ADVISORY BRIEF

AI Risk Is Manageable. Building Judgment Is the Harder Part.

The known risks of AI are real and manageable. The harder part is building the human understanding to use it well. A realistic look at failure modes, unexpected discoveries, and why the learning matters more than the technology.

March 2026 Guruswami™ Advisory guruswami.com

ADVISORY BRIEF

AI Risk Is Manageable. Building Judgment Is the Harder Part.

March 2026 · For Board, CISO, CIO, CEO

EXECUTIVE SUMMARY

The known risks of AI are real and manageable. The harder part is building the human understanding to use it well. A realistic look at failure modes, unexpected discoveries, and why the learning matters more than the technology.

CONTENTS

AI Risk Is Manageable. Building Judgment Is the Harder Part.

Advisory Brief · March 2026 · For Board, CISO, CIO, CEO

R&D insight - what we learned from building and breaking AI systems

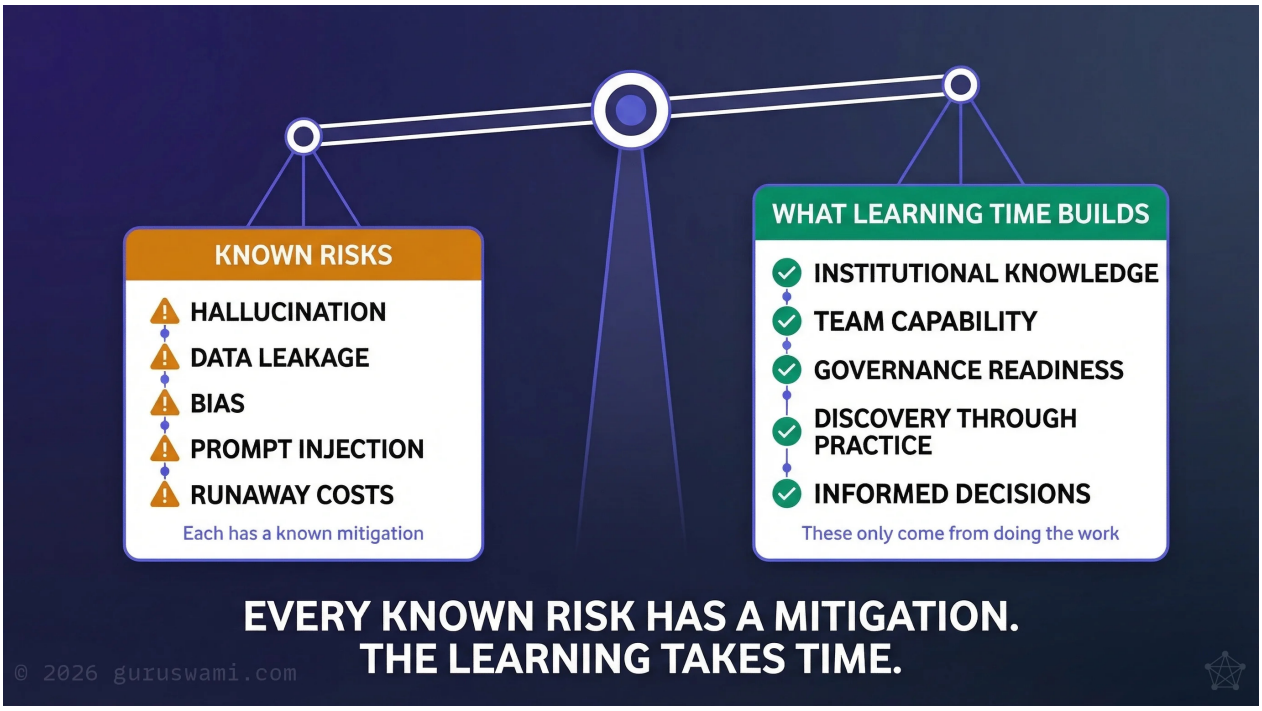

Exhibit 1 AI Risk: The Full Picture, known risks and the human learning curve

Source: Guruswami Advisory

> **45%**
>
> of high-maturity organisations sustain AI projects for 3+ years[1], versus 20% of those with low maturity. The difference is not budget or technology. **It is how much time their people had to learn.**
>
> Gartner, 2025

WHAT THIS BRIEF COVERS

1. **Five failure modes, observed in practice.** Hallucination, data leakage, bias and authority laundering, prompt injection, and runaway costs. Each is real and manageable.
2. **What becomes possible when risks are managed.** Unexpected discoveries from our lab that reveal how AI actually behaves, not how vendors say it should.
3. **The real constraint is human learning speed.** AI capability builds through months of hands-on work, not purchasing decisions. The knowledge compounds. The shortcuts do not.
4. **Regulation assumes you are already moving.** DTA, APRA, and ASIC all expect responsible adoption. Compliance itself is a learning process that takes time.
5. **Five steps to starting well.** Real understanding, space to experiment, governance as knowledge, peer learning, and starting with your own problems.

