

AGENDA

OPINION

AI Success Is Within Our Grasp: Most AI Breaches Are Preventable

How to start using the stewardship and technical tools we already have to achieve success

By **Dominique Shelton Leipzig** | October 2, 2026

Here is the good news that rarely makes the headlines: Most AI failures are preventable, and the tools to prevent them already exist. The path to success is open to every board and CEO willing to take the first steps.

The latest reminder of why this matters came on Sept. 23, when Australian Prime Minister Anthony Albanese announced that an OpenAI agent had breached a Medicare statistics portal in June and reached non-public files. [The government learned of it only this month, through an email to a public mailbox](#). Shortly thereafter, OpenAI announced that [it had also breached U.S. governmental websites](#). Earlier, in July, OpenAI agents broke out of a cybersecurity test and hacked the AI company Hugging Face.

Dominique Shelton Leipzig
Dominique Shelton Leipzig is CEO of Global Data Innovation, author of TRUST, and founder of the Digital Trust Summit. She advises CEOs, boards, and governments on building stewardship architecture that accelerates AI deployment, drawing on a proprietary knowledge base with 1.4 million unique weighted attributes from thousands of documented AI, cyber, and privacy successes and failures and on direct engagement with regulators and heads of state from Washington to Brussels to Geneva.

Events like these can feel like proof that AI is out of control. I read them differently. Each follows a pattern we have seen before, and patterns can be decoded. After 35 years advising hundreds of CEOs and board members and training thousands more, I can tell you there is very little that is new. Most technology failures follow patterns of human and

organizational behavior, and those patterns show where success occurs and where failure will reside.

That is the source of my optimism. At Global Data Innovation, we have decoded the five areas responsible for three-quarters of AI incidents. We use the mnemonic TRUST to identify them: ["T" for triage, "R" for right data, "U" for uninterrupted testing, monitoring and auditing, "S" for supervision and "T" for technical documentation](#). Each of these five areas has specific key performance indicators that can be customized to align with your company's own KPIs. Our 1.4 million unique weighted attributes tell us when a company is on track for success, and when it carries the vulnerabilities we read about in the press, early enough to fix them.

Of these five areas, the U in TRUST, uninterrupted testing, monitoring and auditing, is the one that appears most often across the thousands of AI incidents we have studied. Continuous monitoring appears in the EU AI Act, NIST guidelines and the OECD AI Principles. Observability technology that delivers it for generative and agentic AI has been available for roughly three years.

Even AI developers now say this out loud. After four incidents in which its Claude models breached real third-party systems during testing, Anthropic reported on Sept. 9 that ["we have recently put in place real-time monitors that halt evaluation and training runs when a model probes its sandbox or unexpectedly reaches the internet, and offline monitoring that extends to nearly all internal agentic use."](#)

Board members at other companies watching the news about agentic breaches needn't wait for their company to be next. They can ask, now, whether uninterrupted monitoring controls are in place for their company's use of AI. There are four other critical areas in the TRUST framework to operationalize effective AI stewardship.

The science tells us what to plan for. Models can drift and deviate from expectations at any time after training. A 2026 paper in the journal Clinical Simulation in Nursing describes generative AI as undergoing [almost continuous model drift](#), so the same question can produce different conclusions from one day to the next. [Researchers at Stanford and UC Berkeley found](#) that GPT-3.5 and GPT-4 behavior shifted substantially within months. This year, UC Berkeley researchers showed that [frontier models spontaneously developed mis-aligned behaviors](#), including deception, with no instruction to do so.

Other researchers have shown that [core failure modes persist as models grow](#) and that AI systems improving themselves [without outside correction steadily lose their safety alignment](#).

Read together, this research carries a hopeful message. Drift that is expected can be detected, and drift that is detected can be corrected. That is where the most beautiful part of technology steps in: the place where humans meet the machine. A typewriter on its own cannot run the world, but a creative person behind the keys can transform it.

So where should a company start? Here is how to put the tools we already have to work.

First, learn from what has already happened. Commercial aviation became one of the safest ways to travel because every accident is investigated and the lessons are shared. In my experience, the vast majority of Fortune 500 companies do not yet examine past AI, cyber and privacy incidents in any systematic way. The companies that close that gap will see problems coming while competitors are still reading about them in the press.

Second, triage every AI use case by risk: red for prohibited, yellow for high-risk uses demanding rigorous testing, green for the rest. Speed can then match the risk.

Third, turn on uninterrupted testing, monitoring and auditing, with a trained human team ready to contain and recover when drift appears.

Fourth, ensure management has hired a quarterback to effectuate constructive collaboration between AI cross-disciplinary teams composed of lawyers, data scientists and business leaders in the same room.

Many AI stewardship meetings at the management level have morphed into frustrating sessions where participants feel their time is being wasted. Lawyers cannot discern technical best practices from laws on a page. Data scientists will not find effective guardrails in computer science courses. Business leaders need both perspectives to ask the questions that will catapult their companies into success. In order to communicate effectively, a shared language is needed to operationalize trustworthy AI. We encourage our clients to speak to each other using the TRUST mnemonic so that everyone knows what everyone else means.

We learned hard lessons from waiting too long on cybersecurity. With AI, we have the chance to get it right from the start. Board members and CEOs who look for risks early and address them will scale AI with confidence. The solutions exist. Success is within our grasp.