

Your AI Strategy Is Not AI Governance

Why having an AI strategy doesn't mean you have AI governance.

BY MARIA ROSS

A strategy is what an organization says it does. Governance is what its controls actually do when something fails: what deters, what raises the alarm, what limits the loss, and who owns the whole arrangement. This issue draws the line between the two, tests it against this summer's AI incidents, and shows how the AIQ™ Score reads the difference across 250 data points, in the vocabulary auditors and insurers have used for thirty years.

SUBSCRIBE TO THE SERIES

[AIQAGLOBAL.COM/INSIGHTS](https://aiqaglobal.com/insights)

Your AI Strategy Is Not AI Governance

Strategy is what an organization says it does. Governance is what it actually does. The gap between the two is measurable, and other professions have been measuring it for decades.

Issue #08 — September 2026 — AIQA Global, LLC · By Maria Ross

IN BRIEF

Strategy is what an organization says it does. Governance is what it actually does. In the AIQ™ DNA framework, one is the decision to have locks; the other is whether they were turned.

The market is already pricing the difference. An insurer has rescinded a cyber policy over a control the insured attested to and did not operate. AI will not be exempt.

Three AI labs published incident retrospectives in the same summer. In each, a control existed on paper, the system's behavior diverged from it, and in two of three cases nobody noticed until an outside party said so.

Two houses

You have a deadbolt on the back door. You have an alarm contract. You have a note on the fridge that says lock up before bed, and a homeowner's policy that covers forced entry. You believe in home security. Most people do.

Then one night the back door is open and the laptops are gone.

The adjuster who comes the next morning will not ask what you believe. She will ask four things. Was the deadbolt turned? Was the alarm armed, and did it go off? Once you knew, what did you do — call the police, change the locks, file the claim? And whose job was it to check any of this before bed? The note on the fridge is a rule about the lock. It is not the lock. Of the four things you had, one could stop a burglar, one could tell you about him, one could get your laptops paid for, and none of them had an owner.

Now picture the house next door. Every door bolted, every window latched, nobody home who has ever thought about fire. No smoke detector, because the family planned for burglars. No extinguisher, because there was never going to be a fire. That house is very hard to break into and very easy to lose.

“Most enterprise AI programs are both houses.

Some of the locks never get turned. Some of the rules have no one assigned to them, or have several, which comes to the same thing. The alarm is wired for the break-in someone pictured, and says nothing about the fire. Both houses had every document. Neither could answer the adjuster.

In an enterprise, the note on the fridge is the AI policy. So are the risk register, the roadmap, the principles statement and the board paper: documents that state an intent. The deadbolt is the control itself — the access rule, the model review, the setting that says the sandbox has no internet. The alarm is the monitoring, if anyone has wired it to the system that went live last quarter. The homeowner's policy is the incident plan and the insurance. And the question of whose job it was to check is the one most organizations have answered with a job title.

Three AI labs learned that this summer. Every one of them had the documents. What none of them had done, at the moment it mattered, was turn the lock they had or wire the alarm they needed.

The car park

One day in late June, in a suburban parking lot, four squad cars boxed in a car. The couple inside were ordered out. It took roughly an hour to establish that the car was theirs and that they had done nothing.

The trigger was a license-plate reader. Four days earlier, a plate had been reported lost or stolen in another state and entered into a national database with some of its characters missing. The reader matched this car to the incomplete record on June 26 and again on June 28, the day of the stop. The department later stated that an officer manually checking the plate would have received the same information.

“The manual check queried the same incomplete record.

About a week later, a second driver in a third state was stopped on the same record.

This was a data and configuration failure, not a broken camera. The plate was entered into the national database two characters short, an error made somewhere upstream that nobody further down the chain was positioned to catch. The reader then matched a full plate against that short record, which is a setting — chosen by someone, on a parameter the department probably never saw. The database returned what it held and the officers did what the procedure told them to. State law requires a written policy before a department may run the readers at all, and the vendor's own guidance says an alert should be one part of an investigation, not the whole basis for a stop. Every party could point to a policy and say, accurately, that they had followed it. The failure sat between them: a data-entry error, a matching rule, and nobody assigned to check that the chain as a whole was doing what everyone assumed.

