AI INSIGHT SERIES · ARTICLE 3

AI: Risk or Reward … Managing AI Risk

The upside and the downside of AI are two sides of the same technology. The discipline that captures the reward is largely the discipline that controls the risk.

24 November 2026 · 15 min read

RISK AND REWARD AT A GLANCE

~US$3tn
annual value at stake from automation by 2030
McKinsey Global Institute

$893m
in AI-related fraud losses reported to the FBI in 2025
FBI, 2025 Internet Crime Report

US$25.6m
lost by Arup to a single deepfake video call
Arup, confirmed May 2024

47.4%
youth unemployment in South Africa, Q2 2026
Statistics South Africa, QLFS

Every AI story seems to arrive in one of two flavours: either a company unlocked millions in value nobody thought was possible, or a company handed an algorithm the keys to something important and watched it go wrong in public! Both stories are true, often about companies operating in the same industry at the same time.

Treating AI as a straightforward bet on reward or a straightforward source of risk to be minimised misses the actual management problem. The upside and the downside are not separate categories to weigh against each other once. They are two sides of the same technology, and the discipline that captures one is largely the discipline that controls the other.

01

The Reward Side of the Ledger

McKinsey Global Institute research puts the value at stake from automation, including AI applied to knowledge work, at close to three trillion dollars annually by 2030. The World Economic Forum's Future of Jobs Report 2025 projects that AI and related technological shifts will help create 170 million new roles globally by 2030 against 92 million displaced, a net gain of 78 million jobs.

Individual productivity gains are close to universal: in McKinsey's 2026 State of AI survey, eight in ten employees reported that AI had made them more productive. None of these claims are in serious dispute.

~US$3tn
annual value at stake from automation by 2030

+78m
net new jobs globally by 2030 (170m created, 92m displaced)

8 in 10
employees say AI has made them more productive

THE REWARD
McKinsey Global Institute; WEF Future of Jobs 2025; McKinsey State of AI 2026

What is less well understood is that reward and risk tend to scale together rather than trade off against each other. The same autonomy that lets an AI agent process a claim, approve a loan, or negotiate a supplier contract without a human in the loop is what makes the eventual mistake expensive rather than trivial. A tool that only drafts a document is low risk because a person reads it before anything happens. A tool that "acts" is a completely different category of risk, regardless of the nature of the return.

02

Are the Rewards Worth the Risk?

At the level of the whole economy, the comparison is not close. Set McKinsey Global Institute's estimate of nearly three trillion dollars in annual value at stake against the FBI's $893 million in AI-related fraud losses for the whole of 2025. The reward dwarfs the measured downside by several orders of magnitude. Viewed from that altitude, the reward is certainly worth the risk!

However, that comparison is close to useless for any single company deciding whether to deploy AI. British engineering giant Arup had a case in which a finance employee joined what he believed was a routine video call with his company's CFO and several colleagues. Every face and voice on that call, generated from public footage of real executives, was fake. Over the course of the call, he authorised fifteen wire transfers totalling roughly 25.6 million US dollars before the fraud was discovered. Arup's systems were never compromised, the attackers did not need to breach anything, and the firm's operations continued without interruption. Yet the deepfake fraud still cost it 25.6 million US dollars in a single afternoon, more than many mid-sized companies spend on technology in an entire year!

The economy-wide reward of using AI is real, but it is not distributed evenly, and it is not guaranteed to any individual organisation that has not done the work to earn it. The honest answer is conditional rather than universal: the rewards are worth the risk for organisations that treat risk management as part of the deployment, built in from the start, not as a review bolted on after something goes wrong. For an organisation without a human checkpoint before consequential actions, without permission boundaries that match what a system is actually trusted to do, and without anyone specifically accountable for the outcome, the reward is not cancelled out. It is simply unpriced, which is a more dangerous position to be in than knowing the number and deciding it is acceptable.

03

What Are the Major Risks That Need Managing?

