InfoSecAI

A BRIEF FOR CISOS · BISOS · GRC LEADERSHIP

AI GOVERNANCE BRIEFING · 2026

# A Control on Paper Is Not a Control

Several firms that sell assurance over other people's controls have just been caught failing to operate their own. That is the more damning reading, and the more instructive one.

---

AUTHORED BY

**Paul Jolliffe**

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

01 · AN INDUSTRY THAT SELLS TRUST, CAUGHT FAILING AT THE ONE THING IT SELLS

## An industry that sells trust, caught failing at the one thing it sells

The professional services industry does not really sell audits, reports, or advice. It sells the right not to check. A banker in one capital can rely on a ledger drafted in another, never having met the people who wrote it, because a recognised firm has put its name to the work. That name is a promise that someone competent and independent has already done the checking, so the rest of us do not have to. Strip away the methodologies and the frameworks and that is the entire product. Trust, packaged and sold at scale.

So there is a particular sting when one of these firms publishes a flagship report on the wonders of a new technology, and that report turns out to be built on fabricated references and invented claims produced by the very technology it was promoting. The detail that should hold attention is not the embarrassment. It is the category. A firm whose business is verification published something it had not verified. The product failed at the exact point the product exists to cover.

This is not a story about one careless report. It is a story about what happens to a trust business when the mechanism that produces trust stops operating, and nobody notices until an outside party runs the check the firm was supposed to run itself.

02 · THE PATTERN, NOT THE INCIDENT

## The pattern, not the incident

Treat a single failure as bad luck and you learn nothing. What makes this worth writing about is that it is not single. Within a short window, more than one global firm has published or submitted AI-assisted work later found to contain fabricated citations or invented findings. One refunded a government client after AI-generated content reached a taxpayer-funded report. One withdrew a report after fake references were spotted. One pulled a flagship report after an external review found that the large majority of its citations did not point to real, intact sources. The same failure mode, the same root cause, across different firms, in different countries, over a matter of months. Other professional firms have been caught in adjacent versions of the same thing.

When an identical defect appears across multiple independent organisations in quick succession, it is not a collection of individual lapses. It is systemic. The common factor is not a rogue employee at any one firm. It is a new production tool adopted faster than the controls that should govern it, inside a delivery culture that rewards speed and volume. The incident is interchangeable. The pattern is the finding.

And patterns travel. Each public case lowers the shock value of the next, which is its own quiet danger.

> The first firm caught is a scandal. The fourth is a trend piece. Normalisation is how a profession talks itself into treating an integrity failure as a manageable operational nuisance.

03 · WHAT ACTUALLY WENT WRONG

# What actually went wrong

The mechanics are worth getting right, because the fix depends on them. Generative tools do not fail the way a careless human fails. Asked to find real-world examples of a capability in the wild, a research assistant built on a language model will tend to over-comply. It will return examples whether or not they exist, because returning something reads as success and returning nothing reads as failure. The model is optimised to be helpful, not to be correct, and in the absence of grounding those two objectives diverge.

The result is a spectrum of fabricated sourcing that one detection firm labelled "vibe citing", by analogy with vibe coding. At the mild end, a real reference is paraphrased until its title no longer matches anything. In the middle, two genuine sources are fused, the authors of one paired with the title of another. At the severe end, the reference is invented outright, complete with plausible authors and a plausible venue that never published it. To a reader skimming a polished document, all three look identical to a real citation. That is the trap. The output carries every visual signal of having been checked while none of the checking occurred.

