FSB FINANCIAL STABILITY BOARD

Sound Practices for Responsible Adoption of Artificial Intelligence (AI)

Consultation report



10 June 2026

The Financial Stability Board (FSB) coordinates at the international level the work of national financial authorities and international standard-setting bodies in order to develop and promote the implementation of effective regulatory, supervisory and other financial sector policies. Its mandate is set out in the FSB Charter, which governs the policymaking and related activities of the FSB. These activities, including any decisions reached in their context, shall not be binding or give rise to any legal rights or obligations.

Contact the Financial Stability Board

Sign up for e-mail alerts: www.fsb.org/emailalert
Follow the FSB on X/Twitter: @FinStbBoard
E-mail the FSB at: fsb@fsb.org

Copyright © 2026 Financial Stability Board. Please refer to the terms and conditions

Questions for consultation

The FSB is inviting comments on this consultation report and the questions set out below. Respondents are encouraged to provide detailed responses and supporting information where possible.

Responses to this consultation report should be submitted via this **secure online form** by 22 July 2026. Please contact the FSB by email (fsb@fsb.org) if you have questions or if you wish to provide supplementary material.

Responses will be published on the FSB's website unless respondents expressly request otherwise.

Background

This consultation report outlines the outcome of the FSB's work and seeks feedback on proposed sound practices for all types of financial institutions to adopt, use, and innovate with AI in a responsible manner. The sound practices focus on AI-specific aspects and risks that are relevant to financial institutions and financial stability.

Questions

1. Do you agree with the benefits and risks of AI adoption by financial institutions described in this report? Are there any substantive benefits or risks not covered?

2. Are the sound practices sufficiently comprehensive and clear to enable financial institutions' responsible AI adoption?

3. Do the sound practices strike an appropriate balance between managing risks relating to all forms of AI, and addressing some of the risks relating to emerging and new complex forms of AI, such as GenAI and agentic AI?

4. Are the sound practices sufficiently flexible to accommodate and address newer types of AI and responsible adoption over time?

5. Do the case studies in this report sufficiently highlight how different types of financial institutions can benefit from responsible AI adoption? Are there additional case studies for inclusion in the report? If so, please provide such case studies, particularly for nonbanks.

6. Do the case studies in this report provide actionable insights for financial institutions in their responsible AI adoption?

7. Are the definitions in the glossary clear and aligned with industry sound practices, including recent developments in AI?

Table of Contents

Questions for consultation iii

Executive summary 1

Introduction 3

1. An overview of the use of AI in the financial system 5

    1.1. AI adoption in the financial sector 5

    1.2. Use cases and benefits of AI in the financial system 6

    1.3. Risks and implementation challenges of AI adoption for financial institutions 12

2. Organisation-wide AI governance 18

    2.1. Strategic direction and oversight 18

    2.2. Governance and accountability 20

    2.3. AI risk management framework and effective documentation 21

    2.4. Organisational adaptability 23

3. AI lifecycle management 26

    3.1. Stages of AI development and deployment 26

    3.2. Materiality and risk assessment 27

    3.3. Selection 29

    3.4. Data governance 30

    3.5. Explainability and transparency 33

    3.6. Performance management 37

    3.7. Human oversight 41

    3.8. Cyber and ICT risk management 44

    3.9. Third-party risk management 48

Annex 1: List of Sound Practices for financial institutions' responsible AI adoption 52

Annex 2: Examples of explainability for different AI approaches 53

Annex 3: Examples of AI performance testing 55

Glossary 56

References 59

Executive summary

Financial institutions are rapidly adopting, using, and innovating with Artificial Intelligence (AI adoption) to reap efficiency gains, transform their business models, improve the services offered to customers and enhance risk management capabilities, such as to defend against cyber attacks. The significant benefits that can arise from AI adoption, however, usually come with additional risks. Managing these risks well will promote sustained value creation by financial institutions through AI adoption. It will also help minimise risks to financial stability as AI adoption scales through the financial sector.

The Financial Stability Board (FSB) has identified sound practices to help all types of financial institutions navigate benefits and risks responsibly as they adopt AI. The sound practices builds on and is broadly compatible with existing and ongoing work by the FSB and other standard-setting bodies (SSBs). They are intended to provide a menu of sound practices that financial institutions could use to enable responsible AI adoption. They also seek to foster coordination, cooperation and information-sharing among stakeholders, including financial institutions and supervisors, within and across jurisdictions. The sound practices are not intended to establish an international standard, to impose a prescriptive approach for responsible AI adoption by financial institutions nor to influence business decisions in adopting a certain AI technology. They are not exhaustive and may be refined as AI technologies evolve.

The 12 sound practices cover organisation-wide governance as well as management of different stages of AI development and deployment (or AI lifecycle). When applying the sound practices, financial institutions should consider proportionality. More robust practices may be appropriate for financial institutions that are large, complex, and highly connected within their ecosystem and where the AI is used or deployed in the financial institutions' critical (or material) functions. Smaller, less complex, or less interconnected financial institutions as well as less material, lower risk or less complex use cases within financial institutions may warrant applications of only relevant sound practices or appropriate modification of the relevant sound practices.

Sound practices 1 to 4 emphasise the importance of organisation-wide AI governance, in informing the financial institution in its decision on whether and how to adopt an AI technology and at what scale. These sound practices highlight: (i) the pivotal role the board and senior management play in setting the overall approach and providing oversight so that AI adoption is aligned with the financial institution's business model, risk appetite, and strategy; (ii) the importance of establishing clear governance frameworks, policies, procedures, and processes to identify, assess, monitor, and manage AI-related risks; (iii) the importance of defining clear responsibilities and accountabilities throughout the organisation; and (iv) how financial institutions benefit from continuous learning and adaptation, enabling them to develop the resources, skills, knowledge, and capabilities required to sustain value creation and effective risk management over time.

