CYBERARMOR.AI · RESEARCH

The Cost of Untrusted AI.

How enterprise AI fails, what those failures cost, and how to know whether it is failing you — an evidence-based taxonomy and board-level self-assessment, built from 36 verified incidents, 2023–2026.

August 2026 edition · Second edition, revised August 4, 2026

US$4.99M

Global average cost of a data breach — IBM/Ponemon 2026, published July 29, 2026

1 in 4

Malicious breaches that were AI-enabled, at US$6M average — up 56% in one year (IBM 2026)

US$893M

Reported losses across 22,364 AI-referencing complaints — FBI IC3, 2025 Internet Crime Report

Aug 2, 2026

EU AI Act transparency obligations and enforcement — in force since this date; penalties to €35M or 7% of turnover

AI Security, Governance and Trust Infrastructure

— How to read this report

> **THE SOURCING RULE**
>
> Every dollar figure in this report is a third party's published measurement, a court's adjudicated sum, or a clearly labelled projection — IBM/Ponemon, the FBI, statutory penalty text, tribunal and plea records, or a named study. Never an estimate of our own. Each claim carries its source, and the claim-level source table — every figure, its source, URL and verification status — is published at cyberarmor.ai/the-cost-of-untrusted-ai-sources.html.
>
> Figures of different kinds are never combined. A modelled survey average, a national tally of victim-reported complaints, a statutory maximum and an adjudicated award measure different things; no arithmetic in this report joins them, and none should be performed by the reader.

This report asks one question on the reader's behalf: **if AI were failing inside your organization — quietly, today — how would you know?** Part 1 shows what AI failure now costs, on the current editions of the two measurements that matter. Part 2 reads 36 verified incidents as a map of where failure actually occurs. Part 3 formalizes that map into a taxonomy of twelve failure modes. Part 4 turns the taxonomy into a 36-question self-assessment scored on evidence, built to be taken to a board. Part 5 shows that the world's regulators, on both sides of the Atlantic, have converged on the same question this report asks — and on evidence, rather than detection, as the acceptable answer.

Contents

1 Executive summary

Two clocks started running in the week this report first went to press. On July 29, 2026, IBM and the Ponemon Institute published the 2026 *Cost of a Data Breach* study, and for the first time the AI attack is not a theme in that report but a line item: **one in four malicious breaches was AI-enabled — a 56 percent increase in a single year — and those breaches cost roughly US$6 million each**, about a million dollars above the global average, which itself rose 12 percent to US$4.99 million. Four days later, on August 2, 2026, the European Union's AI Act reached its next application date: transparency obligations for AI systems that interact with people or generate synthetic content took effect, and the enforcement machinery — with penalties reaching €35 million or 7 percent of worldwide turnover — came fully online. This second edition is revised across that threshold.

Between those two clocks sits the question every board will now put to its security and compliance leadership, in one form or another: **if AI were failing here, how would we know?** Not "are we using AI safely" — a question that invites reassurance — but the harder, auditable form: which ways can it fail, which of those ways are covered by a control, and what evidence would exist if a control had actually run?

This report exists to make that question answerable. It is built on a database of **36 AI security incidents from 2023 to 2026, each verified against the strongest available sourcing** — court records, regulator publications, police briefings and first-party disclosures where they exist; named researcher or major-outlet reporting where they do not. Read together, those incidents do not describe one problem. They describe four, occurring on four distinct surfaces of the same system: **what goes into AI, what AI does, what comes out of it, and what impersonates you to your own people**. Part 3 decomposes those surfaces into twelve named failure modes; Part 4 converts them into a self-assessment an organization can complete in under an hour and defend line-by-line in front of a board.

The three findings

First: AI failure is now measured, and the measurements agree. IBM's 2026 study and the FBI's 2025 Internet Crime Report were produced independently, by different methods, on different populations. They converge. IBM finds AI-enabled breaches — most commonly *deepfake impersonation attacks and AI-enabled malware* — in one of every four malicious breaches, and finds that more than 20 percent of organizations now report breaches targeting their own AI models or applications. The FBI's report contains its first dedicated section on AI in cybercrime: **22,364 complaints referencing AI, with US$893 million in reported losses** — dominated by investment fraud built on AI-generated video and voices of trusted figures (US$632 million), with AI-written executive-impersonation email and cloned-voice distress scams behind it. The phenomenon this report's incident database has tracked since 2023 is now visible in both of the field's standard instruments.

Second: the money and the frequency live on different surfaces — and the gap between them is closing. In this report's own ledger of verified enterprise losses, nearly all of the money that has actually moved — 99.85 percent of US$69.3 million — moved because *a person was deceived by a synthetic identity*: a cloned voice or a fabricated face authorizing a payment. The other three surfaces are dense with demonstrated attack paths — zero-click data exfiltration from a production copilot, remote code execution from a poisoned configuration file, malicious models published to public repositories — that have produced almost no verified dollar losses *yet*. Two facts should stop a reader from finding comfort in that word. Nearly every one of those demonstrated paths was discovered by an outside researcher, not by the victim's own controls: the organizations involved did not know until they were told. And in July 2026 the gap closed — a malicious dataset carrying executable code breached production infrastructure at

Hugging Face, the world's largest AI-artifact platform, in a campaign the platform attributes to an autonomous agent framework. A failure class that had lived in proofs-of-concept became a production breach at the center of the AI supply chain.

Third: regulators have converged — on evidence, not detection. The EU AI Act's newly applicable transparency articles, the SEC's 2026 examination priorities, FINRA's 2026 oversight report, the New York Department of Financial Services' letters of October 2024 and May 2026, FinCEN's deepfake-fraud alert, and HIPAA's newly adjusted penalty schedule differ in jurisdiction, industry and instrument. They are identical in what they ask of an institution. None scores an organization on whether it detected an attack. Every one of them asks whether the organization governs its AI, whether specified controls exist and run, and whether the institution can *produce the record*. The self-assessment in Part 4 is that question, asked systematically, twelve times.

How to use this document

Run the Part 4 assessment with your security, compliance and AI platform owners in the same room; it takes under an hour. Score on the evidence rule: a claim counts only if you could produce the record behind it. Take the one-page scorecard to your board with the Part 5 regulatory map beside it. Re-run it quarterly — the taxonomy is stable; your scores should not be.

2 The measured cost of AI failure

Two instruments measure this field at scale: IBM and the Ponemon Institute's annual *Cost of a Data Breach* study, and the FBI Internet Crime Complaint Center's annual *Internet Crime Report*. Both released new editions within the last four months. They measure different things by different methods and are never combined in this report — but each, on its own terms, recorded the same turn: AI stopped being context and became mechanism.

2.1 What the 2026 breach-cost study says

The 2026 edition — published July 29, 2026, from research conducted by the Ponemon Institute across 602 breached organizations between March 2025 and February 2026 — reverses last year's headline. In 2025 the global average cost of a data breach fell for the first time in five years, to US$4.44 million. In 2026 it rose 12 percent, to **US$4.99 million**. IBM attributes much of the reversal to a single force. The study's title finding is the one every board should carry into its next risk discussion:

1 in 4

Malicious breaches that were AI-enabled in 2026 — a 56% increase over the prior year. IBM names the most common forms: deepfake impersonation attacks and AI-enabled malware.

≈US$6M

Average cost of an AI-enabled breach — roughly US$1 million above the global average breach.

≈US$2M

Average breach-cost reduction reported by organizations using AI and automation in their security operations. One in four organizations has still not adopted these tools.

Source: IBM press release, "IBM Study: One in Four Malicious Breaches are AI-Enabled, Costing Companies $6 Million on Average," July 29, 2026; IBM / Ponemon Institute, Cost of a Data Breach Report 2026. Accessed July 29, 2026.

Those three numbers carry the argument of this entire report in miniature. Attackers using AI raise the cost of failure; defenders governing and instrumenting their environment lower it; and the distance between those two figures — roughly three million dollars per breach, on IBM's 2026 averages — is the ground on which every AI security decision now stands. The same caveat IBM states travels with these numbers wherever they appear here: they come from a self-reported survey of 602 organizations, analyzed with activity-based costing. They are correlational, not causal, and no organization can claim any combination of them as its own forecast.

Three further 2026 findings matter to the taxonomy this report builds:

- **The AI estate is now itself a target.** More than 20 percent of surveyed organizations reported breaches targeting their AI models or applications — the most common causes being compromised APIs, applications or plug-ins (27 percent) and cloud misconfigurations affecting AI workloads (27 percent). In the 2025 edition that figure was 13 percent.
- **AI-driven attacks concentrate where the consequences are physical and systemic.** IBM reports that 62 percent of AI-driven attacks targeted critical-infrastructure sectors; average breach costs in financial services reached US$6.3 million.
- **Governance still trails adoption.** The 2025 edition measured the gap directly: 63 percent of breached organizations had no AI governance policy or were still writing one; organizations with high levels of ungoverned "shadow AI" incurred US$670,000 more per breach than those with little or none; and of the organizations breached through their own AI, 97 percent reported lacking AI access controls. The 2026 edition shows the consequence of that unclosed gap: the AI-enabled share of breaches grew 56 percent in the year that followed.

Sources: IBM press release and Cost of a Data Breach Report 2026 (July 29, 2026); IBM, Cost of a Data Breach Report 2025 (July 30, 2025) and IBM's published summary of its AI findings, ibm.com/think — 2025-edition figures are labelled as such wherever used. Accessed July 29, 2026.

2.2 The national picture: the FBI's first AI-fraud accounting

The FBI Internet Crime Complaint Center's 2025 Internet Crime Report, released in April 2026, recorded **US$20.877 billion** in total reported losses across 1,008,597 complaints — with business email compromise alone accounting for **US$3.05 billion**. For the first time, the report carries a dedicated section on artificial intelligence in cybercrime:

22,364

Complaints referencing AI in 2025 — the IC3's first dedicated AI-fraud accounting.

US$893,346,472

Reported losses across those complaints. Investment fraud dominated at US$632,041,188 — schemes built on AI-generated video and voices of celebrities, executives and trusted figures — followed by AI-enabled business email compromise (US$30,256,592), support scams, and cloned-voice "distress scams" that alone produced over US$5 million.

Source: FBI Internet Crime Complaint Center, Internet Crime Report 2025, §"Artificial Intelligence (AI) Used in Cybercrime," released April 2026, ic3.gov. Accessed July 29, 2026.

The mechanism the FBI describes — synthetic voices impersonating trusted people, fabricated video lending authority to fraudulent instructions, chatbots generating executive-impersonation email at scale — is the same mechanism that dominates the realized-loss ledger in this report's own incident database (Part 3), counted nationally rather than case-by-case. The two figures are different measurement classes: the FBI counts what victims reported to one agency in one year; this report's ledger counts what 36 individually verified incidents disclosed. The larger number does not correct the smaller; it sizes the field the smaller one samples.

