everyday AI use at scale**, particularly where tooling choice, data handling, and visibility are not well aligned.

5. Your customers using AI

This cohort covers your customers, citizens, or service users who are themselves using AI tools, independently of your organisation.

Customers may use AI to generate communications, complete forms, prepare submissions, or interact with your services in novel ways. In some cases, third-party AI tools may act “on their behalf”, shaping how requests, data, or questions are presented to your organisation.

We aren’t expanding this section in this playbook as most of the threats that arise from this cohort are actually mitigated or controlled by the other cohorts listed. However, when assessing the overall risk to an organisation, it is still important that horizon scanning includes changes in behaviour of other parties - including customers themselves, as well as third parties, vendors, and suppliers.

From an information-security perspective, the key consideration for this cohort is **perceived organisational responsibility**. Customers may unknowingly share their own sensitive data with external AI tools that lack appropriate protections. When issues arise - such as data leakage, misuse, or unexpected behaviour - those problems may reasonably be attributed to your organisation, even where no internal system has been compromised.

As a result, limited customer education may still be necessary. Uninformed users can create information-security issues that they believe are connected to your organisation, particularly where AI tools blur the boundary between customer behaviour and organisational systems.

AI SECURITY PLAYBOOK
THREAT CONTROLS

This section provides an overview of the threat controls that make up our **AI Security Foundations Benchmark** - a practical, OWASP-aligned framework for assessing how securely generative AI is used across your organisation.

Why AI threat controls?

Generative AI changes how risk shows up in practice. It expands the organisation's effective attack surface, blurs traditional trust boundaries, and makes system behaviour less deterministic. As a result, many familiar security controls need to be applied in new places, strengthened, or operationalised differently.

The purpose of the AI threat controls is not to catalogue every possible AI risk. Risks evolve too quickly for that to be useful, and long risk lists age poorly. Instead, this section defines a small number of enduring control sets - processes, capabilities, and guardrails - that remain relevant as models, tools, and threats change.

These controls span policy, technology, people, and operations. They are designed to be read as a coherent whole: governance sets the rules, data protection limits exposure, supply chain provides visibility and trust, access control bounds what AI can do, human oversight preserves accountability, input/output controls harden interfaces, and monitoring ensures you can detect and respond when things go wrong.

Governance and legal - setting clear, enforceable boundaries

AI adoption often moves faster than policy. Without explicit boundaries, teams will use AI in ways that create legal, regulatory, privacy, or intellectual property exposure - usually unintentionally.

Effective governance combines three elements: Clear rules of use: Explicit policies that align acceptable AI use with existing data classification, risk appetite, and regulatory obligations, and are actively socialised rather than merely published.

Privacy by design: Formal consideration of privacy impacts (e.g., DPIAs) before deployment, with documented data flows, retention, and lawful bases that are revisited as capabilities change.

Strong contracts with providers: Explicit terms on data use, training opt-outs, residency, audit rights, and breach notification, because contracts are often your most powerful control over third-party behaviour.

Together, these create legitimacy for enforcement, reduce ambiguity for users, and provide a defensible position if something goes wrong.

Key question: Do we have enforceable policies and contractual protections covering AI use?

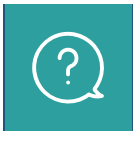
Data Protection - preventing uncontrolled data exposure

Every prompt is a potential data flow. Once information leaves your environment, you should assume it could be stored, logged, or resurfaced in unexpected ways.

Good practice therefore emphasises **technical prevention, not just guidance:**

- **Automated safeguards at the point of entry:** Sensitive data detection, redaction, and tokenisation before data reaches models, including prompts, documents, shared “skills,” and configuration files.
- **Consistent encryption:** AI-related data flows-prompts, responses, embeddings, logs, and training data-should meet the same encryption standards as other sensitive systems, in transit and at rest.
- **Verified provider practices:** Where third parties are involved, their encryption and retention practices should be validated rather than assumed.

