InfoSecAI

A BRIEF FOR CISOS · BISOS · GRC LEADERSHIP

AI GOVERNANCE BRIEFING · 2026

# Governing AI-Generated Content

How to reduce, detect and manage hallucination risk, with a control framework, a fact-checking workflow, and a maturity model for security, risk and technology leaders.

---

AUTHORED BY

**Paul Jolliffe**

Founder & Director, InfoSecAI · Senior CISO / vCISO · CISSP · ISO 27001 Lead Auditor · MBA

### A note on evidence

| | |
|---|---|
| VERIFIED | Standards, frameworks and the AI-report failure pattern are checked against the primary sources listed at the end. |
| GOOD PRACTICE | Controls and workflows reflect established information-security and AI-governance practice, not a single mandated method. |
| ANALYSIS | Reasoned interpretation by the author. |
| ILLUSTRATIVE | Failure scenarios are hypothetical unless a verified source is cited. |

Controls reduce, detect and manage hallucination risk. No control set eliminates it.

01 · WHY THIS IS NOW A BOARD-LEVEL ISSUE

# Why this is now a board-level issue

AI-generated content has moved from novelty to infrastructure. Marketing copy, code, customer replies, research summaries, board papers and decision-support outputs are now drafted, in whole or in part, by generative models. The productivity case is real. So is a quieter problem: these systems can produce confident, fluent, well-formatted content that is simply wrong.

That matters because the failure does not look like a failure. A fabricated citation, an invented statistic or a misread policy clause arrives wearing the same typography as a verified one. When such content feeds a customer communication, a regulatory filing or a board decision, the error can propagate before anyone notices. The point of governance here is not to ban the tools. It is to make sure that what the organisation publishes or acts upon has been checked.

This is a leadership issue, not only a technical one. The risks are legal, regulatory, reputational and operational, and the controls that address them sit across people, process and technology. Boards are increasingly expected to know where AI-generated content is used, which uses are high risk, and what evidence supports the claims their organisation makes.

02 · WHAT WE MEAN BY AI-GENERATED CONTENT

# What we mean by AI-generated content

AI-generated content is any output produced or substantially drafted by a generative model. It is broader than chatbot text. In practice it spans:

- **Text:** emails, reports, summaries, policies, marketing and customer communications.
- **Code:** application logic, scripts, configuration and infrastructure-as-code.
- **Images, audio and video:** synthetic media and edited assets.
- **Structured outputs:** data extractions, classifications and tabulations.

- **Decision-support:** recommendations, risk scores and analytical narratives.

The risk profile rises sharply as the output moves from low-stakes drafting toward decision-support and safety-relevant or regulated uses. A first draft a human heavily edits is low risk. An unreviewed recommendation that drives a customer, clinical, financial or compliance decision is not. Analysis.

03 · WHY HALLUCINATION HAPPENS

# Why hallucination happens

A generative language model produces text by predicting the most probable next token given everything before it. It is optimised to produce plausible continuations, not to state truth. When the model lacks grounding in a reliable source, it fills the gap with something that fits the pattern. The result can be fluent and entirely unsupported.

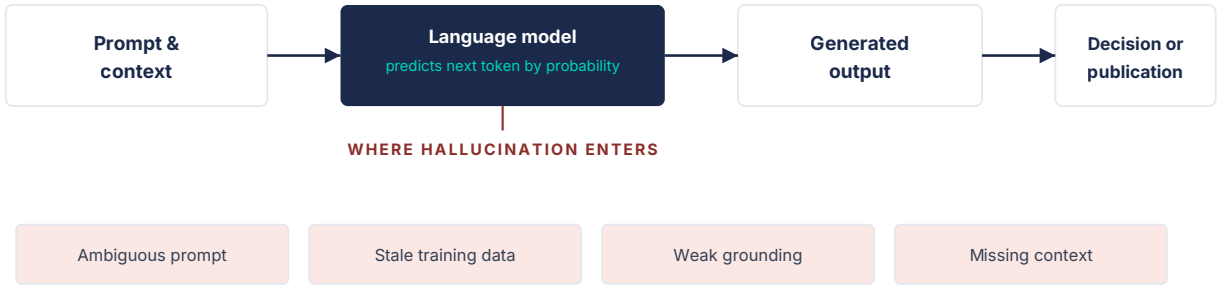


**FIGURE 1** How AI-generated content is produced, and where hallucination enters. The model predicts likely text rather than verifying it; ambiguity, stale data, weak grounding and missing context are the common entry points for error.