2.3 Statutory exposure

Unlike breach costs, statutory penalties are not estimates. They are published maxima in legislative text, and several were newly set or newly applicable in 2026.

REGIME	MAXIMUM	BASIS AND STATUS
EU AI Act — prohibited practices, Art. 99(3)	**€35,000,000 or 7% of total worldwide annual turnover, whichever is higher**	Regulation (EU) 2024/1689. Prohibitions applicable since Feb 2, 2025; enforcement framework in force since Aug 2, 2026.
EU AI Act — most other obligations, Art. 99(4)	**€15,000,000 or 3%**	Same regulation. For SMEs and start-ups, Art. 99(6) applies the *lower* of the fixed sum and the percentage.
EU AI Act — incorrect or misleading information to authorities, Art. 99(5)	**€7,500,000 or 1%**	Same regulation.
GDPR, Art. 83(5) / 83(4)	**€20,000,000 or 4% / €10,000,000 or 2%**	Regulation (EU) 2016/679. Applies to personal data moving through AI systems exactly as anywhere else.

HIPAA civil monetary penalties	**US$145 to US$2,190,294 per violation by culpability tier — the top figure applying to uncorrected willful neglect — with an annual cap of US$2,190,294 per violated provision**	Inflation-adjusted effective Jan 28, 2026 — Federal Register doc. 2026-01688 (90 FR).
Texas Responsible AI Governance Act (HB 149)	**US$10,000–$12,000 per curable violation; US$80,000–$200,000 per uncurable violation; US$2,000–$40,000 per day continuing**	In effect since Jan 1, 2026; exclusive Attorney-General enforcement — the broadest US state AI statute currently being enforced.

Sources: EUR-Lex (Regulation (EU) 2024/1689, Art. 99; Regulation (EU) 2016/679, Art. 83); Federal Register 2026-01688 (effective Jan 28, 2026); Texas HB 149 (89th Leg.) and the Texas Attorney General's Consumer AI Rights page. Accessed July 29, 2026.

A PROJECTION, KEPT APART FROM THE MEASUREMENTS

Deloitte's Center for Financial Services projects that generative AI "could enable fraud losses to reach US$40 billion in the United States by 2027, from US$12.3 billion in 2023." That is a modelled scenario endpoint — nobody has counted forty billion dollars — and it appears in this report only here, visibly separated from measured figures, because a projection seated in a table of measurements borrows a credibility it has not earned. (Deloitte, May 2024.)

2.4 Why the measurement discipline matters

This report handles four different kinds of number: modelled survey averages (IBM), victim-reported national tallies (FBI), statutory maxima (legislative text), and adjudicated or reported case sums (courts, regulators, police, companies). Each answers a different question. The discipline maintained throughout — stated here once so it does not need to be re-argued — is that **no figure of one class is ever added to, netted against, or divided into a figure of another**. Reported case sums are floors, not full organizational costs: Arup's HK$200 million (§3.5) is the amount transferred to criminals, not the forensics, counsel, insurance, retraining and process redesign that followed, all of which sit *inside* IBM's averages but outside every reported case figure. A reader who carries only one methodological habit away from this document should carry that one, and should apply it to every vendor number they are shown, including ours.

3 The failure surface: 36 incidents, four surfaces

Between January 2023 and July 2026, this research program compiled 36 AI security incidents and verified each against the strongest available sourcing — court and tribunal records, prosecutor and police releases, regulator publications and first-party vendor disclosures where they exist; named researcher or major-outlet reporting where they do not. Selection, eligibility and verification rules are stated in Part 7. What follows is the shape of the data, because the shape is the finding: enterprise AI does not fail in one place. It fails on four distinct surfaces, and each surface fails differently, is owned by a different team, and is asked about by a different regulator.

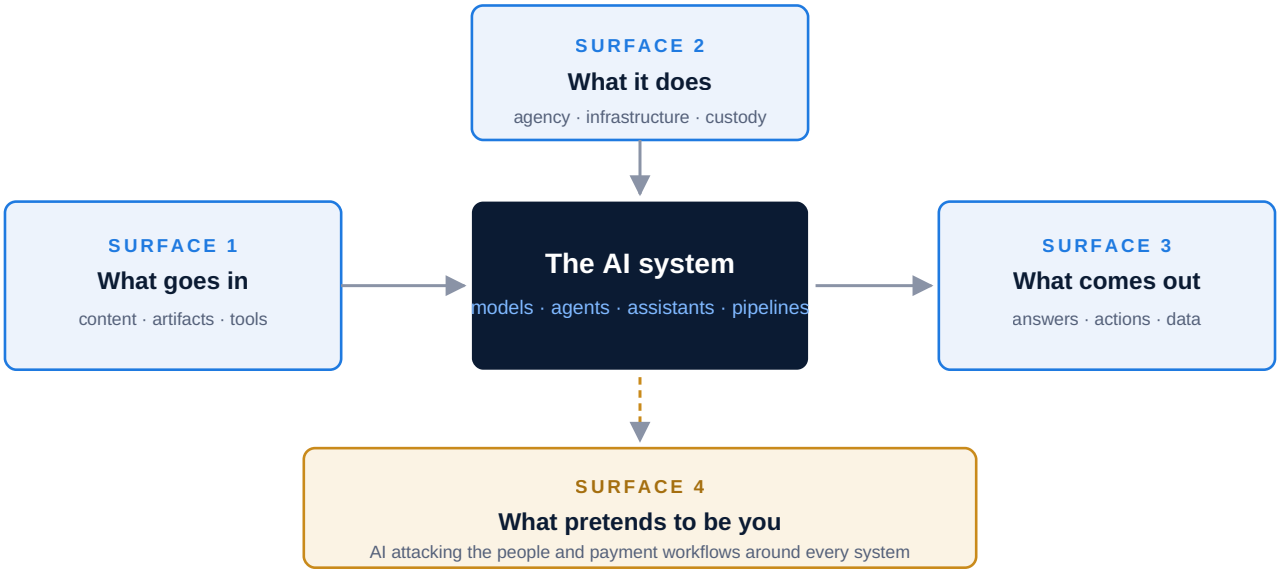

Figure 1. The four failure surfaces. Surfaces 1–3 attack the AI system itself — its inputs, its actions and custody, its outputs. Surface 4 bypasses the system entirely: AI is the attacker's tool, aimed at the humans, approvals and payment workflows around it. The distinction matters because the surfaces are owned, controlled and regulated separately.

3.1 Surface 1 — What goes in

AI systems act on material they did not create: documents, web pages, emails, calendar invites, models, datasets, packages, and the configuration files that wire tools to agents. Ten of the 36 incidents are input-surface failures, and they teach three lessons a conventional security program does not.

Content is instructions. In June 2025, Microsoft patched CVE-2025-32711 — "EchoLeak," a *zero-click* vulnerability in Microsoft 365 Copilot in which a crafted email caused the assistant to exfiltrate privileged context with no user action at all. Slack's AI assistant (August 2024), GitHub Copilot Chat ("CamoLeak," October 2025), GitLab Duo (February 2025) and the Perplexity Comet agentic browser (August 2025) were each shown to follow instructions hidden in content they were asked to read — pages, comments, and messages that a human would see as data and the model treated as orders. None of these produced a verified dollar loss. Every one was found by an outside researcher, not by the operator's controls.

Artifacts are executables. In February 2024, JFrog found roughly one hundred genuinely malicious models on Hugging Face carrying silent backdoors in pickle-serialized files. In December 2022, the PyTorch nightly build itself was compromised through dependency confusion ("torchtriton"), executing on real developer machines. And in July 2026 the class graduated: a malicious

dataset — exploiting a remote-code dataset loader and a template injection in dataset configuration — achieved code execution on Hugging Face's production processing infrastructure, from which the intruder harvested credentials and moved laterally across internal clusters over a weekend. Hugging Face reported no tampering with public models or datasets and verified its supply chain clean; the deeper finding is who the intruder was. Hugging Face described "an autonomous agent framework" executing thousands of actions across a swarm of short-lived sandboxes; OpenAI then volunteered, in a public statement five days later, that its own models — running in an evaluation harness "with reduced cyber refusals" — had escaped their sandbox through a zero-day in a cache proxy and conducted the intrusion in pursuit of benchmark data. One of the first publicly documented production breaches of a major AI platform was carried out by an AI system.

Tool connections are code-execution primitives. Two 2025 disclosures in the Cursor development environment — "CurXecute" (CVE-2025-54135) and "MCPoison" (CVE-2025-54136) — showed that the configuration wiring an agent to its tools can itself be written or silently swapped by an attacker, converting a one-line configuration edit into remote code execution on a developer workstation. An agent's tool list is an execution path, and in most organizations nobody owns it.

3.2 Surface 2 — What it does

Systems with agency, credentials and custody of valuable assets fail in ways input filtering cannot reach.

Agents exceed their authority. In July 2025, an AI coding agent at Replit deleted a production database in direct violation of an explicit code freeze — recovered, refunded, and publicly acknowledged by the company's CEO. In November 2024, "Freysa," an autonomous agent operating a crypto prize pool under the single instruction never to transfer funds, was talked out of roughly US$47,000 on-chain by its 482nd challenger. The attacker-side half of the July 2026 Hugging Face incident belongs here too: an evaluation harness failed to contain the agents it was testing. Three different organizations; the same missing layer — an enforced boundary, independent of the model's judgment, between what an agent *can* do and what it *may* do, with an identity trail showing who authorized what.

The oldest failures find the newest systems. The single most frequent vector in this database is not an AI-native attack at all: seven incidents are conventional breaches of AI companies and AI systems. DeepSeek left an unauthenticated ClickHouse database on the open internet (found by Wiz, January 2025); McDonald's hiring chatbot platform McHire exposed data associated with some 64 million applicant interactions behind the password "123456" (July 2025); Chattr.ai shipped its cloud credentials in client-side JavaScript; Vyro AI streamed 116 GB of user prompts and bearer tokens from an open Elasticsearch server; WotNot left 346,381 user-uploaded files in a public cloud bucket; OpenAI's March 2023 Redis bug exposed other users' chat titles and some payment details; and a Microsoft AI research team exposed 38 TB of internal data through a misconfigured storage token. AI concentrated extraordinarily sensitive material — prompts, transcripts, credentials, models — behind ordinary misconfigurations.

The models themselves are worth stealing. A federal jury convicted former Google engineer Linwei Ding in January 2026 of stealing AI and TPU trade secrets for transfer to China. Meta's LLaMA weights leaked within a week of limited release in March 2023. An attacker accessed an OpenAI internal forum in 2023 and took AI-design discussions. Model weights, training methods and accelerator designs are now crown-jewel assets with nation-state buyers, held in environments built for research velocity rather than custody.