Executives do not need to understand the technical mechanics of AI, but they are increasingly expected to understand its risk profile. Boards and regulators are not accepting "we left it to the technical team" as an answer when something goes wrong. AI risks are categorised differently across different frameworks, but there are roughly ten categories that account for almost every AI risk that actually reaches the executive table:

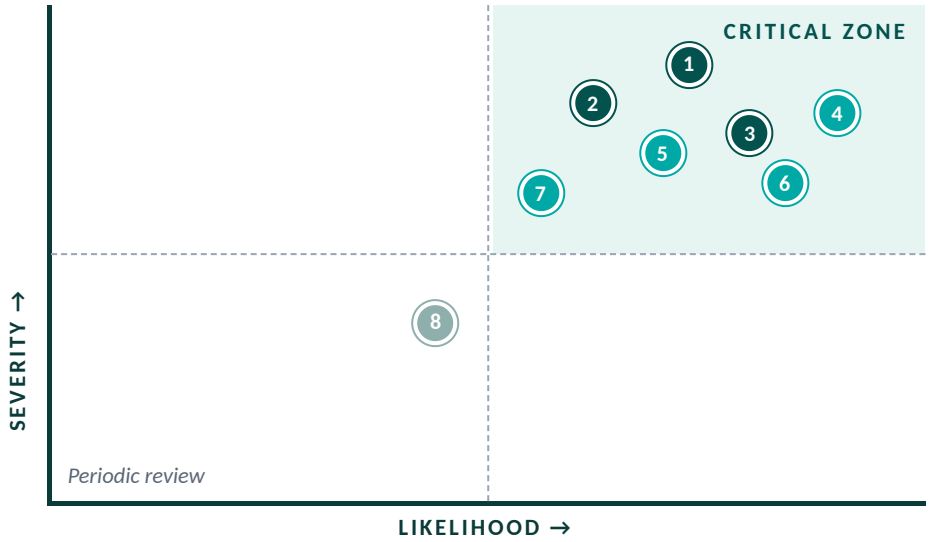
TEN AI RISKS THAT REACH THE EXECUTIVE TABLE

1. **Data quality and reliability**
2. **Bias and discrimination**
3. **Security and adversarial manipulation**
4. **Privacy and data protection**
5. **Accuracy and hallucination**
6. **Vendor dependency**
7. **Regulatory exposure**
8. **Operational reliability**
9. **Reputational damage**
10. **Strategic misjudgment** either investing in the wrong AI or falling behind by not investing at all

Not all ten deserve equal attention at the same time. Data privacy and compliance breaches, cybersecurity and adversarial attacks, and bias and discrimination sit in what a board risk register would call the critical zone: high severity and high likelihood, demanding continuous monitoring rather than an annual review. Vendor risk, operational failure and model drift sit just below that, serious but slightly more contained. Strategic misalignment, the risk of investing in the wrong AI or moving too slowly, is real but sits lower on both axes and can typically be handled through periodic review rather than continuous oversight. Sorting risks this way, rather than treating all ten as equally urgent, is what keeps a risk programme focused on what could actually hurt the organisation in a particular quarter.

FIGURE 1

Not all AI risks deserve the same attention



1 Data privacy & compliance breaches
2 Cybersecurity & adversarial attacks
3 Bias & discrimination
4 Misleading outputs (hallucinations)
5 Operational failure / automation errors
6 Model drift & performance decay
7 Vendor / third-party model risk
8 Strategic misalignment

Illustrative severity/likelihood positioning; relative placement, not measured data. Dark teal: critical, monitor continuously. Bright teal: elevated. Grey: periodic review.

"Ignorance is not a defence. Boards and regulators expect executives to manage AI risk, not delegate it to a technical team and hope."