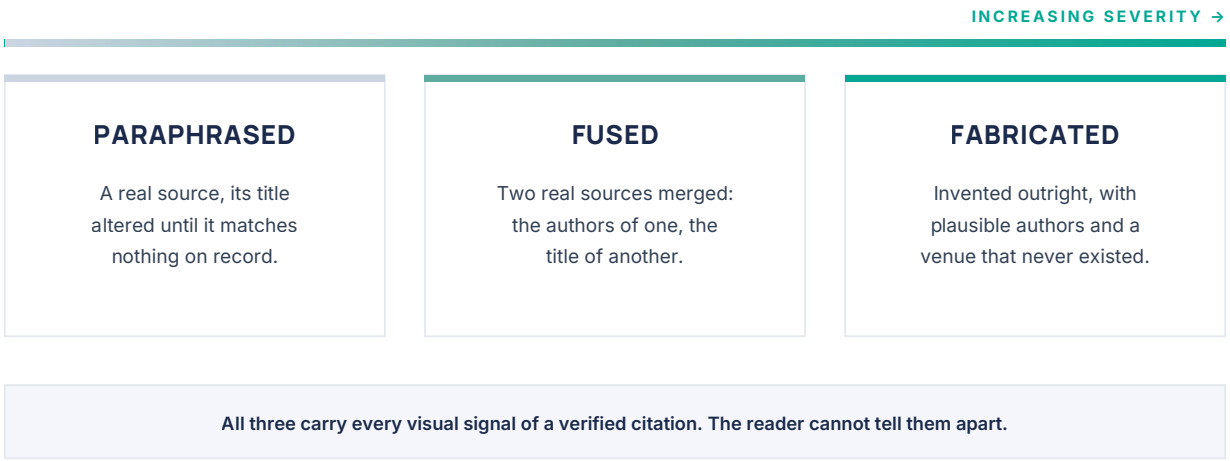


**FIGURE 1** The fabrication spectrum. A failed citation can be a mangled real source, a fusion of two, or a pure invention, and on the page all three are indistinguishable from a verified reference.

The same failure hits factual claims, not just references. In one such report, a large share of the substantive claims were found to be false or misattributed. Real organisations were described as operating systems they did not operate, or as having capabilities they did not have. In at least one instance the report contradicted the firm's own separately published figures on the same subject in the same period. The tool did not know the firm's other numbers, and nobody reconciled them. None of this is mysterious once you understand

that an ungrounded model fills gaps with fluent invention. The mystery is only how it reached publication.

04 · AN OPERATING-EFFECTIVENESS FAILURE, NOT A DESIGN GAP

## An operating-effectiveness failure, not a design gap

Here is the part the profession should sit with. Ask the firms involved whether their policies required human verification of AI-assisted content, and the answer is yes. Public statements after the fact pointed to existing guidelines on responsible AI use, human oversight of output, and verification of independent sources. The control was designed. It was written down. On paper, the firm was covered. It simply did not run.

This is precisely the distinction these firms apply to everyone else. In any serious assurance engagement, a control is assessed on two axes. Is it designed to address the risk, and does it operate as designed across the relevant period. A control can pass the first test and fail the second completely. A documented policy that no one followed, with no evidence of operation, is not a mitigating control. It is a deficiency dressed as a control. A Type 2 attestation exists specifically to test operation over time rather than existence at a point in time, because the profession learned long ago that the gap between the two is where losses live.