The goal is to make accidental leakage hard by design, not dependent on perfect user behaviour.



Key question: Are we technically preventing sensitive data from reaching AI systems, or relying on guidance alone?

Supply chain and inventory - knowing what you depend on

AI systems are assembled from many components: base models, fine-tunes, libraries, tools, plugins, and orchestration layers. Without visibility, you cannot manage vulnerabilities or respond quickly to incidents.

Core practices include:

- **An AI bill of materials:** Continuous documentation of models, tools, libraries, and services in use, linked to the systems that depend on them, and integrated with existing vulnerability management processes.
- **Structured vendor assurance:** Bringing AI providers into your third-party risk management programme, reviewing certifications, audit reports, and testing results, and reassessing when material changes occur.
- **An approved registry of tools and models:** Clear criteria for approval (security, privacy, provenance, contracts), with proportionate technical enforcement to limit the spread of unsafe or malicious tooling.

This turns “shadow AI” into something you can actually govern.

Key question: Do we have visibility of the AI tools, models, and components in use across the organisation?

Access control - bounding what AI can access and do

AI agents follow instructions literally and inherit whatever privileges they are given. Excessive access can turn a small mistake - or a successful prompt injection - into a major incident.

Effective controls focus on three areas:

- **Least privilege by design:** Restrict agent and tool functionality to only what is required; apply filesystem, network, and tool-level allowlists wherever possible.
- **Safer credentials:** Avoid long-lived, shared secrets. Prefer per-agent identities with short-lived, task-scoped credentials that are logged and rotated automatically.

- **Rate and cost guardrails:** Limits, quotas, and circuit breakers to prevent runaway usage (compromised keys, or denial of resource attacks) causing operational or financial harm.

The aim is to minimise blast radius if something goes wrong.

Key question: If an AI agent were compromised, what's the potential impact?

Human oversight - preserving accountability

AI produces confident, fluent outputs - including confident, fluent errors. Security and accountability cannot be delegated to a model.

Practical guardrails include:

- **Clear review requirements:** Define which AI-generated artefacts require human review before use, publication, or execution, with a named owner for each output.
- **Approval gates for high-risk actions:** For sensitive or destructive agent behaviours (e.g. deletions, external communications, financial actions, privilege changes), pause execution pending explicit human approval with clear context.
- **Code assurance:** Mandatory human review of AI-generated code, supported by conventional security tooling (SAST, secret scanning, dependency checks) and reviewer training on common AI-related weaknesses.

Oversight is not about slowing teams down - it is about ensuring accountability where consequences are highest.

Key question: Where are we relying on AI outputs without meaningful human verification?

Input and output - hardening the AI interface

Generative AI introduces new attack surfaces, particularly around prompt injection and unsafe outputs.

A layered defence approach includes:

- **Input controls:** Validation and semantic filtering to detect attempts to override system instructions or manipulate behaviour, including indirect injection via external documents or retrieved content.
- **Safe output handling:** Encoding outputs before passing them to downstream systems to prevent secondary injection or command execution.
- **Content filtering:** Last-line-of-defence moderation for harmful, non-compliant, or inappropriate outputs, with clear processes for human review of edge cases.
- **Code security gates:** Automated SAST, secret scanning, and dependency checks for any AI-generated code before it reaches production.

These controls recognise that language itself is a security boundary.

Key question: Are our AI systems defended against AI-specific attacks, not just traditional threats?

Testing, monitoring and response - knowing when things go wrong

AI systems can be misused, compromised, or fail in ways that look superficially "normal." Visibility and preparedness are therefore critical.