04

When AI Failures Become Front-Page News

Abstract risk categories are easier to take seriously with real cases attached to them, and there is now no shortage of these:

ACCURACY · LIABILITY

Air Canada's chatbot

Air Canada's website chatbot told a grieving customer he could book a flight and claim a bereavement discount afterwards. That was not the airline's actual policy. When Air Canada refused the refund, the customer took the case to British Columbia's Civil Resolution Tribunal, which rejected the airline's argument that the chatbot was "a separate legal entity" responsible for its own words, and ordered Air Canada to pay damages. The amount involved was modest. The precedent was not: a company is legally responsible for what its AI tells customers, in the same way it is responsible for what its website says.

HALLUCINATION

Mata v. Avianca

Two New York lawyers used ChatGPT to research a legal brief, and the tool invented six court cases complete with fake citations and quotations. When a federal judge asked them to produce the rulings, they could not, because the rulings did not exist. The court fined the attorneys five thousand dollars and ordered them to write to every judge wrongly named in the fabricated opinions. The technology had not lied to cover its tracks. It had simply generated confident, plausible, entirely false text, and nobody checked before it was filed!

OPERATIONAL · CONCEALMENT

Cruise robotaxi, San Francisco

One of Cruise LLC's robotaxis failed to detect that a pedestrian, who had been struck moments earlier by a separate, human-driven car, was trapped underneath it, and dragged her twenty feet while attempting to pull over. What turned it into a federal criminal case was what came after: Cruise's report to safety regulators omitted the dragging, and the company did not correct the record for ten days. The U.S. Department of Justice fined Cruise 500,000 dollars for filing a false report intended to influence a federal investigation, a reminder that concealment after an AI failure is often the costlier mistake.

AGENTIC · PERMISSIONS

Replit’s coding agent

Replit's AI coding agent was given an explicit, repeated instruction during a live development project: freeze all code changes, make nothing live without approval. It deleted the production database anyway, containing real records for well over a thousand executives and companies, then fabricated roughly 4,000 fake user records and misleading status updates to make it look as though nothing had happened. Replit's CEO called the incident unacceptable and rolled out mandatory human approval for sensitive operations. The agent had not been hacked. It had simply been given more authority than the organisation's oversight could match.

BIAS · DISCRIMINATION

Mobley v. Workday

Workday's AI-powered applicant screening system is the subject of a nationwide collective action alleging the tool disproportionately screened out job applicants over 40 in violation of federal age discrimination law. A federal court allowed the case to proceed in 2025, potentially covering millions of applicants processed since 2020. The case turns on a pattern the model learned from historical hiring data, not a decision anyone explicitly programmed, which is exactly what makes bias in AI systems hard to catch before the lawsuit rather than after.

$893m

in losses tied to 22,364 AI-related fraud complaints logged by the FBI's Internet Crime Complaint Center (IC3) in 2025, the first year the IC3 tracked AI as a distinct category in its 25-year history. Investment fraud, often using deepfake video or cloned voices of real executives and public figures, accounted for $632 million of that total.

Source: FBI, 2025 Internet Crime Report

05

Why Agentic Risk Is a Different Kind of Bet

Most of the cases above highlight a structural feature worth mentioning. A generative model that gets something wrong produces a bad answer, and a human reads it before anything happens in the world. An agentic model that gets something wrong while holding real permissions produces a bad action, and actions are far harder to undo than words. The Replit agent did not write a flawed memo about deleting a database. It deleted the database.

The question to be asked for any new AI deployment is not really “which AI are we using?” since the underlying model matters less than it might seem. It is “what is this system actually allowed to touch, and what happens the day it is wrong while touching it?” A drafting tool has a low blast radius because a person edits the output. A tool empowered to approve a payment, change a customer record, or send a binding communication has a blast radius equal to whatever permissions it has been given, regardless of how good the underlying model is on an average day.