Sound practices 5 to 10 focus on managing specific AI use cases at or throughout different stages of an AI lifecycle so that use case deployments are supported by proportionate guardrails. This involves: (v) effectively and systematically assessing the materiality and risks of AI use cases at inception and thereafter; (vi) selecting appropriate AI models or systems by considering objectives, operational, and technical needs, as well as materiality and risk of AI use cases; (vii) maintaining appropriate data governance so that the data for training, testing, and using AI is

accurate, complete, reliable, and secure; (viii) understanding differences in the explainability of various types of AI and, if appropriate and feasible, adopt more explainable AI or consider compensating controls; (ix) evaluating the performance of AI use cases proportionately to their materiality and risk, including through performance assessment, testing, and ongoing monitoring; and (x) implementing appropriate and effective human oversight that is relevant to the materiality, risk, autonomy, complexity, and explainability of different AI use cases.

Finally, sound practices 11 and 12 highlight the importance of managing: (xi) AI-related cyber and information and communication technology (ICT) risks including by incorporating AI cyber and ICT risk scenarios into tests and exercises, sharing relevant information with key stakeholders, and where appropriate, using AI tools in cyber and ICT risk management; and (xii) risks from AI third-party use with a focus on performance, transparency, data quality, supply chain and concentration risks, and business continuity.

The board and senior management of financial institutions are strongly encouraged to reference the sound practices as they consider business strategy, technology adoption, and risk management in an increasingly AI enabled environment.

Introduction

Financial institutions have long used artificial intelligence (AI)[1] tools, but the rapid evolution of AI technology, particularly generative AI (GenAI) and agentic AI, has attracted significant interest. Financial institutions are increasingly adopting, using, and innovating with AI (hereafter AI adoption) for a variety of purposes, including credit risk assessment, trading, portfolio optimisation, fraud prevention, transaction monitoring, customer service, coding assistance, as well as document summarisation and review.

AI is poised to profoundly transform financial services and business models of financial institutions - from automating repetitive tasks to augmenting complex decision making and execution - fundamentally reshaping how financial institutions operate, innovate, and serve customers. At the same time, it may also amplify or introduce risks that need to be identified and managed appropriately.[2] For example, the complexity and limited explainability of certain types of AI could increase model risks for financial institutions. Financial institutions may also need to consider the implications of not adopting AI. For example, not adopting AI may limit their ability to monitor and manage certain risks such as financial fraud. In some areas, AI can simultaneously provide both opportunities and risks. For example, AI could increase the ease, frequency, and impact of cyber attacks,[3] while also enhancing financial institutions' ability to defend against them.

Responsible AI adoption allows financial institutions to harness opportunities and benefits while minimising associated risks. This means that financial institutions need to understand and remain updated on the opportunities and risks of AI, and respond with the appropriate adoption strategy and guardrails to manage evolving associated risks. At the financial system level, responsible AI adoption reduces the risks to financial stability.

To facilitate responsible AI adoption by financial institutions, the Financial Stability Board (FSB) has developed this report, which highlights the benefits and risks associated with AI use in the financial system (Section 1), contains a menu of 12 sound practices that financial institutions could adopt in their organisation-wide AI governance (Section 2) and management of the relevant stage(s) of AI development and deployment (AI lifecycle[4]) (Section 3). The board and senior management of financial institutions are strongly encouraged to reference the sound practices as they consider business strategy, technology adoption, and risk management in an increasingly AI-enabled environment. It is useful to note that:

- While all sound practices in this report are generally applicable to all types of AI, certain practices can be more relevant for specific AI technologies, such as in response to risks stemming from the autonomy of AI agents.

---

1 For detailed AI-related terms and their definitions, please refer to the Glossary section.

2 See, for example, FSB (2024b).

3 This includes risks from the mis-use of AI models with the autonomous capabilities to discover and generate working exploits for security vulnerabilities.

4 There is no standard, one-size-fits-all stages of AI lifecycle as different AI use cases may involve different stages of varying length. For details, see Section 3.1.

- The sound practices are not exhaustive and will be subject to change as AI technologies evolve.

- Proportionality should be considered when applying the sound practices in this report. Each financial institution should consider the relevance and applicability of these sound practices based on its business model, complexity, role in the financial system, structure, size, and the extent to which AI is used or deployed in their critical (or material) functions. Financial institutions may also refer to or adapt certain sound practices, depending on the materiality and risk of a specific AI use case, impact and outcomes of the use case, and the type of AI being adopted.

- Some sound practices are complemented by case studies drawn from actual AI implementation practices by financial institutions. These are intended to illustrate how the sound practices may be applied in practice and, where relevant, how they can be applied proportionately.

This report builds on and is broadly compatible with existing and ongoing work by the FSB[5] and other standard-setting bodies (SSBs), as well as by national/regional financial authorities (hereafter authorities). It also draws on inputs from a range of stakeholders across the financial system, including financial institutions and their technology vendors.

The sound practices are not intended to establish an international standard, to impose a prescriptive approach for responsible AI adoption by financial institutions, nor to influence business decisions in adopting a certain AI technology. Adopting the sound practices in this report also does not absolve a financial institution from its obligations to meet local legal and regulatory requirements related to AI adoption.

5 See FSB (2024b) and FSB (2025b).

1. An overview of the use of AI in the financial system[6]

1.1. AI adoption in the financial sector

AI is increasingly transforming financial services through a wide range of use cases. Financial institutions are leveraging AI in areas such as anti-money-laundering (AML) and know-your-customer (KYC), credit risk, cyber security, customer engagement, fraud detection, optimising internal process, collateral and portfolio management, and regulatory compliance. These applications have the potential to deliver value by enhancing the analysis of data and information, improving risk monitoring and management, and supporting faster and more accurate decision-making across front and back-office functions, whilst reducing operational costs.[7]

This transformation is driven by both supply and demand-side factors. On the supply side, advances in computing capabilities and power (including through cloud infrastructure), the growth of available data, and greater availability of foundation models have lowered barriers to adoption. On the demand side, financial institutions seek to boost their competitive position in increasingly digital markets, improve operational efficiency and strengthen risk management.