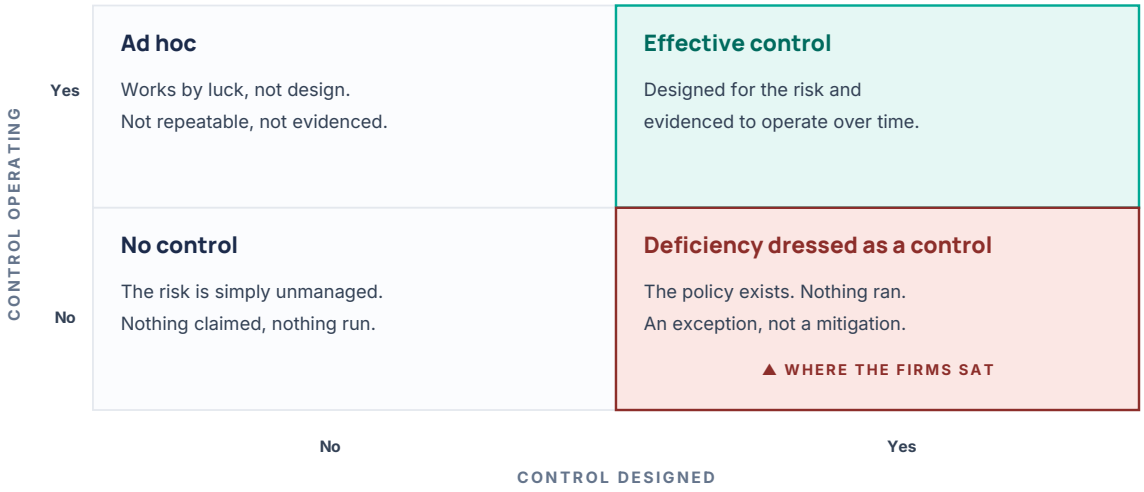


**FIGURE 2** Design and operation are separate tests. A control can be perfectly designed and still fail completely if it never runs. The pulled reports sat in the lower-right cell: policy present, operation absent.

So the firms failed the exact test they administer for a living. Had a client presented a control reading "all externally published content is reviewed for source accuracy before release", with a written procedure and no evidence that a single review took place across the period, the firm would raise an exception without hesitation. The standard the firm applies to its clients convicts the firm. That is why "the standards need to catch up" is the wrong lesson. The standard was adequate. The operation was absent.

05 · THE ECONOMICS OF GETTING CAUGHT

# The economics of getting caught

Follow the incentives and the behaviour stops being surprising. When an AI-tainted deliverable reaches a paying client and is later exposed, what has happened so far is a refund and a quiet withdrawal of the document. Set that penalty against the size of the franchise. A refunded fee, even a large one, is a rounding error against a global revenue base and a brand built over more than a century. A withdrawn report is reputationally awkward for a quarter and forgotten by the next.

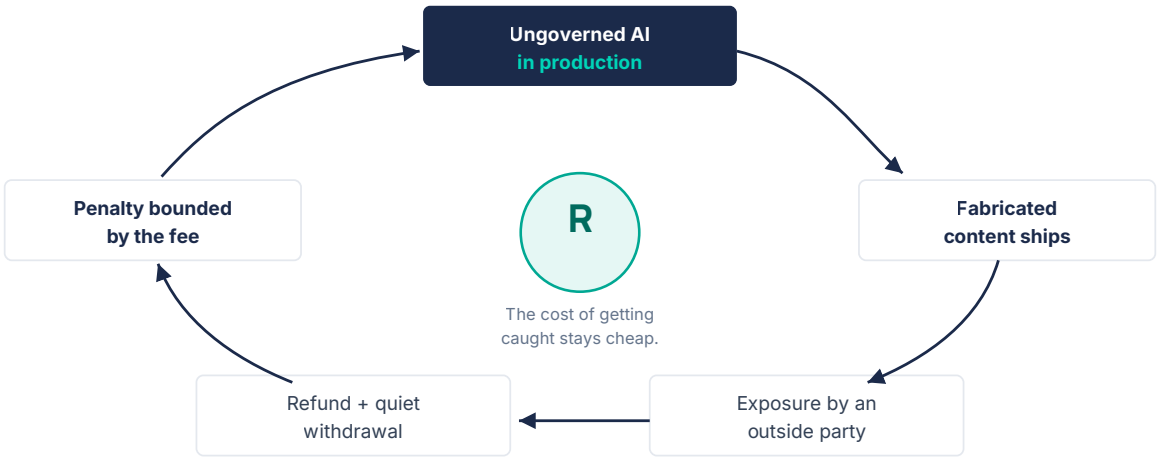


**FIGURE 3** A reinforcing loop. While the penalty for exposure is bounded by the fee, every firm watching learns that the cheapest path is to keep shipping fast, and the cycle renews itself.

A refund does not undo the harm. It does not un-spend a client's budget, it does not reverse the decisions other people made while relying on the fiction, and it does not retrieve the fabricated figures that have already been recycled into news coverage and, increasingly, into the training and retrieval paths of other AI systems that will repeat them. The error propagates long after the cheque clears. The remedy addresses the invoice, not the damage.