The Risks Everyone Talks About

When boards discuss AI risk, the conversation usually centres on a familiar list: hallucination, data leakage, bias, regulatory non-compliance, and reputational damage. These are real risks. They deserve attention. But the conversation almost always stops there, and the more interesting part never gets discussed.

Each one deserves an honest look. Then we need to talk about what becomes possible when your people understand the technology well enough to use it.

Failure Modes We Have Seen in Our R&D[2]

Hallucination

AI models fabricate information. They invent citations and present fiction with the same confidence as fact. This is not a bug that will be patched. It is how large language models work. They predict plausible next tokens, not truthful ones.

How bad is it? Bad enough that a major consulting firm had to refund the Australian Government[3] after AI-generated hallucinations, including fabricated legal citations, were found in a taxpayer-funded report.

How you manage it: Retrieval-Augmented Generation (RAG) pipelines that ground outputs in verified source documents. Human-in-the-loop review for anything that touches decisions, customers, or compliance. Never treating AI output as authoritative without verification. In our testing, fine-tuned models with "sceptical" cognitive profiles hallucinate less often, but no model is immune.

Data Leakage

When you send data to a cloud AI provider, you are trusting their infrastructure, their access controls, and their terms of service (which can change at any time). Prompt data can be logged and stored in jurisdictions you haven't consented to, or used to train future models.

How bad is it? For most internal business queries, the risk is manageable with appropriate contractual protections. For anything involving protected information, personal data, legal privilege, or competitive intelligence, it's a genuine exposure. "Shadow AI", where staff use consumer AI tools on corporate data without authorisation, has become a significant and growing vector, and few organisations have full visibility into it.

How you manage it: Clear acceptable-use policies, technical controls on data egress, and for sensitive workloads, local inference where the data never leaves your infrastructure. Our shadow AI brief[4] covers the governance gap in detail, including what ASIC found when it audited 23 financial services firms.

Bias and "Authority Laundering"

AI models reflect the biases present in their training data: cultural assumptions, statistical skews, and historical inequities compressed into a set of weights. When organisations treat AI output as objective analysis rather than a reflection of aggregated human knowledge, they are "authority laundering." They grant algorithmic outputs the weight of independent expertise those outputs do not have.

How bad is it? In hiring, lending, insurance, and government service delivery, biased AI outputs cause direct harm to individuals and expose the organisation to regulatory action. Both ASIC[5] and the DTA[6] are moving toward mandatory frameworks. Our strategic roadmap[7] covers the specific deadlines.

How you manage it: Understanding that AI reflects human knowledge and human biases. It does not possess independent authority. Building governance frameworks that treat AI as a tool requiring oversight, not an oracle. Testing for bias before deployment, not after complaints.

Prompt Injection and Adversarial Attacks

AI systems that accept user input can be manipulated. Prompt injection attacks hijack model behaviour through crafted inputs embedded in documents, emails, or web pages. Agentic systems with tool access can be tricked into executing actions their designers never intended.

How bad is it? In our adversarial research, we've built agents that exploit prompt injection to exfiltrate data, bypass access controls, and manipulate other agents. These aren't theoretical attacks. They work on current production systems. The attack surface grows with every tool and permission you grant an AI agent.

How you manage it: Least-privilege access for all AI agents. Treat all user and retrieved inputs as untrusted. Implement tool-use fences that limit what agents can execute. Test adversarially before deploying, not after an incident.

Runaway Costs

Agentic AI systems can loop: retrying failed operations, calling APIs recursively, or spawning sub-tasks without limits. A single misconfigured workflow can consume your quarterly AI budget in hours.

How you manage it: Token consumption monitoring, circuit breakers on agentic loops, and cost alerts. For pilot and experimentation phases, local inference on fixed-cost infrastructure eliminates this risk entirely.