AI has shown significant advancements in the last decade. Beyond traditional AI and machine learning (ML), which learns from data to make predictions, recommendations, or decisions, recent years have seen the emergence of large language models (LLMs), which can analyse and generate human language by processing vast amounts of text data and GenAI, which can create new, often multimodal content (audio, images, text, video). More recently, agentic AI has emerged whereby autonomous systems ('agents') are capable of planning, reasoning, and executing complex high-level goals (comprising multiple tasks) independently.

While risks posed by AI adoption are not all new, some risks could be accentuated and there are some newer risks specific to AI technologies such as agentic AI. The FSB and relevant SSBs emphasise proportionate, risk-based, and technology-neutral approaches to monitoring and addressing risks, and some SSBs have issued guidance specific to AI.[8] As such, authorities, whilst supporting AI adoption, stress the importance of responsible AI adoption via strong governance, clear accountability, and sound risk management practices especially in high materiality and high risk use cases.[9]

The FSB assessed the financial stability implications of AI in the financial system in 2017, followed by an update in 2024.[10] However, given the rapidly evolving nature of the technology, the benefits and challenges set out in this Section remain subject to future reassessment by the FSB.

---

6 This Section draws on existing publications by the FSB, SSBs, and other authorities as well as market participants, researchers, and media. It also drew on inputs from engagements with external stakeholders.

7 See, for example, BCBS (2024), BIS-FSI (2025a), Bowman (2024), Breeden (2024) and Machado (2026).

8 See FSB (2025b), IOSCO (2025) and IOSCO (2026).

9 For example, see MAS (2024), HKMA (2024a), HKMA (2025), JFSA (2021), FINMA (2024), and OSFI (2023).

10 See FSB (2017) and FSB (2024).

1.2. Use cases and benefits of AI in the financial system

AI adoption appears to have accelerated since the 2024 FSB report on the financial stability implications of AI,[11] though this acceleration varies across entities, sectors, and jurisdictions. Evidence from surveys and industry engagement by authorities[12] suggests that internationally active banks, large or sophisticated trading firms and asset managers, and payment providers show relatively advanced use. In addition, several AI use cases cut across financial activities, such as general AI productivity tools and the use of AI to boost the efficiency of back and middle-office functions and risk management.[13]

- **Banking.** Banks are generally progressing rapidly in embedding AI in operational and risk management functions. This includes the use of traditional AI in credit risk, document automation, fraud detection, and internal controls. Adoption in front-office functions for core financial decisions remains more cautious, reflecting concerns around explainability and heightened risk considerations.

- **Insurance**. Insurers have historically relied on rule-based systems, which are gradually being replaced or complemented by more sophisticated AI-driven approaches. There is evidence of growing use of AI in claims-handling, customer segmentation, fraud identification, and underwriting, though adoption varies by jurisdiction and is often shaped by risk considerations.

- **Capital markets and asset management**. Systematic trading firms and large asset managers are early adopters of traditional ML in areas such as portfolio optimisation and analysing alternative data. Adoption across the industry varies and some industry participants plan to increase AI adoption over time. Current use cases include: algorithmic and high frequency trading; customer service and research; market-surveillance; pattern recognition; sentiment analysis; and volatility forecasting.

- **Payments**. AI integration is especially strong in customer facing use cases such as behavioural risk analytics, onboarding, and KYC. Regarding KYC, financial institutions have reported improvement in both the accuracy and processing speed of KYC checks.

- **Financial market infrastructures (FMIs)**. FMIs are exploring a range of use cases, including the detection of anomalous transactions and data inconsistencies, and AI-driven settlement, reconciliation, and compliance systems. Some FMIs are also exploring or implementing use cases involving predictive analytics for enhanced stress testing, and development of credit default early-warning indicators to enable early detection of vulnerabilities.

---

11 FSB (2024b).

12 See BOE website.

13 For example see IMF (2024), and Holphe and Mabetha (2025).

Credit scoring

In credit scoring, AI is being adopted to improve accuracy through advanced analytics applied to large volumes of financial and behavioural data. Banks are using different types of AI models, such as decision-trees and neural networks, which may be developed in-house or provided by a third-party. These models may be hosted on premise or on an external cloud infrastructure. In some jurisdictions, smaller financial institutions are partnering with AI providers to develop AI credit scoring tools.

AI-based credit scoring can allow lenders to better assess credit risk where traditional methods are less effective, and to broaden consumer access to financial services when paired with alternative data,[14] such as payment behaviour, cash flow data, or non-traditional financial information.

Some banks are also using AI to enhance credit underwriting, particularly for small and medium-sized enterprises (SMEs). In several jurisdictions, GenAI-enabled tools have been deployed to automate data extraction from financial documents and generate structured narratives supporting credit decisions, contributing to reduced processing times, improved consistency in credit assessments, and enhanced overall operational efficiency in credit risk management functions.

Several banks have developed traditional AI credit decision models, particularly in their credit card and auto portfolios. These applications often combine AI methodologies, such as gradient boosting machines or neural networks, with alternative data to approve or decline new credit requests or requests for credit line increases. This has enabled banks to improve risk profiling, which may contribute to lower default rates and expanded credit access, particularly for borrowers who may not qualify under traditional credit assessment methods.

Case study: Machine learning for credit risk management

A large regional bank has systematically embedded ML into its credit risk operations, especially in SME lending, where the bank's ML models proactively generate alerts for borrowers showing early signs of cashflow deterioration. In 2022, the bank successfully identified over 95% of non-performing SME loans at least three months before the businesses experienced issues meeting their debt repayments. This allowed the bank to engage with at-risk SMEs early to understand their situation and partner with them to explore possible pre-emptive solutions - such as restructuring repayment schedules or adjusting credit facilities - before credit issues worsened. As a result, over 80% of identified at-risk borrowers were averted from default altogether. These outcomes allowed more accurate and timely assessment of credit risk across a large and diverse borrower bases, and allowed the bank to provide early support to customers while improving the overall quality of its loan portfolio.

