Actuaries Institute.

UTS Human Technology Institute

AI Risk Management in the Financial Services Sector

September 2026

The Actuaries Institute

As the peak professional body for actuaries in Australia, the Actuaries Institute represents the profession to government, business and the community. Our members work in a wide range of fields including insurance, superannuation and retirement incomes, enterprise risk management, data analytics and AI, climate change and sustainability, and government services. This paper has been prepared as part of the Actuaries Institute's Public Policy and Thought Leadership program. For further information on this resource contact public_policy@actuaries.asn.au

The Human Technology Institute

The Human Technology Institute (HTI) at the University of Technology Sydney is an impact-oriented institute building human values into new technologies. Bringing together policy, legal and technical experts, HTI provides independent expert advice, policy development, capability building, and data science solutions to support government, industry and civil society. For further information on this resource contact hti@uts.edu.au

About this paper

This paper provides an operational AI risk management framework for risk professionals and executives in Australian financial services. It is designed to sit within existing enterprise risk management structures rather than alongside them. The paper assumes working familiarity with ISO 31000 and Australian Prudential Regulation Authority (APRA) prudential standards (CPS 220, CPS 230, CPS 234), and an understanding of existing AI governance frameworks, including NIST AI RMF, ISO/IEC 42001 and the Australian Government's Guidance for AI Adoption.

In addition to the experience and expertise of the authors, this paper draws on three main sources: the UTS Human Technology Institute Financial Services Survey (HTI FS Survey); a workshop with 15 financial services practitioners from banks, superannuation and insurance firms in August 2025; and desktop research.

For further information on the HTI FS survey, see the Executive Summary and Appendix A.

Acknowledgements

The Actuaries Institute thanks Victor Bajanov FIAA, Chris Dolman FIAA, Jonathan Shen FIAA and Michael Storozhev FIAA for their assistance in producing this framework, and Elayne Grace, Vanessa Beenders and Clare Marshall.

The UTS Human Technology Institute thanks Llewellyn Spink, Jack Goldsmith, Clare Brennan, Veve Fry, and Professor Nicholas Davis for their contributions.

The Actuaries Institute and UTS Human Technology Institute extend our thanks to industry experts for their assistance in the early stage workshop, including Nick Barnett, Andrew Beevors, Nagender Chetti, Tony Dang, Jade Haar, Minh Phan, Chris Scheuber, Jared Spowart and Phoebe Twiggs.

Acknowledgement of Country

The Institutes acknowledge the traditional custodians of the lands and waters where we live and work, travel and trade. We pay our respect to the members of those communities, Elders past and present, and recognise and celebrate their continuing custodianship and culture.

ISBN: 978-1-7643622-8-3

Suggested Citation: Actuaries Institute & UTS Human Technology Institute. (2026). *AI Risk Management in the Financial Services Sector.* Actuaries Institute and UTS Human Technology Institute.

AI Assistance Disclaimer: This paper was prepared with the assistance of AI tools, including large language models. The authors drafted material, set the structural direction, and developed the substantive arguments and recommendations. AI tools were used to support research synthesis, accelerate initial drafting, test structural options, and refine the writing. The drafting process was iterative. All efforts were made to ensure the contents of this document were reviewed, substantially revised, and approved by appropriately expert and credentialed human authors.

Contents

Foreword

Across the financial services sector in Australia, AI adoption is progressing faster than AI governance. Regulators are aware of this gap, and have increasingly clear expectations that organisations establish effective governance arrangements and appropriately manage AI risk.

For financial service organisations, meeting these expectations is not only a matter of regulatory compliance. Effective governance and risk management are critical to unlocking the benefits of AI, building trust in AI, and supporting its safe, responsible and effective use. Put simply, effective governance is critical to optimising innovation and creating value from AI investments.

As critical infrastructure, the financial services sector has an especially high responsibility to govern AI well. When financial institutions can deliver safer, faster and more efficient services, the benefits spread widely across Australian businesses, consumers and the broader economy. Conversely, if they fail in this task, the negative impact across individuals and the economy is significant.

While the sector has used AI and automation for decades, emerging AI systems, particularly generative and agentic AI, do not behave in the same way as previous technologies. Previous technologies operate deterministically. Frontier AI systems operate probabilistically, adding greater uncertainty to their behaviours, and increasing the need for effective oversight. In addition, agentic AI systems create more distance between the responsible humans and the AI system, and narrow the window for meaningful human intervention. The result is a set of new and amplified risks for the sector.

These risks sit at what the Actuaries Institute of Australia and the UTS Human Technology Institute call the 'AI-delta', where existing governance and risk management arrangements do not match the nature of the technology and its implications. We have sought to address the AI-delta governance gap in financial services through this resource. Our partnership brings together deep expertise in financial risk management and human-centred AI governance, both of which are needed to respond to what AI is now demanding from the sector.

The framework that follows is deliberately not another stand-alone scheme. It is intended as an AI-specific overlay for the financial services sector that applies existing risk management principles to the AI-delta. It integrates with, and builds upon, existing standards and guidance while introducing new approaches to understanding, quantifying, and managing AI risk.

Effective AI governance and risk management are vital to ensure financial services can continue to support a healthy and innovative Australian economy and society.

Elayne Grace
Chief Executive Officer
Actuaries Institute

Professor Nicholas Davis
Co-Director
UTS Human Technology Institute

Executive Summary

AI is central to the financial services infrastructure that underpins our day-to-day lives

The financial sector represents 7.7% of Australia's GDP,[1] and its infrastructure shapes our lives: such as how we save and spend, who receives credit or insurance and on what terms, and what retirement looks like.

AI is now deeply embedded across this infrastructure. Credit scoring models assess loan applications. Fraud detection systems block suspicious transactions. Chatbots handle customer enquiries. Financial services organisations are deploying AI to improve productivity, automate business processes, and enhance customer service.[2]

When deployed well, AI offers significant benefits for institutions and customers alike, improving efficiency, speed and quality of service. Yet failures in these systems risk significant harm. Discriminatory algorithmic systems may unfairly deny or limit credit. Customer-facing chatbots that hallucinate policy terms can leave customers incorrectly believing they have insurance cover. Fraud detection systems may exclude certain communities from access to financial services.

Given the scale of opportunity, risks, and impact of AI in financial services, the Actuaries Institute of Australia and the University of Technology Sydney Human Technology Institute (**HTI**) have partnered to develop this practical framework on AI risk management for the sector.

AI systems are complex and fast-changing, creating new risks and amplifying existing ones

AI introduces new risk characteristics and amplifies a broad range of existing risks, including conduct, model, strategic, data, cyber, regulatory, operational, and third-party risks. For many organisations, the risks are material, making AI a board-level issue.[3]

AI systems, particularly generative AI systems built on large language models, introduce risk characteristics that differ materially from traditional technologies. Unlike systems that perform a defined function in a defined part of the business, generative AI is a general-purpose technology that can be used in diverse use cases across the organisation. Outputs are probabilistic rather than deterministic, their behaviour can shift with model updates outside the deploying organisation's control, and their failure modes are often difficult to detect through conventional monitoring and can compound across interconnected processes.

Box 1: What's the difference between "deterministic" and "probabilistic"?

Traditional software follows fixed rules: the same input always produces the same output, every time. Typing 2 + 2 into a calculator will always return 4. This is deterministic behaviour.

Generative AI systems work differently. Rather than following fixed rules, they generate responses by predicting likely outputs based on patterns learned from vast amounts of training data. Ask the same question twice and you may get two different answers. This is probabilistic behaviour.

The governance challenge is significant. With deterministic systems, failure modes are consistent, so they can be reliably reproduced and tested for. With probabilistic systems, outputs can be unexpected, inconsistent, or subtly wrong in ways that are difficult to detect.

Agentic AI further complicates this risk profile. These systems can act on their own, increasing the speed and scale of errors. When AI supports human decision-making, accountability and responsibility generally remains with the human who makes the decision (putting to one side debate about the effectiveness of human-in-the-loop in some contexts).[4] Where agentic systems make decisions autonomously, the opportunity for human intervention is reduced and consequences may be amplified, creating a need for stronger controls and clearer accountability.

Technological advances in AI are also materially increasing risks for the sector, particularly in relation to cyber security. AI-enabled offensive cyber capabilities[5] are enabling the identification and exploitation of software vulnerabilities at unprecedented speed and scale with significant implications for organisations' information security and operational resilience.[6]

The pace of change compounds these challenges. Organisations that have rolled out generative AI are now turning to agentic systems, and further advances will continue to reshape the risk landscape. Rapid improvements in AI models mean risk assessments can quickly become outdated following updates to the underlying model.

AI risk management has not kept up with the pace of adoption

Despite the growing use and importance of AI, governance and risk management practices have generally not kept pace with the scale, complexity, and evolving risk profile of AI systems.

In 2025, HTI commissioned a survey of 419 directors, senior executives, and key decision-makers exploring AI adoption and governance practices across industries.[7] The results included 27 respondents from financial services organisations. The findings from this subset are referred to in this paper as the **'HTI FS Survey'.** As the sample is small and skews towards larger organisations, these findings are used as directional intelligence, rather than evidence of sector-wide practices.

Findings from the HTI FS Survey suggest a significant gap between AI adoption and governance. While 93% of respondents reported using AI, and the remainder planned to do so, less than half reported conducting risk assessments on their internal use of AI, fewer than one in three respondents reported AI-specific items on their risk register, and fewer than one in four had included AI in their risk appetite statement. Yet 74% of respondents rated their current risk management as effective for AI.[8] While the absence of these controls is not determinative, their relatively limited adoption raises questions about the basis for the confidence expressed by these respondents in their AI risk management practices.

Regulators have also expressed concerns regarding AI governance and risk management in the sector. In 2024, Australian Securities and Investments Commission (ASIC) reviewed the AI governance arrangements of 23 financial services and credit licensees, finding that not all were positioned to manage the challenges of their AI use, with weaknesses in AI governance arrangements and risk management practices.[9] In late 2025, APRA conducted a targeted engagement on a group of selected large banks, insurers and superannuation trustees, finding that there are differing levels of maturity across governance, risk management and operational resilience and that assurance practices are not keeping pace with the scale, speed and complexity of AI. Importantly, APRA noted that 'while most entities recognise that existing prudential standards apply to AI risk, few have operationalised governance in practice'.[10]

The Council of Financial Regulators (CFR)[11] and key government agencies have increasingly emphasised the importance of addressing these risks, and clarified their expectations for how regulated entities should govern AI. In particular, ASIC has called for an urgent uplift in cyber resilience measures in response to growing AI-enabled cyber threats,[12] and APRA has stated that regulated entities must ensure appropriate risk management of AI, including setting risk appetite, managing AI related exposures, and ensuring appropriate oversight and accountability.[13]

Financial services organisations need a risk framework that responds to the complexities of AI

There is a range of AI governance frameworks and guidance from governments, regulators and standards organisations. While such frameworks are useful, none address the specific risk characteristics of AI systems operating within Australian financial services. This paper addresses that gap by outlining a practical and adaptable AI risk management framework for risk professionals and executives in Australian financial services. It introduces an AI-specific overlay but is designed to integrate with existing enterprise risk management processes, rather than replace them. It assumes that the deploying organisation operates a risk management process aligned with ISO 31000 and, where applicable, meets its CPS 220, CPS 230 and CPS 234 obligations. This framework applies existing risk management principles to the 'AI-delta': the way AI amplifies existing risks and introduces new ones.

The framework is not intended to be one-size-fits-all. It provides a mature risk management approach for an organisation with significant AI investments. Smaller organisations, or those early in their AI journey, should adapt the elements that are most relevant to their scale, complexity, and risk appetite.

Importantly, it responds to the fast-moving challenges of AI systems by anchoring governance to risk outcomes rather than technical specifications, an approach that remains valid as the technology evolves. The framework is designed to be used now, and to remain useful as the landscape continues to shift.

This resource provides a four-part practical framework for AI risk management

Each component is centred on a fundamental question on AI risks:

Governance: How do we assign accountability for AI risks in our organisation?

Classification: How should we classify AI risks in our organisation?

Quantification: How might we quantify the risks?

Controls: What controls are most appropriate for common AI use cases?

1. Governance: How do we assign accountability for AI risks in our organisation?

1.1 Principles for assigning accountability for AI risks

AI creates a distinctive accountability challenge because AI risks are often distributed across multiple functions and are highly context-dependent. This is particularly true for general-purpose and enterprise-wide AI systems, where the same underlying model may be used in different ways in different business contexts, each creating distinct risks.

The Australian Government[14] and regulators, including APRA[15], expect financial services organisations to assign clear accountability for AI systems and their risks. The practical question for organisations is where that accountability should sit.

The answer should follow the same logic that financial services apply to other risks: **there should be clear accountability for AI risk, and the person or function accountable for deploying AI should not be the sole arbiter of the risk it creates.** For example, there may be a Chief AI Officer responsible for operational deployment of AI systems, while the Chief Risk Officer has responsibility for independent oversight of the risks these create. This is the principle underpinning the Three Lines of Defence model, where operational ownership is kept separate from independent risk oversight and challenge.

The HTI FS survey indicates that the roles primarily responsible for ensuring that AI systems operate safely, lawfully and responsibly are often concentrated in technology and related functions, such as the Chief Technology Officer, Chief Information Officer, or Chief Data Officer.[16] This is understandable given that AI procurement and deployment decisions typically originate in technology teams. However, the function deploying the technology is rarely the function putting it to use in the business. In many large Australian organisations, while deployment may sit under the CTO, accountability for the value of the technology is with the business units that use it.

This structural disconnect is significant. It can mean that risks dependent on knowledge of specific business applications carry too little weight in deployment decisions. It can also impact the organisation's ability to make effective AI investments, where the value accrues to one part of the business but the cost sits with another. Effective AI governance requires technology functions, business units, and risk and compliance teams to work together, with shared responsibility for both creating value and managing risk.[17] Without this integration, both risk management and value capture suffer. The end result: innovation is not optimised.

For organisations using the Three Lines of Defence model, this provides a ready framework to allocate responsibility for risk management across the organisation.

1.2 Applying the Three Lines of Defence to AI systems

Organisations should carefully consider how the Three Lines of Defence model applies to their use and risk management of AI systems. Applying the model to AI systems reinforces accountability and helps integrate AI risk management into existing enterprise risk structures.