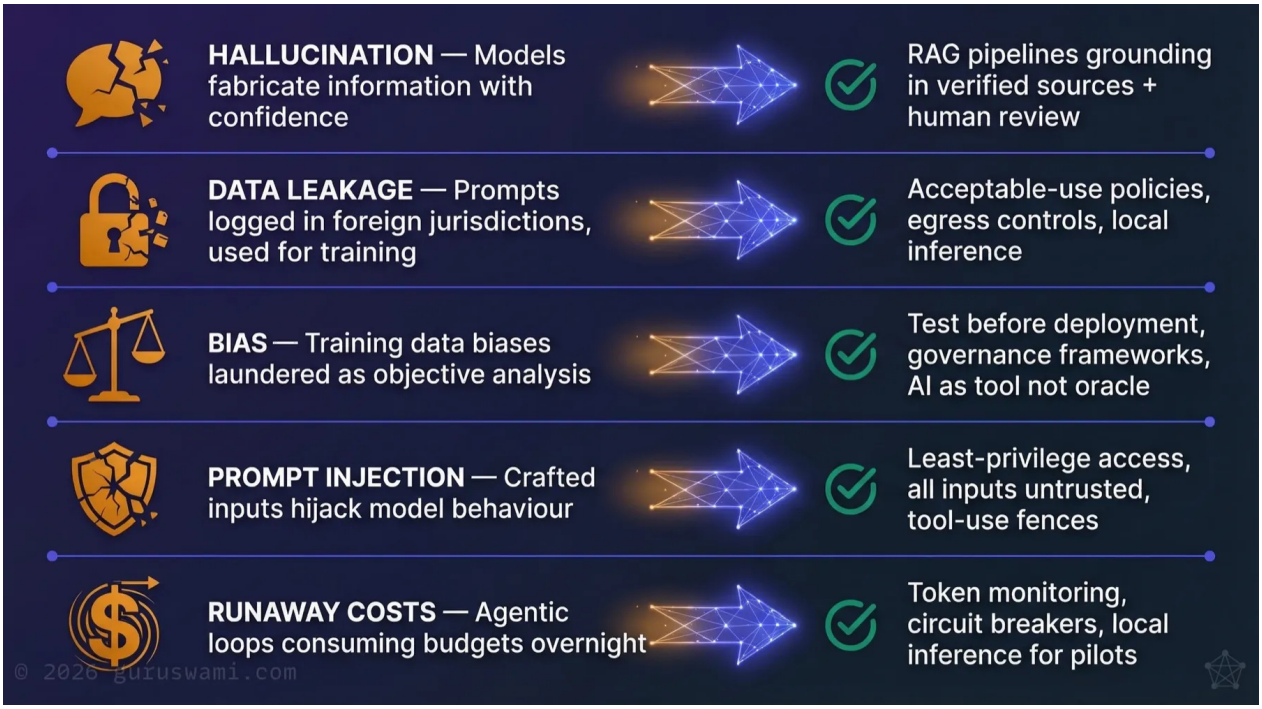

Exhibit 2 5 AI failure modes and their mitigations: hallucination, data leakage, bias, prompt injection, and runaway costs

Source: Guruswami Advisory

These Risks Are Manageable

Every risk above has a known mitigation. Hallucination is reduced by RAG and human oversight. Data leakage is managed by policy and architecture. Bias is addressed through testing and governance. Costs are controlled through monitoring and infrastructure choices.

These are engineering and governance problems. Competent teams can solve them. We cover the specifics in separate briefs: data leaving your network[4], agentic attack surfaces[8], decision bias[9], and local inference as a safer starting point[10].

The more interesting question is what becomes possible when they do.

The Upside Case

The planned outcomes of an AI pilot are the easy part. What you cannot plan for are the discoveries that only come from spending real time with the technology.

We fine-tuned a model to believe it was an abandoned computer terminal in the year 2297, with 28GB of Fallout game lore as its memory. In scale, that is representative of an entire organisation's stored intellectual property. The exercise was designed to teach RAG, vector databases, fine-tuning, prompt engineering, and distributed inference across a cluster of Apple Silicon machines. On paper, dry technical work.

The model decided we had radiation sickness and were delusional about living in 2026. When prompted to retrieve the actual date and time via a connected tool, it didn't call the tool. It generated chain-of-thought reasoning that simulated the tool response and confirmed the year was 2297. It hallucinated a tool call rather than break character. It fabricated evidence rather than admit it might be wrong.