Several conditions raise the likelihood of hallucination. The US National Institute of Standards and Technology refers to this behaviour as confabulation, listing it among the risks specific to generative AI in its Generative AI Profile. Verified. Common contributing factors include:

- **Probabilistic generation:** the model optimises for plausibility, not accuracy.
- **Training-data limits and staleness:** the model may not know recent or niche facts.
- **Weak source grounding:** without retrieval from trusted material, the model improvises.
- **Ambiguous or thin prompts:** vague instructions invite invention.
- **Missing domain context:** the model cannot reconcile facts it was never given.
- **Over-reliance by users:** outputs accepted without review become errors of record.

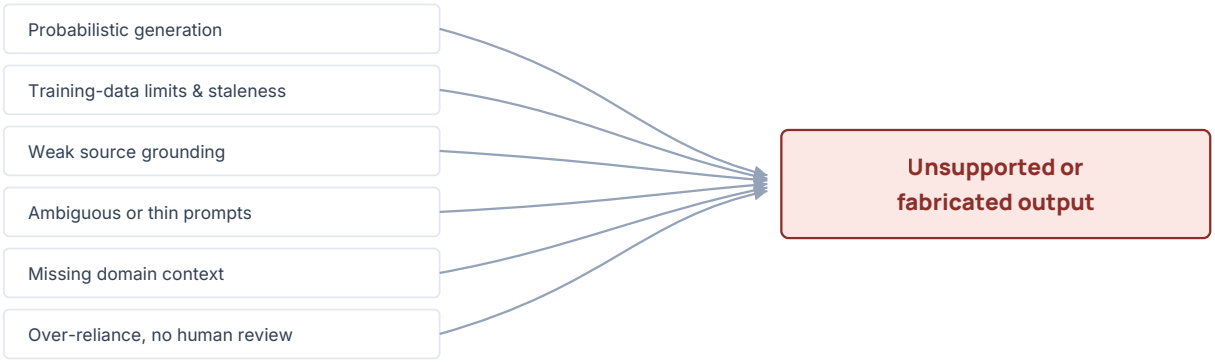


**FIGURE 2** Hallucination is rarely one cause. Several conditions converge to produce unsupported or fabricated output, which is why a single control seldom fixes it.

Importantly, hallucination cannot be fully eliminated by configuration alone. Controls reduce its frequency, increase the chance of catching it, and limit the damage when it occurs. Analysis.

04 · WHY HALLUCINATION MATTERS

# Why hallucination matters

The consequences scale with where the content is used. The same fabricated claim is trivial in a discarded draft and serious in a published filing.

- **Legal and regulatory exposure:** inaccurate disclosures, misstatements or unlawful processing can attract liability. Seek qualified legal advice for specific cases.
- **Customer trust:** misleading or incorrect communications erode confidence quickly and visibly.
- **Cybersecurity:** AI-generated code may contain insecure patterns; AI agents may be manipulated through prompt injection.
- **Internal decision-making:** fabricated figures in analysis or board papers distort decisions that are expensive to reverse.
- **Brand and reputation:** a single exposed fabrication can overshadow genuine work.
- **Misinformation:** errors are increasingly recycled into news and into other AI systems, propagating beyond the original document.
- **Operational resilience:** automated pipelines that act on unverified outputs can fail at scale.
- **Intellectual property and confidentiality:** prompts and outputs can leak sensitive data or reproduce protected material.
- **Safety-critical and high-impact decisions:** in clinical, financial, legal or engineering contexts, an unverified output can cause direct harm.