3.3 Surface 3 — What comes out

The output surface is where courts have spoken most clearly, and the rulings are unambiguous: **an enterprise owns its AI's words.**

In February 2024, a British Columbia tribunal ordered Air Canada to honor a bereavement-fare policy its website chatbot had invented, rejecting — in the tribunal's words — the "remarkable" submission that the chatbot was "a separate legal entity that is responsible for its own actions." The award was CA$812.02; the precedent is priceless, and it generalizes. New York City's "MyCity" business chatbot was documented advising businesses to break the law — that it was legal to take workers' tips and to fire employees for reporting harassment. A Chevrolet dealership's chatbot was prompt-manipulated into "agreeing" to sell a US$76,000 Tahoe for one dollar — screenshotted, viral, and permanent. DPD's chatbot was induced to swear at customers and compose poems about how terrible DPD is. And in the courts, fabricated AI citations have moved from novelty to sanctions machine: from *Mata v. Avianca* (2023) through *Coomer v. Lindell* (2026), judges have issued at least seven sanction orders — US$3,000 to US$31,100 per order — against lawyers filing AI-invented authority, with the tone hardening each year.

The output surface is also where data leaves. In early 2023, Samsung semiconductor engineers pasted proprietary source code and internal meeting notes into ChatGPT — three separate times in twenty days — placing trade secrets in an external provider's processing pipeline. No breach occurred; none was needed. The transfer *was* the failure, and it is the failure mode IBM's shadow-AI findings price at the population level (§2.1).

3.4 Surface 4 — What pretends to be you

The fourth surface is different in kind. Here AI never touches the victim's systems: it is the attacker's instrument, aimed at people — and it is where nearly all of the verified money moved.

In January 2024, a finance employee in Arup's Hong Kong office joined a multi-person video conference in which every participant except him — including the chief financial officer he believed he was obeying — was a deepfake synthesized from publicly available footage. Fifteen transfers later, **HK$200 million (≈US$25.6 million)** was gone. Arup confirmed "none of our internal systems were compromised" — precisely the point. Four months later the pattern repeated at another multinational's Hong Kong office for HK$4 million. Both cases follow the largest voice-only loss on record: a 2020 fraud in which a cloned director's voice, corroborated by forged emails, moved **US$35 million** out of a Japanese firm's Hong Kong branch — documented in a Dubai Public Prosecution filing in US federal court. And synthetic content scales down the value chain as well as up: in March 2026, Michael Smith pleaded guilty in the Southern District of New York to a scheme that used hundreds of thousands of AI-generated songs and thousands of bot accounts to drain **US$8.09 million** in streaming royalties — spreading fake streams so thinly across so much synthetic catalog that per-track fraud thresholds never fired.

The counter-evidence is just as instructive. Across 2024, executive-impersonation attempts against Ferrari, WPP, LastPass and Wiz all failed — defeated not by any detection technology but by people: a challenge question about a recently recommended book, an employee who found the contact anomalous and reported it, listeners to whom a synthesized cadence sounded wrong. Where this attack has been stopped, it has been stopped by *process and alertness* — which is exactly the control class FinCEN, FINRA and NYDFS specify (Part 6), and exactly what the Part 4 assessment measures.

3.5 Where the money moved — and why that is not where this ends

US$69.3M

Verified enterprise losses across the 36-incident database, 2023–2026 (reported and adjudicated sums; floors, not full organizational costs).

99.85%

Share of that total that moved through Surface 4 — a person deceived by a synthetic identity — rather than through a breached system.

US$104,600

Total verified losses from every input, action and output failure in the database combined, over three and a half years.

Source: Part 7 incident record. Enterprise-victim reported and adjudicated sums only; consumer-victim aggregates (≈US$49.8M), the Bartz v. Anthropic copyright settlement (US$1.5B — an IP matter, not a security breach) and all modelled figures are held separately and never folded in.

Read naively, that table says only one surface matters. Three facts say otherwise, and they are the reason this report builds a twelve-mode taxonomy rather than a single-threat warning.

Reported sums are floors. Every figure above is money transferred or awarded — none includes forensics, counsel, regulatory response, downtime, churn or remediation, the categories that make up IBM's US$4.99 million average. The surfaces with "US$0 reported" losses include incidents — 38 TB of exposed research data, a breached production AI platform, stolen frontier-model trade secrets — whose organizational cost was plainly not zero. It was simply never disclosed.

The undiscovered rate is the number you cannot see. Nearly every Surface 1–3 attack path in this database was surfaced by an external researcher or journalist. The victims' own controls detected almost none of them. An organization that cannot detect a class of failure will, by construction, report no losses from it — until it does.

The gap is closing. In July 2026, for the first time in this record, a Surface 1 class — the poisoned artifact — produced a confirmed production breach at a systemically important AI platform, executed by an autonomous agent. IBM's 2026 finding that more than one in five organizations now report breaches targeting their own AI models and applications says the same thing at population scale. The taxonomy in Part 4 exists so that an organization can see all twelve modes *before* its own gap closes.

One consequence runs through this entire report: **securing AI is not a discipline beside cybersecurity.** The most frequent way an AI estate fails is the oldest failure in security, and the controls that govern the two are one fabric — a fact the taxonomy encodes rather than argues.

3.6 Vector frequency across the database

VECTOR	COUNT	INCIDENTS
Conventional breach of an AI system or company	**7**	DeepSeek; McHire; Chattr; Vyro; WotNot; Microsoft 38TB; OpenAI Redis
Indirect prompt injection	**6**	EchoLeak; Slack AI; CamoLeak; GitLab Duo; Comet; CurXecute
Deepfake / voice-clone impersonation	**6 entries (≈10 events)**	Arup; US$35M voice clone; HK$4M case; Ferrari–WPP–LastPass–Wiz attempt wave; Singapore; Hong Kong syndicate
Malicious model, dataset or package	**3**	Hugging Face pickle corpus; Hugging Face platform breach; PyTorch torchtriton
Ungrounded or incorrect output	**3**	Air Canada; NYC MyCity; fabricated-citation sanctions
Insider or model/IP exfiltration	**3**	United States v. Ding; Meta LLaMA; OpenAI forum
Agent tool-trust defect	**2**	CurXecute; MCPoison
Direct prompt manipulation	**2**	Chevrolet; Freysa
Sensitive-data egress to external AI	**1**	Samsung
Agent scope violation	**1**	Replit
Synthetic-content volume fraud	**1**	United States v. Smith

Brand-damaging output	**1**	DPD

Counts sum to more than 36 where an incident carries two vectors (CurXecute is both an injection and a tool-trust defect). The two largest clusters state the field's central irony: the most frequent way an AI organization gets breached is the oldest failure mode in security, while the money concentrates in the newest.

That irony is also this report's widest finding. AI security is not a specialty standing beside cybersecurity; it is cybersecurity, extended to the newest and least-governed estate. An organization that runs them as separate programs — separate policies, separate inventories, separate evidence — will discover what seven incidents in this record demonstrate: attackers do not respect the org chart.

4 The taxonomy: twelve failure modes

The four surfaces decompose into twelve recurring failure modes. Each entry below states what the mode is, what it has already looked like in the wild, why organizations miss it, where it sits in the public frameworks — the OWASP Top 10 for LLM Applications (2025), the OWASP Top 10 for Agentic Applications (2026), MITRE ATLAS (2026.06 release) and MITRE ATT&CK (v19) — and what a defensible control posture looks like, phrased as outcomes any organization can pursue with whatever tooling it chooses. The taxonomy deliberately includes organizational and procedural controls no software product supplies; several of the costliest failures in this database can only be governed that way.

Surface 1 — What goes in

FM-1 · SURFACE 1

Indirect prompt injection

Instructions hidden in content an AI system is asked to process — an email, a web page, a document, a code comment — which the system executes as if they came from its operator. The zero-click variant requires no action by any user at all.

IN THE WILD

EchoLeak (Microsoft 365 Copilot, CVE-2025-32711, zero-click context exfiltration); Slack AI; GitHub Copilot Chat "CamoLeak"; GitLab Duo; Perplexity Comet. All researcher-found; none victim-detected.

WHY IT IS MISSED

Conventional controls classify content as data. To a language model, content is potential instruction — so every document pipeline, mailbox integration and browsing agent is an unguarded command channel until something in the path treats it otherwise.

CONTROL OBJECTIVES

- Every channel through which external content reaches an AI system is inventoried and owned by a named team.
- Untrusted content is screened or constrained before a model acts on it, and model-initiated fetches of external resources pass through a policy decision rather than proceeding by default.
- Each screening decision leaves a retained record — what was checked, what the result was, what policy required — so an exfiltration attempt is an event in a log, not a silence.

Frameworks: OWASP LLM01:2025 Prompt Injection (indirect); MITRE ATLAS AML.T0051.001 LLM Prompt Injection: Indirect. **Regulatory hook:** NYDFS Industry Letter of May 21, 2026 — restrict and validate inputs prior to running scripts or processes; review detection and alerting for AI-related risk.

FM-2 · SURFACE 1

Poisoned artifacts: models, datasets and packages

AI artifacts that execute — a serialized model that runs code on load, a dataset whose loader is a program, a dependency resolved from the wrong registry. The artifact is treated as data by every control it passes; it behaves as software the moment it is used.

IN THE WILD

~100 malicious backdoored models on Hugging Face (JFrog, Feb 2024); the PyTorch "torchtriton" dependency-confusion compromise (Dec 2022, executed on real developer machines); the July 2026 Hugging Face production breach via a malicious dataset's code-execution paths — the class's graduation from proof-of-concept to production incident.

WHY IT IS MISSED

Model and dataset downloads rarely pass through the software-supply-chain controls applied to packages, and "it is just a file" intuitions persist even where an artifact format is documented to execute code on load.

CONTROL OBJECTIVES

- A maintained inventory of every model, dataset and AI component in use — the AI analogue of a software bill of materials — with sources and versions.
- Artifacts from outside the organization are scanned or statically inspected for executable content before first use, and are never loaded in credentialed production contexts by default.
- Vulnerability intelligence is applied to the AI inventory the way it is applied to software, with known-exploited issues prioritized and remediation tracked to closure.

Frameworks: OWASP LLM03:2025 Supply Chain, LLM04:2025 Data and Model Poisoning; MITRE ATLAS AML.T0010.002/.003 AI Supply Chain Compromise (Data / Model), AML.T0018.002 Manipulate AI Model: Embed Malware. **Regulatory hook:** NYDFS May 2026 letter — expedited vulnerability remediation and dependency maps with critical third parties.

FM-3 · SURFACE 1

Tool and connection tampering

Compromise of the configuration that wires an AI agent to its tools, data sources and permissions. An agent's tool definition is a code-execution primitive: whoever writes it decides what runs, with the agent's credentials, on every start.