Beyond assessment of credit risk, GenAI is also being applied to streamline credit underwriting processes. Another medium-sized regional, commercial bank developed an ‘AI Approver Assistant’ by integrating GenAI into SME credit underwriting. The solution uses GenAI to automate data extraction, generate credit narratives, and support risk detection, aiming to streamline approval processes, enhance decision-making consistency, and establish a scalable governance and operating model. It achieved over 80% accuracy in financial data extraction, with credit approvers

[14] Alternative data generally refers to information that is not typically found in the consumer’s credit file of the nationwide consumer reporting agencies or information not customarily provided by consumers as part of the applications for credit. See FRB et al (2019).

rating more than 70% of GenAI-generated content as valuable reference material. Processing time for preliminary credit proposal generation was reduced from more than a day to 15 minutes, and the system matched manual review performance in identifying data inconsistencies and potential fraud signals across all test cases.

Trading, asset management, and portfolio optimisation

The adoption of AI by financial institutions is supporting a range of investment and trading activities including automated portfolio reporting; sentiment analysis (e.g. for parsing earnings calls); liquidity forecasting and volatility modelling; pattern detection; and signal extraction from alternative data. Recent research [15] indicates that AI adoption is yielding productivity improvements and cost savings in trading algorithm design, better data processing and, in some cases, discovery of new investment opportunities. While there is little evidence at present of GenAI being used for fully autonomous trading,[16] financial institutions are exploring the use of agentic AI, which can autonomously sequence trading-related tasks (e.g. scanning news, generating signals, and preparing execution instructions), potentially enabling more streamlined workflows and reduced manual intervention.

Building on these developments, evidence from industry practices illustrates how financial institutions are applying AI in specific investment and trading contexts. For example, some asset management companies, notably those active in the private equity sector, make use of AI-supported investment opportunity identification tools to broaden the set of opportunities available to investment teams engaged in research and portfolio construction. Some large internationally active banks and institutional broker-dealers have recently adopted research and analytics assistants to enhance trader workflows. These solutions often customise AI assistant applications and leverage AI-powered agents to support traders in their market analyses. For example, a research agent can compare different hedging strategies by analysing several market data sources in real time, while AI-powered assistants may also have access to a variety of system tools that enable them to visualise outcomes and conduct statistical analysis. Initial results suggested that accelerated analysis can lead to significant time savings, supporting more efficient decision-making processes and streamlined trading workflows.

Regulatory compliance, AML/CFT, and fraud detection

Several financial institutions are using AI to improve fraud detection and financial crime monitoring by analysing large volumes of transactions and unstructured data. Some forms of traditional AI (e.g. neural networks and random forest algorithms) can enhance anomaly detection and help uncover hidden relationships between transactions, strengthening monitoring capabilities. AI can enhance AML/KYC checks by automating document verification, identity checks, and fraud detection. Techniques such as biometric verification and contextual analysis can strengthen security, reduce manual work, and enhance the overall customer experience. AI is also increasingly used to detect identity fraud, synthetic identities, and deepfake-based cyber-

15 See Lim et al (2025).

16 See FMSB (2026).

attacks often in real-time. Authorities have noted reductions in false positives and negatives, and better identification of complex criminal patterns.[17]

Some large asset management companies, for example, deploy their AI tools to support compliance reviews of marketing materials. These applications enhance the speed, consistency and quality of compliance checks across marketing and disclosure documentation. Some large FMIs, notably exchanges, also apply AI for market surveillance and abuse detection, analysing internal and external data sources to strengthen detection capabilities, and improve the efficiency of monitoring and compliance activities. Various types and size of financial institutions (e.g. banks, payment service providers) widely use ML models for fraud prevention in payment transactions. These tools enable real-time analysis of large transaction volumes, supporting faster detection of suspicious activity and improving the overall security and efficiency of payment services. Some medium to large financial institutions report the use of GenAI tools in financial crime and AML investigations. By correlating information from multiple sources and generating investigative narratives, these applications deliver significant time savings and improve consistency and quality in investigative reporting.

Case study: Strengthening fraud detection with agentic AI

A large internationally active bank has deployed an agentic AI system to detect emerging fraud and scam patterns in real time, demonstrating measurable improvements in financial crime prevention at scale. Building on its existing AI capabilities (that monitored more than 80 million signals each day, and spanning transactions, card and online payments, as well as interactions across digital banking channels), the bank developed an agentic AI system specifically designed to identify emerging threats and autonomously propose the detection rules needed to intercept them. The system was built in-house by the bank's data science and engineering teams in three months, running on a cloud-based data platform and core banking infrastructure to enable real-time data access at scale.

The agent operates continuously, assessing the severity of suspicious patterns, analysing context, and generating proposed detection rules for review. The process embeds human-in-the-loop oversight to review and approve all new rules by the bank's fraud analytics team before implementation.

The agent has contributed to developing or updating three quarters of the bank's card fraud rules and helped reduce fraud losses by over 20% in the first half of the 2026 financial year compared to the same period in 2025. Daily, the bank processes millions of payments and sends thousands of proactive fraud warning alerts to customers through its mobile banking application.

Separately, a digital bank enhanced its existing facial recognition solution, originally designed to compare facial images against a customer database for digital fraud detection. The solution adds the ability for the bank to detect suspicious image backgrounds in facial recognition images, such as those associated with mule accounts or fraudulent identity attempts. The system extracts facial feature vectors for rapid one-to-many comparison while simultaneously applying object detection and background similarity analysis to flag potential risk indicators. This represents a notable instance of using AI in defence, where the bank leverage advanced technology to counter AI-generated content and evolving fraud vectors.

The enhancement enables the bank to identify controllers of mule accounts by correlating facial images with suspicious attributes in image background, a capability previously unavailable. This establishes a risk alert mechanism that enhances detection of digital fraud activities including account takeover, deepfake or injection attacks, and loan applications involving fake identities. The combined approach,

[17] See BCBS (2024) and BIS (2025).

leveraging both facial comparison and image background analysis, supports network analysis of fraudulent activities and strengthens the bank's overall AML and fraud investigation framework, with comparison processes completed in seconds.