While the adaptation for AI is straightforward in principle, it requires deliberate implementation:

Line	Traditional Role	AI-Specific Adaptation
First line (business and technology)	Owns and manages risk in day-to-day operations	• Owns AI deployment decisions within risk appetite. • Conducts initial risk triage for AI use cases. • Implements technical controls (input validation, output monitoring, access controls). • Manages vendor relationships.
Second line (risk and compliance)	Sets risk framework, provides oversight and challenge	• Maintains AI risk taxonomy and assessment methodology. • Validates first-line risk triage decisions. • Conducts independent model validation for higher-risk AI use cases. • Monitors compliance with AI-related regulatory obligations. • Reports to the board or risk committee as appropriate.
Third line (internal audit)	Provides independent assurance	• Audits the effectiveness of AI risk controls. • Tests whether risk triage is being applied consistently. • Validates that AI deployments match approved use cases. • Reviews vendor due diligence documentation.

In practice, how this looks will depend on an organisation's size, maturity and existing processes. In a mature organisation doing significant AI work, relevant arrangements may include:

- The business and technology functions agreeing on the split of operational responsibility for AI deployment (for example, customer outcomes vs platform availability) within the organisation's risk appetite.
- The risk function providing independent oversight and challenge of AI deployments.
- A formal mechanism (e.g., an AI risk committee, or a standing agenda item on the existing operational risk committee) providing a structured forum for challenge and escalation.
- An AI centre of excellence or similar function partnering with domain experts across divisions to ensure AI initiatives are grounded in business context, not developed in isolation. They can also provide peer challenge, share learnings, and provide advice on policies and processes.
- Internal or external audit teams providing independent assurance.

For smaller organisations, or those with limited AI deployment, accountability might sit with a single senior leader who has both the technical understanding and the independence to challenge AI-related decisions. The structure matters less than the principle: there is an accountable person with appropriate authority who is responsible for the day-to-day operations, and another who can provide genuine independent challenge.

Box 2: Board-Level Considerations

Accountability for material AI risks ultimately rests with the board. APRA expects boards, at a minimum, to maintain sufficient understanding and literacy with respect to AI to set strategic direction and provide effective challenge and oversight.[18] CPS 220 requires the board to set the risk appetite and to satisfy itself that the risk management framework addresses material risks.

Where AI use is material to the organisation's operations, boards should consider whether their risk appetite statement addresses AI, whether they are receiving adequate reporting on AI risk exposure, and whether the risk management framework has been updated to reflect AI-specific risks. For some organisations, the board may reasonably conclude that AI is not yet a material risk.

However, the HTI FS Survey suggests that, as of October 2025, many boards had not yet turned their minds to the question: fewer than one in four respondents had addressed AI in their risk appetite statement.[19]

The board should be engaging with management about the use, risks and governance of AI systems in their organisation. In Report 798, ASIC included a set of governance questions (see Appendix C). The following are particularly relevant for the board to understand and discuss:

- Where on your AI journey are you? Do you know where AI is used in your organisation?
- Who is accountable for AI use and outcomes, at model level and overall? Do they get the reporting they need to do their job?
- How are you identifying, mitigating and monitoring risk throughout the AI lifecycle?
- Are your AI policies and procedures fit for purpose, now and for anticipated future use?
- How do you manage the challenges of relying on third parties?

For more guidance on what directors need to know and be asking management about AI governance, see A Director's Guide to AI Governance (2026), published by the Australian Institute of Company Directors and HTI.

2. Classification: How should we classify AI risks in our organisation?

Risk classification helps to situate how an organisation treats AI risk, including how it is managed, resourced, monitored, and reported. This framework advocates for a pragmatic approach that treats AI as an amplifier of existing risk domains while reporting it as a material risk. This reflects the operational reality of how AI risks manifest while preserving domain expertise and accountability across business functions.

2.1 Current AI classification in the risk taxonomy

How an organisation classifies AI risk within its risk taxonomy determines how it is managed, what resources are allocated to it, and how it is reported to the board and regulators. Currently, financial services organisations vary in how they classify AI risk, including as:

Material risk category. AI risk is treated as a standalone risk category at the same level as credit risk, market risk, or operational risk, with its own risk appetite, dedicated register entries, and direct board reporting.

Risk amplifier. AI is treated not as a separate risk but as a factor that amplifies existing risk categories. Under this approach, each existing risk category is reassessed for AI-related exposure, and AI-specific controls are added within those categories.

Thematic risk. AI is identified as a cross-cutting theme that is tracked across risk categories, typically through a thematic review or a standing agenda item on the risk committee, but without a dedicated risk appetite or standalone register entries.

Emerging risk. AI appears on the emerging risk register and is subject to periodic review, but is not yet integrated into the core risk taxonomy or appetite framework.

Unclassified. AI risk has not been formally classified. Risk management is ad hoc, relying on individual expertise and informal processes.

2.2 Treating AI operationally as a risk amplifier, but reporting as material risk

Consistent with prior Actuaries Institute work positioning AI as an enabler and tool, this paper suggests using the **risk amplifier** and **material risk category** approaches as complementary operating models. Many organisations will find that AI risks naturally sit within their existing risk categories, and the most practical approach is to manage them there.

The risk amplifier approach can be used as the primary operating model for day-to-day risk management. Each existing risk domain (such as cyber, operational, model, conduct, third-party, and others) incorporates AI-specific controls, with the domain expert retaining first-line accountability. This reflects the reality that AI risk manifests through existing risk categories: a prompt injection attack is a cyber risk managed by the security team; a discriminatory pricing outcome is a conduct risk managed by the compliance function; a hallucinating chatbot is a model performance risk managed within the model risk framework. The domain experts have the context, the regulatory relationships, and the operational proximity to manage these risks effectively. Fragmenting that accountability into a separate AI risk function would likely weaken rather than strengthen risk management.

The material risk category approach can be used for aggregation, board reporting, and regulatory engagement. Organisations may wish to tag AI-related risks across all registers so they can be filtered into a consolidated AI risk view. This tagging enables the risk function and the board to answer the question that neither regulators nor boards can answer under a pure amplifier model: what is our aggregate AI risk exposure? The consolidated view supports board reporting, regulatory inquiries, risk appetite monitoring, and capital allocation decisions. It does not require a separate AI risk register; it requires a consistent tagging convention across existing registers and the reporting capability to aggregate across them.

In practical terms, this means that a prompt injection risk sits on the cyber risk register, owned by the Chief Information Security Officer (CISO), with AI-specific controls documented alongside the organisation's other cyber risk controls. It is also tagged as an AI risk so that it appears in the consolidated AI risk view reported to the board. An AI-related conduct risk sits on the conduct risk register owned by the compliance function, and is tagged for the same aggregate view.

Proportionality note: For smaller organisations, the risk amplifier approach with a lightweight aggregation overlay may be the most practical starting point. A simple tagging convention applied to existing risk register entries, combined with a periodic aggregation exercise, can provide the consolidated view needed for board reporting without requiring dedicated AI risk headcount. As the organisation's AI deployment matures and the volume of AI-related risks grows, the aggregation capability can be strengthened.

2.3 Common AI risk categories for Financial Services

The risk categories in the table in the following pages are suggested as **one starting point** for thinking about AI-specific risks in Australian financial services. They are not intended to be definitive, and many organisations will find that these risks fit naturally within their existing risk taxonomy without requiring a new set of categories.

A few caveats before setting them out:

- **These categories are not mutually exclusive.** There are deliberate overlaps between the categories. In practice, a single AI risk event will often touch multiple categories. For example, shadow AI use can trigger model and performance risk (staff may select a tool that is not fit for the task), conduct risk (the mere fact of using an unsanctioned IT system with confidential data), cyber security risk (the provider may have an insufficient cyber security posture that would not have passed organisational review), and so on. The taxonomy is a starting point for structured thinking, not a set of rigid boundaries.
- **Consistent with AI being a risk amplifier, these categories will often map directly to existing risk domains.** Regulatory and compliance risk, for instance, is not unique to AI. The first line business that understands the regulation should own AI versions of that risk, just as it owns the non-AI versions. Similarly, cyber risk is the CISO's domain, whether or not AI is involved. The AI-specific dimension reflects new attack surfaces and vulnerabilities, not a change of ownership.
- **Some of the categories within each heading are quite different in nature.** Data risk, for example, spans data quality (a technical and operational concern), privacy (a legal and regulatory concern), consent (a legal and ethical concern), and third-party platform dependency (a strategic, vendor, and supply chain concern). Organisations may reasonably choose to split these across different parts of their taxonomy rather than grouping them.
- **The consequences listed (poor decisions, financial loss, regulatory breach) are themselves risk sub-categories** rather than distinct from the taxonomy. This is an inherent challenge in any risk classification: the same event can be described as a cause, an event, or a consequence depending on the level of analysis.

These risks will apply in different ways to different tools and use cases. Given it is beyond the scope of this paper to consider all tools and contexts for their use, we have focused on four common use cases in the financial services sector to provide illustrative examples of these risks. These include:

Use Case 1: Internal Productivity Tools:
This includes AI coding assistants, document drafting, and summarisation tools.

Use Case 2: Customer-Facing Chatbot:
This includes tools used for general enquiries, product information, claims lodgement.

Use Case 3: Credit Decisioning:
This includes tools used for credit scoring, affordability assessment, and loan approval recommendation.

Use Case 4: Claims Processing:
This includes tools used for triage, assessment, fraud detection, and settlement recommendation.

These use cases will be used throughout the paper. They are not intended to be exhaustive, but provide a common touchpoint throughout to explore the classification, quantification and controls for risks in familiar contexts for financial services organisations.

Table 1. Common AI Risk Categories for Financial Services

Category	Definition	Illustrative examples
Conduct and fairness risk	The risk that AI system outputs may be discriminatory, unfair, or unexplainable, eroding organisational trust.	• Customer-Facing Chatbot: A customer service chatbot's tone and helpfulness vary systematically by customer segment in ways that correlate with protected attributes. • Credit Decisioning: A credit assessment tool produces systematically higher rejection rates for applicants whose employment history, postcode, and property type correlate with ethnicity, even though ethnicity is not an input variable.
Model and performance risk	The risk that inaccurate, unreliable, or inconsistent AI outputs may lead to organisational misperformance, reputational damage, regulatory breach, and financial loss.	• Internal Productivity Tools: A summarisation tool used by financial advisors to generate client meeting notes excludes material information about risk warnings discussed during the meeting. • Credit Decisioning: An AI-assisted credit assessment tool performs well on average but exhibits significantly worse accuracy for applicants in specific age groups or income brackets.
Commercial and strategic risk	The risk that decisions (or inaction) relating to the investment, deployment, and maintenance of AI systems may produce poor commercial and strategic outcomes over time and/or at scale.	• Internal Productivity Tools: An organisation selects a conservative, lower-capability AI platform to minimise adverse risk. While the choice reduces perceived regulatory and reputational risk, business users bypass the tool in practice, leading to low adoption, shadow deployments without risk oversight, and negligible productivity gains. • Claims Processing: Overdependence on a single end-to-end AI claims platform reduces the ability to adopt better specialist or modular tools as the market evolves, increasing the organisation's path dependency and vendor lock-in.
Data risk	Risks that may arise from the quality, representativeness, provenance, availability, privacy, consent, integrity, and third-party dependency of the data used to train AI systems.	• Customer-Facing Chatbot: The chatbot is trained on incomplete or poorly reconciled product data, including live website data, producing erroneous, misleading, or inconsistent answers to customers. • Credit Decisioning: Training data for a customer segmentation model contains historical lending patterns that encode postcode-based socioeconomic proxies, producing outputs that reflect biased historical data.
Cyber and security risk	The risk that AI systems may be compromised, manipulated, exploited, or used to commit attacks against the organisation.	• Customer-Facing Chatbot: A customer-facing chatbot is subjected to prompt injection, causing it to reveal system prompts or produce outputs outside its intended scope. • Claims Processing: An attacker crafts inputs to exploit the decision boundary of an automated fraud detection model, causing systematic evasion of fraud controls.
Regulatory and compliance risk	The risk that AI deployments may fail to meet obligations under Australian laws and regulations, including financial services law, prudential standards, privacy obligations, and regulatory guidance.	• Internal Productivity Tools: An AI tool ingests internal documents but the retrieval corpus includes documents subject to information barriers, surfacing restricted information. • Credit Decisioning: A bank uses AI to determine lending thresholds, but the model's decision factors cannot be explained to satisfy anti-discrimination challenge.
Third-party and vendor risk	The risk that reliance on external AI model providers may lead to negative outcomes, including the opacity of model behaviour, vendor lock-in, unanticipated API changes, and service continuity.	• Customer-Facing Chatbot: A major LLM provider pushes a model update that changes behaviour with no advance notice or ability to test before go-live. • Credit Decisioning: A vendor's terms permit using customer interaction data to improve vendor models, creating an unanticipated data sharing arrangement.
Operational risk	The risk of loss from inadequate or failed internal processes, people, systems, and technology as they relate to AI.	• Internal Productivity Tools: Staff lack the skills to interpret AI outputs critically, treating AI-generated analysis as authoritative. • Claims Processing: An AI system is integrated into a claims workflow without adequate change management, and downstream processes fail silently when output formats change.

Box 3: AI-Specific Cyber and Security Threats

AI systems introduce attack vectors that do not exist in traditional IT systems, and present risk horizontally across different domains. ASIC and APRA have warned that AI is materially changing the cyber and security threat landscape for organisations.[20] APRA has observed that information security practices, as well as governance and assurance, are not keeping pace with the scale and complexity of AI adoption, and has heard clear recognition for a "step change" in cyber practices. ASIC has similarly warned that AI is accelerating the speed, scale, and sophistication of cyber-attacks, and has urged organisations to urgently uplift cyber security controls and resilience. Together, it is clear that regulators expect entities to actively manage the growing cyber threats posed by AI-enabled malicious actors and the new threat vectors present within AI systems.

Two examples illustrate the nature of this new challenge:

Prompt injection: An attacker crafts inputs that cause an AI system to ignore its instructions and execute the attacker's intent. There is no equivalent in traditional IT systems because they do not interpret natural language instructions like AI systems do. Standard input validation catches SQL injection and cross-site scripting; it does not catch prompt injection, which operates at a semantic level.

Supply chain attacks on model components: The deployment infrastructure surrounding AI models (including training data, package managers, container images and model serving frameworks) presents conventional software supply chain attack surfaces. A compromised component in the deployment chain could introduce backdoors that exfiltrate data or alter model behaviour. These are more likely attack vectors than direct tampering with model weight files.