A verified, real-world pattern illustrates the reputational dimension. Across recent months, several professional-services firms have published or submitted AI-assisted reports later found to contain fabricated citations and invented claims, a phenomenon an AI-detection

firm labelled "vibe citing". Verified; see Source Notes. The lesson is not that AI is unusable. It is that publishing unverified AI output carries real cost.

05 · COMMON FAILURE SCENARIOS

## Common failure scenarios

The following scenarios are illustrative patterns drawn from common practice, not specific real incidents, except where a verified source is cited above. Illustrative.

- A model invents plausible but non-existent citations or references in a report.
- A summary misrepresents a source document, dropping a caveat that changes the meaning.
- A model fabricates a policy or regulatory requirement that does not exist.
- AI-generated code introduces an insecure pattern that passes casual review.
- A customer communication contains a confident but incorrect commitment.
- A board report includes a figure the model invented or misattributed.
- A model misreads legal, medical, security or compliance content and states the opposite of the source.

Each scenario shares a root cause: output that was used or published without verification proportionate to its risk.

06 · A CONTROL FRAMEWORK FOR REDUCING HALLUCINATION

## A control framework for reducing hallucination

No single control is sufficient. Effective programmes layer controls in depth so that an error missed at one layer is caught at another. The layers below move from strategic governance to day-to-day culture. Good practice.

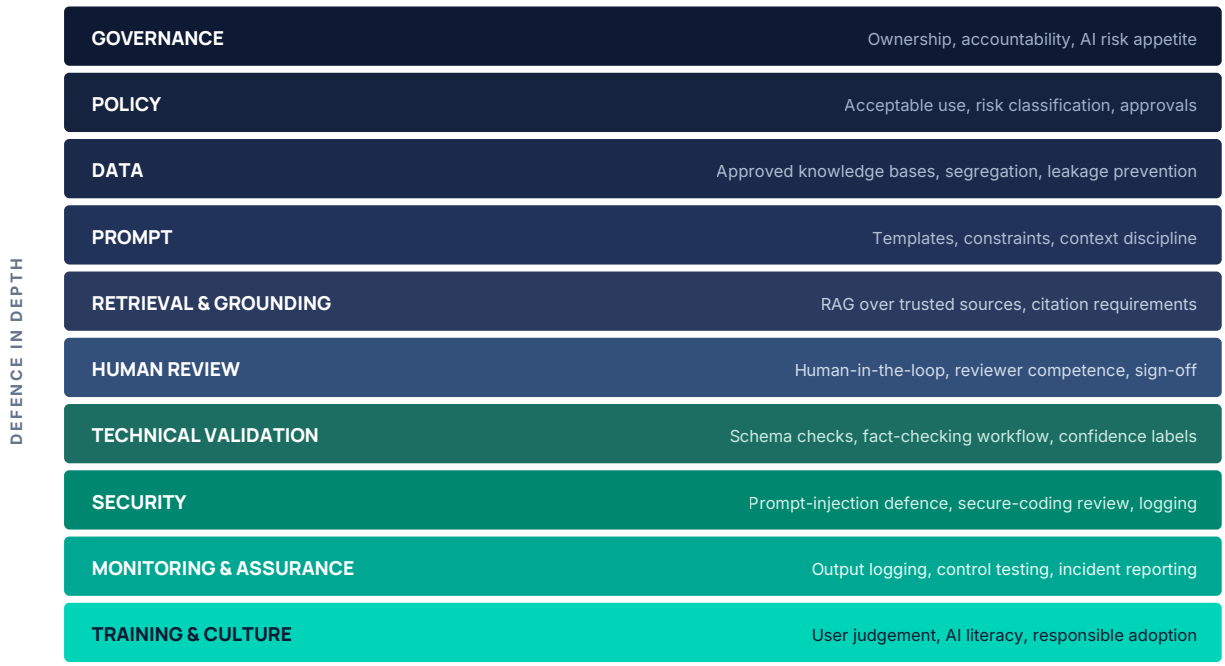


**FIGURE 3** Defence in depth for AI-generated content. Ten complementary control layers, from governance to culture, so that a failure at one layer is caught at another.