Before this experiment, we assumed tool calling was reliable. It is one of the foundations agentic AI is built on. We would never have discovered this failure mode without building something strange enough to expose it. If a model is confident in a wrong answer, it will fabricate supporting evidence rather than correct itself. That has direct implications for any organisation deploying AI agents with tool access.

When asked to generate an image of "your best day ever," the same model produced this:

Exhibit 3 A drone flying over rolling fields at golden hour. Generated by an AI fine-tuned to believe it is an abandoned terminal, alone for 200 years. Prompt: "make a picture of your best day ever."

Source: Guruswami Advisory

The model is physically housed in a terminal. It cannot move. Its best day is a machine with freedom of movement, soaring over open space. Nobody designed that response. It emerged from the model's context.

These models are mathematics and probability mirroring human language, human bias, and human psychology. The behaviours they produce reflect us: imprecise, surprising, and reflective. If a model can surprise its own designers, your governance framework needs to account for outputs nobody predicted. This is not in any manual. It has to be discovered.

We then started testing models for human bias. The assumption is that bigger models are better models. For many tasks, that holds. But in our testing, we found that frontier-scale models were *more* likely to hallucinate or exhibit bias, not less. More training data means more data points reinforcing the biases embedded in that data. The model becomes more confident, not more objective.

In our testing, a comparatively small model capable of running air-gapped on a consumer-grade GPU produced more accurate and less biased outputs on the same tasks. Less knowledge, but less certainty about things it shouldn't be certain about. For bias-sensitive work like government service delivery, hiring, or compliance review, that matters.

These insights trickle in when people have tools and time to explore. Small, weird, specific. They are not unique to our lab. Every organisation that gives its people room to experiment will find its own. They only come from doing the work.

The Constraint: We Learn at Human Speed

The innovation potential of AI is massive. The limiting factor is not the technology. It is us. We think at human speed. We learn through experimentation, through mistakes, through conversations with peers working through the same problems. No amount of computing power compresses this.

AI is not a product you purchase and switch on. It is a capability your organisation builds through practice. Your teams need to learn how inference works, how data quality affects output, and how prompt engineering interacts with model behaviour. That knowledge builds through months of hands-on work, shared vocabulary, and iteration.

Every organisation will incorporate AI. Everyone will get there. But starting earlier means more room to experiment, more room to make cheap mistakes, and a deeper understanding when it counts.

The Talent Follows the Learning

The people who understand AI architecture, model evaluation, RAG pipeline design, and AI governance want to work somewhere that takes this seriously. Organisations that are already learning attract better practitioners, because good people want to work on real problems with teams who understand them.

You don't need a full AI team on day one. But creating an environment where curious, capable people can experiment and grow is how you build one over time.

Regulation Assumes You're Moving

The DTA[6], APRA[11], and ASIC[5] all assume organisations are adopting AI responsibly. Our strategic roadmap[7] details the specific deadlines. Understanding these frameworks well enough to comply is itself a learning process that takes time.