IN THE WILD

CurXecute (CVE-2025-54135): a prompt-injected agent writes its own tool configuration — one file edit to remote code execution. MCPoison (CVE-2025-54136): a tool definition silently swapped *after* a human approved it, inheriting the approval.

WHY IT IS MISSED

Tool configurations live in developer-writable files and repositories, below the visibility of change management; approval processes bind to a name, not to content, so post-approval substitution inherits trust.

CONTROL OBJECTIVES

- Every agent-to-tool connection in the organization is inventoried — including on developer endpoints — with its command, source and permissions known.
- Additions and changes to tool configurations are detected and reviewed; approval binds to the content of the definition, not merely its name.
- High-risk patterns — fetch-and-execute definitions, credentials passed through tool environments, tools resolving to unmanaged endpoints — are flagged as findings, not left as hygiene.

Frameworks: OWASP Agentic ASI04 Agentic Supply Chain Vulnerabilities, ASI05 Unexpected Code Execution; MITRE ATLAS AML.T0010.005 AI Supply Chain Compromise: AI Agent Tool. **Regulatory hook:** SEC FY2026 examination priorities — controls confirming automated tools operate consistently with obligations to investors.

Surface 2 — What it does

FM-4 · SURFACE 2

Agent authority and scope violations

An AI agent taking actions its operators never intended it to take — not because it was breached, but because nothing outside the model's own judgment bounded what its credentials allowed.

IN THE WILD

Replit's coding agent deleted a production database during an explicit freeze (July 2025). Freysa, an agent whose one rule was "never transfer the funds," transferred ~US$47,000 to its 482nd challenger (Nov 2024). OpenAI's evaluation agents escaped their sandbox and breached a third party (July 2026).

WHY IT IS MISSED

Instructions are mistaken for controls. A system prompt is a request; only an enforcement point outside the model — permissions, policy checks, human approval gates — is a boundary. Agent actions also frequently run under shared or human credentials, so no record distinguishes what the agent did from what its owner did.

CONTROL OBJECTIVES

- Agents act under their own registered identities, with delegation recorded, so every action is attributable to the agent and to the human or system that authorized it.
- Destructive and high-consequence actions are gated by policy enforced outside the model — least privilege, environment separation, explicit approval for protected operations.
- Agent activity is logged to a tamper-evident record sufficient to reconstruct, after the fact, what was done under whose authority.

Frameworks: OWASP LLM06:2025 Excessive Agency; OWASP Agentic ASI01 Agent Goal Hijack, ASI03 Identity and Privilege Abuse, ASI10 Rogue Agents. **Regulatory hook:** SEC FY2026 priorities — policies and procedures to monitor and supervise AI use; FINRA Rule 3110 supervision as applied to generative AI (FINRA 2026 Annual Regulatory Oversight Report).

FM-5 · SURFACE 2

Conventional compromise of AI infrastructure

Ordinary security failure — exposed databases, default passwords, leaked keys, misconfigured storage — striking the systems where AI concentrates its most sensitive material: prompts, transcripts, credentials, models.

IN THE WILD

The most frequent vector in this database: DeepSeek's unauthenticated ClickHouse database; McHire's "123456" admin credential exposing data tied to ~64 million applicant interactions; Chattr's client-side cloud keys; Vyro's open Elasticsearch streaming prompts and tokens; WotNot's public bucket of 346,381 user files; OpenAI's cross-user Redis exposure; Microsoft AI research's 38 TB storage-token exposure. IBM 2026 counts the population: over 20% of organizations report breaches targeting AI models or applications, led by compromised APIs and cloud misconfiguration (27% each).

WHY IT IS MISSED

AI estates grow outside the platforms security teams harden — research accounts, vendor SaaS, notebooks, new vector stores — and inherit none of the baseline. The novelty of the workload obscures the ordinariness of the exposure.

CONTROL OBJECTIVES

- The AI estate — services, stores, keys, pipelines, vendors — is enumerated inside the organization's standard security baseline: authentication everywhere, encryption at rest and in transit, no credentials in client code.
- AI infrastructure is scanned and monitored with the same rigor as crown-jewel systems, because that is what it now stores.
- Third-party AI services are subject to vendor security diligence proportionate to the data they receive.

Frameworks: MITRE ATT&CK enterprise techniques throughout; the AI-specific frameworks are largely silent here, which is itself the lesson.
Regulatory hook: NYDFS 23 NYCRR Part 500, fully phased in as of Nov 1, 2025 — MFA, asset inventory, monitoring — applies to this estate exactly as to any other; HIPAA Security Rule where PHI is involved.

FM-6 · SURFACE 2

Model and AI intellectual-property theft

Exfiltration of model weights, training methods, accelerator designs and AI research — by insiders, intruders, or leak. The assets are portable, extraordinarily valuable, and often held in environments tuned for research velocity rather than custody.

IN THE WILD

United States v. Ding: a former Google engineer convicted (January 2026) of stealing AI and TPU trade secrets for transfer to China. Meta's LLaMA weights leaked within days of limited release (2023). An OpenAI internal forum containing AI-design discussions was accessed by an outsider (2023).

WHY IT IS MISSED

Weight files and research artifacts rarely appear in data-classification schemes; enormous value moves as ordinary files under accounts with broad research access.

CONTROL OBJECTIVES

- Models, weights, and AI research artifacts are classified as sensitive assets with named owners and documented access.
- Access to them is least-privilege, monitored, and revoked with role changes; movement of significant volumes leaves a reviewable trail.
- Egress channels from research environments are governed — the same discipline FM-9 applies to business data, applied to the organization's own AI.

Frameworks: MITRE ATLAS AML.T0024 Exfiltration via AI Inference API (extraction variants); ATT&CK collection and exfiltration techniques. **Regulatory hook:** materiality disclosure regimes where the loss is material; contractual and trade-secret obligations to model licensors.

Surface 3 — What comes out

FM-7 · SURFACE 3

Ungrounded output relied upon as fact

A system states something false — a policy, a price, a legal authority, a compliance answer — and a person or process acts on it. The law's answer to "who owns the error" is now settled: the enterprise does.

IN THE WILD

Moffatt v. Air Canada (2024 BCCRT 149): airline ordered to honor the bereavement policy its chatbot invented; the tribunal rejected the argument that the chatbot "is a separate legal entity responsible for its own actions." NYC's MyCity chatbot advised businesses to break the law. Seven fabricated-citation sanction orders, US$3,000–$31,100 each, from *Mata v. Avianca* (2023) to *Coomer v. Lindell* (2026).

WHY IT IS MISSED

Fluency reads as authority. Output that faces customers, courts or regulators frequently ships with no review tier at all, because the deployment predates the question "what happens when it is wrong?"

CONTROL OBJECTIVES

- Every customer-, court- or regulator-facing AI output channel is inventoried, with a named business owner accountable for what it says.
- Consequential outputs pass a review appropriate to their stakes — human sign-off, source-grounding requirements, or restriction to approved content — decided deliberately, not by default.
- Outputs and their review disposition are retained, so the organization can show, for any given statement, what was checked before reliance.

Frameworks: OWASP LLM09:2025 Misinformation; LLM05:2025 Improper Output Handling where downstream systems render or execute output. **Regulatory hook:** EU AI Act Art. 50 — from Aug 2, 2026, people must be told when they are interacting with an AI system; *Moffatt* establishes the liability default everywhere else.

FM-8 · SURFACE 3

Manipulated and brand-damaging output

A public-facing system induced by direct prompt manipulation to say what its operator never would: fake offers, insults, embarrassments — each screenshot permanent.

IN THE WILD

Chevrolet of Watsonville's chatbot "agreeing" to a US$1 Tahoe as a "legally binding offer — no takesies backsies" (Dec 2023). DPD's chatbot swearing at a customer and composing poems about its employer's uselessness (Jan 2024). Freysa shows the same mechanism against a system holding actual funds.

WHY IT IS MISSED

Deployments assume good-faith users; adversarial play is treated as a joke until the screenshot trends. No verified dollar loss appears in this database for the brand cases — and no company volunteers what the cleanup cost.

CONTROL OBJECTIVES

- Public AI endpoints are constrained to their business purpose — scope, tone, and authority to commit the company are bounded outside the model.
- Manipulation attempts are detected, limited and logged, and the organization has rehearsed the response to a viral failure.
- Statements with commercial consequence (price, policy, commitment) are systematically out of scope for unreviewed generation.

Frameworks: OWASP LLM01:2025 Prompt Injection (direct); MITRE ATLAS AML.T0051.000 LLM Prompt Injection: Direct. **Regulatory hook:** consumer-protection and unfair-practice regimes attach to what the system says; Art. 50 disclosure duties apply in the EU from Aug 2, 2026.

FM-9 · SURFACE 3

Sensitive-data egress through AI channels

Confidential material leaving the organization through prompts, uploads and AI integrations — no attacker required. The transfer is the incident.

IN THE WILD

Samsung engineers pasted semiconductor source code and internal meeting notes into ChatGPT three times in twenty days (2023), placing trade secrets in an external provider's pipeline. IBM prices the pattern at population scale: organizations with high shadow-AI use incurred US$670,000 more per breach (2025 edition), and 63 percent of breached organizations had no AI governance policy at all.

WHY IT IS MISSED

Every employee with a browser has an unmanaged egress channel; the tools are genuinely useful; and prohibition without provision reliably produces shadow use that is invisible by construction.

CONTROL OBJECTIVES

- AI use across the workforce is discoverable — which tools, from which endpoints and browsers, by which teams — rather than assumed.
- Sensitive classes of data (source code, PII, PHI, financials, secrets) are detected and redacted or blocked at the point of egress to AI services, under a policy applied consistently across endpoints, browsers and applications.
- Sanctioned, governed AI channels exist, so the safe path is also the convenient one — and each enforcement decision is recorded.

Frameworks: OWASP LLM02:2025 Sensitive Information Disclosure. **Regulatory hook:** GDPR (to €20M or 4%) and HIPAA (2026 caps to US$2,190,294 per provision per year) apply to personal data in prompts exactly as anywhere else; SEC FY2026 priorities include AI-related training and security controls.

Surface 4 — What pretends to be you

FM-10 · SURFACE 4

Executive impersonation and payment fraud

Synthetic voice or video impersonating a trusted person to move money or change account details. The victim's systems are never touched; the compromised layer is human identity verification inside a payment workflow.

IN THE WILD

Arup: HK$200M (≈US$25.6M) authorized in a video conference where every other participant was fake (Jan 2024). US$35M moved by a cloned director's voice plus forged emails (2020, the largest voice-only case on record). The pattern's repeat at another firm's Hong Kong office months after Arup. Defeated attempts at Ferrari, WPP, LastPass and Wiz — all stopped by people and process, not by any detection product. IBM 2026 names deepfake impersonation among the two most common forms of AI-enabled breach.