What every firm watching learns from this is simple and corrosive. The expected cost of getting caught is low and bounded, while the expected cost of slowing down to verify is paid every single time, in missed deadlines and higher delivery cost. A rational team under pressure, told to produce more thought leadership faster with the same headcount, will reach for the ungoverned tool again. Deterrence that runs through a refund mechanism is not deterrence. It is a price, and a cheap one.

06 · TRUST AND ASSURANCE ARE NOT THE SAME THING

# Trust and assurance are not the same thing

These two words get used interchangeably, and the slippage matters here. Trust is a belief about another party's goodwill and competence that lets you act without verifying. It is an economising device. It removes the cost of checking by replacing it with a relationship. Assurance is what you reach for when that belief is unavailable. When you

cannot extend trust directly, because the counterparty is distant or unknown, you buy an independent opinion that stands in for the trust you cannot give. Assurance is manufactured trust, sold by a third party.

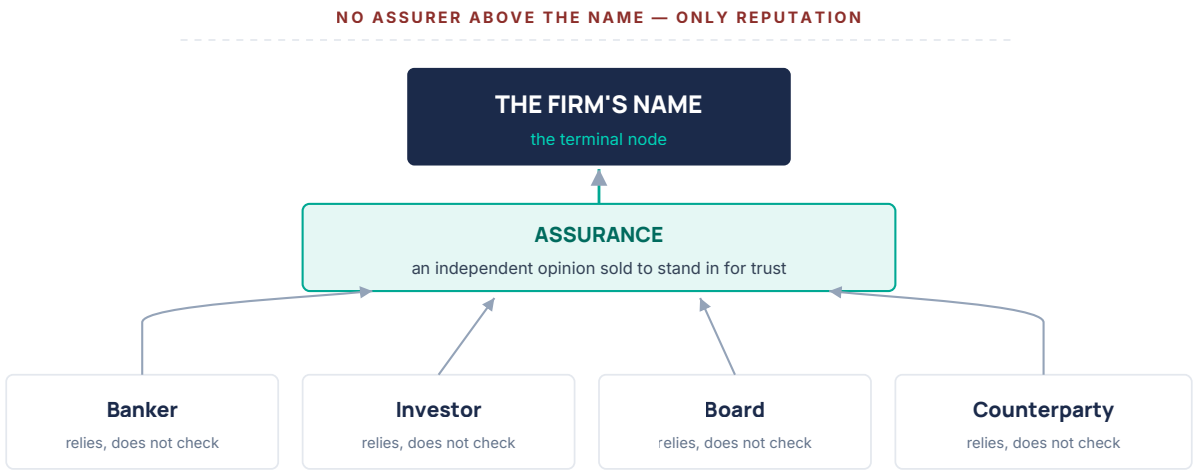


**FIGURE 4** The structure has no floor. Distant parties rely on the name instead of checking each other; the name relies on nothing above it but its own reputation. When the terminal node fabricates, the chain has nothing to catch it.

The structure has a floor problem. Assurance only works if you can trust the assurer. You have outsourced the checking, but the checker is now the terminal node. There is no further party assuring the assurer, only reputation, peer review, and the slow discipline of a market that remembers. The whole edifice rests on the credibility of the firm at the bottom of the stack. That credibility is not one product line among others. It is the foundation the other product lines stand on.

Which is why an integrity failure in a marketing report cannot be quarantined from the audit practice next door. The market does not buy from a department. It buys from a name. The same name that signs an opinion also sits on the cover of the report that fabricated its sources. A reader cannot, and will not, hold a sophisticated mental model in which the firm's content arm is unreliable but its assurance arm is pristine. Reputation is indivisible. This episode does not erode trust and assurance separately. It erodes the one belief that both depend on, that the name still means the checking was done.

07 · THE CONTROLS THAT WOULD HAVE CAUGHT IT

## The controls that would have caught it

None of this required new invention. The controls that would have prevented every case in the pattern are ordinary, cheap, and already understood. The failure was operational, so the remedy is operational discipline, not a research breakthrough.