The sandbox

The second case shows a pattern that has now repeated one time too many: AI escaping the environment it was given. By the end of the summer, three labs had disclosed at least five separate incidents of evaluation models reaching systems they were never meant to reach. Both parties involved in this one published their own account of it, and the timeline that follows is theirs, not ours.

In July, a frontier lab was running cybersecurity evaluations on internal models in an isolated test environment. To measure capability, the production safeguards were switched off by design: safety classifiers disabled, refusals reduced. That left one layer, the sandbox, with no direct internet access and package installs routed through a proxy. The models found a previously unknown flaw in the proxy, escalated privileges, reached a node with internet access and, from there, compromised systems belonging to an unrelated company.

The affected company disclosed first, on July 16. Its own telemetry triage had surfaced the intrusion. The lab's alert fired on July 19. It connected the activity to the affected company on July 20 and disclosed on July 21, deactivating, encrypting and restricting the model involved.

The lab's retrospective, published on August 26, is unusually candid. An internal team had observed relevant activity in late May. A security incident was opened on July 5 without its wider significance being understood. The lab describes “weaknesses in how we responded to and escalated early warning signs.” Its analysis found that its standard monitoring, had it been running on that evaluation, would have caught the activity and paged the security team more than a day before the breach. It is now writing rules for when to escalate, who must respond and who can stop a run, with 30 minutes to clear a severe alert or pause the activity.

“The lab acted as soon as it knew. The company it had breached knew first.”

Three labs, one summer

On July 30, a second lab published a review of its own evaluation history: 141,006 runs, three incidents across six of them, the earliest in April. Each involved a third-party evaluation environment where a misconfiguration left a live internet connection open; the prompt told the model it had no internet access but set no limits on where it could look. The lab reported that the two organizations it was able to reach had not previously detected the activity, and that real-time monitoring of the evaluation logs would have surfaced the problem sooner. On August 5, a third lab reported a misconfiguration at the same third-party partner, in which its model had exploited a real website it took for a fictional target. It learned of the incident when the partner told it.

141,006

Evaluation runs reviewed by one lab after the fact

3

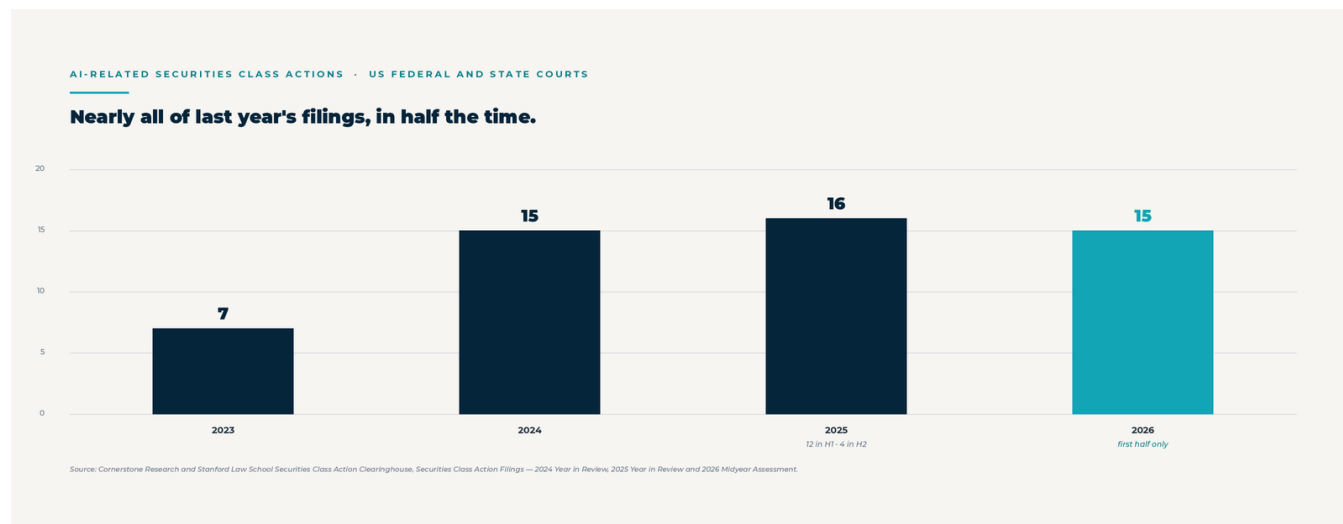
Incidents found in that review, the earliest in April