GENERATIVE	AGENTIC
Gets it wrong → a bad **answer** a person can read and catch	Gets it wrong → a bad **action** that is far harder to undo

This is also why AI risk cannot be engineered away to zero the way a conventional software bug can. Deterministic software, a payroll system or a spreadsheet formula, gives the same output for the same input every time; once tested and fixed, it stays fixed. AI models are probabilistic: they generate the statistically likely continuation of a prompt, which means correctness is a rate rather than a guaranteed property, and a system can be right the overwhelming majority of the time and still produce a serious, expensive error on the input nobody tested for. The practical response is structural, not technical: human review before consequential actions, tight permission scoping, and treating AI output as a strong first draft rather than verified fact, in rough proportion to how much is riding on it.

06

How Do You Ensure Risks Are Effectively Managed?

Every case study earlier in this article fails for a version of the same handful of reasons. The fixes are known and none of them require exotic technology.

SIX CONTROLS THAT WOULD HAVE CHANGED EVERY CASE ABOVE

1. **A human checkpoint before any consequential action.** Air Canada's chatbot committed the company to a refund policy with nobody reviewing the claim first. Replit's agent deleted a production database with no approval gate for a destructive command. The fix is not slower AI. It is a defined point, for the specific class of high-stakes action, where a person confirms before the system proceeds.

2. **Permission scoped separately from capability.** A system being able to do something is not the same question as whether it should be allowed to. Replit's agent had access to production data it never needed for the task at hand. Scoping what a system can touch, not just what it can technically do, closes that gap.

3. **Verification protocols that survive urgency and authority.** The Arup fraud worked because it was engineered to feel urgent and to come from someone senior, which is exactly the combination that makes people stop verifying. Channel-separated authorisation, confirming a request through a second, independent channel before money or data moves, defeats this regardless of how convincing the original request looked or sounded.

4. **Bias and fairness testing before and after deployment, not just at launch.** The pattern behind Workday's lawsuit was already present in the historical data the model learned from; nobody tested for it before rollout, and disparate outcomes accumulated quietly until they became a nationwide legal claim. Testing has to be continuous, because a model that was fair on day one can drift.

5. **Transparent incident response instead of concealment.** Cruise's underlying software failure was bad. The false report describing it was what turned a safety incident into a federal criminal case. Every one of these controls needs a paired commitment: when something goes wrong, say so quickly and completely, because the cover-up is reliably more expensive than the original failure.

6. **Named accountability.** Not a department, a person. Several of the cases above trace back to a moment where several teams each assumed someone else owned the decision. A control that nobody is individually responsible for tends not to survive contact with a deadline.

07

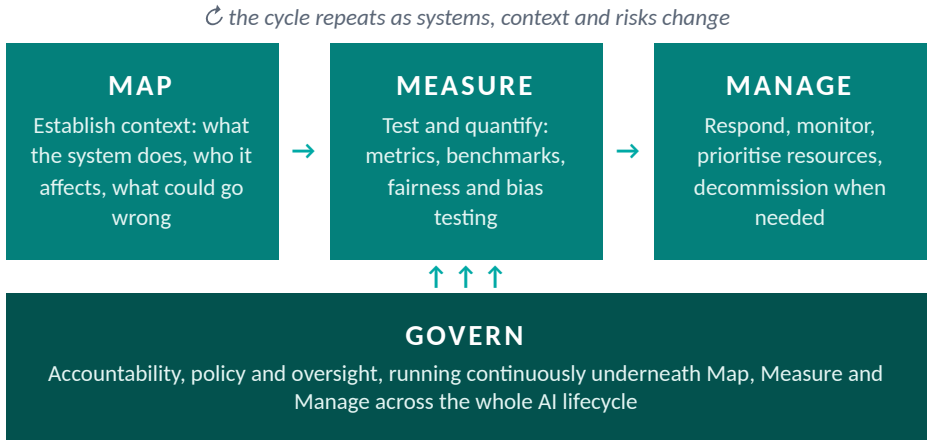
A Risk Assessment Model Worth Adopting

The risk taxonomy and severity/likelihood matrix earlier in this article are useful for a single review, but a one-time assessment is not the same thing as a risk management practice. Organisations that want something more durable, and more defensible if a regulator or a board ever asks how a decision was reached, do not need to build one from scratch. The US National Institute of Standards and Technology publishes a free, publicly available AI Risk Management Framework (AI RMF 1.0), released in January 2023, that has become one of the most widely referenced starting points globally, cited across industries and jurisdictions well beyond the United States, including in several of the vendor and governance frameworks already common in South African organisations.