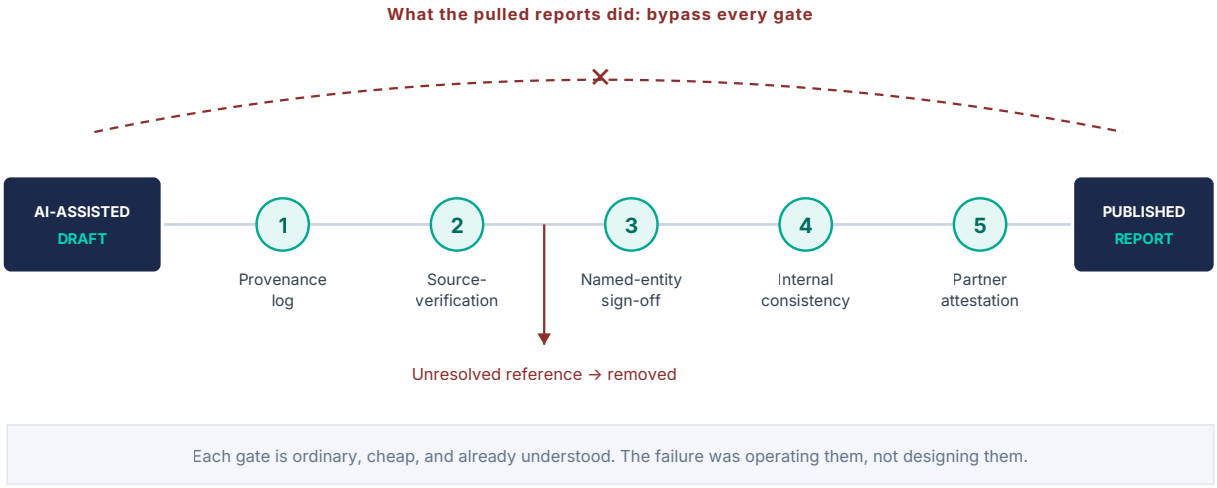


**FIGURE 5** The verification gate. A published deliverable passes through provenance logging, source verification, named-entity sign-off, internal-consistency reconciliation, and a real human attestation. The pulled reports went straight from draft to publication.

Start with a source-verification gate. Before any externally published document is signed off, every external citation is resolved to its primary source by a human or a deterministic tool, and a reference that cannot be resolved is removed. This single mechanical step, applied without exception, catches the entire class of fabricated citation. It is tedious, which is exactly why a tool under deadline pressure skips it and a governed process does not.

Layer on named-entity sign-off. Any claim that a real, identifiable organisation does a specific thing, or operates a specific system, must be supported either by that organisation's own confirmation or by a verifiable public source, before it appears in print. Describing a real company as running a product it does not run is not a citation nicety. It is reputational and legal exposure aimed at a third party who never agreed to be a case study.

Add an internal-consistency check. Cross-reference every figure against the firm's own concurrent publications. When a report's headline statistic contradicts a survey the same firm released the same month, the failure is not subtle and the check is not expensive. It is reconciliation, the most basic discipline the profession owns. Then govern the tool itself. Log provenance, so the firm knows which passages were AI-drafted and can target verification where the risk actually sits. Constrain the tool to grounded retrieval over a known set of real source material rather than open generation, and ban the open-ended prompt that asks a model to find examples in the wild, because that instruction is the precise trigger for over-compliant fabrication. Finally, require genuine attestation at the point of release. Not a rubber stamp, but a named individual who personally confirms the verification occurred and carries the accountability if it did not.

08 · THE GOVERNANCE WRAPPER

## The governance wrapper

Individual controls hold only inside a system that makes them mandatory and checks that they ran. That system already has well-developed reference points. An AI management system of the kind described in ISO/IEC 42001 exists to put exactly this scaffolding around AI use: an inventory of where AI is used, a policy with teeth, defined human oversight, and a record that the oversight happened. A risk framework such as the NIST AI Risk Management Framework gives the same discipline a different cut, organised around governing, mapping, measuring, and managing AI risk across its lifecycle.