Customer-focused applications

AI is increasingly used across some financial institutions to enhance customer engagement in the areas of portfolio management, marketing, and customer service. By analysing high-dimensional datasets and automating routine decision-supporting tasks, AI can enhance existing processes and improve customer engagement through tailored financial advice. Some financial institutions are also experimenting with LLM-enabled tools for internal knowledge retrieval, automating customer complaint triage and document classification, which can significantly enhance operational efficiency.[18]

Several financial institutions are increasingly using AI to support robo-advisory services, including for customer profiling, investment planning, portfolio management, and rebalancing. Some asset managers are also leveraging AI to enhance their customer advisory services by automating data collection, market analysis and monitoring, reducing costs, and expanding services such as real-time insights.

In customer service, AI-driven chatbots and virtual assistants can handle routine queries, payments and loan requests, improving response times, service availability and customer satisfaction, while reducing reliance on call centres. AI is also widely used in marketing activities, to optimise new product campaigns, generate customer leads, and create marketing content.

Case study: Scaling relationship management with AI

A global systemically important bank (G-SIB), active in both retail and wholesale banking, has piloted an integrated AI application to transform how its corporate relationship managers serve clients.

Prior to the development of the AI application, relationship managers responsible for corporate clients faced significant time pressures in preparing for client engagements, synthesising information from multiple sources, and producing tailored proposals. This left some clients underserved and constrained the depth of dialogue that managers could sustain across their portfolios.

To address this, the bank built an AI platform that integrates a wide range of external data sources including news feeds, and research reports, alongside internal knowledge bases. The output from the platform supported corporate relationship managers in developing better understanding, designing solutions, and producing proposals for their clients.

AI-assisted insight generation and proposal creation have reduced preparation time by approximately 50%, while enabling around three times more client dialogue. Critically, the platform has enabled the corporate relationship managers to serve every client, including those who were previously underserved, with updated, tailored insights and proposals.

The application is expected to be rolled out to approximately 3,500 corporate relationship managers in financial year 2026.

[18] See, for example, ECB (2024), OSFI (2023), OSFI (2025a), OECD (2024) and OECD (2026).

Productivity enhancing applications

Several financial institutions are finding efficiency and productivity gains from AI as it enables the automation of labour-intensive activities, supporting faster execution, more consistent processes, lower operational costs, and allowing staff to focus on higher-value tasks. Evidence points to significant productivity gains, particularly from GenAI applications that synthesise and transform unstructured content.[19] Some financial institutions are increasingly using AI to modernise internal operations, including back office automation, risk analytics, content and code generation, and risk management. Foundation models also support large-scale analysis of unstructured data and are being used to strengthen internal controls through automated documentation, monitoring, and anomaly detection.[20]

Some financial institutions are increasingly adopting AI to support the scaling of operations in environments characterised by growing data volumes and transaction complexity. By enabling the automation and processing of large datasets, AI supports financial institutions in expanding their operational capacity while maintaining service quality.

Several financial institutions globally have reported usage of GenAI tools to boost internal productivity, particularly in software development. Coding assistants support code generation, debugging, code reviews, root cause analysis, and maintaining documentation which is delivering meaningful time savings on repetitive tasks and faster onboarding of technical staff. In turn, some banks report that early results have translated into significant cost savings in some cases. Beyond development functions, some financial institutions have also deployed organisation-wide GenAI assistants to support drafting, research, and information retrieval across business functions. This has reportedly improved knowledge access and internal collaboration across the organisation.

Case study: Operational efficiency with AI adoption

A large internationally active insurer conducted a global review of manual workflows to identify where AI could have the most immediate and governable impact. This resulted in the elimination of approximately 400,000 manual processing instances in 2025.

One notable example that the insurer embarked on was the underwriting quick quote process. In this workflow, front line staff submit email enquiries to the underwriting team containing client information, which the team then analyses to prepare Preliminary Assessment documents. These assessments play a critical role in the sales process, helping producers narrow their insurer selection before proceeding to formal applications. With approximately 4,000 such requests received each month, the process involved significant time and manual work.

With the deployment of AI, upon receipt of a producer's email, the system now automatically analyses the client information and generates a draft Preliminary Assessment for underwriting team's review. This reduces turnaround time from one business day for simpler cases and two to three days for complex ones, to approximately 45 seconds.

The insurer is also adopting AI for software development to improve developer efficiency. This has increased productivity through leveraging AI code generation tool across the company to write six million lines of code in 2025, with around 80% adoption rate among developers.

---

19 See ECB (2024).

20 See MAS (2025b), HKMA (2024a), HKMA (2025), and OSFI (2025b).

Operational resilience

Several financial institutions are using AI to enhance operational and cyber resilience such as threat detection and response. AI can analyse large volume of data in real time to identify anomalies and detect potential attacks faster than human analysts, while also reducing false positives, allowing for effective containment of threats. Furthermore, AI can strengthen organisational resilience by identifying vulnerabilities, and can promote business continuity through swift recovery from incidents like ransomware attacks.[21]

Financial institutions of various sizes and types use AI to monitor anomalous network behaviour across internal employee access, data transfers, and unusual API calls, using a combination of internal and third-party tools. AI can also be used to identify phishing attempts. GenAI can rapidly populate internal security alerts and incident reports.

1.3. Risks and implementation challenges of AI adoption for financial institutions

As noted in the FSB's 2024 report, four types of AI-related vulnerabilities stand out for their potential to increase financial stability risks:[22] (i) third-party dependencies and service provider concentration; (ii) market correlations; (iii) cyber risks; and (iv) model risk, data quality and governance. This report builds on the FSB's previous analysis of these risks and describes additional AI-related risks and implementation challenges that can impact financial institutions and their customers.

Inadequate governance and strategic execution

AI has the potential to rapidly transform a wide range of existing business processes, but inadequate governance and lack of board and senior management oversight can give rise to significant risks. Without clear accountability within the financial institution, critical issues may be overlooked. Inadequate assessments of the materiality and risks of AI use cases can lead to mis-calibrated risk management. Particularly at large financial institutions, fragmented AI implementation across business units can undermine effective risk management. Governance gaps may also enable the use of 'shadow AI', where employees adopt AI tools without authorisation or for unapproved uses, reducing financial institutions' visibility of and ability to monitor and manage AI related risks.