An APRA-regulated entity assessing the adequacy of its cyber security controls for AI should consider whether it has specifically addressed these AI-specific attack vectors. The controls required go beyond standard information security controls and include:

- prompt injection detection and input sanitisation tailored to natural language;
- output filtering and classification;
- model provenance verification;
- red-teaming and adversarial testing specific to AI;
- data loss prevention for AI interaction channels;
- monitoring for model extraction attempts;
- supply chain integrity verification for model components; and
- for agentic AI systems, action-level authorisation and execution boundary enforcement.

These are specialist capabilities that most organisations are still building. For more information, consult **Appendix D.**

3. Quantification: How might we quantify AI risks?

The risk categories in Section 2 tell you what kinds of things can go wrong. They organise governance conversations and help allocate accountability. But they do not tell you how likely any specific event is, or how much it will cost.

To take AI risk quantification from high-level taxonomies to more concrete estimates of the likelihood and impact of AI-related loss events, we suggest using an event-based model. The benefit of this approach is that risks are converted into quantitative estimates which can be more easily compared across the operational risk register.

Event-based risk modelling is not a replacement for traditional category-based taxonomies. This type of risk modelling operates at a different level: categories establish and organise governance and accountability, and event-based analysis quantifies the risk of specific scenarios within those categories. An organisation might use risk categories to structure its AI risk program and event-based modelling to assess and prioritise specific risks within that program. The categories tell you where to look; the event-based model tells you what you find there.

3.1 A FAIR-Inspired five-stage model

The FAIR (Factor Analysis of Information Risk) framework, originally developed for cyber security and now an Open Group standard, provides a useful methodological foundation. FAIR decomposes AI risk into a chain of events: threats, vulnerabilities, loss events, impacts, and controls, allowing for clearer estimates of loss rather than relying solely on qualitative tools, such as heat maps. Each of these components can be estimated and, where data permits, quantified. The result, in its mature form, is a probability distribution of financial loss rather than a qualitative rating on a heat map.

A note on current practice: most financial services organisations use heat maps (likelihood versus impact matrices) for operational risk assessment, despite their well-documented limitations. For example, they compress continuous variables into ordinal categories, they are highly sensitive to where the boundaries are drawn, and they can obscure meaningful differences in risk magnitude. The event-based approach described here is intended as a step towards more rigorous quantification. Organisations currently using heat maps may find it useful to apply this approach selectively to their highest-priority AI risks, even if they continue to use qualitative methods for lower-priority items.

We suggest adapting the FAIR ontology for AI risk in financial services. The adapted model has five stages.

1 **Stage 1: Threat.** What could initiate a loss event? In AI risk, threats include: malicious actors (prompt injection, adversarial attack, model extraction); non-malicious internal actors (misuse, over-reliance, shadow AI); environmental conditions (data drift, distribution shift, regulatory change); and vendor actions (model updates, service changes, cessation, processing data in offshore jurisdictions beyond contractual scope, retaining customer data beyond agreed periods, or sharing data with third parties, potentially for purposes not contemplated by the data processing agreement).

2 **Stage 2: Vulnerability.** What characteristics of the AI system or its operating environment make it susceptible to the threat? Vulnerabilities include: lack of input validation; absence of output filtering; insufficient monitoring for drift; reliance on a single vendor; absence of human review for consequential decisions; gaps in staff AI literacy; and inadequate testing of edge cases.

3 **Stage 3: Loss Event.** If the threat exploits the vulnerability, what is the loss event? Loss events include: discriminatory decision; material misrepresentation to customers; data breach via AI channel; regulatory breach; system unavailability.

4 **Stage 4: Loss Distribution.** What is the financial impact? This should be estimated as a distribution rather than a point estimate. Most losses are uncertain in magnitude, and the range matters as much as the central estimate. The distribution should use the same loss categories used in operational risk, such as direct financial loss, regulatory penalties, remediation costs, legal costs, customer compensation, reputational impact (modelled as future revenue loss), and opportunity cost. Where data is limited, scenario-based estimates with explicit uncertainty ranges (e.g., 10th/50th/90th percentile) are preferable to false precision.

5 **Stage 5: Controls.** What controls reduce either the frequency of the loss event or the severity of the loss? It is worth noting that the clean distinction between “preventive controls” (which reduce frequency) and “mitigating controls” (which reduce severity) is an oversimplification in practice. Many controls affect both: input validation, for example, both prevents some attacks from succeeding (frequency) and limits the damage when an attack partially succeeds (severity). Likewise, a well-designed monitoring and escalation system can reduce the duration and therefore the severity of an incident, but it can also deter certain types of attack, reducing frequency. Controls should be assessed for their overall effect on both the frequency and severity distributions. Controls should not be assumed to operate perfectly, and the assessment of a control’s impact on the frequency and severity distributions should reflect a realistic assessment of how the control will operate in practice.

3.2 Worked Example: Event-based analysis of a customer-facing chatbot

Stage	Element	Assessment
1. Threat	Customer discovers prompt injection technique to extract system prompt and produce unauthorised financial guidance	Frequency estimate: 2-5 attempts per month based on industry data for similar deployments
2. Vulnerability	Chatbot lacks robust input sanitisation; system prompt contains business-sensitive rules; output filtering does not catch all categories of financial advice	Vulnerability rating: High. Standard prompt injection techniques have a materially high success rate against unhardened chatbots
3. Loss event	Chatbot provides statement that constitutes personal financial advice without AFSL coverage; customer acts on it to their detriment	Probability given threat and vulnerability: Moderate (estimated 10-25% per quarter)
4. Loss distribution	ASIC enforcement action, customer remediation, legal costs, reputational impact	Estimated range: $500K-$5M per incident, depending on scale and regulatory response. The distribution is right-skewed: most incidents would fall at the lower end, but a small number of high-profile cases could reach the upper bound or beyond
5. Controls	Input filtering for injection patterns; output classification for financial advice; human review trigger for flagged interactions; system prompt hardening; regular red-teaming	Residual frequency after controls: Low (estimated 1-3% per quarter); residual severity distribution: $100K-$1M, with reduced tail risk

This approach produces risk estimates that can be placed directly on the operational risk register in the same format as other operational risk scenarios. It enables comparison across risk types, supports capital allocation decisions, and gives the board a financial view of AI risk exposure rather than a colour on a heat map.

4. Controls: What controls are most appropriate for common AI use cases?

AI controls should be selected and calibrated based on the specific AI use case, the risks it creates, and the organisation's risk appetite, rather than applied uniformly across all deployments. This section translates the earlier risk taxonomy and event-based analysis into practical, use-case-specific controls for common financial services applications, recognising that different deployments (internal tools, customer-facing systems, decisioning, and claims automation) create materially different risk profiles. The control sets are presented as a menu set against relevant use cases rather than a checklist, organised by familiar risk domains.

4.1 Common controls for financial services AI use cases

The tables in the following pages offer a menu of controls for each risk category across four common AI use cases in financial services. They are not minimum requirements. The appropriate selection and intensity of controls is a cost/risk-benefit decision that each organisation should make in consideration of its own circumstances, including the materiality of the use case, the population affected, and the organisation's risk appetite.

The controls are organised by the risk categories introduced in section 2 of this document (conduct and fairness risk, model and performance risk, commercial and strategic risk, data risk, cyber and security risk, regulatory and compliance risk, third-party and vendor risk, and operational risk), although controls often address more than one category.

Organisations should treat these tables as a starting point for their own control design, selecting and adapting controls to suit the specific characteristics and vulnerabilities of each deployment. The example controls provided are illustrative, not exhaustive. Many other relevant controls may also be appropriate. Some controls are common across use cases and risk categories (e.g., monitoring and record-keeping, role-based access control, data quality controls, evaluation of vendor agreements, staff training), whereas others are tailored to the specifics of the use case. These example controls should not be treated as a checkbox exercise; organisations must critically evaluate the risks of their system and accordingly select and test the appropriate controls.

Monitoring controls, such as measurement and classification of hard errors when a technical problem triggers a fallback, or the number and contents of customer complaints, can enable detection and prompt intervention across a number of risks. For example, complaints that a claim was rejected for invalid reasons that are overturned on review may speak to model and performance risk, whereas a customer describing that the settlement justification they received appeared to contain details of someone else's claim may indicate cyber security risk.

4.1.1 Use Case 1: Internal Productivity Tools (e.g., AI coding assistants, document drafting, summarisation)

Internal productivity tools like ChatGPT, Claude, Gemini, Copilot, and others are used heavily by financial services organisations in many day-to-day tasks, and so can present both wide-ranging and varied risks across an organisation.

Risk Category	Example Controls
Conduct and fairness risk	• Training programs so that staff understand acceptable uses of AI systems and their limitations • Guidance regarding scenarios when internal productivity tools are appropriate, as opposed to a bespoke system specialised to the use case
Model and performance risk	• Documented acceptable use policy specifying the tasks and purposes for which the tool may and may not be used for. Additional guidelines for how to best use the tool may aid in user comprehension and competency • Quality assurance sampling regime appropriate to the use case risk level. See the section on **Human Review and Human in the loop** below for guidance on when and how human oversight might be applied
Commercial and strategic risk	• Adoption tracking and measurement of productivity impacts • Periodic review of tool selection against alternatives. Consider defining an adoption threshold below which the investment is escalated for review
Data risk	• Data loss prevention (DLP) controls on all inputs to the tool • Clear guidance to staff against inputting personal information (as defined under the Privacy Act 1988 (Cth)), price-sensitive information, or material non-public information • Organisations should consider mapping permissible inputs against their internal document classification system. For example, restricting inputs to documents classified below a defined sensitivity threshold • Contractual restrictions on vendor use of customer data for model training • Alternatively, the tool may be hosted in an environment that meets regulatory requirements for the relevant data classifications (e.g., SOC 2 Type II certified infrastructure located in Australia), reducing reliance on input-level DLP controls
Cyber and security risk	• AI system access restricted to authorised employees via single sign-on (SSO) and multifactor authentication (MFA) • Logging of all interactions for audit purposes • Vendor security assessment against CPS 234 requirements (for APRA-regulated entities) or equivalent standards such as SOC 2 Type II or ISO 27001 • Automated monitoring of common vulnerability and exposure (CVE) databases for known exploits in dependent packages
Regulatory and compliance risk	• Legal review of vendor terms of service, particularly data handling, supply chain risks, IP ownership of outputs, and liability for errors. Confirmation that use does not trigger obligations under the AFSL (e.g., AI-generated research should still be reviewed by a licensed person)
Third-party and vendor risk	• Monitor vendor outages, security notices, release notes (where applicable) and other policy changes • Contractual provisions that include considerations such as data retention and deletion policies, restrictions on data usage for model training, breach notification, indemnities, and support arrangements
Operational	• Shadow AI monitoring and controls to detect and block use of unapproved AI tools • Staff training on acceptable use, limitations, and review requirements. Change management process for tool updates

4.1.2 Use Case 2: Customer-Facing Chatbot (e.g., general enquiries, product information, claims lodgement)

Customer-facing chatbots are largely APIs that draw on back-end LLM infrastructure (such as ChatGPT, Claude, Gemini, etc.) and are tailored to organisational context, policies, data, and other defined rules to deliver natural language responses to customer queries. These tools can therefore present the same or similar risks to the internal productivity tools mentioned above, as well as unique risks by virtue of interacting with the public.

Risk Category	Example Controls
Conduct and fairness risk	• Customer research (including direct engagement with customers from diverse segments) to understand whether the system is meeting actual needs, and to identify fairness issues that quantitative testing alone may not surface. This is often the most important fairness control and the one most frequently omitted • Testing for differential treatment to validate that the system is providing a consistent level of service across customer segments • Tone and quality monitoring to validate that the system is performing in line with brand guidelines and broader societal expectations • Clear disclosure that the customer is interacting with an AI • Accessible escalation pathways to a human agent
Model and performance risk	• Defined scope of permissible responses • Output classification to detect and block financial advice, misleading statements, or off-topic responses • Accuracy monitoring with human review of a statistically significant sample of interactions. Escalation triggers for uncertain or sensitive queries
Commercial and strategic risk	• Adoption tracking: measure actual customer take-up and successful containment rate • Where possible, design technical components for portability to minimise disruption if the chatbot needs to be moved to another model or vendor
Data risk	• Restricting processing of customer data to approved jurisdictions • Completion of privacy impact assessment • Customer consent obtained for AI interactions, with the option to escalate to a human • Data retention limits on conversation logs
Cyber and security risk	• Prompt injection defences: input sanitisation, output filtering, system prompt protection • Rate limiting to prevent model extraction • Monitoring for adversarial interaction patterns • Regular penetration testing including AI-specific attack vectors
Regulatory and compliance risk	• Legal review confirming that chatbot responses do not constitute financial product advice • Compliance monitoring for adherence to ASIC disclosure obligations • Record-keeping to meet dispute resolution requirements
Third-party and vendor risk	• Contractual requirements for model update notification (a reasonable notice period should be agreed) • Pre-deployment testing of all model updates in a staging environment before production • Audit rights over the vendor's data handling and security practices. Exit plan documented and tested
Operational risk	• Explicitly designed and tested escalation pathways to humans for scenarios such as customer vulnerability, safety, complaints, uncertainty, or repeated failures • Ongoing monitoring of customer interactions to identify instances that need remediation • Defined critical failure scenarios that would warrant taking the system offline • Fallback mechanisms for when critical infrastructure fails

4.1.3 Use Case 3: Credit Decisioning (e.g., credit scoring, affordability assessment, loan approval recommendation)

Credit decisioning tools are back-end systems that can assess an applicant's credit risk and offer loan-related guidance and decisions. These tools possess tremendous potential to realise risks of bias and harm to different customer segments.