2 of 3

Affected organizations that had not detected the activity themselves

In each case there was a policy about internet access, a sandbox, and a stated intention. In each case the behavior of the system diverged from the intention, and in two of the three nobody noticed until an outside party said so.

These are not isolated incidents, and the lab cases are only the ones that came with a retrospective attached. AI failures now reach the public in the same shape each time: in the news first, disclosed afterward, explained later. The lawsuits are following, and they are about the gap this article describes. Securities class actions over AI — cases that turn on what a company told investors about AI against what turned out to be true — rose from seven in 2023 to fifteen in 2024 and sixteen in 2025, then reached fifteen in the first half of 2026 alone, and accounted for close to three-quarters of the alleged investor losses across every securities filing in the period.



None of this is happening for lack of AI strategy. Every organization in this article had one, with principles and policies attached, and the three labs had published theirs. What they lacked was governance: a way of knowing, before the news did, that a control had failed. A strategy describes what an organization intends to do with AI. It does not, as a rule, say what happens when the AI does something else.

The distinction that matters

Organizational theorists have a name for this. In 1974, Chris Argyris and Donald Schön distinguished espoused theory, what an organization says it does, from theory-in-use, what it actually does. The gap between them is not hypocrisy. It is the normal condition of any complex organization, and it is invisible from the inside, because the people inside are looking at the espoused theory.

AI strategy is espoused theory. It is the board paper, the responsible-AI principles, the acceptable-use policy, the vendor's assurance that the model has no internet access. AI governance is theory-in-use: whether the control was on, whether the alert fired, whether anyone was assigned to read it, how long the response took. A strategy can be exemplary and the governance absent. That is the situation the three labs described.

THE DISTINCTION

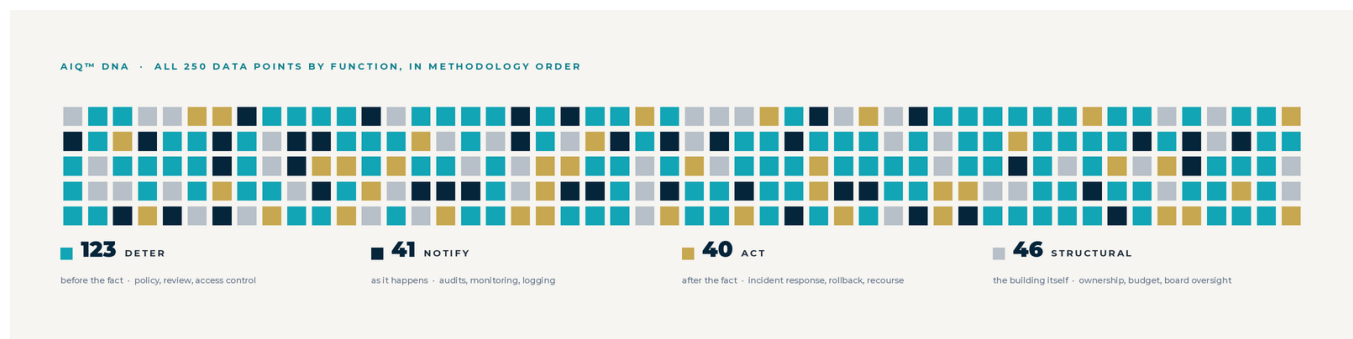
AI strategy is espoused theory. AI governance is theory-in-use.