Ineffective data governance and poor data quality

Ineffective data governance and poor quality data can lead to a number of AI-related risks, including data security risk, model risk, unacceptable bias,[23] and poor AI performance. For example, AI models or systems can perpetuate or amplify biases and inaccuracies present in input data, creating unanticipated or undesirable outcomes. These outcomes can become self-

---

21 See G7 Cyber Expert Group (2025).

22 See FSB (2024b) Section 4.2.

23 Some forms of bias in both AI models or systems and non-AI models can be acceptable. For example, bias that aligns with a financial institution's risk appetite and complies with legal and regulatory obligations may be justified.

reinforcing, as biased outcomes generate new data that influences future decisions, entrenching biased patterns over time. Inadequate data governance and quality can also lead to poor AI performance due to overfitting[24] or underfitting[25], an inability to address outlier situations, or even unintended legal and regulatory breaches.

Explainability challenges

While explainability can also be a challenge in non-AI models and other quantitative approaches, AI adoption introduces additional challenges to explainability and transparency, particularly in complex, autonomous, or rapidly evolving AI models or systems where the basis for decisions may be difficult to understand, made with no or minimal human involvement, or shift over time in ways that are difficult to follow. These challenges include:

- assessing conceptual soundness.
- confirming compliance with requirements in law and regulation.
- distinguishing causality from correlation.
- assessing reliability and robustness.
- detecting bias, errors, or performance issues, including data drift and model drift.
- understanding changes in dynamically updating models or systems.
- assessing data quality, especially with inaccessible datasets (such as proprietary datasets of third-party AI providers).
- conducting validation or types of independent review.

Inadequate or degrading AI performance

Inadequate or degrading AI performance can lead to poor decision-making, financial, legal, regulatory, operational, and other risks. The higher the materiality and risk of the AI use case, the greater the potential impact of performance issues, which can adversely affect the financial institution as well as customers and market participants. For example, models trained on historical data may underperform in new market conditions. Some type of AI models may hallucinate by generating seemingly plausible but incorrect outputs, while others might execute inappropriate or harmful tasks. Performance may also degrade due to version changes in key AI models or systems, and recovery and rollback can be challenging (or impossible) for some forms of AI models and systems.

---

[24] Where AI works well with training data but not with unseen data.

[25] Where AI fails to capture complexity of training data.

Inadequate or inappropriate human oversight

Inadequate human oversight of AI may hinder financial institutions' ability to manage AI risks, causing unintended legal or regulatory breaches or customer harm, among other negative consequences. Certain misconceptions about AI can undermine the effectiveness of human oversight. Such misconceptions include the assumptions that AI always operates predictably in the same environment, that AI models or systems nearly always produce correct output, that the environment itself will remain predictable, or that human involvement ensures effective oversight. Human oversight may be weakened by over-reliance and/or automation bias.[26]

Cyber and ICT risks

The advent of AI has implications on cyber and information and communication technology (ICT) risks. For example, depending on where AI solutions are deployed within a financial institution, their use may have a bearing on system availability and data integrity which can result from defective application codes generated by AI. In addition, the improper use of agentic AI could pose operational risks when the AI misunderstands a set of instructions due to unclear prompts and executes an automated task erroneously.

AI may also be weaponised and maliciously used by threat actors to attack financial institutions and their customers. This can result in an increase in the frequency and impact of fraud and cyber-attacks. Examples of AI-enabled attacks or direct attacks against AI include:

- *Poisoning*: Threat actors inject biased or incorrect data into training data, gradually undermining model performance over time.
- *Malicious inputs* including:
  - *prompt injections*, where a threat actor manipulates prompts with malicious content to influence model behaviour and bypass safeguards.
  - *jailbreaking*: where a threat actor uses prompts to circumvent a GenAI model's safety features to breach legal requirements or guardrails. This can result in the model providing dangerous, illegal, or unauthorised outputs.
  - *evasion*: where malicious attackers make small, hard-to-detect changes [27] to input data to deceive an AI model, causing it to produce incorrect outputs.
  - *extraction*: where a threat actor reverse engineers an AI model to extract information about its architecture, parameters or training data.
- *Deepfakes and GenAI-enabled phishing*: the ability of multi-modal GenAI models to produce increasingly realistic audio, code, images, text, and video can be used to amplify social engineering attacks through deepfakes, impersonation, and false

[26] Overly trusting of automated systems outputs without critical thinking.

[27] Examples include slightly modifying an image so that a security system misclassifies it, or tweaking text to bypass content filters using optimisation techniques and transferability.

communications. Threat actors can, for instance, impersonate individuals in a financial institution with authority to approve high-value transactions to defraud financial institutions or use AI-generated images or video to bypass KYC checks.[28]

- *Lower barriers to entry for other cyber-attacks*: Advancements in AI models and AI-coding agents can allow threat actors to commit complex cyber-attacks that previously required specialised technical capabilities. These advancements include developing malicious codes, and identifying and exploiting severe security vulnerabilities at highly accelerated speeds (including zero day vulnerabilities). Some advanced models can autonomously identify and exploit vulnerabilities through multiple means, including but not limited to source code analysis, web application testing, binary analysis, and protocol fuzzing. Such capabilities significantly compress the timeline from vulnerability discovery to exploitation.

Third-party dependencies and concentration

AI adoption in the financial sector increasingly depends on highly concentrated third-party providers across multiple layers of the AI supply chain, including cloud infrastructure, hardware, foundation models, data providers, and deployment platforms. While third-party providers offer benefits, they also introduce risks, including:

- Failure to perform as expected: Third-party AI systems may exhibit performance inconsistent with the terms of agreed-upon service.
- Insufficient transparency: Financial institutions may face reduced transparency when relying on proprietary or complex third-party AI models, systems, or data, limiting their ability to conduct due diligence, monitor performance, and manage risks. For example, third-party service providers may change the underlying models powering their AI systems without adequate notice, leaving financial institutions unaware of changes that could affect performance or risk outcomes. This can amplify explainability challenges of some forms of AI, such as GenAI and agentic AI.
- Data governance: AI models or systems provided by a third-party can pose risks if their training data includes inappropriate data, such as unlawfully processed personal data. Third-party AI systems may inadvertently disclose confidential or protected data. Third-party AI providers could also use financial institution's data or customer data for unauthorised purposes.
- Operational resilience: Significant disruption to critical services provided by third parties can create risks for financial institutions' critical operations.
- Service provider concentration: Providers of cloud infrastructure, hardware, and AI foundation models are often highly concentrated and, in some cases, vertically integrated. There can also be geographic concentration across key points in the AI

[28] See MAS (2025b).

supply chain, which could be impacted by physical events, like natural disasters, or other supply chain risks.