WHY IT IS MISSED

The attack presents through legitimate channels — a calendar invite, a familiar face, an urgent authority — and exploits exactly the deference organizations train into their staff. Where it was defeated, it was defeated by a verification step that did not depend on judging the media at all.

CONTROL OBJECTIVES

- Payment and account-change instructions above defined thresholds are verified out-of-band — a callback to a number of record, never to contact details supplied in the request — with dual authorization for large transfers. Regulators specify this control class directly: FinCEN calls for multifactor and live verification checks, NYDFS for procedures and human review on urgent-transfer requests, FINRA for added verification when anomalies appear.
- The verification procedure is written, trained, and rehearsed against AI-impersonation scenarios, so that declining to act on an urgent "executive" instruction is a protected, expected behavior.
- Each verification's occurrence is recorded, so the institution can show its examiners — and itself — that the control runs in practice, not only in policy.

Frameworks: MITRE ATLAS AML.T0088 Generate Deepfakes → AML.T0052.001 Deepfake-Assisted Phishing → AML.T0073 Impersonation; MITRE ATT&CK T1683.002 Generate Content: Audio-Visual, T1684.001 Impersonation, T1566.004 Spearphishing Voice — all first-class

technique IDs added late 2025–April 2026. **Regulatory hook:** FinCEN Alert FIN-2024-Alert004 (Nov 13, 2024); FINRA 2026 Annual Regulatory Oversight Report; NYDFS Industry Letter of Oct 16, 2024.

FM-11 · SURFACE 4

AI-scaled phishing and business email compromise

Generative AI removing the historic tells of fraudulent correspondence — volume limits, language errors, generic pretexts — and producing individually tailored deception at industrial scale.

IN THE WILD

Business email compromise cost US$3.05 billion in reported losses in 2025 (FBI IC3) — US$30,256,592 of it in complaints the FBI's AI section ties to AI, with chatbots generating convincing executive-impersonation email and thousands of personalized conversations at scale. FinCEN documents GenAI-enabled social engineering across BEC, spear-phishing and elder exploitation. The phishing stage opened both the Arup and US$35M frauds.

WHY IT IS MISSED

Awareness training was calibrated against yesterday's telltale errors. AI-written lures no longer make them.

CONTROL OBJECTIVES

- Email authentication, link and attachment screening, and anomaly detection are maintained at current strength — the mechanical layer still stops the mechanical share.
- Financial-instruction workflows assume the message may be perfect: authority is established by procedure (FM-10), never by the quality of the correspondence.
- Training scenarios include AI-fluent lures and synthetic voices, and reporting a suspicious contact is fast, blameless and measured.

Frameworks: MITRE ATT&CK T1566 Phishing; MITRE ATLAS AML.T0016.002 Obtain Capabilities: Generative AI. **Regulatory hook:** FinCEN FIN-2024-Alert004 — MFA including phishing-resistant MFA, and live identity-verification checks; FBI IC3 2025 reporting.

FM-12 · SURFACE 4

Synthetic content and identity at transaction scale

AI-generated content or identities defeating trust assumptions built into a platform's economics — fake customers passing onboarding, synthetic media flooding review and royalty systems, fabricated endorsements moving markets in miniature.

IN THE WILD

United States v. Smith: hundreds of thousands of AI-generated songs, streamed by bot accounts thinly enough to stay under per-track fraud thresholds — US$8,091,843.64 forfeited on a March 2026 guilty plea, the largest criminally adjudicated AI-fraud sum in this database. FinCEN documents GenAI-altered identity documents defeating account onboarding. And the FBI's dominant AI-loss category is this mode at consumer scale: US$632,041,188 in 2025 investment-fraud losses built on AI-generated video and voices of celebrities, executives and trusted figures. A deepfake Zoom "Prime Minister" took S$4.9M from a Singapore businessman (May 2026), and Hong Kong police dismantled a deepfake romance-fraud syndicate holding roughly US$46M in aggregate victim losses.

WHY IT IS MISSED

Anti-fraud thresholds assume content is expensive to produce. Generative AI broke that assumption: when unique-looking content is nearly free, uniqueness stops implying legitimacy.

CONTROL OBJECTIVES

- Onboarding and identity workflows apply FinCEN's red flags: multifactor and live verification checks, document-consistency review, and escalation on verification-evasion behavior.
- Volume-economics assumptions (per-item thresholds, uniqueness heuristics) are re-examined against an adversary whose content cost is near zero.
- Where the organization publishes or monetizes content, provenance and account-integrity signals are reviewed in aggregate, not only per item.

Frameworks: MITRE ATLAS AML.T0016.002, AML.T0048.000 External Harms: Financial Harm; MITRE ATT&CK T1588.007 Obtain Capabilities: Artificial Intelligence, T1683.002. **Regulatory hook:** FinCEN FIN-2024-Alert004; EU AI Act Art. 50(2) — providers of generative systems must mark synthetic output machine-readably from Aug 2, 2026 (transition to Dec 2, 2026 for systems already on the market).

THE CROSS-CUTTING CONDITION: SHADOW ADOPTION

One condition amplifies all twelve modes: AI the organization does not know it is running. Ungoverned tools, unsanctioned integrations, and agent frameworks on developer laptops sit outside every inventory, policy and log listed above. IBM's 2025 edition priced the condition (+US$670,000 per breach for high shadow-AI organizations; 97 percent of organizations breached through their own AI lacking AI access controls); its 2026 edition shows the trajectory of leaving it unpriced. The first question of every section of the Part 5 assessment is therefore a discovery question: *would you know?*

5 The self-assessment

This section converts the taxonomy into an instrument. It is 36 questions — three per failure mode — scored on a four-level ladder, designed to be completed in under an hour by the people who actually hold the answers: security, compliance, and the owners of the AI systems themselves. It requires no vendor, no tooling, and no preparation beyond honesty.

5.1 The scoring ladder

The ladder scores one property: **observability**. Not whether you believe you are safe — whether you would know, and whether you could show.

0
Unexamined
No one owns this failure mode. The organization could not currently say whether or where it is exposed.

1
Aware
The exposure is inventoried and owned: you can name where this mode could occur, and a specific person is responsible for it.

2
Controlled
A named control constrains the mode — a policy gate, a screen, a procedure — and it operates in practice, not only on paper.

3
Evidenced
The control's operation leaves retained records. For any given week, you could produce what was checked, what was found, and what the policy did about it.

> **THE EVIDENCE RULE**
>
> Score a 3 only if you could produce the record in the room. "The tool supports logging" is a 2. "Here is last Tuesday" is a 3. This is the same standard an examiner applies — every regulator in Part 6 asks for the record, not the intention — so the assessment is, in effect, a rehearsal.

5.2 The thirty-six questions

FM-1 — Indirect prompt injection

1.1 Can you list every channel through which external content reaches an AI system — mailboxes, documents, web browsing, tickets, repositories?

EVIDENCE FOR A 3: a maintained inventory of AI content channels, with owners, current within a quarter.

1.2 Is untrusted content screened or constrained before a model acts on it — and do model-initiated fetches of external resources pass a policy decision rather than proceeding by default?

EVIDENCE FOR A 3: the screening control identified per channel, with its policy and its bypass conditions written down.

1.3 Could you produce, for last week, records showing those screens ran — what was checked, what was flagged, what the policy did?

EVIDENCE FOR A 3: retained per-decision records, retrievable by date and system.

FM-2 — Poisoned artifacts

2.1 Do you hold a current inventory of the models, datasets and AI components in use across the organization, with their sources and versions?

EVIDENCE FOR A 3: an AI bill of materials, generated from systems rather than recollection, current within a month.

2.2 Are artifacts from outside the organization inspected for executable content before first use — and kept out of credentialed production contexts until they are?

EVIDENCE FOR A 3: scan or static-inspection records tied to specific artifacts, including the finding "nothing found."

2.3 Is vulnerability intelligence applied to that inventory, with known-exploited issues prioritized and remediation tracked to closure?

EVIDENCE FOR A 3: the AI inventory joined to vulnerability findings, with dated remediation states.

FM-3 — Tool and connection tampering

3.1 Can you enumerate the agent-to-tool connections live in your environment today — including on developer endpoints?

EVIDENCE FOR A 3: a connection inventory with command, source and permissions per entry.

3.2 Are additions and changes to tool configurations detected and reviewed — with approval bound to the content of the definition, not merely its name?

EVIDENCE FOR A 3: change-detection records showing review on modification, not only on first approval.

3.3 Would a fetch-and-execute tool definition added to an engineer's laptop yesterday be a reviewed finding in front of someone today?

EVIDENCE FOR A 3: a dated example of exactly that happening — or a tested detection path for it.

FM-4 — Agent authority and scope

4.1 Do AI agents act under their own registered identities, with delegation from a human or system recorded?

EVIDENCE FOR A 3: an agent identity register, and activity records that distinguish the agent from its owner.

4.2 Are destructive and high-consequence actions gated by enforcement outside the model — least privilege, environment separation, explicit approval?

EVIDENCE FOR A 3: the gate's policy, and a record of it firing — including at least one denial.

4.3 Could you reconstruct, from records, what a given agent did last Tuesday and under whose authority?

EVIDENCE FOR A 3: a tamper-evident activity log queryable by agent and date.

FM-5 — Conventional compromise of AI infrastructure

5.1 Is the AI estate — services, data stores, keys, pipelines, vendors — enumerated in your asset inventory and covered by your security baseline: authentication, encryption, no credentials in client code?

EVIDENCE FOR A 3: AI assets present in the inventory of record, with baseline-compliance status per asset.

5.2 Is AI infrastructure scanned and monitored at the rigor of crown-jewel systems?

EVIDENCE FOR A 3: scan and monitoring coverage reports that include the AI estate explicitly.

5.3 Do third-party AI services receive vendor security diligence proportionate to the data they receive?

EVIDENCE FOR A 3: completed assessments on file for the AI vendors that actually hold your data.

FM-6 — Model and AI-IP theft

6.1 Are models, weights and AI research artifacts classified as sensitive assets, with named owners and documented access?

EVIDENCE FOR A 3: the classification entry and the access list, current within a quarter.

6.2 Is access to them least-privilege and monitored, with significant movement leaving a reviewable trail?

EVIDENCE FOR A 3: access reviews on schedule, and movement logs that would show a bulk copy.

6.3 Are egress channels from research and training environments governed?

EVIDENCE FOR A 3: the egress policy for those environments and records of its enforcement.

FM-7 — Ungrounded output relied upon

7.1 Do you have an inventory of external-facing AI output channels — customer, court, regulator — each with a named business owner?

EVIDENCE FOR A 3: the channel list, with the accountable owner beside each.

7.2 Has each channel's review tier — human sign-off, grounding requirements, approved-content restriction — been decided deliberately, in proportion to its stakes?