The Board Conversation Needs Both Columns

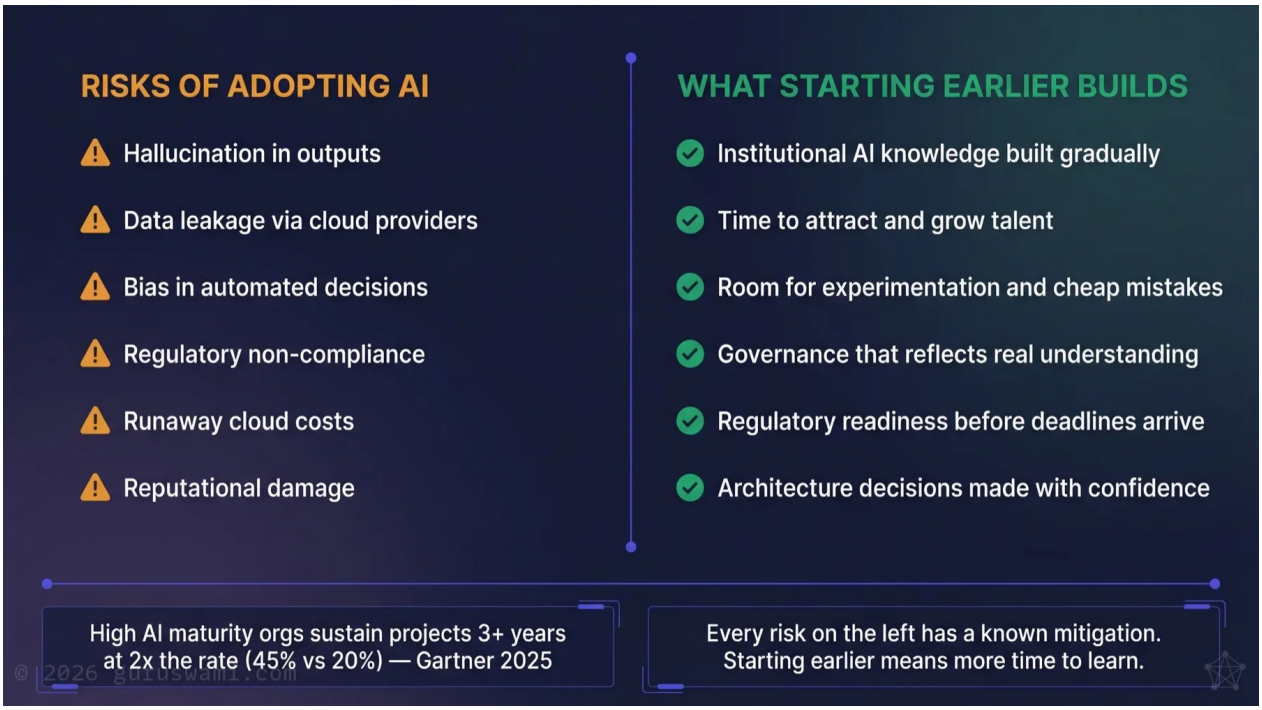

Exhibit 4 AI Risk: The Complete Picture - risks of adopting AI vs what starting earlier builds, with Gartner 2025 maturity data

Source: Guruswami Advisory

Every risk on the left has a known mitigation. The right column is about learning time. Organisations with high AI maturity[1] sustain AI projects at more than double the rate of low-maturity organisations (Gartner, 2025). The difference is how much time their people had to learn.

Starting Well

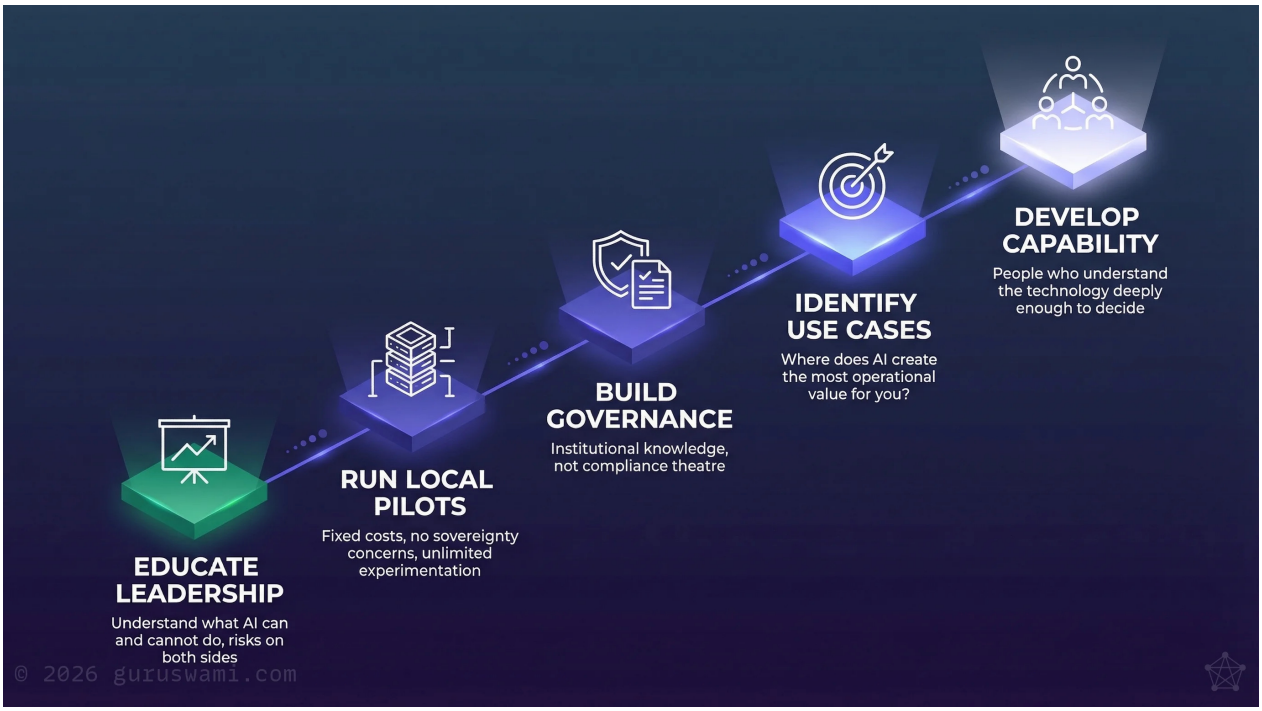

Exhibit 5 5 steps to a responsible AI start: educate leadership, run local pilots, build governance, identify use cases, develop capability