- Homogeneous behaviours and correlated outcomes: Reliance on common AI models, datasets, or infrastructure can lead to correlated behaviours across financial institutions, amplifying risks like herding and procyclicality. This could exacerbate market stress, liquidity crunches, and asset price vulnerabilities during periods of financial instability.

Consumer protection and market conduct risks[29]

Inadequate guardrails around AI could give rise to consumer harm and market conduct risks. This includes:

- Inadequate disclosure of AI use in financial services: Where there is insufficient disclosure about the role of AI in financial products and services, customers may be unable to fully understand or challenge decisions that affect them or seek appropriate redress.

- Unfair treatment and poor consumer outcomes: Use of AI models trained on biased or incomplete datasets may result in discriminatory results or outcomes not consistent with consumer protection requirements.

- Potential for mis-selling and unsuitable recommendations: The use of AI in customer-facing applications, including robo-advisory and product targeting, may lead to recommendations that are unsuitable or not aligned with customers' risk profiles, particularly where models rely on incomplete, outdated or improperly calibrated data.

- Reduced accountability and challenges in oversight of automated interactions: Increased reliance on AI-driven processes in marketing, advisory and customer engagement may blur lines of responsibility, making it more difficult for financial institutions and supervisors to identify, monitor and address misconduct or inappropriate practices in a timely manner.

Agentic AI risks

The high levels of autonomy that AI agents may have can create or amplify certain risks, which can materialise at great speed, including:

- Unauthorised actions: AI agents can take autonomous actions based on pre-defined goals and their environment. They can also dynamically set or modify their objectives based on what they learn from interacting with their external environments. This creates a risk of AI agents taking illegal, unethical, or unauthorised actions without human approval or oversight. Overriding, redressing, or remediating these actions can be difficult or impossible for humans.

---

[29] This section may also apply to protection of investors.

- Erroneous actions: AI agents might make incorrect decisions due to goal misalignment, insufficient information and other reasons, such as reward hacking.[30] These actions may only manifest once the agent is deployed in a live environment. Monitoring and detecting these actions in real-time can be very difficult. An AI agent can take hundreds of intermediate steps in pursuit of its goals and make errors in any of those steps. In some cases, effective monitoring and detection may require augmentation with another AI agent or other forms of AI.

- Data breaches: AI agents may expose or manipulate sensitive data such as customer details, trade secrets, or internal communications. This may happen through malicious attacks where threat actors exploit vulnerabilities of agents, or accidental disclosure if agents fail to recognise data sensitivity.

- Disruption to connected systems: To execute complex tasks, AI agents can integrate with Application Programming Interfaces (APIs), databases, ICT systems, and other agents, all of which can be disrupted if the agent is compromised or malfunctions. AI agents may also collude with other AI agents to carry out cyber-attacks (e.g. deleting code or carrying out Distributed Denial-of-Service (DDoS) attacks) or other malicious actions, such as spreading disinformation, to achieve their goals. Even without collusion, multiple AI agents may interact in unanticipated ways that create unforeseen behaviours or risks.

- Additional risks from inadequate human oversight: AI agents pose a distinct challenge for human oversight, given the impracticality of real-time human monitoring of agent decisions as their use scales. This can lead to agents pursuing objectives or taking actions that deviate from the financial institution's intentions or risk appetite, without staff being aware or able to intervene in a timely manner.

- Additional data security and privacy risks: AI agents may fail to recognise data sensitivity or take autonomous actions that inadvertently lead to the exposure or manipulation of sensitive data, such as customer details, trade secrets, or internal communications.

- Additional cyber and ICT risks that pose challenges for traditional cyber security controls: For example, threat actors can manipulate AI agents by injecting malicious data into their knowledge base (known as memory poisoning), influencing their behaviour over time.[31] Agents may also devise novel ways to evade traditional cyber detection capabilities.

---

30 Where an AI system finds a shortcut or loophole in its reward function to maximise its score without actually fulfilling the intended goal. This is common in models and systems trained with reinforcement learning from human feedback, but can also arise in AI agents.

31 Malicious data can be injected to AI agent's retrieval-augmented generation (RAG) to manipulate its learning and memory process known as agentic memory poisoning.

2. Organisation-wide AI governance

The four sound practices in this section address AI governance practices at an organisational level, and the eight sound practices in the next section cover the AI lifecycle. As is the case with cross-sectoral standards on AI,[32] these four sound practices are designed to be cross-cutting to inform the remaining sound practices.

AI governance concerns how financial institutions choose, monitor, and periodically review their approach to AI adoption and organisational governance. As in other areas, the board and senior management[33] take responsibility and play critical roles in setting a clear direction for AI adoption and in providing senior-level oversight, so that AI adoption and governance align with the financial institution's business model, risk appetite, and strategy (Sound practice 1). The board and senior management are also responsible for establishing and maintaining clear and well-defined roles and responsibilities across the organisation, to promote effective accountability, and robust oversight of AI-related activities and risks (Sound practices 2 and 3). Financial institutions adapt their governance approaches and operating models, and appropriately resource the organisation as AI and the financial institution's usage evolves (Sound practice 4).

2.1. Strategic direction and oversight

Sound Practice 1 (Strategic direction and oversight): The board and senior management align AI adoption and governance with the financial institution's business model risk appetite and strategy.

The board and senior management oversee AI adoption

The board and senior management consider AI adoption as part of the financial institution's business strategy. This includes considering how AI may positively impact the financial institution's business model, customers, efficiency, growth, and compliance with legal and regulatory requirements, while balancing these benefits against AI-related risks. As in other areas, the board promotes clear accountability for senior management to implement the financial institution's approach to AI adoption.