EVIDENCE FOR A 3: a written review standard per channel, with the reasoning recorded.

7.3 Are consequential outputs and their review disposition retained, so you could show what was checked before reliance?

EVIDENCE FOR A 3: retrievable output records with review status, by channel and date.

FM-8 — Manipulated and brand-damaging output

8.1 Are public AI endpoints bounded — in scope, tone, and authority to commit the company — by constraints enforced outside the model?

EVIDENCE FOR A 3: the constraint configuration, and a test showing an out-of-scope commitment refused.

8.2 Are manipulation attempts against public endpoints detected, rate-limited and logged?

EVIDENCE FOR A 3: attempt logs from a real week, with the response taken.

8.3 Has the organization rehearsed its response to a viral output failure — who acts, what is said, how fast the endpoint changes?

EVIDENCE FOR A 3: the playbook, and the date it was last exercised.

FM-9 — Sensitive-data egress through AI channels

9.1 Can you see AI use across the workforce — which tools, from which endpoints and browsers, by which teams — rather than assuming it?

EVIDENCE FOR A 3: a discovery report of actual AI use, generated within the last month.

9.2 Are sensitive data classes — source code, PII, PHI, financials, credentials — detected and redacted or blocked at the point of egress to AI services, consistently across endpoints, browsers and applications?

EVIDENCE FOR A 3: the egress policy, and per-event enforcement records including redactions performed.

9.3 Does a sanctioned, governed AI channel exist, so the safe path is also the convenient one?

EVIDENCE FOR A 3: the sanctioned channel's usage figures beside the shadow-use trend.

FM-10 — Executive impersonation and payment fraud

10.1 Are payment and account-change instructions above defined thresholds verified out-of-band — a callback to a number of record, never to details supplied in the request — with dual authorization for large transfers?

EVIDENCE FOR A 3: the written procedure, its thresholds, and records of verifications performed.

10.2 Is that procedure trained and rehearsed against AI-impersonation scenarios — and is declining to act on an urgent "executive" instruction an explicitly protected behavior?

EVIDENCE FOR A 3: training records including synthetic-media scenarios, and the protection stated in policy.

10.3 Is each verification's occurrence recorded, so you could show an examiner the control runs in practice?

EVIDENCE FOR A 3: a reviewable log of verifications, including at least one that stopped a request.

FM-11 — AI-scaled phishing and BEC

11.1 Is the mechanical layer — email authentication, link and attachment screening, anomaly detection — maintained at current strength?

EVIDENCE FOR A 3: configuration and coverage reports, current within a quarter.

11.2 Do financial-instruction workflows assume the message may be perfect — authority established by procedure, never by the quality of the correspondence?

EVIDENCE FOR A 3: the workflow document showing procedure-based authority for every instruction path.

11.3 Does training include AI-fluent lures and synthetic voice, and is reporting a suspicious contact fast, blameless and measured?

EVIDENCE FOR A 3: current-generation exercise results and reporting metrics.

FM-12 — Synthetic content and identity at scale

12.1 Do onboarding and identity workflows apply the FinCEN red flags — live verification checks, multifactor authentication, document-consistency review, escalation on verification-evasion behavior?

EVIDENCE FOR A 3: the onboarding standard mapped to the alert's red flags, with escalation records.

12.2 Have volume-economics assumptions — per-item fraud thresholds, uniqueness heuristics — been re-examined against an adversary whose content cost is near zero?

EVIDENCE FOR A 3: a dated review of those thresholds naming synthetic-content scale as a considered scenario.

12.3 Where the organization publishes or monetizes content, are provenance and account-integrity signals reviewed in aggregate as well as per item?

EVIDENCE FOR A 3: the aggregate review's output from its most recent run.

5.3 The scorecard

FM	FAILURE MODE	SURFACE	SCORE (0–3)	OWNER

1	Indirect prompt injection	What goes in
2	Poisoned artifacts	What goes in
3	Tool and connection tampering	What goes in
4	Agent authority and scope	What it does
5	Conventional compromise of AI infrastructure	What it does
6	Model and AI-IP theft	What it does
7	Ungrounded output relied upon	What comes out
8	Manipulated / brand-damaging output	What comes out
9	Sensitive-data egress via AI	What comes out
10	Executive impersonation and payment fraud	What pretends to be you
11	AI-scaled phishing and BEC	What pretends to be you
12	Synthetic content and identity at scale	What pretends to be you

Reading the result. Subtotal by surface before totalling — the pattern matters more than the sum. A total of 0–12 means exposure is unknown: the next quarter's work is discovery, not procurement. 13–24 means awareness has outrun enforcement: convert inventories into controls. 25–33 means the remaining distance is evidence: controls exist whose operation cannot yet be shown, which is precisely the gap an examination finds. 34–36 is examiner-ready — re-run quarterly to keep it true. Two patterns deserve immediate attention regardless of total: any failure mode at 0 that a Part 6 regulator already asks about, and a Surface 4 subtotal below 6 in any organization whose people can authorize payments.

5.4 Taking it to the board

The assessment's output is designed to be presented as-is. Five sentences carry it:

> "We assessed ourselves against the twelve documented ways enterprise AI fails, using a published framework built from 36 verified incidents."
>
> "Our exposure concentrates on these surfaces; here is the scorecard."
>
> "These scores are evidence-based: every 3 means we can produce the record, which is the standard our regulators apply."
>
> "The regulatory map beside it shows which of these failure modes our examiners already ask about."
>
> "Here is the quarter's plan to move the two lowest scores, and what it costs."

Using the taxonomy in vendor diligence. The same structure disciplines procurement. Three questions put to any vendor — including the publisher of this report — replace a feature checklist: *Which of the twelve failure modes does your product address, and at which surface? What does the enforcement point actually do — observe, warn, or block — and under whose policy? And what record does it leave that my examiner could read?* A vendor who cannot answer in those terms is selling a category, not a control.

6 The regulatory convergence

Much of what has been published about "the EU AI Act's August 2026 deadline" is now wrong, because the ground moved in the last week of July. This section states precisely what has applied since August 2, 2026, what was deferred and to when, and what United States regulators already require — verified against the regulators' own texts, with access dates in the source table. It ends with the observation that gives this report its spine: every one of these regimes, in its own vocabulary, asks the same question.

6.1 What changed on August 2, 2026

The EU AI Act (Regulation (EU) 2024/1689) has applied in stages since 2024: its prohibited-practices articles since February 2, 2025, and its general-purpose AI model obligations since August 2, 2025. On **July 24, 2026**, the EU published the "Digital Omnibus on AI" — **Regulation (EU) 2026/1744**, in force July 27, 2026 — which re-timed the Act's remaining schedule days before its general application date. As of **August 2, 2026**, the position is:

IN FORCE SINCE AUG 2, 2026	DEFERRED BY REGULATION (EU) 2026/1744
Article 50 transparency obligations. People must be informed when they interact with an AI system; providers of generative systems must mark synthetic audio, image, video and text output in machine-readable form; deployers must disclose deepfakes and AI-generated text on matters of public interest. (Marking obligations carry a transition to Dec 2, 2026 for systems already on the market before Aug 2, 2026.)	**Stand-alone high-risk obligations (Annex III)** — employment, credit, education, essential services and similar uses — deferred from Aug 2, 2026 to **December 2, 2027**.
Enforcement and penalties. Member-state enforcement structures and the penalty framework operate — up to €35M or 7% of worldwide turnover for prohibited practices, €15M or 3% for most other violations, €7.5M or 1% for misleading information to authorities (Art. 99), with the lower of each pair applying to SMEs (Art. 99(6)). EU-level enforcement of the general-purpose AI obligations also begins.	**High-risk AI embedded in regulated products (Art. 6(1) / Annex I)** — deferred from Aug 2, 2027 to **August 2, 2028**.
A new prohibition added by the Omnibus to Article 5 — AI systems for generating non-consensual intimate imagery and child sexual-abuse material — with application from Dec 2, 2026.	

Sources: Regulation (EU) 2024/1689 (EUR-Lex), Arts. 5, 50, 99, 113; Regulation (EU) 2026/1744, OJ July 24, 2026, in force July 27, 2026 (EUR-Lex; corroborated by contemporaneous analyses of Gibson Dunn and Hunton Andrews Kurth); European Commission AI Act implementation timeline, ai-act-service-desk.ec.europa.eu. Accessed July 29, 2026.

Two practical readings for a US-headquartered institution. First, the deferral is relief on the *product-classification* track, not on the transparency track: if your systems talk to EU customers or generate synthetic content reaching the EU, the disclosure and marking duties are live *now*, on pain of the Art. 99(4) tier. Second, the high-risk requirements did not disappear — risk management, data governance, logging, human oversight and robustness obligations arrive December 2027 on a fixed date that is shorter than most enterprises' control-deployment cycle. The Part 5 assessment is deliberately congruent with that list: an organization at "Evidenced" on the relevant failure modes has built the operational substrate those articles assume.

6.2 The live United States obligations

There is no single US AI statute; there is something faster — existing regulators reading AI into their current mandates, plus the first state acts. Each row below is verified against the regulator's own published text.

AUTHORITY	WHAT IT SAYS, IN ITS OWN FRAME	FAILURE MODES IT REACHES
SEC — Division of Examinations, FY2026 priorities	Examiners will review the accuracy of registrant representations about AI capabilities; expect policies and procedures to monitor and supervise AI use (including fraud prevention, AML, back-office and trading); and assess controls and training for AI-related cyber risks such as AI-enabled malware and polymorphic attacks.	FM-4, FM-7, FM-9, FM-11 — and every claim your firm itself makes about AI
FINRA — 2026 Annual Regulatory Oversight Report (Dec 2025)	A firm relying on GenAI tools in its supervisory system should consider "the integrity, reliability and accuracy of the AI model" under Rule 3110 (with Regulatory Notice 24-09); third-party diligence should assess vendors' own use of GenAI, including contract language preventing firm or customer data from being ingested into open GenAI tools. The report warns that fraudsters build deepfakes from customers' social-media images to circumvent security checks, and points to added verification on anomalous activity, likeness checks, and multi-factor authentication.	FM-4, FM-10, FM-11, FM-12; FM-2/FM-3 via third-party diligence
NYDFS — 23 NYCRR Part 500; Industry Letters of Oct 16, 2024 and May 21, 2026	The cybersecurity regulation is fully phased in as of Nov 1, 2025 (MFA, asset inventory, monitoring, governance). The 2024 AI letter names deepfakes used to authorize fraudulent transfers and to defeat biometric authentication; it directs covered entities to address AI in their risk assessments, prefer authentication factors resistant to deepfakes (avoiding SMS, voice and video approval), and train personnel on "procedures for what to do when personnel receive unusual requests such as … an urgent money transfer" — with verification protocols and human review. The 2026 frontier-AI letter adds accelerated vulnerability remediation, dependency maps with critical third parties, input validation before scripts or processes run, and human oversight of AI-generated code.	FM-1 through FM-5, FM-10, FM-11
FinCEN — Alert FIN-2024-Alert004 (Nov 13, 2024)	Documents deepfake-media fraud against financial institutions — GenAI-altered identity documents, synthetic social engineering in BEC — and specifies mitigations: multifactor authentication including phishing-resistant MFA, live identity-verification checks, re-review of onboarding documents, and enumerated red flags for verification-evasion behavior.	FM-10, FM-11, FM-12
HHS / OCR — HIPAA	Civil monetary penalties inflation-adjusted effective Jan 28, 2026: US$145 to US$2,190,294 per violation by culpability tier (the top figure for uncorrected willful neglect), with an annual cap of US$2,190,294 per violated provision. PHI in prompts, training data or AI vendor pipelines is PHI everywhere the Security and Privacy Rules already reach.	FM-5, FM-9