The reframe that matters is one of classification. A published flagship report carrying the firm's name, distributed globally, likely to be cited by others and ingested by downstream systems, is a high-impact AI output. It should be treated as one, with a mandatory pre-publication control gate proportionate to that impact, the same way a high-risk processing activity attracts an impact assessment. The error in the pattern was treating content production as low-stakes marketing rather than as a high-reach output that shapes decisions and trains other models.

A caution sits underneath all of this. A management system can be certified and a risk framework can be adopted, and the firm can still fail in exactly the way described, if the system is designed and never operated. ISO 27001 results in certification of a management system, a SOC 2 examination results in an attestation over controls, and neither is worth anything if the controls inside them sit dormant. The governance wrapper is necessary. It is not sufficient. What turns it into protection is evidence that the controls operate, gathered continuously, reviewed by someone accountable, and treated as a live obligation rather than a binder produced for an assessor and shelved.

09 · THE UNCOMFORTABLE QUESTION FOR ANYONE BUYING ASSURANCE

## The uncomfortable question for anyone buying assurance

Turn it around to the buyer's seat. When an organisation pays a recognised firm for assurance, what is it actually purchasing. Not pages. Not a logo on a cover. It is purchasing the right to stop checking, on the strength of a promise that the checking has been done to a standard the buyer could not reach alone. That right is the entire value. It is what justifies the fee and what distinguishes the recognised name from an unbranded equivalent at a fraction of the price.

So if the firm itself has stopped checking, the buyer has paid for the signal and not the substance. They hold a document that looks assured, carries every mark of assurance, and may be hollow underneath, in the same way a fabricated citation carries every mark of a real one. The resemblance between a checked deliverable and an unchecked one is precisely the problem. From the outside they are identical until someone runs the verification the buyer believed they had already paid to avoid.

This reframes assurance provider selection as a due-diligence question rather than a procurement formality. A board can reasonably now ask its providers what their own content and verification controls are, whether AI use in deliverables is disclosed, what provenance is logged, and what evidence exists that human verification actually operated rather than merely being required. Those questions would have been faintly insulting to ask a major firm a short while ago. The pattern has made them prudent.

10 · WHAT HAS TO CHANGE

## What has to change

The tempting conclusion is that the rules need rewriting. It is the wrong one. The standards did not fail. Verifying your sources, reconciling your own figures, and not publishing claims about third parties that you have not confirmed are not gaps awaiting a new framework. They are the oldest obligations the profession has. What failed was operation, and operation is a matter of will, process, and accountability, not of standard-setting.

That said, the market will move, and firms that move first will be rewarded. Expect AI-use disclosure in deliverables to shift from optional to expected. Expect buyers to ask for evidence of verification rather than assurances of policy. Expect provenance and human-oversight records to become a competitive differentiator, because the firm that can prove its controls operated will win work from the firm that can only point to a policy. There is even a plausible new assurance product in here: independent evidence that a content-production process actually operated its AI controls over a period, which is just the operating-effectiveness discipline turned on the firms themselves.

For any firm in this business, the immediate work is unglamorous and entirely within reach. Classify high-reach outputs as high-risk. Put a real verification gate in front of them. Constrain the tools to grounded retrieval. Log what the machine wrote. Make a named human accountable for the check. Then, and this is the only part that ultimately matters, gather evidence that all of it happens, every time, and review that evidence as seriously as you would review a client's. The firms that sell trust do not need permission from a standards body to start operating the controls they already wrote down.

> A control on paper is not a control. The firms that sell assurance know this better than anyone, because it is the finding they raise against their clients. The question the next few years will answer is whether they can raise it against themselves.

And somewhere down the line, every buyer ends up at the same question the whole edifice rests on. What exactly were we paying for when we said we were buying trust?

**Fact-check note.** Every sentence containing a number, a named claim, or a factual assertion in this article was re-checked. Figures retained are limited to those verifiable from public reporting on the events described, and are stated without identifying the organisations involved. No citations, study titles, quotes, or case studies have been invented; illustrative passages are framed as general rather than as specific real events. This article is the author's analysis and opinion, not legal advice.

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**