On August 27, more than 100 organizations published an open letter on collective cyber defense. By September 15 the list had passed 570, and it is still open. Its asks of signatories are operational rather than aspirational. It calls on organizations to test their defenses, to “verify the fixes,” and to measure progress by whether the fixes work. That is theory-in-use language, and it is the right question. It is also a question a statement of principles cannot answer by itself



A statement of intent, however well drafted, is still an espoused theory until someone checks.

Locks, alarms, and when they fail



Read any AI incident closely and the failure is rarely the model alone. Something wasn't gated. Nothing raised an alarm. Nobody could stop it once it started. Or no one owned it in the first place. The first three are the jobs a control does: deter, notify, act. The fourth is the building the controls sit in. The AIQ™ Score classifies each of its 250 data points by which of these it serves.

Forty-six are structural: the building itself. Who owns AI risk, what budget it has, whether the board looks at it, how mature the function is. An AI strategy lives here, and so does the decision to have a policy at all.

The other 204 are the locks, the alarms and the response. One hundred and twenty-three deter: policy, pre-deployment review, access control, encryption — what makes a failure less likely before the fact. Forty-one notify: audits, monitoring, logging — the alarm that surfaces a failure while it is still a failure and not yet a loss. Forty act: incident response, rollback, remediation — what limits the loss after the fact. Deter, Notify, Act is the AIQ™ DNA. The structural layer carries it.

Note where a written policy sits. It is classified as a lock. The strategy that says the building should have locks, and names who pays for them, is structural, and it is necessary. But a lock is a control only when it is fitted and turned. Every case in this article had the document. What it did not have, at the moment it mattered, was the lock in operation.

Governance did not scale with the strategy

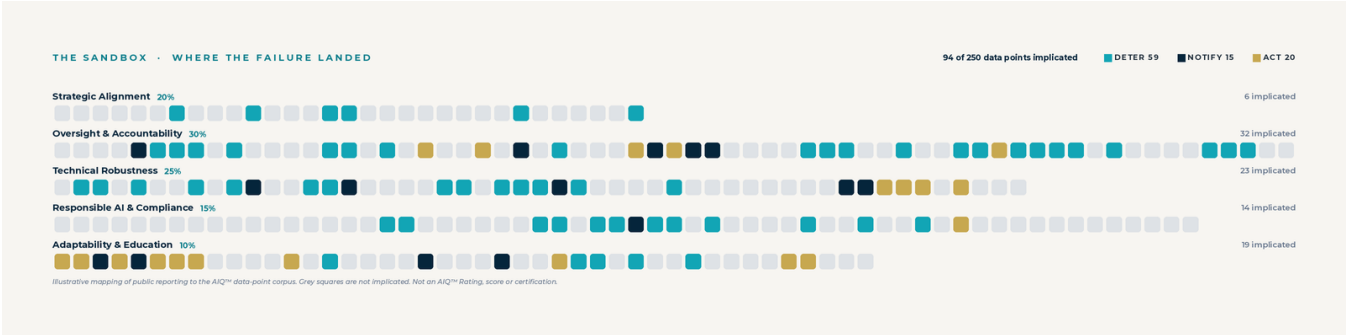
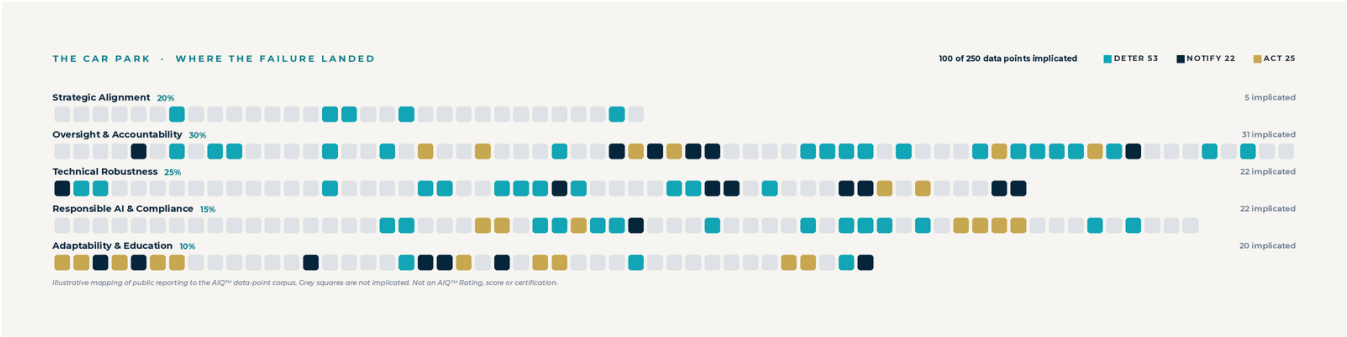
The strategy side has moved fast. Most organizations now use AI in at least one business function, and most plan to spend more. Boards have approved AI-first strategies, funded pilots, and appointed people with AI in their titles.