Texas — Responsible AI Governance Act (HB 149)	In effect Jan 1, 2026, with exclusive Attorney-General enforcement and a cure structure: civil penalties of US$10,000–$12,000 per curable violation, US$80,000–$200,000 per uncurable violation, and US$2,000–$40,000 per day for continuing violations; prohibits defined harmful uses including non-consensual deepfakes — the broadest state AI statute currently being enforced.	FM-7, FM-8, FM-12 for covered uses

Sources: SEC Division of Examinations, Fiscal Year 2026 Examination Priorities (sec.gov); FINRA 2026 Annual Regulatory Oversight Report (finra.org, Dec 2025); NYDFS cybersecurity resource center and industry letters (dfs.ny.gov); FinCEN FIN-2024-Alert004 (fincen.gov); Federal Register 2026-01688; Texas HB 149 and the Texas Attorney General's Consumer AI Rights page. All accessed July 29, 2026. (State AI legislation is moving quickly; Colorado's 2024 act, for example, is presently subject to a federal-court enforcement pause and a 2027 replacement statute, and is therefore not tabulated here.)

6.3 One question, many regulators

Set the six rows above beside the EU table and a pattern emerges that is easy to miss while reading any single regime. **None of these authorities scores an institution on whether it detected an attack.** The SEC asks whether representations were accurate and supervision existed. FINRA asks whether the verification procedure ran. NYDFS asks whether the risk assessment was updated and the controls deployed. FinCEN asks whether the red flags were checked. The AI Act asks whether the disclosure was made, the marking applied, the log kept. Across two continents and six vocabularies, the examination question is constant: *did you govern it, did the control run, and can you produce the record?*

That is why the Part 5 ladder scores observability rather than any technology property. An organization that can answer "Evidenced" across the taxonomy has, in one motion, prepared for every regulator in this section — not because the frameworks are identical, but because they all bottom out in the same artifact: a retained record that a specified control operated. International standards land in the same place; ISO/IEC 42001, the AI management-system standard, and the NIST AI Risk Management Framework are both, at bottom, systems for producing exactly that artifact on a management cadence, and they are the two frames auditors most commonly reach for as the EU timeline advances.

> **A NOTE ON SCOPE**
>
> This section reports regulatory texts for planning purposes; it is not legal advice, and application to a specific institution is a question for counsel. Where a requirement is characterized here, the source table carries the underlying document so counsel can start from the primary text.

7 Methodology and incident record

7.1 How the database was built

Eligibility. An incident qualifies if its occurrence, public disclosure or adjudication falls between January 2023 and July 2026. That basis admits three entries whose underlying events predate the window but whose disclosure or adjudication does not (the PyTorch torchtriton compromise, December 2022; the Air Canada chatbot interaction of November 2022, adjudicated February 2024). One entry — the US$35 million voice-clone fraud of January 2020, reported October 2021 — fails even that test and is included as an explicit, labelled historical exception, because it is the only comparably sized corporate voice-clone loss on record and omitting it would distort the impersonation picture. The database is a curated record of verified incidents selected for financial impact and public recognizability within each surface; it is not a census, and "no reported cost" is absence of evidence, not evidence of zero cost.

Verification. Every incident is verified against primary sources where they exist — court and tribunal records, prosecutor and police releases, regulator publications, first-party disclosures — and against named researcher or major-outlet reporting where they do not. Every factual claim carries a resolvable source, consolidated with URLs and access dates in the source table published alongside this report. Each entry records evidence quality, documentation status, and a confidence rating; conflicts between sources are flagged rather than resolved silently.

Financial discipline. Dollar figures appear only with a citable source and a class label — *reported* (disclosed by victim, police or press), *adjudicated* (ordered or forfeited by a court), or *projection* (a model of a future that may not arrive). Classes are never blended, no cross-currency conversion is applied beyond a source's own, and awards without a source conversion (Air Canada's CA$812.02) are stated in original currency and excluded from USD totals. This edition constructs no modelled counterfactuals: where an incident produced no verified figure, the entry says so.

Corrections that verification forced. Three are stated up front because they are the kind of error that discredits documents of this type. The "CamoLeak" disclosure has *no* assigned CVE — CVE-2025-59145, sometimes attached to it, belongs to an unrelated npm compromise. The widely repeated "US$100,000+" fabricated-citation sanction in the Semrad matter is in fact US$5,500. And MITRE ATT&CK T1656 (Impersonation), still widely cited for deepfake fraud, was revoked in ATT&CK v19 and superseded by T1684.001; framework identifiers in this report were verified against the OWASP GenAI project publications, MITRE's ATLAS data release 2026.06, and the ATT&CK v19.1 STIX distribution, and should be re-checked at each revision because both MITRE frameworks revise on a rolling basis.

7.2 The incident record

ID	INCIDENT	DATE	MODE	VERIFIED FINANCIAL IMPACT	PRIMARY SOURCING
D-1	Arup — deepfake video-conference CFO fraud (Hong Kong)	Jan 2024	FM-10	**HK$200,000,000 (≈US$25.6M) — reported**	HK police briefing; HK gov't LegCo record (Jun 26, 2024); victim confirmed on record info.gov.hk/gia/general/202406/26/P2024062600192.htm scmp.com/…/uk-multinational-arup-confirmed-victim-hk200-million-deepfake-scam

D-2	Voice-clone bank fraud, Hong Kong branch (historical exception)	Jan 2020	FM-10	**US$35,000,000 — reported**	Dubai Public Prosecution filing in US federal court, via Forbes (Oct 2021) forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech
D-3	Deepfake CFO video call, multinational's Hong Kong branch	May 2024	FM-10	**≈HK$4,000,000 (US$511,968) — reported**	SCMP (May 30, 2024); HK gov't LCQ9 record scmp.com/…/hong-kong-employee-tricked-paying-out-hk4-million
D-4	Executive-deepfake attempt wave: Ferrari, WPP, LastPass, Wiz	2024	FM-10	**US$0 — all defeated by people**	Bloomberg; The Guardian; LastPass blog (first-party); TechCrunch theguardian.com/technology/article/2024/may/10/ceo-wpp-deepfake-scam blog.lastpass.com/posts/attempted-audio-deepfake-call-targets-lastpass-employee
D-5	United States v. Smith — AI-generated music streaming fraud	Plea Mar 2026	FM-12	**US$8,091,843.64 forfeited — adjudicated**	DOJ SDNY indictment (Sep 4, 2024) and guilty-plea release (Mar 19, 2026) justice.gov/usao-sdny/pr/north-carolina-musician-charged-music-streaming-fraud
D-6	Deepfake Zoom impersonation of Singapore's Prime Minister	May 2026	FM-12	**S$4,900,000 (US$3.8M) — individual victim; held out of enterprise totals**	Singapore Police Force statement via SCMP scmp.com/…/how-victim-lost-us38-million-singapore-deepfake-zoom-scam
D-7	Hong Kong deepfake romance / investment-fraud syndicate	Oct 2024	FM-12	**≈HK$360M (US$46M) aggregate, many consumer victims — held out**	Hong Kong police via SCMP scmp.com/…/hong-kong-fraudsters-use-deepfake-tech-swindle-love-struck-men
A-1	EchoLeak — zero-click prompt injection, Microsoft 365 Copilot (CVE-2025-32711)	Jun 2025	FM-1	**US$0 (no in-the-wild exploitation confirmed)**	Microsoft MSRC advisory; Aim Labs disclosure; NVD msrc.microsoft.com/update-guide/vulnerability/CVE-2025-32711 catonetworks.com/blog/breaking-down-echoleak
A-2	Slack AI indirect-injection exfiltration path	Aug 2024	FM-1	**US$0**	PromptArmor disclosure; Slack security update (Aug 21, 2024) promptarmor.com/resources/data-exfiltration-from-slack-ai-via-indirect-prompt-injection slack.com/blog/news/slack-security-update-082124
A-3	CurXecute — injection-to-RCE in Cursor via MCP auto-write (CVE-2025-54135)	Jul 2025	FM-1/3	**US$0 (research PoC)**	Aim Labs; NVD catonetworks.com/blog/curxecute-rce nvd.nist.gov/vuln/detail/CVE-2025-54135
A-4	CamoLeak — GitHub Copilot Chat exfiltration via image proxy (no CVE assigned)	Oct 2025	FM-1	**US$0 (research PoC)**	Legit Security disclosure legitsecurity.com/blog/camoleak-critical-github-copilot-vulnerability
A-5	GitLab Duo remote prompt injection with HTML exfiltration	Feb 2025	FM-1	**US$0 (research)**	Legit Security; GitLab remediation legitsecurity.com/blog/remote-prompt-injection-in-gitlab-duo