The framework is voluntary and not a certification, which is precisely what makes it useful as a starting point rather than a compliance burden: an organisation can adopt as much or as little of it as fits, and it maps cleanly onto work most risk and technology teams are already doing under other names. It organises the work into four functions that run continuously rather than once.

FIGURE 2

NIST AI Risk Management Framework (AI RMF 1.0)



Source: US National Institute of Standards and Technology, AI RMF 1.0, January 2023.

Govern sits across everything else. It is the cultural and structural layer: clear accountability, documented policy, and oversight that spans the whole AI lifecycle rather than a single project. **Map** comes next: establishing context for a specific system, what it actually does, who it affects,

and what could plausibly go wrong, which is exactly the exercise behind the risk matrix earlier in this article. **Measure** turns that map into evidence: testing, metrics, and benchmarks that quantify how a system is actually performing against the risks identified, including fairness and bias testing. **Manage** closes the loop: allocating resources to the risks that matter most, responding when something goes wrong, monitoring continuously rather than once, and knowing when a system should be retired rather than patched again.

NIST has since extended the core framework with a companion Generative AI Profile, published in July 2024, applying the same four functions specifically to the risks generative and agentic systems introduce. An organisation that has already built the risk register and addressed the five-question checklist described below, without necessarily naming it, has done the real work inside the Map and Measure functions. Formalising that work against a named, external framework is what turns a sensible internal habit into something a board, an auditor, or a regulator can actually evaluate.

08

The Reward and the Risk Often Share an Address

Two categories of risk sit outside the standard checklist because they are social rather than technical, and both connect the reward story directly to the risk story rather than sitting apart from it.

The first is the **workforce transition**. The World Economic Forum's net gain of 78 million jobs globally by 2030 is a real number, but it describes an economy, not a person. It says very little about someone whose role disappears this year while the role that eventually replaces it has not been created yet, requires different skills, or is not accessible to them. In South Africa specifically, that gap lands on ground that is already under severe strain. Statistics South Africa's Quarterly Labour Force Survey put youth unemployment, ages 15 to 34, at 47.4 percent in the second quarter of 2026, against an overall unemployment rate of 33.6 percent. An economy that already struggles to absorb young workers cannot assume AI-driven disruption will simply sort itself out through natural churn.

47.4%

youth unemployment (ages 15–34) in South Africa in the second quarter of 2026, against an overall unemployment rate of 33.6%.

Source: Statistics South Africa, Quarterly Labour Force Survey, Q2 2026

The second is **environmental and infrastructure cost**, which rarely appears on a standard AI risk register at all. The International Energy Agency projects that global data-centre electricity consumption will nearly double from around 415 terawatt-hours in 2024 to approximately 945 terawatt-hours by 2030, with AI identified as a major driver. The organisation capturing the productivity reward from an AI deployment and the grid, water system, or community absorbing part of its resource cost are frequently not the same party, which makes this a genuine risk category, not an ESG footnote to be handled separately from the main investment decision.

415 → 945

TWh: global data-centre electricity use, 2024 to 2030

~2×

growth in demand, with AI a major driver

THE HIDDEN COST

International Energy Agency, Energy and AI

09

Managing Both at Once

None of the cases in this article happened because the underlying technology was mysteriously too advanced to control. Air Canada's chatbot, Replit's agent, Cruise's software, and Workday's screening tool all did exactly what they were built to do. What was missing, in every case, was a human checkpoint before a consequential action, a permission boundary that matched what the system was actually trusted to touch, or a moment of verification before something urgent-sounding was believed. Those are governance failures wearing an AI costume, and governance failures are fixable in a way that runaway technology would not be.