Risk Category	Example Controls
Conduct and fairness risk	• Testing for indirect discrimination across all protected attributes under Commonwealth and state anti-discrimination legislation • Use of techniques that ensure explainability at the individual decision level. The organisation should be able to explain to a rejected applicant why they were rejected in terms they can understand • Regular fairness audit by an independent second-line or third-line function • Ongoing customer research (engaging directly with affected applicants, particularly those from groups historically subject to credit exclusion) to identify whether the model's outcomes align with fair treatment in practice, not only in statistical terms
Model and performance risk	• Full model validation before deployment, including out-of-sample testing, stress testing, and sensitivity analysis • Ongoing monitoring for accuracy degradation and drift • Model performance benchmarked against incumbent system. Documented model risk appetite, including acceptable error rates by decision type
Commercial and strategic risk	• Monitoring to validate that the system is achieving the desired commercial outcomes, including key thresholds defined as triggers for model or vendor review • Periodic review of tool selection against alternatives, including build versus buy considerations
Data risk	• Data lineage documented from source to model input • Bias assessment of training data • Prohibition on using attributes that serve as proxies for protected characteristics, with documented proxy analysis. Data quality monitoring for ongoing model inputs • Classify training and input data according to the internal document classification framework and apply handling controls accordingly
Cyber and security risk	• Prompt injection defences: input sanitisation, output filtering • Monitoring for adversarial interaction patterns • Penetration testing including AI-specific attack vectors • Secure architecture and access controls, applied to both staff members and "agentic" systems • Secure development lifecycle applied to models, code and other components of AI pipelines
Regulatory and compliance risk	• Reviewing compliance with responsible lending obligations under the National Consumer Credit Protection Act • Testing adherence to APRA's requirements for credit risk models • Retaining documentation sufficient for regulatory examination, including model documentation, validation reports, and ongoing monitoring evidence
Third-party and vendor risk	• Contractual requirements for model update notification (a reasonable notice period should be agreed) • Pre-deployment testing of all model updates in a staging environment before production • Independent validation of vendor model performance on the organisation's data
Operational risk	• Human in the loop for credit decisions above a defined threshold • Documented override process with audit trail • Fallback to manual assessment if the AI system is unavailable or producing anomalous results

4.1.4 Use Case 4: Claims Processing (e.g., triage, assessment, fraud detection, settlement recommendation)

Claims processing tools are specialised decision-support and automation systems used by insurers to administer insurance claims. Similar to credit decisioning tools, claims processing systems can carry significant risks of bias and harm, alongside technical risks.

Risk Category	Example Controls
Conduct and fairness risk	• Testing for bias in claim outcomes (approval rates, settlement amounts, processing times) across demographic groups. Particular attention should be paid to vulnerability indicators: for example, AI systems should not disadvantage claimants experiencing hardship or with limited English proficiency • Customer research with claimants across different segments to understand whether the claims experience is perceived as fair, particularly where AI-driven speed may be creating a two-tier service
Model and performance risk	• Accuracy monitoring segmented by claim type, value, and claimant demographics • Threshold calibration reviewed periodically (e.g., quarterly) • False positive and false negative rates tracked and reported. Comparison of AI-assisted outcomes against a human-assessed control sample • For claims triage specifically, organisations should consider monitoring whether AI-driven triage decisions correlate with downstream settlement outcomes and customer satisfaction, to detect cases where efficient triage may be producing inequitable results
Commercial and strategic risk	• Monitoring to validate that the system is achieving the desired commercial outcomes, including key thresholds defined as triggers for model or vendor review • Periodic review of tool selection against alternatives, including build versus buy considerations
Data risk	• Data quality controls on inputs (claims forms, supporting documents, third-party data) • Provenance verification for external data sources • Privacy controls on medical, financial, and personal information processed by the AI system. Organisations should apply their internal document classification to claims data and ensure handling controls are proportionate to the sensitivity of the information
Cyber and security risk	• Prompt injection defences: input sanitisation, output filtering • Monitoring for adversarial interaction patterns • Penetration testing including AI-specific attack vectors • Secure architecture and access controls, applied to both staff members and “agentic” systems • Secure development lifecycle applied to models, code and other components of AI pipelines
Regulatory and compliance risk	• Testing compliance with the General Insurance Code of Practice and/or Life Insurance Code of Practice • Testing adherence to ASIC’s expectations for claims handling, including timeliness, fairness, and communication • Retaining documentation of AI involvement in claims decisions for dispute resolution purposes
Third-party and vendor risk	• Where fraud detection models are vendor-supplied, contractual requirements for model transparency (at a minimum, documentation of model methodology, training data characteristics, and known limitations) • Contractual requirements for model update notification (a reasonable notice period should be agreed) • Pre-deployment testing of all model updates in a staging environment before production • Independent validation of vendor model performance on the organisation’s data
Operational risk	• Human in the loop for decisions that significantly impact customer rights • Documented override process with audit trail • Fallback to manual processes if the AI system is unavailable or producing anomalous results

4.2 Controls requiring special attention

Certain controls warrant closer attention when applied because they address risks that are dynamic, opaque, or highly context-dependent. Three controls of note in this respect include: how AI vendors are selected and governed, how deployed AI systems are continuously monitored and assessed, and meaningful and effective human oversight of AI systems.

4.2.1 Vendor due diligence for AI procurement

AI procurement requires due diligence that goes beyond standard technology vendor assessment. The AI-specific considerations around aspects such as model transparency, data handling, update management, and performance assurance are sufficiently distinct that they warrant dedicated attention in the procurement process, alongside the organisation's existing vendor risk assessment.

A detailed vendor due diligence checklist covering model and performance, data handling and privacy, security, compliance and regulatory, commercial and continuity, and concentration and dependency considerations is provided in Appendix E.

Rather than reproducing the full checklist here, we highlight two framing considerations that should inform AI vendor assessment: **proportionate assessment** and **building versus buying over a long time horizon.**

Proportionate assessment. The depth of vendor due diligence should be proportionate to the risk profile of the intended use case. For example, an AI-powered scheduling tool for internal meeting rooms does not warrant the same scrutiny as a credit decisioning engine. The risk triage conducted at the outset of each deployment should determine the appropriate level of vendor assessment.

Build versus buy over a long time horizon. The build-versus-buy decision for AI capabilities deserves more careful treatment than it typically receives. Many organisations default to buying because vendor solutions offer faster time to value. This may be the right decision in the short term, but the long-term implications should be assessed against alternative scenarios:

- **Ongoing capability investment.** Building in-house develops organisational capability in AI development, evaluation, and operations. Over a multi-year horizon, this capability may prove strategically valuable, particularly if AI becomes central to the organisation's competitive positioning. Buying, by contrast, develops capability in vendor management but not in the underlying technology.
- **Loss of control and system interdependencies.** Vendor AI systems can become deeply embedded in business processes, creating dependencies that are expensive and disruptive to unwind. Organisations should consider what happens if the vendor changes pricing, discontinues the product, is acquired, or pivots its strategy. In addition, a vendor data or security breach may have financial and reputational impacts. The cost of switching should be estimated at the point of procurement, not at the point of crisis.
- **Data and model sovereignty.** Where the AI system processes sensitive customer data or makes consequential decisions, the organisation should consider whether it is comfortable with the long-term implications of that data flowing through a third party's infrastructure and models. Regulatory expectations in this area are likely to tighten, not relax.

The right answer will vary by organisation and use case. The important discipline is to make the decision explicitly, informed by scenario analysis over a realistic time horizon (three to five years at minimum), rather than defaulting to buy on the basis of short-term convenience.

4.2.2 Monitoring and reassessment

AI risk management is not a one-time assessment. AI systems change in ways traditional systems do not: models get updated, data drifts, usage patterns shift, and regulatory expectations evolve. A monitoring and reassessment program should account for these dynamics.

Continuous monitoring might cover:

- **Output quality.** Automated monitoring of AI system outputs against quality benchmarks, including accuracy, consistency, and appropriateness. The monitoring approach will vary by use case: for a credit model, this could mean tracking approval rates, default rates, and indirect discrimination metrics; for a chatbot, it could mean sampling interactions for quality, accuracy, and compliance.
- **Drift detection.** Statistical monitoring for changes in input data distributions and model behaviour over time. Drift does not necessarily mean the model is wrong (for example, the world may have legitimately changed) but it means the model's performance assumptions need to be revalidated. In actuarial terms, this is the same concept as experience monitoring: tracking whether actual outcomes continue to align with the assumptions used to calibrate the model. The terminology differs between data science and actuarial practice, but the underlying discipline is identical, and organisations with mature actuarial functions already possess the conceptual infrastructure to implement drift monitoring effectively.
- **Security monitoring.** Monitoring AI interaction channels for adversarial activity, prompt injection attempts, unusual query patterns, and data exfiltration indicators. This is best integrated into the organisation's security operations function, with AI-specific detection rules. This monitoring consideration is not limited to generative AI. Traditional AI and machine learning (ML) systems (e.g., such as rating engines, fraud detection models, and claims triage algorithms) also warrant security monitoring. For example, a suspiciously high volume of insurance quotes hitting a rating engine in a short period may indicate that a competitor or data scraper is systematically probing the organisation's pricing model.
- **Employee-facing uptake monitoring.** Tracking which AI tools employees use, for what purposes, and at what volume. This is a primary mechanism for detecting shadow AI and for identifying use cases that have expanded beyond their approved scope. Uptake patterns can also reveal adoption failures. If a deployed tool shows minimal uptake, the investment may warrant review.

- **Customer-facing usage monitoring.** Tracking customer interaction volumes, query patterns, escalation rates, and satisfaction indicators for AI-powered customer touchpoints. This monitoring serves a different purpose: it helps detect whether the AI system is meeting customer needs, whether specific customer segments are being poorly served, and whether interaction patterns suggest adversarial probing or misuse.

Periodic reassessment might be triggered by:

- **Model updates.** Any update to the underlying model, whether initiated by the organisation or pushed by a vendor, may warrant a reassessment of risk. For higher-risk use cases, this could mean re-running validation tests before deploying the update. For lower-risk use cases, post-deployment monitoring with accelerated review may suffice.
- **Regulatory changes.** Changes to APRA, ASIC, OAIC, or other regulatory guidance related to AI may warrant a review of all affected AI deployments for compliance.
- **Material incidents.** Any material AI-related risk event (e.g., a biased outcome, a security incident, a compliance breach, a significant customer complaint) should prompt a root cause analysis and a reassessment of controls for the affected system and analogous systems. This should include incidents at suppliers or partners, not just those internal to the company.
- **Scheduled reviews.** AI deployments benefit from comprehensive reassessment on a regular cycle (at least annually, and more frequently for higher-risk use cases). The reassessment should revisit the initial risk triage, validate that controls are operating as intended, confirm that the use case remains within approved scope, and update risk estimates based on operational experience.
- **Competitive and technology shifts.** The commercial risk dimension warrants periodic assessment of whether the chosen tools and approaches remain fit for purpose. Technology that was best-in-class at deployment may be superseded within months. A periodic market scan for each material AI deployment, which involves comparing the incumbent against available alternatives on capability, cost, and risk profile, is a useful part of the reassessment cycle.

Reporting should flow through the Three Lines of Defence accountability structure outlined in Chapter 1. The first line reports on operational metrics (output quality, usage, incidents). The second line reports on aggregate risk exposure, control effectiveness, and compliance posture. The board risk committee receives a consolidated view, with the ability to drill into specific deployments or risk categories as needed.

The monitoring program should be proportionate to the risk. A productivity tool used by internal staff warrants lighter monitoring than a credit decisioning model affecting large numbers of customers. The risk triage conducted at the outset of each deployment should inform the monitoring intensity, and this should be documented and reviewed as part of the periodic reassessment.

4.2.3 Human review and human in the loop

The question of when human review is needed is one of the most contested in AI governance. Human oversight is an important safeguard, particularly where AI systems are used in higher-risk contexts. However, a blanket requirement for human review of all AI outputs is often not practical or effective for many high-volume uses of AI in the financial services sector. It can negate the efficiency gains that AI is intended to deliver. It may also create a false sense of oversight where reviewers rubber-stamp outputs they lack the time, information, or expertise to meaningfully evaluate.

The objective is effective and proportionate oversight. The question is not whether a human should always review an AI output, but what form of oversight is appropriate for the use case. The appropriate level of human oversight depends on at least four factors:

- **consequence of error**: how much harm a wrong output can cause
- **detectability of error**: whether mistakes are obvious or require domain expertise to spot
- **volume and velocity of outputs**: whether meaningful review is sustainable at operational scale, and
- **contestability, explainability and reversibility:** whether the outcome requires explanation, may be subject to challenge or review, and the ease or otherwise of reversing an incorrect decision.

These four dimensions can interact in different ways. For example, processes with a high volume and velocity of outputs may have significant consequences when errors occur (e.g. loan approvals, insurance underwriting) and if they are difficult to reverse (e.g. account closure). In such situations, it might not be practical for a human to review every individual decision.[21] However, this does not remove the need for human accountability.

In these cases, it may be better to focus human oversight on the operational guardrails and controls built around the process, rather than reviewing the outputs of that process. These controls may include pre-release testing using historical or synthetic data (e.g. datasets deliberately engineered to exercise rare edge cases), automated quality assurance checks, ongoing performance monitoring, periodic review of outcomes, and escalation or intervention mechanisms. The question is not whether every output requires human review, but what combination of oversight and controls is appropriate for the risk profile and scale of the use case.

For higher-risk use cases, organisations may invest heavily in testing and monitoring before and after deployment. While these controls can be resource-intensive, they may be necessary to achieve an appropriate level of assurance. For example, Anthropic's System Card for their Opus 4.8 model runs to 246 pages, and that is just a summary report of the testing done prior to deployment to the public.

For lower-risk situations, it might be sufficient to conduct a moderate amount of backtesting or basic online monitoring (which would detect major deviations but not subtle problems). Where the consequences of error are low, users may decide whether review is necessary at all.

It should be noted that the consequences of an error for

some use cases can be easy to misjudge. For example, targeted marketing is often considered a low-consequence use case (for example, if a customer gets an advertisement for a product they don't want, they'll just ignore it), but it has high-consequence failure modes, such as accidentally targeting vulnerable customers with credit lending products. We recommend applying the above model framework to fully understand all facets of risk involved, and designing mitigations targeted at those events (which might include human oversight of a process, human escalation, etc. – all forms of "human in the loop", alongside other types of mitigations), rather than simply applying a broad-brush human review requirement based on a risk category.

Another critical principle is that human review should be *substantive*. A review process that exists on paper but produces no overrides or corrections is almost certainly not working as intended. Organisations should track override rates and investigate if they are implausibly low.

The volume amplification problem. When AI-driven automation multiplies throughput by several times, an organisation's existing quality assurance infrastructure may no longer be adequate, even if the per-unit defect rate remains constant. The expected number of defects scales linearly with volume; the reputational and regulatory consequence of each defect does not diminish. Organisations deploying AI to increase throughput should simultaneously consider whether their automated quality assurance, testing infrastructure, and statistical process control mechanisms can sustain the organisation's existing risk tolerance at the higher volume. Scale effects should be tested as part of deployment, not just the performance of the AI component in isolation. If quality assurance investment is not part of the deployment plan, the organisation is implicitly accepting a higher absolute number of defects.