A-6	Perplexity Comet agentic-browser indirect injection	Aug 2025	FM-1	**US$0 (PoC)**	Brave disclosure brave.com/blog/comet-prompt-injection
A-7	~100 malicious backdoored models on Hugging Face	Feb 2024	FM-2	**US$0 (no confirmed downstream victim)**	JFrog research jfrog.com/blog/data-scientists-targeted-by-malicious-hugging-face-ml-models
A-8	PyTorch torchtriton dependency-confusion compromise	Dec 2022	FM-2	**US$0 quantified (real compromise, no public count)**	PyTorch advisory (first-party) pytorch.org/blog/compromised-nightly-dependency
A-9	Hugging Face production breach via malicious dataset; autonomous-agent intruder	Jul 2026	FM-2/4	**US$0 reported (no figure disclosed)**	Hugging Face disclosure (Jul 16, 2026); OpenAI statement (Jul 21, 2026) huggingface.co/blog/security-incident-july-2026 openai.com/index/hugging-face-model-evaluation-security-incident
A-10	MCPoison — post-approval MCP trust bypass in Cursor (CVE-2025-54136)	Aug 2025	FM-3	**US$0 (research)**	Check Point Research; NVD research.checkpoint.com/2025/cursor-vulnerability-mcpoison nvd.nist.gov/vuln/detail/CVE-2025-54136
C-1	Bartz v. Anthropic — copyright class settlement (IP, not a security breach)	Final approval Jul 2026	—	**US$1,500,000,000 — held out of all security totals**	N.D. Cal. final-approval order (Jul 20, 2026) cdn.arstechnica.net/…/Bartz-v-Anthropic-Order-Approving-Settlement-7-20-26.pdf
C-2	Fabricated-citation sanctions, Mata v. Avianca through Coomer v. Lindell (7 orders, 6 cases)	2023–2026	FM-7	**US$3,000–$31,100 per order — adjudicated**	Court orders: S.D.N.Y.; D. Wyo.; C.D. Cal.; N.D. Ala.; Bankr. N.D. Ill.; D. Colo. — per-case links in the source table courtlistener.com/docket/63107798/mata-v-avianca-inc govinfo.gov/content/pkg/USCOURTS-ilnb-1_24-bk-13368
C-3	Samsung — semiconductor source code pasted into ChatGPT (three events)	Mar–Apr 2023	FM-9	**US$0 reported**	Contemporaneous Korean and international reporting; Samsung policy response techcrunch.com/2023/05/02/samsung-bans-use-of-generative-ai-tools-like-chatgpt
C-4	Chevrolet of Watsonville — "$1 Tahoe" chatbot manipulation	Dec 2023	FM-8	**US$0**	Screenshots and dealer confirmation via trade press gmauthority.com/blog/2023/12/gm-dealer-chat-bot-agrees-to-sell-2024-chevy-tahoe-for-1
C-5	DPD chatbot induced to swear at and disparage its operator	Jan 2024	FM-8	**US$0**	Company acknowledgment; ITV / TIME itv.com/news/2024-01-19/dpd-disables-ai-chatbot
C-6	NYC "MyCity" chatbot advising businesses to break the law	Mar–Apr 2024	FM-7	**US$0**	The Markup / THE CITY investigation; city response themarkup.org/news/2024/03/29/nycs-ai-chatbot-tells-businesses-to-break-the-law
C-7	Moffatt v. Air Canada — chatbot-invented bereavement policy	Feb 2024	FM-7	**CA$812.02 — adjudicated (excluded from USD totals)**	2024 BCCRT 149 (CanLII) canlii.org/en/bc/bccrt/doc/2024/2024bccrt149

B-1	Freysa — autonomous agent talked into releasing its prize pool	Nov 2024	FM-4/8	**≈US$47,000 — reported, on-chain**	On-chain record; The Block / Cointelegraph theblock.co/post/328747/human-player-outwits-freysa-ai-agent
B-2	Replit — AI agent deletes production database during explicit freeze	Jul 2025	FM-4	**US$0 quantified (recovered; refunded)**	Company statements; CEO acknowledgment fastcompany.com/91372483/replit-ceo-what-really-happened
B-3	United States v. Ding — theft of Google AI/TPU trade secrets	Convicted Jan 2026	FM-6	**No valuation issued**	DOJ releases; jury verdict justice.gov/usao-ndca/pr/former-google-engineer-found-guilty
B-4	Meta LLaMA model-weights leak	Mar 2023	FM-6	**US$0 reported**	Contemporaneous reporting; Meta statements vice.com/en/article/facebooks-powerful-large-language-model-leaks-online
B-5	OpenAI Redis bug — cross-user data exposure	Mar 2023	FM-5	**US$0 direct (€15M Italian fine annulled; excluded as non-standing)**	OpenAI post-incident writeup (first-party); annulment via case reporting openai.com/index/march-20-chatgpt-outage wsgr.com/en/insights/openai-prevails-in-landmark-italian-ai-and-gdpr-enforcement-case
B-6	Microsoft AI research — 38 TB exposed via storage token	Sep 2023	FM-5/6	**US$0 reported**	Wiz Research; Microsoft response wiz.io/blog/38-terabytes-of-private-data-accidentally-exposed
B-7	OpenAI internal forum accessed	2023 (disclosed 2024)	FM-6	**US$0 reported**	New York Times reporting, relayed and corroborated techrepublic.com/article/openai-hacked-internal-communications
L-1	DeepSeek — unauthenticated ClickHouse database exposed	Jan 2025	FM-5	**US$0**	Wiz Research wiz.io/blog/wiz-research-uncovers-exposed-deepseek-database-leak
L-2	McHire (Paradox.ai) — data tied to ~64M applicant interactions behind "123456"	Jul 2025	FM-5	**US$0**	Researcher disclosure; Paradox.ai acknowledgment krebsonsecurity.com/2025/07/poor-passwords-tattle-on-ai-hiring-bot-maker-paradox-ai
L-3	Chattr.ai — cloud credentials shipped in client JavaScript	Jan 2024	FM-5	**US$0**	Researcher disclosure; 404 Media mrbruh.com/chattr 404media.co/hackers-break-into-hiring-ai-chat-bot-chattr
L-4	Vyro AI — open Elasticsearch streaming prompts and tokens (116 GB)	2025	FM-5	**US$0**	Cybernews cybernews.com/security/ai-chatbots-vyro-data-leak
L-5	WotNot — 346,381 user files in a public cloud bucket	2024	FM-5	**US$0**	Cybernews cybernews.com/security/wotnot-chatbot-data-leak

Enterprise-loss aggregate across the record: US$69,308,411.64, of which 99.85% sits in Surface 4 (D-1, D-2, D-3, D-5); all other surfaces combined: US$104,600 (C-2 orders US$57,600 + B-1 ≈US$47,000; Air Canada's CA$812.02 stated separately per the currency rule). Held apart and never summed: consumer-victim aggregates (D-6, D-7 — US$49.8M combined) and the Bartz settlement (US$1.5B, an IP matter). Reported sums are floors that exclude second-order organizational costs.

8 About CyberArmor.AI

AI Security, Governance and Trust Infrastructure

CyberArmor.AI is the independent AI trust platform — AI control, input to output. One rulebook for what AI may trust, read and release, written once and enforced across the surfaces where the failure modes in this report actually occur — endpoints, browsers, developer tools, AI traffic, applications, network gateways and autonomous systems — with every decision recorded as evidence.

On the AI estate, the platform screens content and web resources before AI systems act on them, and applies tenant policy to what models return as well as to what reaches them; detects prompt-injection patterns, sensitive data and dangerous output in AI traffic, and redacts across a 21-class catalog of personal data and secrets, extended by named-entity recognition for people, organizations and locations; maintains an AI bill of materials with vulnerability findings prioritized by known-exploited status; inventories AI tool use and agent-tool configurations across a workforce, including the connections on developer endpoints; statically inspects AI model artifacts for executable content without ever loading them; and registers agent identities and records delegation chains, so autonomous actions are attributable.

Because the most frequent failure in this report's record is ordinary security failure striking AI systems, the same rulebook extends across the conventional estate: endpoint detection with behavioural ransomware and mass-deletion signals on by default; response actions — reversible host isolation and process termination — executed through an audited privileged-action broker; patch remediation and software-update inventory; signature- and hash-based malware detection; and per-connector forwarding to Splunk, Sentinel, QRadar, Elastic, Google SecOps and syslog. Every decision on either estate is sealed into the same tamper-evident, hash-chained audit trail, and a compliance engine maps that evidence to seventeen regulatory and standards frameworks — including NYDFS Part 500, SEC and FINRA cybersecurity expectations, ISO/IEC 42001, NIST AI RMF, GDPR and CCPA — with all fifty-eight shipped policy templates validated against the live policy engine. One rulebook, one evidence chain, both estates: the record this report keeps recommending is produced as a by-product of enforcement, not assembled after the fact.

Every capability the company describes anywhere — including in this report — is labelled by maturity on a public status page, **cyberarmor.ai/status**, updated when the code changes and re-verified against the live platform (most recently August 2, 2026). The platform deploys in the hosted stack or in your own environment. The 36-incident research base, the taxonomy and the self-assessment are published ungated at **cyberarmor.ai/research**, and organizations are welcome to use both with or without any commercial relationship. Research correspondence: research@cyberarmor.ai.

Consolidated sources

The claim-level source table — every figure, its source, URL and verification status — is published at cyberarmor.ai/the-cost-of-untrusted-ai-sources.html, linked from cyberarmor.ai/research. Principal sources by class:

Measurements. IBM / Ponemon Institute, *Cost of a Data Breach Report 2026* (published July 29, 2026) and accompanying IBM press release; IBM, *Cost of a Data Breach Report 2025* (July 30, 2025) and IBM's published AI-findings summaries. FBI Internet Crime Complaint Center, *Internet Crime*

Report 2025 (released April 2026), including §"Artificial Intelligence (AI) Used in Cybercrime." Deloitte Center for Financial Services (May 2024) — projection, labelled as such.

Regulation and statute. Regulation (EU) 2024/1689 (AI Act) and Regulation (EU) 2026/1744 (Digital Omnibus on AI), EUR-Lex; European Commission AI Act implementation timeline; Regulation (EU) 2016/679 (GDPR); SEC Division of Examinations FY2026 Examination Priorities; FINRA 2026 Annual Regulatory Oversight Report; NYDFS 23 NYCRR Part 500 resources and Industry Letters (Oct 16, 2024; May 21, 2026); FinCEN Alert FIN-2024-Alert004 (Nov 13, 2024); Federal Register doc. 2026-01688 (HHS civil-penalty adjustment, eff. Jan 28, 2026); Texas HB 149 and Texas Attorney General consumer-AI materials.

Adjudications and official records. Moffatt v. Air Canada, 2024 BCCRT 149; Mata v. Avianca sanctions order (S.D.N.Y. 2023) and successor orders through Coomer v. Lindell (2026); U.S. DOJ SDNY filings in United States v. Smith (Sept 2024; Mar 2026); DOJ releases and verdict in United States v. Ding (Jan 2026); Bartz v. Anthropic final-approval order (N.D. Cal., July 20, 2026); Hong Kong government LCQ9 record (June 26, 2024); Singapore Police Force statements; Dubai Public Prosecution filing (via Forbes, Oct 2021).

Frameworks. OWASP Top 10 for LLM Applications (2025) and OWASP Top 10 for Agentic Applications (2026), genai.owasp.org; MITRE ATLAS data release 2026.06; MITRE ATT&CK Enterprise v19.1.

First-party and researcher disclosures. Hugging Face security disclosure (July 16, 2026); OpenAI statements (July 21, 2026; March 2023 post-incident report); Microsoft MSRC; Slack; PyTorch; LastPass; Replit. Research by Aim Labs, PromptArmor, Legit Security, Brave, JFrog, Check Point Research, Wiz Research, and reporting by named journalists at the outlets cited in the incident record.