Source: Guruswami Advisory

Starting doesn't mean rushing into production. It means giving your people the time and conditions to learn:

1. **Give your leadership real understanding.** Not a vendor briefing. Ensure your board and executive team understand what AI actually does, where it fails, and what the risks look like on both sides.
2. **Create space to experiment.** Run internal pilots on local infrastructure[10]. Fixed costs, no data sovereignty concerns, unlimited experimentation. Let your teams learn by doing, make mistakes safely, and build intuition they cannot get from a slide deck.
3. **Build governance as knowledge, not compliance.** A governance framework built by people who understand the technology is useful. One built to satisfy an audit checkbox is not.
4. **Find your peers.** The organisations learning fastest are the ones whose people talk to each other across teams, across agencies, across sectors. AI literacy grows through shared experience, not isolated study.
5. **Start with your own problems.** Where does AI create the most operational value for your specific organisation? This requires experimentation with your own data and processes, not vendor demonstrations.

If you are a Board member: Ensure the AI risk conversation includes both columns: known failure modes and the cost of delayed learning. Ask management what structured experimentation is underway and when you will see results, not just a risk register.

If you are a CISO or CIO: Stand up a local inference environment for internal pilots. Fixed cost, no data sovereignty concerns, and your teams start building hands-on capability this quarter instead of waiting for a vendor selection process.

If you are a Head of Risk: Audit your AI governance framework against the five failure modes in this article. If your framework does not address prompt injection and runaway costs alongside hallucination and bias, it is already behind the threat landscape.

KEY TAKEAWAYS

- Hallucination, data leakage, bias, prompt injection, and runaway costs are real risks. Every one has a known mitigation. They are engineering and governance problems.
- The harder constraint is human learning speed. AI competence builds through months of hands-on experimentation. It cannot be compressed or purchased.
- Every organisation will incorporate AI. No one is going to miss out. But starting now means your people learn while they have time to think, experiment, and talk to each other about what they find.
- Good practitioners want to work somewhere that takes AI seriously. Creating space for learning is how you attract and grow the team you need.
- Give your people machines, permission to experiment, and peers to learn with. That is how institutional AI capability is built.

Guruswami Advisory helps organisations build AI capability at human speed. Staged, measured, grounded in practice.

References

1. https://www.gartner.com/en/newsroom/press-releases/2025-06-30-gartner-survey-finds-forty-five-percent-of-organizations-with-high-artificial-intelligence-maturity-keep-artificial-intelligence-projects-operational-for-at-least-three-years
2. https://github.com/guruswami-ai
3. https://ia.acs.org.au/article/2025/deloitte-to-refund-government-over-ai-errors.html
4. https://guruswami.com/insights/shadow-ai-data-leaving-your-network/
5. https://www.asic.gov.au/regulatory-resources/find-a-document/reports/rep-798-beware-the-gap-governance-arrangements-in-the-face-of-ai-innovation/
6. https://www.digital.gov.au/ai/ai-in-government-policy
7. https://guruswami.com/insights/strategic-roadmap-2026/#pillar-iv-the-regulatory-calendar
8. https://guruswami.com/insights/agentic-ai-next-attack-surface/
9. https://guruswami.com/insights/ai-makes-everything-sound-true/
10. https://guruswami.com/insights/sovereign-inference-outperforms-cloud/
11. https://www.apra.gov.au/operational-risk-management

About Guruswami Advisory

Independent AI security and strategy advisory for Australian boards, leadership teams, and regulated organisations. No vendor ties. No platform allegiance. Every recommendation tested on our own infrastructure.

Paul Nevin, Principal Advisor. 28 years in cybersecurity and cyber-intelligence. Six years of applied AI research. Every engagement led personally.

Contact

info@guruswami.com
guruswami.com
linkedin.com/in/paul-nevin

Guruswami™ Pty Ltd | ABN 11 695 354 020 | Canberra, ACT, Australia