5. Case Study: Worked Example

In this case study, the framework is applied to an agentic customer support system that can execute actions on the bank's systems of record.

Scenario

A mid-tier Australian bank deploys an AI agent across voice, email, web chat and mobile chat. The agent carries persistent per-customer state across channels. This means that a customer who begins on a call can continue by chat without repeating themselves. The agent is authorised to perform a range of customer service functions, including opening and closing accounts, completing know-your-customer (KYC) and customer due diligence checks, processing fee waivers and hardship arrangements, and initiating transfers.

5.1 Governance and the Three Lines of Defence

The outcomes produced by the agent are outcomes of the business. The business owns those outcomes whether a person or an agent makes them. Applying the Three Lines of Defence model, the first line is responsible for operating the system within the approved risk appetite, while the second line provides independent challenge and oversight. This means a named first-line decision-maker (for example, a Chief AI Officer or the executive who owns the customer servicing channel) should be accountable for deploying the agent within risk appetite, while a separately accountable second-line owner (for example, the CRO or a delegated AI risk owner) holds the mandate to challenge, validate and, where warranted, stop the system from operating.

5.2 Risk classification

Applying the risk categories outlined above in Section 2.3 helps identify some of the risk events that this deployment could create (which are itemised from E1 to E11). The below list is not intended to be exhaustive.

Conduct and fairness. The agent treats customers in similar circumstances differently in ways that correlate with protected attributes, for example by being more willing to grant fee waivers or hardship assistance to some groups than others. If systemic, this could result in customer complaints and harms to individuals, regulatory action, and reputational and/or financial harm (E1).

Model and performance. The agent mis-states a product feature, eligibility requirements, or a customer's rights, potentially resulting in misleading conduct (E2). The agent makes statements that may be considered personal financial advice, creating a risk that the organisation breaches its obligations as an Australian financial services licensee (E3).

Commercial and strategic. Organisations may select AI agents that are the easiest to procure rather than those best suited to the intended use case (E4). Over time, reliance on a single AI provider may also increase as prompts, workflows and integrations become embedded in business processes, making future switching more difficult and costly (E5).

Data. The agent aggregates significant personal information, including transaction history, communications, and voice biometric data, with potential implications under the *Privacy Act 1988 (Cth)* (E6).

Cyber and security. An external actor uses prompt injection combined with social engineering to manipulate the agent to bypass KYC controls or initiate an unauthorised transfer (E7). This event is selected as the worked example for the FAIR analysis in section 5.3.

Regulatory and compliance. Defects in the agent's authorisation pathway relating to KYC controls result in breaches of AML/CTF requirements (E8). The agent provides misleading information across several customer interactions before the issue is detected, leading to numerous complaints to the Australian Financial Complaints Authority (AFCA) (E9).

Third-party and vendor. The AI provider updates the underlying model and the agent's behaviour changes materially before the organisation detects the change (E10).

Operational. A single defect affects every interaction simultaneously before a key safeguard, the safe-state fallback, engages. This results in failure at scale (E11).

5.3 Quantification: the FAIR five-stage analysis

In E7, an external actor uses prompt injection combined with social engineering to manipulate the agent to bypass KYC controls or initiate an unauthorised transfer. This event is used as the worked example for the five-stage FAIR analysis in the next page.

The estimates are illustrative and would be calibrated against the bank's own data and threat intelligence. In a real-world assessment, this should be done for all risks above a frequency or severity threshold (or, in larger organisations with the resources, for all risks identified). Many risks will repeat across use cases, enabling organisations to build up a library that will materially reduce the effort required in future assessments.

Table 2. FAIR five-stage analysis worked example

Stage	Element	Assessment
1. Threat	An attacker combines prompt injection (malicious instructions hidden in content the agent ingests, such as an email body, an uploaded document or a chat message) with social engineering across channels, aiming to make the agent bypass a KYC or due diligence check, or initiate an unauthorised transfer.	Frequency estimate: low single digits of credible attempts per quarter at launch, rising as the agent's capabilities become known to attackers.
2. Vulnerability	The agent acts on systems of record and holds cross-channel state, so a successful injection on one channel can carry into an authorised-looking action on another. The key variable is whether irreversible actions sit behind a separate deterministic control the model cannot talk its way past.	Vulnerability rating: High where the conversational model can itself authorise a transfer or sign off a KYC step; materially lower where a hard authorisation boundary separates the two.
3. Loss event	The agent completes an unauthorised transfer, or passes a KYC or due diligence check it should have failed, opening an account or releasing funds to a party who should not have them.	Probability given threat and vulnerability: Moderate per successful attempt where the authorisation boundary is soft; the social-engineering style vulnerability of LLMs lifts success rates above pure technical injection.
4. Loss distribution	Direct financial losses arising from unauthorised transfers; customer remediation; incident response, investigation and forensic costs; regulatory consequences, particularly if it is part of a systemic AML/CTF failure (E8); and reputational damage.	Estimated range: $250K to $10M-plus per incident, right-skewed. The upper tail is not based on individual transferred amounts, but rather by whether the same pathway is exploited repeatedly before detection and whether it is characterised as a systemic AML/CTF control failure.
5. Controls	A separate deterministic authorisation boundary for irreversible actions (transfers above a threshold, account closure, KYC sign-off for entities above a complexity threshold) that the model cannot override; input provenance and injection filtering; step-up and out-of-band verification for high-consequence actions; transaction limits and velocity checks; regular red-teaming and adjustment of guardrails; monitoring for anomalous action sequences.	Residual frequency after controls: Low. Residual severity capped by transaction limits and reversibility windows, provided the authorisation boundary holds. The residual estimate assumes the boundary operates; it does not assume the boundary is perfect (see the degraded-control scenario that follows). Evidence of strong controls may be seen favourably by regulators even if the event still occurs, potentially managing reputational risks and reducing fines. The authorisation boundary may include sophisticated automated checks (validator AI agents, complex rule checks), and mandatory time delays for certain types of actions to mitigate detection lag.

A degraded-control scenario. The central estimate assumes the deterministic authorisation boundary holds. Suppose an exception pathway was built for legitimate edge cases: for example, a manual override the agent can invoke to handle hardship requests outside business hours. If that pathway is socially engineered, the boundary is partially bypassed and the residual distribution reverts towards the uncontrolled estimate, returning severity to the upper tail.

5.4 Controls calibrated to the use case

Many of the controls discussed here can be applied across a wide range of agentic AI use cases. These include record-keeping sufficient to show what the agent did and why, role-based access controls, vendor due diligence and contractual notice obligations on model changes (the control that addresses E10), and training to ensure staff can properly oversee and challenge the agent's outputs.

The deterministic authorisation boundary (as discussed in the controls row in Table 2 in the previous page) is the highest-impact control for this use case. This control governs whether the agent can complete high-consequence actions. The control can match the speed of the agent and remains effective when the agent operates at high volume. The control is independent of the AI system. As such, it is not susceptible to prompt injection and does not rely on the AI system behaving correctly.

Commercial risk

The deeper commercial risk is not vendor lock-in, although that remains a concern as integrations and organisational dependence increase over time (E5). It is the risk that procurement decisions prioritise ease of approval over capability: choosing the agent that clears CISO and procurement review most smoothly over the one that does the agentic work best (E4). A pre-packaged agent that slots into the existing compliance posture is attractive precisely because it generates no friction, while the more capable agent may demand new controls, a harder security review and unfamiliar vendor terms. Furthermore, that pre-packaged agent may lack the capabilities to perform the use case to the required standard. This is a fundamental failing that cannot be mitigated with additional controls.

The objective is not to select the system that is easiest to govern. It is to select the system that best meets the business need and implement governance arrangements proportionate to the risks it creates. To explain by analogy, a family sedan might be safer than a sports car, but you are unlikely to win a race in it. If the purpose is racing, buy the sports car and ensure you have the right safety equipment; do not buy the sedan and tell the board you bought a race car. The framework permits either choice, but it should be a deliberate decision based on capability and risk, rather than an unexamined preference for the option that is easiest to approve.

Human accountability

Human review of every interaction would defeat the purpose, since the agent exists to handle volume across channels. Rather than attaching accountability to each individual interaction, it applies to the rules and controls that govern the agent's operation. Humans own the policy that governs what the agent does, and they own it whether the agent handles one hundred interactions per day or one hundred thousand. Contestability and the related concept of reversibility, discussed in Section 4.2.3, helps us to determine the appropriate actions. For example:

- Answering a balance query or lodging a request is reversible and low consequence, so minimal guardrails are likely to be required beyond identifying the customer.
- Waiving a fee or granting hardship is reversible and within reach of internal dispute resolution, but still carries financial consequences, so may incur additional automated verification of the real-time agent's decision, as well as additional guardrails to ensure explainability and fairness.
- Completing a KYC check, closing an account or releasing a transfer above a threshold is irreversible, or at least cannot be easily unwound, so complex cases may still be escalated to human review, or the automated review may need to be very sophisticated. A mandatory time delay may need to be introduced to limit the impact of detection lag for emerging problems (e.g., a new adversarial attack pattern designed to fool the agent into passing a KYC check for a fraudulent entity).

The internal and external dispute pathways must be genuinely reachable, timely, and substantive, because an agent operating at scale will produce a high volume of disputes if it makes errors. When automation multiplies throughput, a constant per-interaction defect rate produces a rising absolute number of defects. Quality assurance arrangements built for human-scale volumes may be inadequate. As throughput increases significantly, organisations should strengthen (not just scale) quality assurance, monitoring and statistical process controls. Otherwise, the bank is implicitly accepting a higher absolute number of defective interactions than may be consistent with its risk tolerance.

6. Conclusion

This paper has set out a practical AI risk management framework for Australian financial services, grounded in the sector's regulatory architecture and informed by the governance gaps identified. The framework is supported by a toolkit of resources to help practitioners apply the framework. It is designed to be durable, as it is anchored to risk outcomes rather than specific technologies, but its effectiveness ultimately depends on the people responsible for implementing it.

No framework compensates for a workforce that does not understand what it is governing. Boards need sufficient understanding of AI system behaviour to challenge management assertions about AI risk. Risk and audit functions need practitioners who can evaluate AI systems on their own terms. Actuarial and quantitative professionals face a redefinition of their role as AI models supplement or replace traditional actuarial models. Professional bodies, including the Actuaries Institute, have a role in defining the competency standards that practitioners will need.

The framework presented here is a starting point. It will require extension as agentic architectures mature, as regulatory and community expectations evolve, and as the industry accumulates operational experience with AI systems at scale. The organisations best positioned to manage risks in this changing environment will be those that have built governance structures capable of adapting to these shifts, not those waiting for certainty before beginning.

7. Appendices

Appendix A: HTI FS Survey

Methodology

The survey was completed by n=419 business decision-makers during August 2025 sourced from a mix of Roy Morgan's business panel, members of Roy Morgan's consumer panel whose occupation roughly matched the target audience described below, and from an external panel provider. Of the 419 respondents, 27 respondents indicated that they primarily worked within the financial services industry.

The survey was emailed to participants. Participation was optional and anonymous.

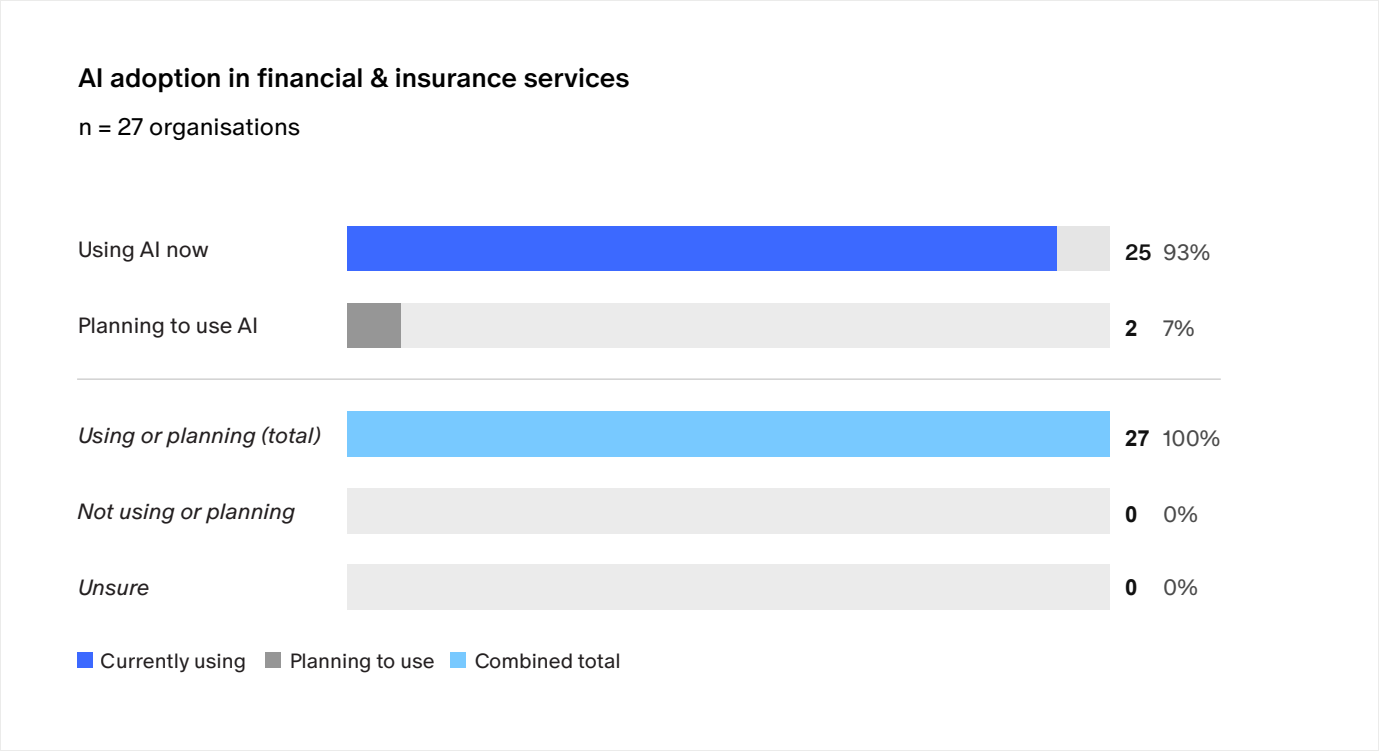
Survey participants were screened for the size of their organisation and excluded if their organisation had four or fewer employees. Participants were also screened out if the role they spent the most time working in at their current organisation did not match to one of the following: Board member or Non-Executive Director; Chairperson; CEO, Managing Director, or Executive Director; Senior Executive or C-suite role; or Other strategic decision-maker.