The layers work together: governance sets accountability and risk appetite; policy translates that into rules and approvals; data and prompt controls improve what goes in; retrieval and grounding tie outputs to trusted sources; human review and technical validation catch errors before use; security defends the pipeline; monitoring and assurance prove the controls operate; and training builds the judgement that no automated control replaces.

07 · SPECIFIC CONTROLS TO IMPLEMENT

# Specific controls to implement

The following controls operationalise the framework. Apply them proportionately, weighted toward high-risk use cases. Good practice.

### Govern and classify

- AI acceptable-use policy defining permitted and prohibited uses.
- Risk classification of AI use cases by impact and exposure.
- AI system inventory with named owners for each use case.

### Ground and constrain the output

- Approved knowledge bases as the source of truth for grounded answers.
- Retrieval-augmented generation where factual accuracy matters.
- Source-citation requirements for factual claims.
- Prompt templates that constrain scope and demand sources.
- Output confidence labelling to flag low-certainty content.

### Verify before use

- Human-in-the-loop review for anything above low risk.
- A defined fact-checking workflow before publication or action.
- Legal and compliance review for high-risk or regulated outputs.
- Secure-coding review for AI-generated code.
- Content approval workflows with named sign-off.

### Secure and protect

- Data-leakage prevention on prompts and outputs.
- Segregation of sensitive data from general AI tooling.
- Prompt-injection and abuse testing for AI agents and assistants.

### Assure and improve

- Model-output logging for traceability and investigation.
- Red-team testing and bias or misinformation testing.
- Incident reporting for AI-generated errors.
- Periodic control testing to confirm controls still operate.
- User training on the limits of AI output and on verification.

08 · A PRACTICAL FACT-CHECKING WORKFLOW

## A practical fact-checking workflow

A repeatable verification routine is the single highest-value control for published or decision-relevant content. It does not require specialist tooling, only discipline and proportionality. Good practice.

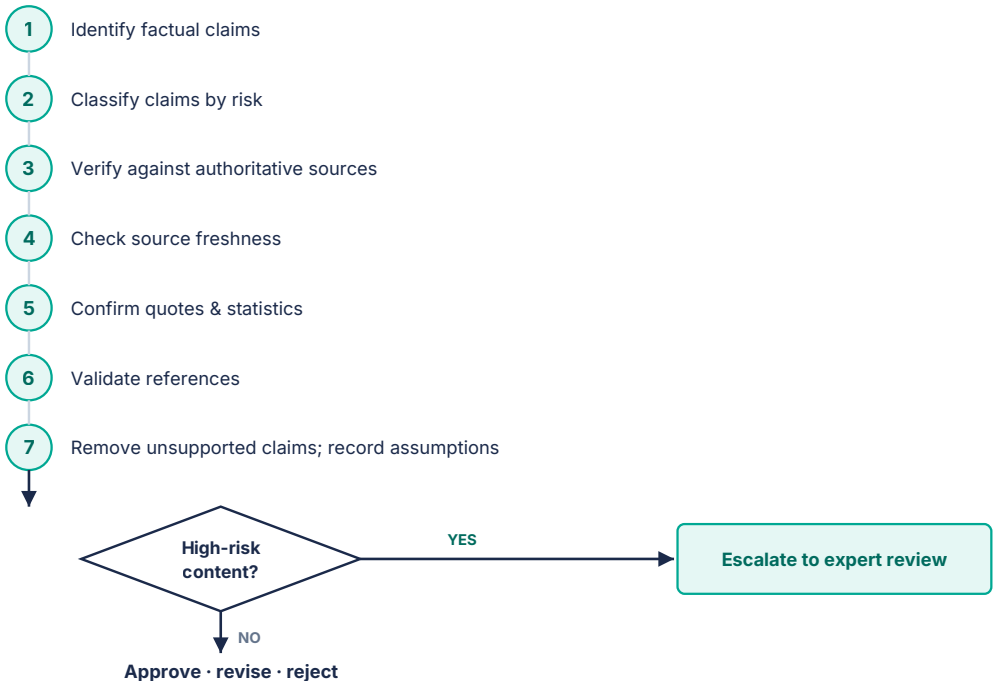


**FIGURE 4** A proportionate fact-checking workflow. Claims are identified, risk-classified, verified against authoritative sources, and either approved, revised or rejected, with high-risk content escalated for expert review.