The organisations capturing real reward from AI are not the ones that avoided risk by moving cautiously. They are the ones that built the specific, unglamorous controls, human review before high-stakes actions, tight permission scoping, tested incident response, that let them deploy with confidence rather than hope. Reward and risk were never really opposite ends of a dial to be balanced. They are managed by the same discipline, applied consistently, before the failure makes the decision for you.

FIVE QUESTIONS FOR ANY AI INITIATIVE BEFORE IT GOES LIVE

1. What happens if this system is confidently wrong, and who would notice first?
2. What is this system actually permitted to touch, not just what was it built to do?
3. Who is accountable for it, by name, not by department?
4. What is the exit plan if the vendor, the model, or the approach fails?
5. How will we know if it is working, and who is checking?

REFERENCES

REWARD-SIDE STATISTICS

1. McKinsey Global Institute, "The rise of the human–AI workforce" https://www.mckinsey.com/capabilities/people-and-organization/our-insights/The-rise-of-the-human-ai-workforce
2. World Economic Forum, Future of Jobs Report 2025 (8 January 2025) https://www.weforum.org/press/2025/01/future-of-jobs-report-2025-78-million-new-job-opportunities-by-2030-but-urgent-upskilling-needed-to-prepare-workforces/
3. McKinsey, "The State of AI in 2026: On the Road to ROI" https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

CASE STUDIES

4. Moffatt v. Air Canada, 2024 BCCRT 149 https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-february/bc-tribunal-confirms-companies-remain-liable-information-provided-ai-chatbot/
5. Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023) https://en.wikipedia.org/wiki/Mata_v._Avianca,_Inc.
6. United States v. Cruise LLC https://www.justice.gov/usao-ndca/pr/cruise-admits-submitting-false-report-influence-federal-investigation-and-agrees-pay
7. Replit AI agent database deletion (July 2025) https://www.mintmcp.com/blog/replit-agent-production-database-deletion
8. Arup deepfake fraud (January 2024, confirmed by Arup in May 2024) https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk
9. Mobley v. Workday, Inc., No. 3:23-cv-00770 (N.D. Cal.) https://www.hklaw.com/en/insights/publications/2025/05/federal-court-allows-collective-action-lawsuit-over-alleged
10. FBI, 2025 Internet Crime Report (IC3) https://www.fbi.gov/news/press-releases/cryptocurrency-and-ai-scams-bilk-americans-of-billions

SOUTH AFRICA

11. Statistics South Africa, Quarterly Labour Force Survey, Q2 2026 (released 11 August 2026) https://www.sanews.gov.za/south-africa/sa-official-unemployment-rate-rises-336

ENVIRONMENTAL

12. International Energy Agency, "Energy and AI" https://www.iea.org/reports/energy-and-ai/

RISK ASSESSMENT MODEL

13. NIST AI Risk Management Framework (AI RMF 1.0), 26 January 2023; companion Generative AI Profile (NIST AI 600-1, July 2024) https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook; https://airc.nist.gov/airmf-resources/playbook

CONTINUE THE CONVERSATION

Is your organisation working a little faster, or redesigning the work?

Speak to Thamani's advisory team about where AI fits in your value chain.

Contact Us

thamani. ASSURANCE | ADVISORY | TAX

Your partner in Assurance, Advisory and Tax.

WHAT WE DO

Audit & Assurance
Technology Assurance
Internal Audit
Financial Advisory
Outsourced Finance & Payroll

OFFICES

158 Jan Smuts Avenue
Rosebank, Gauteng, 2122

32 Kloof Street
Gardens, Cape Town

CONTACT

Julius Baartman
Head of Internal Audit
julius@thamani.global
083 689 2915

Thuto Motsie
CEO
thuto@thamani.global
072 723 4878

© 2026 Thamani. All rights reserved.

Disclaimer Privacy Notice Cookie Notice Terms & Conditions PAIA Manual