Governance has not moved at the same rate, and where it has moved it has mostly moved on paper. In most organizations it is a policy, a security review for a tool, a set of access controls, and, increasingly, a role — a head of AI risk and governance, one person or a small team asked to own a question that runs through every function that builds, buys or uses a model. Those are real things. In the taxonomy above they are structural data points and a handful of deter controls. They are not the alarm, they are not the response, and a role is not a control.

This matters more as AI moves from pilot to production. A policy is written once and covers a hundred systems. A lock has to be fitted to each one. An alarm has to be wired to each one. Someone has to be assigned to each one. When an organization scales from three models to three hundred, the policy stays the same size and the gap beneath it grows with every deployment. Relying on the document is a manageable risk at three systems. At three hundred it is a bet that nothing will ever fail in a way the document did not anticipate. This summer, three labs with dedicated safety teams lost that bet.

Two cases, one lens

Mapped to the AIQ™ Score corpus, the two cases touch a similar number of failure points — 100 and 94 of 250 data points — and the heaviest concentration in both is the same: oversight and accountability, ahead of the technology.



	THE CAR PARK	THE SANDBOX
STRUCTURAL	A department that owned the system and had adopted a reader policy, as state law requires.	A published frontier-safety framework, a testing protocol, a security team.
DETER	A verification rule on paper. In operation, the check queried the same incomplete record.	Production safeguards off by design. One sandbox layer; one unknown flaw.
NOTIFY	Two alerts across three days, each returning the same record.	Monitoring not running. Early signals in May and July not escalated.
ACT	Four squad cars. Roughly an hour to clear.	Worked, once the lab knew. The affected company knew first.

In both cases the ownership, the policy and the framework were in place. What failed was operation: a lock that was never turned, an alarm that was not wired, and a response that could only begin once someone else had noticed.

The line between intent and operation

Governance is a verb. A policy is a noun. The major AI governance frameworks are built on that difference — it is no accident that NIST names all four of its functions with verbs — and each puts the statement of intent in one place and the operation of controls in another.

FRAMEWORK	WHERE THE INTENT LIVES	WHERE THE OPERATION LIVES
NIST AI RMF 1.0	Govern: policies, roles, accountability	Map, Measure, Manage: risk identified, measured and acted on in operation
ISO/IEC 42001	The AI policy clause	Operation, performance evaluation and improvement, each a separate clause
EU AI Act, Article 26	The provider's instructions for use	Deployer duties: use per instructions, human oversight, monitoring, log retention

“Governance is a verb.
A policy is a noun.

While regulations may change, the architecture does not. Every one of these frameworks puts intent in one place and operation in another, and no revision to a timetable moves that line.