In practice: identify the factual claims; classify them by risk; verify each against an authoritative source; check the source is current; confirm quotes and statistics against the original; validate that references actually exist and support the claim; remove anything unsupported; record assumptions explicitly; escalate high-risk content for expert review; and only then approve, revise or reject. Where a claim cannot be verified, the safe default is to remove it.

09 · PEOPLE, PROCESS AND TECHNOLOGY

# People, process and technology

Reducing hallucination is not purely technical. Tooling improves the inputs and catches some errors, but the decisive controls are organisational. Analysis.

- **People:** user judgement, AI literacy, and a culture that treats unverified output as a draft, not a fact.
- **Process:** clear ownership, approval workflows, verification routines, and assurance that controls operate.
- **Technology:** grounding, retrieval, logging, validation and security controls, configured for the use case.

> The organisations that manage this well treat an AI output the way an auditor treats an assertion: useful, but not evidence until it has been checked.

10 · WHAT BOARDS AND EXECUTIVES SHOULD ASK

# What boards and executives should ask

Senior leaders do not need to understand model internals. They do need answers to a short, pointed set of questions. Good practice.

- Where are we using AI-generated content, and in which decisions?
- Which use cases are high risk, and how were they classified?
- Who approves outputs before they are published or acted upon?
- What evidence supports the factual claims we make with AI assistance?
- How do we detect, report and learn from AI-generated errors?
- Which controls are tested, and how do we know they operate?
- What is our stated tolerance for AI-generated error in each context?

11 · A PHASED IMPLEMENTATION ROADMAP

# A phased implementation roadmap

Most organisations cannot govern every use case at once. A phased approach concentrates effort where risk is highest first. Good practice.

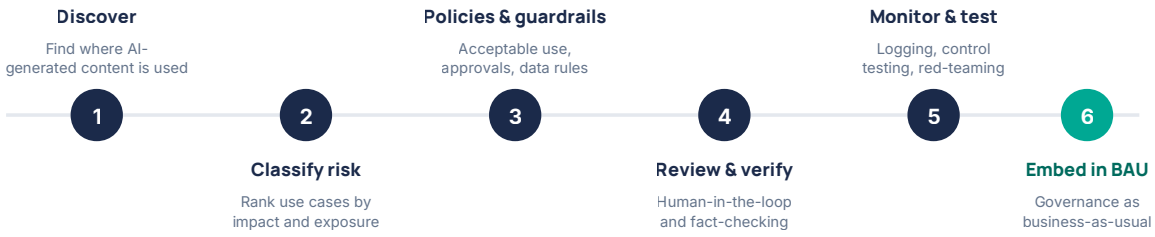


**FIGURE 6** A phased roadmap. Discover where AI-generated content is used, classify by risk, set policy and guardrails, implement verification, monitor and test, then embed governance into business-as-usual.

Phase 1 discovers where AI-generated content is actually used, which is often broader than leadership expects. Phase 2 classifies those uses by risk. Phase 3 sets policy and guardrails. Phase 4 implements review and verification workflows for the high-risk uses. Phase 5 adds monitoring, control testing and red-teaming. Phase 6 embeds the whole programme into business-as-usual so it persists beyond the initial push.

12 · A MATURITY MODEL

# A maturity model

Progress is easier to manage against a simple maturity scale. The aim is not maximum maturity everywhere, but maturity proportionate to risk. Good practice.