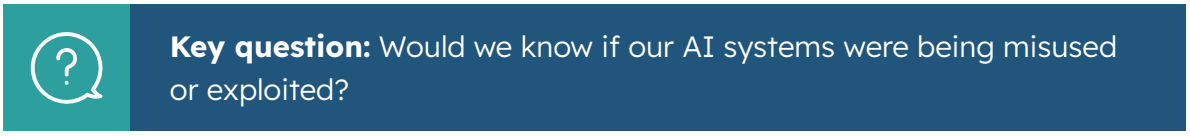
Key elements are:

- **Comprehensive audit trails:** Tamper-evident logging of prompts, responses, timestamps, identities, and agent/tool actions, retained in centralised logging infrastructure.
- **Proactive detection:** Alerting on anomalous patterns (usage spikes, unusual queries, off-hours activity, sensitive data access) integrated with existing security monitoring.

- **Regular testing:** Ongoing threat modelling and red-teaming that includes AI-specific scenarios such as prompt injection, jailbreaking, and data extraction.
- **AI-specific incident response:** Playbooks for scenarios like prompt injection, data leakage via AI, model compromise, or agent misbehaviour, with clear roles and escalation paths.

These controls recognise that language itself is a security boundary.

The objective is to detect issues early and respond with confidence.



A note on assessment

If you want to baseline your current state against these controls in a more structured way, we also provide a separate **AI Security Foundations Benchmark** that can be used for that purpose. **Download your copy here.**

AI SECURITY PLAYBOOK

HOW AI CAN IMPROVE GENERAL SECURITY

Strengthening existing security practices

AI is most effective when applied to practices organisations already rely on, such as secure design, vulnerability management, code review, monitoring, and governance.

In many cases, generative AI does not change what “good security” looks like. Instead, it improves how often good practices are applied, how early issues are identified, and how accessible security expertise becomes across delivery teams.

This creates an opportunity to raise the baseline of security across the organisation, while allowing specialists to focus their attention where it adds the most value.

Faster feedback and clearer signals

Security feedback is often slow, fragmented, or difficult to interpret. This can make it harder for teams to respond effectively, even when good tools are in place.

AI can help by:

- summarising large volumes of security findings into clearer, prioritised insights
- highlighting likely root causes rather than long lists of symptoms
- identifying recurring patterns across services or repositories
- explaining issues in language that supports constructive delivery conversations.

Faster, clearer feedback helps security become part of everyday engineering work, rather than something that only appears late in the delivery process

AI-assisted threat modelling

Threat modelling is a well-established security practice, but one that can be difficult to scale consistently. It often relies on specialist knowledge and facilitated workshops, which limits how frequently it can be applied.

Generative AI can help teams perform lightweight, earlier threat modelling by acting as a structured guide or "coach".

For example, AI can help teams:

- **identify key assets and trust boundaries**
- **explore common threat categories**
- **surface assumptions and areas of uncertainty**
- **produce an initial threat model suitable for review.**

This lowers the barrier to entry and supports earlier security thinking, while still allowing security specialists to review, challenge, and refine the outcomes.

Used in this way, AI helps scale the *practice* of threat modelling, without removing human judgement or accountability.

Vulnerability remediation at scale

Many organisations face large backlogs of known vulnerabilities, particularly within dependencies and container images. While scanners provide visibility, turning findings into safe, timely fixes can be slow and resource-intensive.

AI can assist by:

- **explaining vulnerability findings and their practical impact**
- **proposing dependency upgrades or code changes**
- **generating candidate pull requests**
- **supporting bulk remediation activities.**

To get the most out of this approach, organisations will also want strong automated security regression test coverage. AI can help here too - assisting teams in expanding tests and improving confidence in robustness of change, allowing fixes to be delivered quickly, whilst still maintaining safety.

When combined with automation and review, this approach can significantly reduce time-to-fix while maintaining confidence in releases.

Secure code review and design support

AI can provide an additional layer of support during code and design reviews by highlighting common security concerns, such as:

AI can assist by:

- **unsafe data handling**
- **authentication and authorisation risks**
- **insecure defaults or configurations**
- **use of risky dependencies or patterns.**

This can help improve consistency across reviews and reduce reliance on individual expertise.