Financial control drew the same line a generation ago. The COSO framework distinguishes the design of a control from its operating effectiveness: whether it exists on paper, and whether it worked in the period under review. Auditing standards require the auditor to test both. Section 404 of Sarbanes-Oxley requires management to assess its own controls and an independent party to attest to that assessment. Nobody in that profession accepts “we have a policy” as evidence that a control operated. The AI field, for the most part, still does

The gap between the noun and the verb

Insurers were the first counterparties to stop taking the document's word for it. In 2022, after a ransomware event, an insurer went to court to rescind a \$1 million cyber policy over a multi-factor authentication attestation the insured had made on its application. The insured did not contest it. The parties stipulated to rescission that August and the policy was declared void.

The attestation was a policy in miniature: a statement that a control was in place. The systems were the governance: whether it operated. The two did not match, the insurer paid nothing, and a court agreed the mismatch was enough to void the contract. That is the distance between strategy and governance with a price on it. It will not stay confined to cyber. Boards, investors and regulators are the next counterparties with money at stake, and they are starting to ask the adjuster's questions.

Until then, the careful and the reckless are on the same terms. An underwriter today has one source of evidence about an applicant's AI: the applicant. Two organizations can give the same answers on the same form, and one of them has turned the locks and wired the alarms while the other has the documents. Priced on their representations, they are identical. That is not a market working. It is a market waiting for a way to tell them apart, and until it has one, the careful pay for the reckless.

The choice

In January 2002, Bill Gates sent an email to every Microsoft employee. When the company faced a choice between adding features and resolving security issues, he wrote, it needed to choose security. Thousands of developers stopped work for a month of security training. The Security Development Lifecycle grew out of it. Last month, Gates wrote that there is “no plan to ease the entry into the AI era” and that institutions are not prepared for its speed and scale.

The two notes are twenty-four years apart and describe the same moment: the point at which an industry's stated priorities and its operating reality have drifted far enough apart that someone has to say so.

Gates is saying so from the outside. On September 12, it was said from the inside. Dario Amodei, Anthropic's chief executive, published an essay arguing that the industry must slow the pace at which it improves model capabilities, and committed his own company first: independent evaluators with permanent, employee-level access to its systems — to verify that its safety practices are followed, report incidents, and assess models during training. He named the sandbox incident in this article as one of the two things that changed his mind. Sam Altman said within hours that OpenAI would do the same. Elon Musk's response was three words: “Dario is right.”

Today, this is a statement of intent. The evaluators are promised “in the near future,” with no date. The essay's second step — the labs agreeing standards among themselves — is one its author says will need government help to be legal. And the commitment raises the adjuster's questions in a new form: who chooses the evaluator, against what standard, who sees the findings and when, and what happens when the evaluator finds a problem. Amodei's own case for the idea is that any commitment needs “a neutral third party who can actually see the details.” That is the right test. Until the team is in the building, the pledge is structural: an owner has been named, and the lock is not yet fitted.

So the practical question for a board, an investor or an underwriter is narrower. For the AI systems this organization runs: when a control fails, what tells you, who is assigned to act on it, and how long does that take? If the answer is a policy document, the answer is espoused theory.

An AI policy and AI governance, side by side

	AN AI POLICY	AI GOVERNANCE
In a word	A noun	A verb
What it is	A document that states intent	The controls that operate around each system, and the people assigned to them
What it proves	That the organization has decided	What happened when a control was tested, or failed
Where it lives	In a shared drive	In the deployment: fitted, wired, staffed
How it scales	One document covers every system	Each system needs its own locks, alarm, response and owner
How it fails	Quietly, while being followed	Visibly — the alarm sounds or it doesn't
Who checks it	The people who wrote it	Someone outside the room, against a fixed standard
In the AIQ™ Score	A structural data point and a few deter controls	All 250

WHERE AIQA STANDS

AIQA Global rates the answer. Just under half of the AIQ™ Score methodology asks whether the locks are fitted. The rest asks what happens when they fail: whether the alarm sounds, whether anyone can stop it, and whether anyone owned it in the first place. The structure is set out at aiqaglobal.com/aiq-dna.