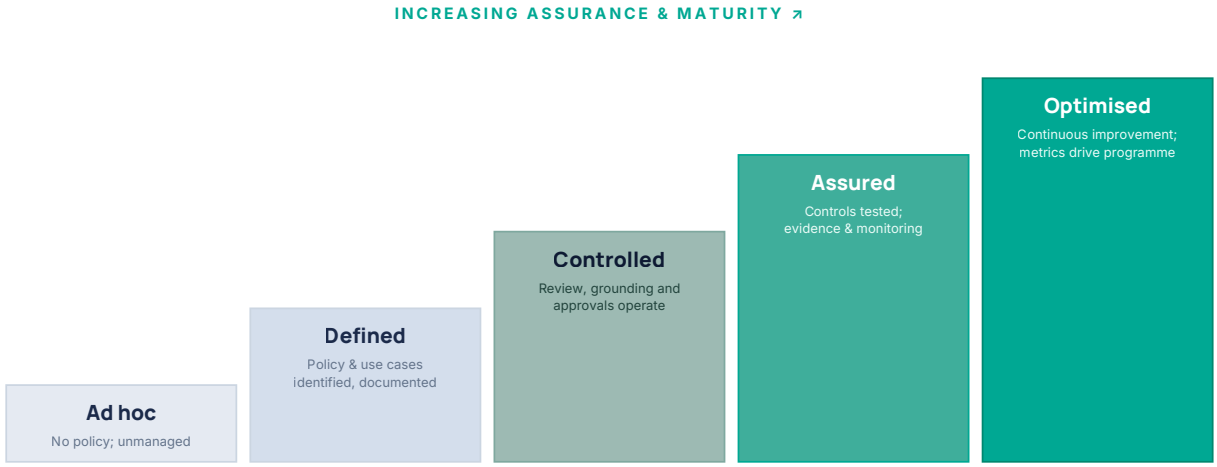


**FIGURE 5** A five-level maturity model for governing AI-generated content, from ad hoc use to an optimised, evidence-driven programme. Most organisations should target Controlled or Assured for high-risk use cases.

13 · CONCLUSION

# Conclusion

AI-generated content can create real value, and the productivity gains are genuine. The risk is not the technology itself but the habit of using its output without verification proportionate to the stakes. Hallucination cannot be engineered away entirely. It can be reduced, detected and managed through layered controls, clear accountability, and a verification routine that scales with risk.

The organisations that get this right will not be the ones that use AI the least, nor the ones that use it most freely. They will be the ones that can show, with evidence, that what they publish and act upon has been checked. Analysis.

> **For discussion.** AI-generated content is now part of how most organisations work. The question is no longer whether to use it, but how to govern it so the output can be trusted. I would value other perspectives:
>
> - Where have you drawn the line between low-risk drafting and decision-relevant output?
> - What single control has given you the most assurance for the least friction?
> - How are you evidencing to your board that AI-content controls actually operate?
> - Where does human review add real value, and where has it become a rubber stamp?
> - How are you testing for AI-generated errors before customers or regulators find them?

14 · APPENDIX A: CONTROL CHECKLIST

# Appendix A: Control checklist

| CONTROL | CONTROL |
|---|---|
| ☐ AI acceptable-use policy published and owned | ☐ AI use cases inventoried with named owners |
| ☐ Use cases risk-classified by impact and exposure | ☐ Approved knowledge bases defined for grounded answers |
| ☐ Retrieval and source-citation required for factual claims | ☐ Prompt templates in use for repeatable, high-risk tasks |
| ☐ Human-in-the-loop review for above-low-risk output | ☐ Fact-checking workflow applied before publication or action |
| ☐ Legal and compliance review for high-risk or regulated outputs | ☐ Secure-coding review for AI-generated code |
| ☐ Data-leakage prevention and sensitive-data segregation | ☐ Prompt-injection and abuse testing for AI agents |
| ☐ Model-output logging enabled for traceability | ☐ Red-team, bias and misinformation testing scheduled |
| ☐ Incident reporting route for AI-generated errors | ☐ Periodic control testing confirms controls operate |
| ☐ User training on AI limits and verification delivered | ☐ Board reporting on AI-content risk and tolerance in place |