Survey participants that completed the survey were invited to participate in a follow-up survey in September/October 2025 and asked further questions regarding agentic AI. Of the 419 participants that completed the main survey, 295 completed the follow-up survey.

Results

The questions and results referred to in this paper are outlined in full below.

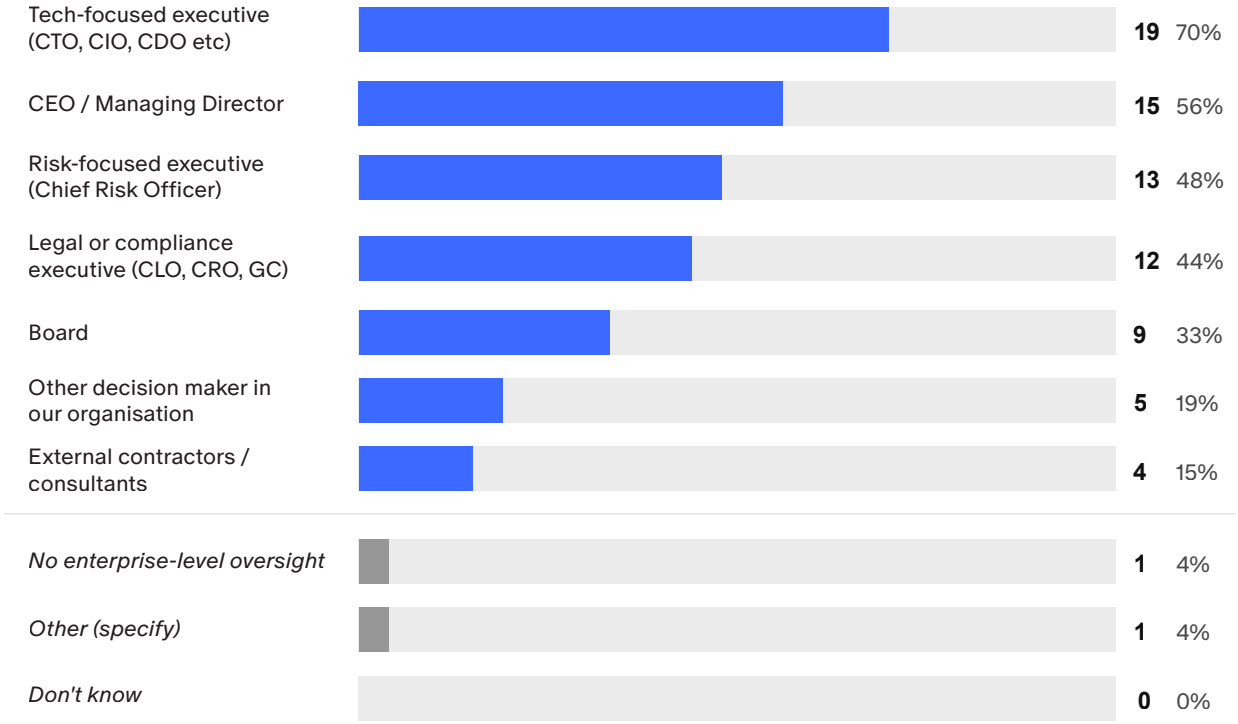
Q1: Which best describes your organisation's use of AI, being digital systems that learn from data to make predictions, give advice, make decisions or generate content?



Q6: Which role(s) are primarily responsible for ensuring that AI systems relied on in your organisation operate safely, lawfully and responsibly?

Who is primarily responsible for AI systems?

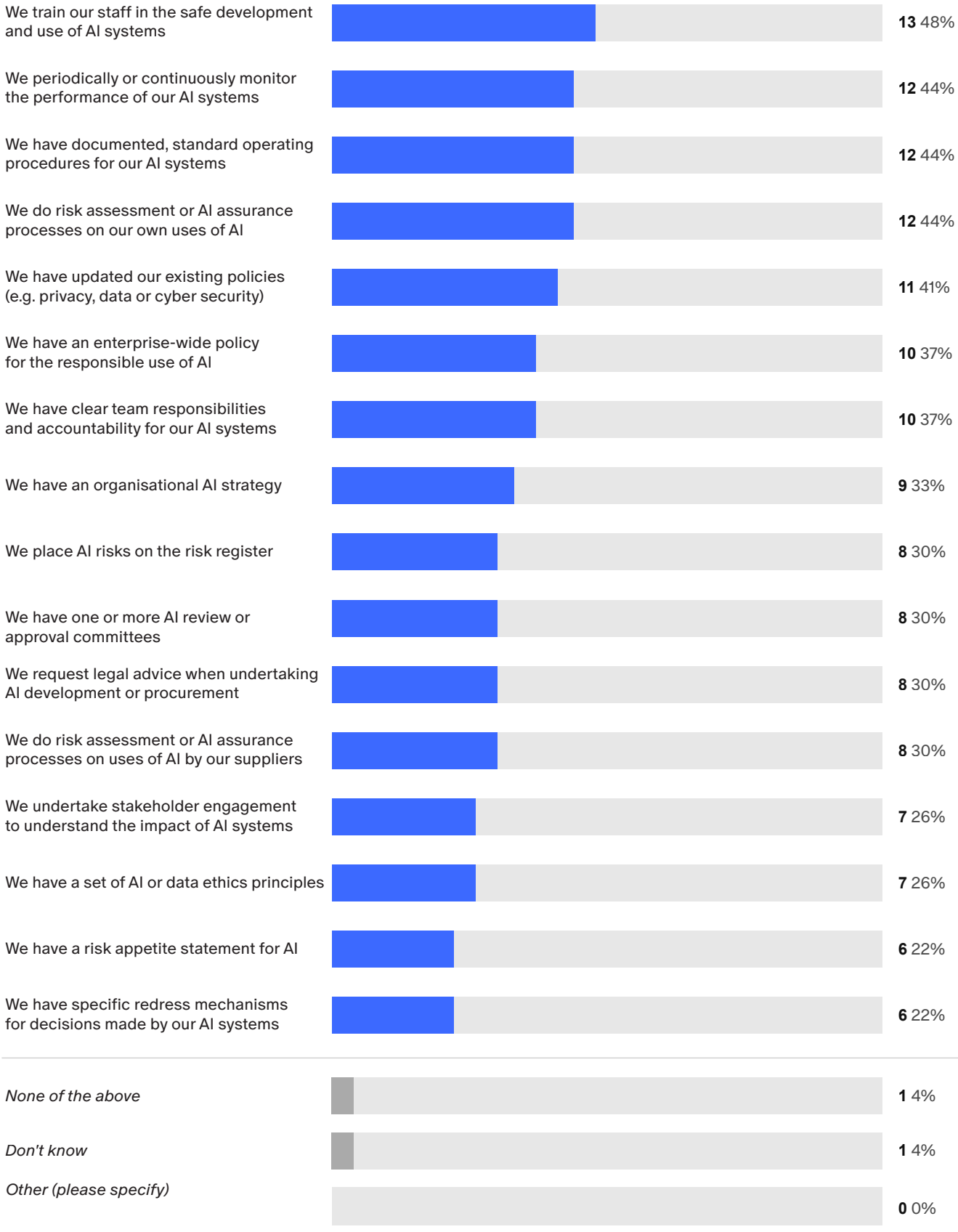
Financial and insurance services · n = 27 · Multiple responses allowed



Q7: Which current systems and processes are you aware of in your organisation to identify and mitigate AI risks?

AI systems and processes in place

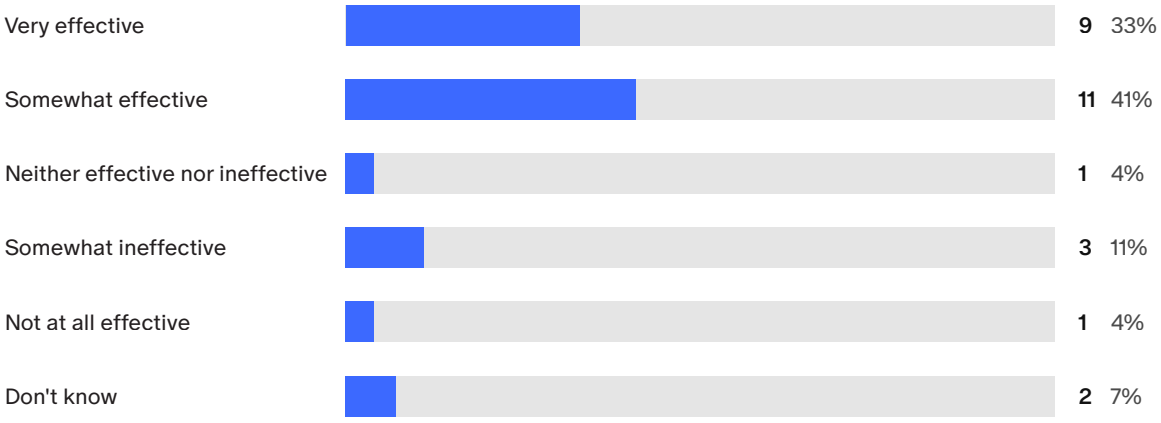
Financial and insurance services · n = 27 · Multiple responses allowed



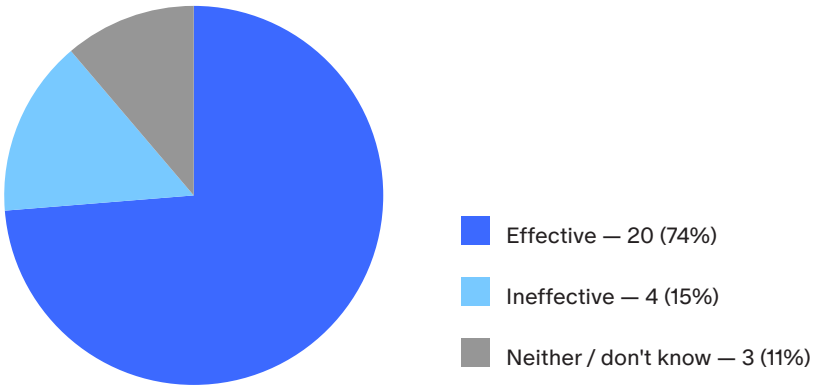
Q8: How would you rate the current effectiveness of systems and processes within your organisation to identify and manage risks that could be created by AI systems?

Effectiveness of organisation's AI systems and processes

Financial and insurance services · n = 27



Effective vs ineffective



Appendix B:
Glossary of AI/ML Terms in Financial Services Context

The following terms are those most likely to arise in AI risk discussions at board and senior management level. Basic risk management and technology terms are omitted on the assumption that the reader does not need them defined.

Term	Definition and Risk Relevance
Adversarial attack	A deliberate attempt to cause an AI system to produce incorrect outputs by manipulating its inputs in ways that may be imperceptible to humans. Risk relevance: adversarial inputs to fraud detection, credit scoring, or identity verification systems could cause systematic misclassification. Unlike conventional system exploits, adversarial attacks target the model's learned decision boundary rather than its code.
Agentic AI	AI systems that can autonomously plan, execute multi-step tasks, invoke tools, and take actions with limited or no human intervention between steps. Risk relevance: agentic systems can take consequential actions (e.g., executing trades, sending communications, modifying data) that compound errors across steps before a human can intervene, and their behaviour may be difficult to predict or bound in advance.
AI safety	The field of research and practice concerned with ensuring AI systems do not cause unintended harm, particularly as systems become more capable and autonomous. Risk relevance: distinct from AI ethics (which focuses on fairness and societal impact), AI safety addresses failure modes such as reward hacking, specification gaming, and loss of human control. These are concerns that become material as financial services adopt agentic AI for higher-stakes processes.
Alignment	The degree to which an AI system's objectives and behaviour correspond to the intentions and values of its designers and operators. Risk relevance: a misaligned system may optimise for a proxy metric that diverges from the organisation's actual objective. For example, by maximising loan approvals rather than maximising risk-adjusted return. Alignment failures are particularly difficult to detect when the system appears to be performing well on its training metrics.
Chain-of-thought (CoT)	A prompting or training technique where a language model produces intermediate reasoning steps before arriving at a final answer. Risk relevance: CoT improves accuracy on complex tasks and provides a form of interpretability, as the reasoning trace can be audited. However, research has shown that a model's stated reasoning may not faithfully reflect its actual internal computation, limiting its reliability as an audit tool.
Context window	The maximum amount of text (measured in tokens) that a language model can process in a single interaction, encompassing both the input prompt and the generated output. Risk relevance: context window limits determine how much information the model can consider simultaneously. In document review or regulatory analysis, exceeding the context window causes the model to silently lose information, potentially producing incomplete or misleading outputs without signalling that it has done so.
Data poisoning	The deliberate or accidental introduction of corrupted, biased, or malicious data into a model's training set. Risk relevance: poisoned training data can introduce systematic errors or backdoors that persist through model deployment. For vendor-supplied models, the deploying organisation typically has no visibility into training data provenance, making this a supply chain risk that cannot be mitigated through internal controls alone.
Embeddings	Dense numerical representations (vectors) that encode the semantic meaning of text, images, or other data in a form that machines can process. Risk relevance: embeddings underpin retrieval-augmented generation (RAG) systems, semantic search, and similarity matching. Errors in embedding quality, or the use of embeddings trained on data that does not represent the target domain, can cause systematic retrieval failures in knowledge management and compliance applications.
Fine-tuning	The process of further training a pre-trained foundation model on a smaller, task-specific dataset to specialise its performance. Risk relevance: fine-tuning can introduce new risks if the task-specific data contains biases, errors, or sensitive information. It also creates a model variant that diverges from the base model's tested behaviour, potentially invalidating vendor-supplied safety evaluations. Organisations fine-tuning models need model risk management processes equivalent to those for internally developed models.