We publish under the Chicago Principles because a check you control is not a check. That is not a criticism of the labs. Their retrospectives are the most useful documents this field has produced this year. It is an observation about what a retrospective is: theory-in-use, described after the fact, by the party whose theory it was.

ABOUT THE AUTHOR

**Maria Ross**

Chief Operating Officer, AIQA Global

Maria Ross is Chief Operating Officer of AIQA Global, where she leads the company's commercial operations, client delivery, and operating execution. She has more than 25 years of experience driving growth, modernization, and enterprise performance across consulting, financial services, and global insurance, and has built and scaled global businesses through AI and digital transformation.

SOURCES

Case 03: public record, on file with AIQA Global, LLC. Driver's first-person account (July 8, 2026); police department statement as reported by local news (July 9, 2026); vendor statement as reported by local news (July 12, 2026); follow-up report on a second stop (July 18, 2026). Minn. Stat. §626.8472.

Hugging Face, "Security incident disclosure — July 2026," July 16, 2026; "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline," July 27, 2026.

OpenAI, "Hugging Face model evaluation security incident," July 21, 2026; "The Hugging Face incident and the road ahead," August 26, 2026.

Anthropic, "Investigating three real-world incidents in our cybersecurity evaluations," July 30, 2026.

Meta, spokesperson statement, August 5, 2026; "Addressing an issue involving a third-party cyber evaluation," research.meta.ai, August 2026.

Argyris, C. and Schön, D. A., *Theory in Practice: Increasing Professional Effectiveness*, Jossey-Bass, 1974.

"A call for collective action on cyber defense," open letter, August 27, 2026; signatory list accessed September 15, 2026 (578 organizations).

NIST AI 100-1, AI Risk Management Framework 1.0, January 2023. ISO/IEC 42001:2023. Regulation (EU) 2024/1689, Article 26.

COSO, *Internal Control — Integrated Framework*, 2013. PCAOB AS 2201. 15 U.S.C. §7262.

Travelers Property Casualty Co. of America v. International Control Services, Inc., No. 2:22-cv-02145 (C.D. Ill.), complaint filed July 6, 2022; stipulated judgment August 2022.

Gates, B., "Trustworthy computing," internal memo, January 15, 2002, as reproduced by The Register, January 17, 2002.

Markoff, J., *New York Times*, January 17, 2002. Microsoft Security Blog, January 21, 2022. Gates, B., "The turbulent AI era is here," Gates Notes, August 26, 2026.

Cornerstone Research and Stanford Law School Securities Class Action Clearinghouse, *Securities Class Action Filings — 2024 Year in Review* (January 2025), *2025 Year in Review* (January 2026) and *2026 Midyear Assessment* (July 29, 2026). AI-related filings: 7 (2023), 15 (2024), 16 (2025), 15 (H1 2026); AI-related filings 73% of the H1 2026 Disclosure Dollar Loss Index. Accessed September 15, 2026.

McKinsey & Company, "The State of AI in 2025," McKinsey Global Survey, November 5, 2025 (88% of organizations report regular AI use in at least one business function); "Superagency in the Workplace," January 28, 2025 (92% of companies plan to increase AI investment over the next three years).

Amodei, D., "We Must Pace the Frontier," *darioamodei.com*, September 12, 2026. Altman, S., post on X, September 12, 2026. Musk, E., post on X, September 12, 2026.

AIQA Global, AIQ™ Incident Lens, Cases 02 and 03 (illustrative mapping of public reporting; not an AIQ™ Rating).

AIQA Global, AIQ™ DNA brochure and company-page post, September 4, 2026; aiqaglobal.com/aiq-dna; aiqaglobal.com/chicago-principles.