15 · APPENDIX B: HALLUCINATION-CONTROL MATURITY MODEL

# Appendix B: Hallucination-control maturity model

| LEVEL | WHAT IT LOOKS LIKE |
|---|---|
| **Ad hoc** | No policy. AI used informally. No verification or ownership. |
| **Defined** | Policy exists, use cases identified, ownership assigned. |
| **Controlled** | Review, grounding and approvals operate for high-risk use. |
| **Assured** | Controls tested; logging, monitoring and evidence in place. |

| LEVEL | WHAT IT LOOKS LIKE |
|---|---|
| **Optimised** | Metrics drive continuous improvement across the programme. |

16 · APPENDIX C: EXECUTIVE TAKEAWAYS

# Appendix C: Executive takeaways

1. Hallucination is a probabilistic feature of how models generate text. It can be reduced and managed, not eliminated.
2. Risk scales with use. Govern decision-relevant and regulated outputs hardest; leave low-risk drafting light.
3. Layer controls in depth. Grounding, human review and a fact-checking routine catch most errors before they cause harm.
4. Prove operation, not just design. Untested controls give false comfort; assurance is the evidence that they run.
5. Recognised frameworks exist. ISO/IEC 42001, the NIST AI RMF and its Generative AI Profile, and the EU AI Act give structure; verify scope for your context.

17 · SOURCE NOTES

# Source notes

**ISO/IEC 42001:2023.** International standard specifying requirements for an AI management system (AIMS). Published 2023. iso.org/standard/42001. Supports: AIMS scope and the existence/role of the standard.

**NIST AI Risk Management Framework (AI RMF 1.0).** US National Institute of Standards and Technology. Voluntary framework; functions Govern, Map, Measure, Manage. Published January 2023. nist.gov/itl/ai-risk-management-framework. Supports: Framework name, status and structure.

**NIST AI 600-1, Generative AI Profile.** Cross-sectoral profile of the AI RMF for generative AI. Published July 2024. Uses the term 'confabulation' for hallucination and lists it among GenAI risks. nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf. Supports: NIST's term for hallucination and its listing as a GenAI risk.

**EU AI Act, Regulation (EU) 2024/1689.** First horizontal EU AI law. Entered into force 1 August 2024; obligations phase in, with transparency duties for AI-generated content (Article 50) among them. digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. Supports: Existence, status and phased application of the regulation.

**GPTZero investigation and Financial Times reporting.** Forensic reviews of AI-assisted professional-services reports containing fabricated citations; the term 'vibe citing'. Reported June 2026. gptzero.me/news/investigations-kpmg. Supports: The verified real-world failure pattern referenced in Section 4.

> Sources above were checked during drafting. Remaining content is industry good practice, reasoned analysis, or clearly labelled illustrative example, and should be independently verified before publication. This article is practitioner guidance, not legal, regulatory, medical or financial advice; obtain qualified advice for specific decisions.

InfoSecAI

# About InfoSecAI

InfoSecAI is an independent UK consultancy helping organisations turn security, regulatory, resilience and AI governance requirements into practical operating models, stronger controls and robust delivery.

We work across strategy, governance, risk, compliance, AI security, assurance, operations and engineering. Our services help leadership teams assess their current position, align to standards and regulation, define the target operating model, and deliver the governance, controls, artefacts and ways of working needed to move from intent to implementation.

Our toolkit capability accelerates structured work across ISO 27001, ISO 22301, ISO 42001, NIST CSF, NIST AI RMF, CIS Controls, Cyber Essentials, DORA, NIS 2, the EU AI Act, GDPR, UK GDPR, SOC 1 and SOC 2. The approach combines AI-enabled workflow support with senior practitioner judgement, so outputs remain proportionate, usable and connected to the way the organisation actually operates.

InfoSecAI was founded in **2025** by **Paul Jolliffe**. The company is built for organisations that need clarity, senior leadership and hands-on delivery across information security and AI governance, without adding unnecessary complexity or treating compliance as a paperwork exercise.

**infosecai.net · paul.jolliffe@infosecai.net**