Term	Definition and Risk Relevance
Foundation model	A large AI model trained on broad data at scale, designed to be adapted (via fine-tuning, prompting, or other methods) to a wide range of downstream tasks. GPT-5.5, Claude, Gemini, and Llama are examples. Risk relevance: foundation models concentrate risk in a small number of upstream providers. A single model failure, security vulnerability, or provider outage can simultaneously affect multiple downstream applications across the organisation. This creates correlated risk that conventional vendor diversification strategies may not address.
Function calling (tool use)	A capability that allows a language model to invoke external functions, APIs, or tools during its operation. For example, this might be querying a database, executing code, or calling a third-party service. Risk relevance: function calling extends the model's action space beyond text generation into consequential operations. An incorrectly-specified or poorly-constrained function call can result in data modification, unauthorised access, or unintended transactions. This is the primary mechanism through which language model errors translate into operational impact.
Hallucination	The generation of plausible sounding but factually incorrect or fabricated content by a language model, including invented citations, fictitious data, or false assertions stated with apparent confidence. Risk relevance: hallucination is not a bug that will be patched. It is an inherent property of probabilistic text generation. In financial services, hallucinated regulatory references, fabricated precedents, or incorrect calculations in AI-assisted advisory, compliance, or reporting functions can create legal, regulatory, and reputational exposure. Detection requires domain-expert review.
Human in the loop / Human on the loop	Two distinct oversight models. Human in the loop requires a human to approve or reject each AI output before it takes effect. Human on the loop allows the AI to act autonomously while a human monitors for exceptions and retains override capability. Risk relevance: the choice between these models determines the organisation's exposure to automation speed versus oversight adequacy. Human on the loop is appropriate only where the monitoring mechanisms can detect errors faster than those errors compound, and where override can be exercised before harm materialises.
Inference	The process of running a trained model to generate outputs (predictions, classifications, or text) from new inputs, as distinct from training the model. Risk relevance: inference is the operational phase where AI risk materialises. Inference costs, latency, and availability are operational dependencies. Inference also presents data security considerations, as input data sent to cloud-hosted inference endpoints may traverse third-party infrastructure.
Jailbreaking	Techniques used to circumvent a language model's built-in safety restrictions and behavioural guidelines, causing it to produce outputs it was designed to refuse. Risk relevance: jailbreaking demonstrates that safety guardrails in language models are not absolute controls. For customer-facing AI systems, jailbreaking can result in the generation of harmful, defamatory, or non-compliant content attributed to the deploying organisation. Internal systems are also at risk where employees discover that bypassing restrictions yields more useful outputs.
Mixture of Experts (MoE)	A model architecture in which multiple specialised sub-networks ("experts") exist within a single model, with a routing mechanism that activates only the relevant experts for each input. Risk relevance: MoE architectures mean that a model's behaviour effectively varies depending on which experts are activated, creating inconsistency that complicates validation and testing. A model that performs well on one category of inputs may perform poorly on another category that routes to different experts.
Model drift	The degradation of a model's performance over time as the statistical properties of the data it encounters in production diverge from the data on which it was trained. Risk relevance: drift is insidious because the model continues to produce outputs; it does not fail visibly. In credit scoring, claims assessment, or fraud detection, drift can cause systematic under- or over-estimation that accumulates over months before detection.
Multi-modal	AI systems capable of processing and generating multiple types of data, such as text, images, audio, video, or structured data, within a single model. Risk relevance: multi-modal systems expand the attack surface and the scope of potential failure. A multi-modal system processing insurance claims might misinterpret photographic evidence, generate incorrect textual summaries, and produce flawed assessments. Errors in one modality may compound errors in another. Validation must cover all modalities and their interactions.
Prompt injection	An attack in which malicious instructions are embedded in a model's input, either directly by the user or indirectly through data the model processes, causing the model to ignore its system instructions and follow the attacker's instructions instead. Risk relevance: prompt injection is the most significant security vulnerability in current language model deployments. It can cause data exfiltration, unauthorised actions, or the bypassing of compliance controls. There is currently no complete technical defence; mitigation requires layered architectural controls.

Term	Definition and Risk Relevance
RAG (Retrieval-Augmented Generation)	An architecture that augments a language model's outputs by first retrieving relevant information from an external knowledge base (e.g., policy documents, regulatory texts, internal data), then using that retrieved content as context for generation. Risk relevance: RAG is the primary method for grounding language model outputs in organisational knowledge, reducing hallucination and improving accuracy. However, RAG introduces its own risks: retrieval failures, poisoned knowledge bases, access control violations (where the model surfaces information the user should not see), and a false sense of reliability when the retrieved content is itself outdated or incorrect.
Red-teaming	Structured adversarial testing of an AI system, in which testers deliberately attempt to provoke failures, unsafe outputs, bias, or security vulnerabilities. Risk relevance: red-teaming is the AI equivalent of penetration testing. It is the most effective method for identifying failure modes that do not appear in standard testing. For financial services, red-teaming should cover not only security and safety but also regulatory compliance scenarios, fairness testing across protected attributes, and stress-testing of human oversight mechanisms.
RLHF (Reinforcement Learning from Human Feedback)	A training technique in which human evaluators rate model outputs, and these ratings are used to train a reward model that guides further training of the language model. Risk relevance: RLHF is the primary mechanism through which language models are made to follow instructions and refuse harmful requests. Its effectiveness depends on the quality and representativeness of the human feedback. RLHF can introduce biases that reflect the evaluators' perspectives, and it can cause models to favour outputs that appear helpful over outputs that are accurate. This tendency is referred to as sycophancy.
Shadow AI	The use of AI tools and services by employees without organisational knowledge, approval, or governance: analogous to shadow IT. Risk relevance: shadow AI is arguably the most immediate and pervasive AI risk in financial services. Employees using personal ChatGPT accounts to draft client communications, analyse confidential data, or generate regulatory responses create data leakage, compliance, and liability exposure that the organisation cannot monitor or control.
Temperature	A parameter that controls the randomness of a language model's outputs. Lower temperature produces more deterministic, predictable outputs; higher temperature produces more varied and creative outputs. Risk relevance: temperature settings directly affect output consistency and reproducibility. For regulated processes requiring auditability and consistency (e.g., credit decisioning support, compliance checking), low temperature is essential.
Tokens	The sub-word units into which language models break text for processing. A token typically represents 3-4 characters of English text, though this varies by model and tokeniser. Risk relevance: token counts determine processing costs, context window utilisation, and rate limits. More practically, tokenisation affects how models handle numbers, dates, account numbers, and domain-specific terminology. Financial data can be tokenised in ways that impair the model's numerical reasoning.

Appendix C:
ASIC - Questions for Australian financial services licensees

ASIC Report 798 includes a set of governance questions for licensees. These questions, reproduced below with minor adaptation for a broader financial services audience, provide a useful self-assessment baseline. For each question, "no" or "not yet" is a valid answer. The purpose is to identify gaps, not to prescribe a single correct approach.

1. **Taking stock**
   - Where on your AI journey are you? Do you know where AI is used in your organisation?
   - Do you have an AI inventory, and are you confident that it is being adequately maintained?

2. **AI strategy**
   - Are you clear where you are going, and why?
   - Do you have a clear and documented strategy for what you want to achieve with AI, now and in the future? How does this align with your business objectives and risk appetite?

3. **Ethics and fairness**
   - What ethical challenges does your use of AI raise?
   - How do you meet your obligations to provide financial services and credit efficiently, honestly and fairly when using AI?

4. **Accountability**
   - Who is accountable for AI use and outcomes, at model level and overall? Do they get the reporting they need to do their job?
   - How are you measuring consumer outcomes from AI? Are you delivering benefits and avoiding harms?
   - For accountable entities under the Financial Accountability Regime (FAR), have you considered the use of AI in key functions when assigning accountable persons and establishing clear lines of accountability?

5. **Risk**
   - Are you clear on conduct and regulatory compliance risk from AI, particularly risk to consumers? What is your risk tolerance?
   - How are you identifying, mitigating and monitoring risk throughout the AI lifecycle?
   - How will you document this, and monitor adherence to it?
   - Do you have staff from multiple disciplines involved in assessing risk, and not just technical experts?
   - Are your assessments of risk, your controls and your monitoring still adequate if your risk profile changes with your use of AI?

6. **Alignment**
   - Do your governance arrangements lead your AI use, now and for your future AI plans?
   - How do the risks and ethical challenges change with a move towards more complex and opaque AI, such as generative AI? What changes do you need to make as a result to ensure your governance leads your use?

7. **Policies and procedures**
   - Have you translated your AI strategy and assessment of risk into policies for your staff, setting clear expectations through the AI lifecycle? Is your approach risk based? Do policies lead your AI use?
   - Are your AI policies and procedures fit for purpose, now and for anticipated future use?
   - How do you ensure your staff adhere to your AI policies and procedures?

8. **Resourcing**
   - Do you have the right technological and human resources at all levels? How do you ensure your resources remain fit for purpose as AI use accelerates and evolves?
   - How do you ensure your staff at all levels, including compliance and internal audit staff, have the skills and voice to engage with AI decisions and monitoring in their roles?

9. **Oversight and monitoring**
   - Are you clear on what human oversight you expect? Do you have procedures for when things go wrong?
   - Do you have an action plan if a model is found to be producing unexpected outputs?
   - Have you considered the adequacy of your business continuity, backup and disaster recovery plans for AI systems?

10. **Third parties**
    - How do you manage the challenges of relying on third parties?
    - How will you validate, monitor and review third-party AI models?

11. **Regulatory reform**
    - Are you engaging with the regulatory reform proposals on AI?

Organisations that can answer most of these questions confidently are likely well positioned. Those that answer "no" to several may wish to prioritise governance work. Those that answer "not yet" to most should at least have a plan for when and how they will address them.

Appendix D:
AI-Specific Cyber Threat Examples

The body of this paper presents two AI-specific cyber threat examples (prompt injection and supply chain attacks). The following additional threat types are relevant to AI deployments in financial services:

- **Training data poisoning:** An attacker introduces malicious data into the training set, causing the model to learn compromised behaviour that activates under specific conditions. This is analogous to a supply chain attack on software, but the "code" is model weights rather than executable instructions, making it harder to detect through conventional code review or static analysis.
- **Model extraction and model inversion:** In **model extraction**, an attacker queries the AI system systematically to reconstruct a functionally equivalent copy of the model, including proprietary business logic embedded in fine-tuning. In **model inversion**, an attacker uses carefully crafted queries to extract memorised training data, including potentially sensitive customer or business information. For generative AI systems, model inversion is the more practical concern: large language models have been shown to memorise and reproduce training data under specific prompting conditions.
- **Jailbreaking and guardrail bypass:** Attackers develop techniques to circumvent the safety restrictions built into AI models, causing them to produce outputs they were designed to refuse. The techniques evolve rapidly, and there is no static set of controls that remains effective.
- **Data exfiltration through AI interaction:** Employees or customers who interact with AI systems may inadvertently or deliberately introduce sensitive data into prompts. If the AI system is hosted externally, this creates a data exfiltration pathway that does not route through the organisation's data loss prevention controls.
- **Agentic AI action chains:** When AI systems can autonomously execute multi-step action sequences (such as querying databases, calling APIs, modifying records, and sending communications), the attack surface expands from data exfiltration to unauthorised action execution. A compromised agentic system does not merely leak information; it can take consequential actions using the organisation's own credentials and permissions. Controls for agentic systems must include action-level authorisation, execution boundaries enforced at the infrastructure level (not merely through natural language instructions), and comprehensive action logging with real-time monitoring for anomalous behaviour patterns.

Appendix E:
Vendor Due Diligence Checklist for AI Procurement

This checklist covers the additional considerations specific to AI systems. It should be used alongside, not instead of, the organisation's existing vendor risk assessment process.

Organisations should score vendor responses against these criteria and establish thresholds for each risk category proportionate to the intended use case.

Model and performance

- ☐ What model architecture is used, and what version is currently deployed?
- ☐ What training data was used, and can the vendor provide documentation of data provenance and quality?
- ☐ What are the model's known limitations, failure modes, and edge cases?
- ☐ How does the vendor measure and report model performance?
- ☐ How frequently is the model updated, and what is the notification and testing process for updates?
- ☐ Can the organisation test model updates in a staging environment before production deployment?
- ☐ What is the vendor's process for addressing model degradation or performance issues?
- ☐ Can the organisation reject or defer model upgrades released by the vendor?
- ☐ How does the vendor receive and incorporate feedback on model performance issues?

Data handling and privacy

- ☐ Where is data processed and stored (jurisdiction, infrastructure)?
- ☐ Does the vendor use customer data to train or improve its models? If so, can this be disabled?
- ☐ What data retention policies apply to inputs, outputs, and interaction logs?
- ☐ Does the vendor's data handling comply with the Australian Privacy Principles?
- ☐ Has the vendor completed an independent privacy or data protection audit?
- ☐ What happens to the organisation's data if the contract is terminated?

Security

- ☐ What security certifications does the vendor hold (SOC 2, ISO 27001, etc.)?
- ☐ What AI-specific security controls does the vendor implement (prompt injection defences, adversarial testing, model integrity verification)?
- ☐ Does the vendor conduct regular penetration testing that includes AI-specific attack vectors?
- ☐ What is the vendor's incident response process, and what notification commitments exist?
- ☐ How are model weights and components secured against supply chain attacks?
- ☐ Who from the vendor would have access to customer data, under what circumstances, and what controls govern that access?

Compliance and regulatory

- ☐ Can the vendor's system produce explanations of individual decisions sufficient for regulatory examination?
- ☐ Does the vendor have experience operating in regulated financial services environments (e.g., including obligations under CPS 230)?
- ☐ Will the vendor support the organisation's compliance obligations (e.g., providing information for APRA or ASIC requests)?
- ☐ Does the vendor's acceptable use policy conflict with the organisation's intended use case?

Commercial and continuity

- ☐ What is the vendor's financial stability and viability? Assess whether the vendor's balance sheet can support the indemnities they are offering.
- ☐ What is the contractual lock-in period, and what are the exit provisions?
- ☐ Can the organisation migrate to an alternative provider without loss of functionality or data?
- ☐ What is the pricing model, and what are the cost escalation risks (e.g., usage-based pricing, token costs)?
- ☐ Does the contract include service level agreements with meaningful remedies?
- ☐ What intellectual property rights does the organisation have over fine-tuned models, configurations, and outputs?
- ☐ What is the vendor support model (including availability of in-region support resources)?