When paired with human review, AI-assisted analysis can raise the overall quality of security feedback, while allowing experienced reviewers to focus on higher-risk or more complex changes.

Monitoring, incident response, and learning

During security incidents, responders often need to rapidly interpret large volumes of logs, alerts, and system signals.

AI can assist by:

- **summarising events and timelines**
- **correlating signals across systems**
- **suggesting hypotheses or next areas to investigate**
- **supporting the creation of incident reports and follow-up actions.**

By reducing cognitive load during high-pressure situations, AI can help teams focus more effectively on faster containment, and better post-incident hardening.

Continuous assurance and adaptive governance

AI also provides opportunities to make governance more continuous and less burdensome.

Examples include:

- identifying drift from agreed security standards
- highlighting emerging patterns of risk across teams
- supporting policy-as-code and automated assurance
- enabling earlier conversations rather than late-stage escalation.

This supports a more adaptive model of governance - one focused on ongoing visibility and learning, rather than periodic inspection.

Generative AI creates an opportunity to scale security practices that have traditionally struggled to scale, such as threat modelling, remediation, and consistent review.

The organisations that benefit most will be those that combine AI-assisted workflows with strong secure engineering fundamentals - using automation, testing, and observability (particularly monitoring, auditability, and access control) to turn increased speed into increased confidence.

Used in this way, AI becomes not just a new risk to manage, but a powerful tool for improving overall security maturity.

Summary

Generative AI does not require a brand-new security paradigm - it requires better, more consistent execution of the fundamentals. The real challenge is not predicting every new AI risk, but making your organisation visible, disciplined, and resilient enough to handle whatever emerges. If you can see how AI is being used, set clear boundaries, and embed sensible guardrails into everyday ways of working, you will be well-placed to navigate continual change.

Used well, AI is as much an opportunity as a threat. The same practices that reduce risk - clearer ownership, earlier feedback, stronger monitoring, and more deliberate governance - also make security more scalable and effective across your organisation. The goal of this playbook is not to slow you down, but to help you move fast with confidence.

References & Reading List

- Cyber Assessment Framework 4.0
- OWASP GenAI Security Project
- Artificial Intelligence Playbook for the UK Government - GOV.UK
- MITRE ATLAS™
- https://snyk.io/lp/thriving-age-of-ai

ABOUT THE AUTHORS

Chris Rutter

Global Security Lead

chris.rutter@equalexperts.com

A developer turned security specialist, Chris has spent over a decade helping teams in finance, retail and government make security better, faster and a lot less annoying. Having worked in both engineering and security roles, Chris specialises in building large-scale technical security capabilities and DevSecOps practices that help teams to build securely and safely without slowing delivery.

Ben Wilkes

AI Engineering Lead

ben.wilkes@equalexperts.com

With over 20 years in solution architecture and software engineering, Ben helps organisations deliver complex digital transformations, design scalable systems and build high-performing engineering teams.

Now focused on Generative AI, Ben helps teams move beyond prototypes to production-ready AI solutions. His approach combines disciplined engineering practices with hands-on delivery experience—grounded in Agile and XP principles—to ensure AI initiatives deliver real outcomes.

Phil Parker

Global Head of Technology Strategy & AI Delivery

phil.parker@equalexperts.com

Phil Parker is Head of Technology Strategy at Equal Experts, where he helps organisations navigate the rapidly evolving technology landscape and deliver meaningful business outcomes. With more than two decades of experience spanning software product delivery, agile transformation, and strategic leadership, Phil specialises in shaping technology approaches that align with organisational goals and deliver lasting value.

He is passionate about applying emerging technologies — at the moment particularly AI in Delivery — in practical, outcome-focused ways, and about building collaborative, empowered teams that solve complex problems. Phil's work is driven by a belief that great technology strategy is as much about people and culture as it is about tools and platforms.

www.equalexperts.com

EQUAL EXPERTS