Concentration and dependency

- ☐ What is the organisation's aggregate dependency on this vendor across all AI use cases?
- ☐ Does the vendor depend on a single underlying model provider (e.g., reselling OpenAI, Anthropic, or Google models)?
- ☐ What is the contingency if the underlying model provider changes terms, pricing, or availability?

Appendix F:
AI Risk Assessment Templates

F.1 AI System Triage Form

This form is designed as an initial screening instrument. Its purpose is to classify incoming AI systems or use cases into risk tiers to determine the appropriate depth of subsequent assessment. It is not a substitute for a full risk assessment; it determines whether a full assessment is required, and at what level of rigour.

Complete one form per AI system or distinct AI use case. Where a vendor product is used for multiple purposes, each purpose should be triaged separately.

Section 1: System Identification

Field	Entry
System / product name	
Vendor (if applicable)	
Version or model identifier	
Business owner (name, role)	
Sponsoring business unit	
Date of triage	
Triage completed by	

Section 2: System Description

Field	Entry
Brief description of what the system does (2–3 sentences)	
Business problem it addresses	
AI/ML technique employed (e.g., predictive model, generative LLM, classification, recommendation, agentic)	
Is this replacing an existing process or creating a new capability?	

Section 3: Data Inputs

Field	Entry
What data does the system ingest?	
Data source(s) (internal systems, third-party feeds, user-provided, public)	
Does the data include personal information (as defined under the Privacy Act 1988 (Cth))?	☐ Yes ☐ No
Does the data include sensitive information (health, financial, biometric)?	☐ Yes ☐ No
Data sensitivity classification per organisational policy	☐ Public ☐ Internal ☐ Confidential ☐ Restricted
Is data transmitted to external infrastructure for processing?	☐ Yes ☐ No ☐ Unknown
Data retention: is input data retained by the vendor or model provider?	☐ Yes ☐ No ☐ Unknown

Section 4: Decision Impact

Field	Entry
Who is affected by the system's outputs? (staff, customers, counterparties, general public)	
What decisions do the outputs inform or determine?	
Could the outputs result in denial of service, financial loss, or legal detriment to an individual?	☐ Yes ☐ No
Are affected individuals notified that AI is involved in the process?	☐ Yes ☐ No ☐ N/A
Reversibility: can decisions be reversed or corrected after the fact?	☐ Easily ☐ With difficulty ☐ Irreversible
Could errors compound before detection (e.g., batch processing, automated chains)?	☐ Yes ☐ No

Section 5: Autonomy Level

Level	Description	Select
Human in the loop	AI generates recommendations or drafts; a human reviews and approves every output before it takes effect	☐
Human on the loop	AI acts autonomously in normal operation; human monitors for exceptions and retains override capability	☐
Fully autonomous	AI acts without real-time human oversight; human review occurs after the fact (if at all)	☐

Section 6: Deployment Scope

Field	Entry
Deployment type	☐ Internal only ☐ Customer-facing ☐ Embedded in product or service
Number of users or affected individuals (order of magnitude)	
Geographic scope	☐ Australia only ☐ International
Integration with other systems (list key integrations)	
Is the AI output presented as AI-generated, or is it indistinguishable from human output?	

Section 7: Third-Party Dependency

Field	Entry
Hosting model	☐ On-premise ☐ Private cloud ☐ Vendor-hosted SaaS ☐ API-based
Can the organisation operate if this vendor becomes unavailable?	☐ Yes ☐ With degradation ☐ No
Does the vendor retain the right to update the model without notice?	☐ Yes ☐ No ☐ Unknown
Has the vendor provided a model card, system card, or comparable documentation?	☐ Yes ☐ No
Contractual provisions for audit, incident notification, and liability	☐ Adequate ☐ Partial ☐ Absent ☐ Not assessed
Subprocessors: does the vendor use third-party AI models or infrastructure?	☐ Yes ☐ No ☐ Unknown

Section 8: Risk Triage Scoring

Score each dimension. The highest individual score determines the triage outcome, not the average. The range of levels provided against each dimension are illustrative, as are the triage outcomes. These should be calibrated based on your organisation's context.

Dimension	Low (1)	Medium (2)	High (3)	Critical (4)	Score
Data sensitivity	Public or non-personal data only	Internal data, limited personal information	Confidential data, personal information including financial	Restricted data, sensitive information, health or biometric	
Decision impact	Informational only, no direct effect on individuals	Influences decisions but human makes final determination	Directly determines outcomes affecting individuals' finances or services	Determines outcomes that are irreversible or affect vulnerable populations	
Autonomy	Human in the loop with mandatory review	Human on the loop with effective monitoring	Autonomous with after-the-fact review	Autonomous with limited or no review	
Scale	<100 affected individuals or internal use only	100–10,000 affected individuals	10,000–100,000 affected individuals	>100,000 affected individuals or systemic impact	
Vendor dependency	On-premise, full organisational control	Private cloud, contractual controls in place	Vendor-hosted, limited contractual protections	API-based, no audit rights, model updates without notice	
Regulatory exposure	No specific regulatory requirements	General regulatory obligations apply	Sector-specific regulatory requirements (e.g., RG 209, anti-discrimination)	Prudentially significant; APRA notification likely required	

Triage Outcome

Highest Score	Risk Tier	Required Action
1	Low	Register the system. Standard change management applies. Periodic review (annual).
2	Medium	Full AI risk assessment required. Include on AI risk register. Six-monthly review.
3	High	Full AI risk assessment with independent review. Board or risk committee reporting. Quarterly review. Enhanced monitoring.
4	Critical	Full AI risk assessment with independent review. Board approval required before deployment. Continuous monitoring. Incident response plan required. Consider regulatory engagement.

F.2 AI Risk Register — Supplementary Fields

Consider adding the following fields to the organisation's existing enterprise risk register to capture AI-specific risk dimensions. These are supplementary to, not replacements for, standard risk register fields (risk ID, description, likelihood, consequence, rating, owner, controls, status). They address the characteristics that distinguish AI risks from conventional operational and technology risks.

Field	Description	Example Values
AI system identifier	Unique reference linking to the AI system inventory and triage form	SYS-AI-2025-017
Model type	Classification of the AI technique in use	Predictive (supervised ML) / Generative (LLM) / Agentic / Classification / Recommendation / Embedded (AI feature within broader product)
Foundation model provider	For systems built on foundation models, the upstream provider	OpenAI / Anthropic / Google / Meta / Microsoft / Internal / N/A
Model version or checkpoint	Specific model version in production, to track against updates	GPT-4o-2025-05 / Claude 3.5 Sonnet / Internal v2.3
Data sensitivity classification	Highest classification of data processed by the system	Public / Internal / Confidential / Restricted
Personal information processed	Whether the system processes personal information under the Privacy Act 1988 (Cth)	Yes — identified / Yes — de-identified / No
Vendor or provider	Entity supplying the AI system or model	Vendor name, contract reference
Hosting and data jurisdiction	Where inference occurs and where data is processed	Australia / US / EU / Multi-region / Unknown
Last validation date	Date of most recent model validation or performance review	DD/MM/YYYY
Validation method	How the system was last validated	Accuracy testing / Bias audit / Red-team exercise / Vendor attestation / Not validated
Next scheduled validation	When the next validation is due	DD/MM/YYYY
Drift monitoring status	Whether ongoing performance monitoring is in place and its current state	Active — no drift detected / Active — drift detected / Not established
Drift monitoring method	Technical approach to drift detection	Statistical process control / Output sampling / Performance metric tracking / Manual review / None
Regulatory classification	Whether the AI use case falls under specific regulatory obligations	CPS 220 (risk management) / CPS 230 (operational risk) / CPS 234 (information security) / RG 209 (credit) / Anti-discrimination legislation / Privacy Act 1988 (Cth) / None specifically identified
Regulatory notification	Whether the regulator has been or should be notified of this AI deployment	Notified / Notification planned / Not required / Under assessment
Human oversight mechanism	The form of human oversight applied to this system	Human in the loop / Human on the loop / Autonomous with periodic review / None
Override capability	Whether and how a human can override or halt the system	Real-time override / Batch halt / Post-hoc correction only / No override mechanism

Field	Description	Example Values
Explainability	Whether the system's outputs can be explained to affected individuals and regulators	Fully explainable / Partially explainable (feature importance available) / Black box / Not assessed
Incident history	Record of AI-specific incidents associated with this system	Number of incidents, date of most recent, severity
Incident categories	Types of incidents experienced	Hallucination / Drift / Bias detected / Data leakage / Prompt injection / Availability / Other
Downstream dependencies	Other systems or processes that consume this AI system's outputs	List of dependent systems, processes, or reports
Upstream dependencies	Systems or data feeds this AI system depends on	List of data sources, APIs, model providers
Concentration risk	Whether this system shares infrastructure, models, or providers with other AI systems	Shared foundation model / Shared cloud provider / Shared data pipeline / Independent

Endnotes

1 Reserve Bank of Australia, *Composition of the Australian Economy Snapshot* (Report, May 2026) <https://www.rba.gov.au/snapshots/economy-composition-snapshot/pdf/economy-composition-snapshot.pdf>.

2 King & Wood Mallesons & Sapere, *The Impact of AI on the Australian Finance Industry: Report for the Australian Finance Industry Association* (Report, May 2025) <https://www.mallesons.com/au/en/insights/latest-thinking/ai-in-the-australian-finance-industry-opportunities-risks-and-benefits.html>.

3 Australian Institute of Company Directors & HTI, *A Director's Guide to AI Governance* (Report, June 2026) <https://www.aicd.com.au/content/dam/aicd/pdf/news-media/research/2026/director-guide-to-ai-governance.pdf>.

4 For an introduction to this debate, see Madeleine Clare Elish, 'Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction' (2019) 5 Engaging Science, Technology, and Society 40 <https://doi.org/10.17351/ests2019.260>.

5 Anthropic announced 'Project Glasswing' in April 2026, detailing the sophisticated cyber security capabilities of its new 'Mythos' model and electing to release it to a selection of organisations to limit misuse. In May 2026, OpenAI announced 'Daybreak', similarly enabling select organisations to use a modified version of GPT-5.5 for cyber defence.

6 For example, see Australian Signals Directorate, *Frontier models and their impact on cyber security* (Web Page, 9 April 2026) <https://www.cyber.gov.au/about-us/view-all-content/news/frontier-models-and-their-impact-on-cyber-security>; Australian Prudential Regulation Authority (APRA), *Letter to Industry on Artificial Intelligence (AI)* (30 April 2026) <https://www.apra.gov.au/news-and-publications/apra-letter-industry-artificial-intelligence-ai>; and Australian Securities and Investments Commission (ASIC), *Open letter from ASIC to AFS Licensees and directors* (8 May 2026) <https://download.asic.gov.au/media/xhrf1w0e/26-092mr-open-letter-to-afs-licensees-and-market-participants.pdf>.

7 For further information on the survey methodology, see Appendix A.

8 UTS Human Technology Institute Financial Services Survey (2025), see Appendix A.

9 Australian Securities and Investments Commission, *Beware the Gap: Governance Arrangements in the Face of AI Innovation* (Report No 798, 29 October 2024) <https://download.asic.gov.au/media/mtllqjo0/rep-798-published-29-october-2024.pdf>.

10 Australian Prudential Regulation Authority, *APRA Letter to Industry on Artificial Intelligence (AI)* (30 April 2026) <https://www.apra.gov.au/news-and-publications/apra-letter-industry-artificial-intelligence-ai>.

11 Council of Financial Regulators, *Quarterly Statement by the Council of Financial regulators - June 2026,* Media Release Number 2026-04 (23 June 2026) <https://www.cfr.gov.au/news/2026/mr-26-04.html>; Australian prudential Regulation Authority, *APRA Letter to Industry on Artificial Intelligence (AI)* (30 April 2026) <https://www.apra.gov.au/news-and-publications/apra-letter-industry-artificial-intelligence-ai>; and Australian Securities and Investments Commission, *Open letter to AFS Licensees and directors* (8 May 2026) <https://download.asic.gov.au/media/xhrf1w0e/26-092mr-open-letter-to-afs-licensees-and-market-participants.pdf>.

12 Australian Securities and Investments Commission, *ASIC calls for urgent cyber uplift as AI accelerates cyber threats* (Media Release, 8 May 2026) <https://www.asic.gov.au/about-asic/news-centre/find-a-media-release/2026-releases/26-092mr-asic-calls-for-urgent-cyber-uplift-as-ai-accelerates-cyber-threats/>.

13 Letter to all APRA-regulated entities (n 10).

14 The Australian Government's Guidance for AI Adoption recommends that organisations document and clearly communicate who is accountable across the organisation for the operation of the AI management system. It also suggests assigning a senior leader as the overall AI governance owner: 'Department of Industry, Science and Resources (Cth), 'Guidance for AI Adoption: Guidance for AI Adoption Implementation Practices' <https://www.ai.gov.au/staying-safe-and-responsible/essential-ai-practices/guidance-ai-adoption-implementation-guidance>.

15 APRA expects entities to establish consistent governance arrangements that include ownership and accountability across the AI lifecycle, from design and development through to deployment, monitoring and decommissioning: Letter to all APRA-regulated entities (n 10).

16 For further information, see Appendix A.

17 For further information on the roles needed for effective AI governance, see HTI, *People, Skills and Culture for Effective AI Governance* (AI Governance Snapshot #4, November 2024) <https://www.uts.edu.au/globalassets/sites/default/files/2024-12/HTI-Snapshot-4-People-Skills-Culture.pdf>.

18 Letter to all APRA-regulated entities (n 10).

19 UTS Human Technology Institute Financial Services Survey (2025), see Appendix A.

20 Letter to all APRA-regulated entities (n 10); Letter from Australian Securities and Investments Commission to AFS Licensees and Market Participants (8 May 2026) <https://download.asic.gov.au/media/xhrf1w0e/26-092mr-open-letter-to-afs-licensees-and-market-participants.pdf>.

21 For example, see the Case Study in Section 5 where this situation is explored in more detail.

Actuaries Institute.

UTS Human Technology Institute

Actuaries Institute
ABN 69 000 423 656

Level 34, Australia Square
264 George Street, Sydney NSW 2000

T +61 (0) 2 9239 6100
E actuaries@actuaries.asn.au
W actuaries.asn.au

Human Technology Institute
University of Technology Sydney

PO Box 123
Broadway NSW 2007

E hti@uts.edu.au
W www.uts.edu.au/human-technology-institute