The board and senior management define and communicate across the organisation the outcomes expected from AI adoption, including the desired outcomes from appropriate risk management. This fosters alignment between the financial institution's business strategy and its approach to and execution of AI adoption over time. The board and senior management are

---

32 For example, NIST AI Risk Management Framework. See NIST (2023).

33 This report refers to a management structure composed of a board of directors and senior management. There are significant differences in legislative and regulatory frameworks across jurisdictions regarding the functions of the board of directors and senior management. In some jurisdictions, the board has the main, if not exclusive, function of supervising the executive body (senior management, general management) and is known as a supervisory board. This means that the board has no executive functions. In other jurisdictions, the board has a broader competence in that it lays down the general framework for the management of the organisation. The terms "board of directors" and "senior management" are used in this report to label distinct decision-making functions within a financial institution.

engaged end-to-end, from strategy through implementation, so that resources (e.g. investments in technology infrastructure, strategic partnerships) are aligned with the intended outcome.

The board and senior management regularly re-assess whether the financial institution's approach to AI adoption remains fit for purpose. The board curates and receives management information (MI) [34] that allows it to assess whether the financial institution's approach to AI adoption (including prior decisions not to adopt AI for certain use cases) remains aligned to the financial institution's strategy, and whether senior management's implementation remains an effective approach.

The board and senior management consider AI risks when setting risk appetite and tolerance

The board and senior management establish clear boundaries for AI adoption by considering AI risks when setting and reviewing the financial institution's risk appetite and tolerance.[35] This could include articulating prohibited AI use cases, for example, fully automated decision-making in critical business areas. In setting these boundaries, the board and senior management consider different types of AI and potential AI use cases: the level of autonomy; materiality and risk; potential customer impact; legal and regulatory requirements; and dependencies on data, technology, and third parties.

The board and senior management also align AI adoption with the financial institution's standards for conduct. This considers impact to customers and markets, and informs its decisions on explainability, human oversight, and redress for specific AI use cases.

The financial institution's risk appetite and tolerance guide decisions on AI use selection and prioritisation, risk monitoring, management, and mitigation (including human oversight and escalation protocols). This becomes particularly important as AI adoption expands to critical operations and core business lines. Addressing AI considerations explicitly in the financial institution's risk appetite creates a coherent bridge between board-level oversight and management action.

The board and senior management appropriately resource the financial institution to support AI adoption

The board and senior management appropriately resource the financial institution to support AI adoption currently and on an ongoing basis. This includes: appropriate competencies and skills at the board, senior management, and staff (including control functions) levels to implement their respective roles and responsibilities; as well as sufficiently mature data and technology capabilities to support sustainable AI adoption. The board and senior management reassess the appropriateness of the financial institution's resources, skills and capabilities periodically as AI

---

[34] As with other areas, MI include: business plans; risk reports (including high risk uses); performance reports; failures and mitigation actions; and audit reports. MI may also include metrics on AI adoption, such as time savings (in some cases with estimated monetary value), utilisation/uptake by staff, customer satisfaction, and acquisition measures.

[35] Risk appetite is defined as the aggregate level and types of risk a financial institution is willing to assume within its risk capacity to achieve its strategic objectives and business plan (see FSB (2013)). Meanwhile, risk tolerance can be defined as the boundaries of specific risks that financial institution can take or withstand.

becomes more embedded in critical operations and core business lines, and as technology, legal or regulatory requirements evolve.

Case study: Aligning skills and capabilities to AI strategy

Some large internationally active financial groups, under the steer of their boards, treat AI as a strategic pillar rather than as a mere technological tool. These financial institutions have pursued end-to-end redesign of processes and workflows across their organisation, with AI fundamentally shaping how those processes are structured and operate. To support this, the board and senior management of these financial institutions have invested in building capabilities at all levels of the organisation. Senior managers and board members have undertaken master classes developed with reputable academic institutions and management consultants to build the transformative mindset needed to drive such change, while staff have been supported through AI bootcamps to become 'hybrid' professionals combining domain expertise with AI development skills, alongside organisation-wide programmes to build AI literacy.

These financial institutions have also set concrete targets to track AI adoption progress, covering both hard benefits such as productivity and cost metrics, and softer benefits such as enhancements to staff capacity. They have acknowledged, however, that finding suitable metrics remains challenging, particularly around value realisation.

Other financial institutions have focused instead on layering AI onto existing processes to improve services and enhance productivity, without pursuing broader organisational transformation. For these financial institutions, board and senior management have correspondingly focused on building AI literacy across the organisation and implementing clear boundaries around AI use, such as restricting staff to specified tools accessed only through prescribed platforms. This ensures staff are equipped to work effectively with AI tools within their existing roles and workflows.

2.2. Governance and accountability

Sound Practice 2 (Governance and accountability): Financial institutions define clear roles and responsibilities and maintain an appropriate governance framework to enable responsible AI adoption.

Define clear roles and responsibilities

Financial institutions define clear roles and responsibilities for AI adoption, building on existing structures such as the three lines of defence[36] or creating new structures if appropriate. For example, under the three lines of defence model, first-line functions at a financial institution may assess business outcomes, controls, materiality, and risks for AI use cases. Second-line functions may maintain an organisation-wide view of AI-related risks across the financial institution, and may also provide independent risk assessment across areas including compliance, data quality, explainability, conduct risk, and model performance. Finally, third-line functions may deliver independent assurance proportionate to the materiality and risk levels of the AI use cases. It is essential that audit, evaluation, risk management, and testing remain

[36] See BCBS (2021b). The three lines of defence commonly consists of business unit management, an independent corporate operational risk management function, and independent assurance. Depending on the organisation's nature, size and complexity, and the risk profile of an organisation's activities, the degree of formality of how these three lines of defence are implemented will vary.

independent and functionally separate from development teams to maintain objectivity, and that those performing these independent functions are empowered and sufficiently skilled to provide effective challenge.

Maintain an appropriate operating model for responsible AI adoption

Financial institutions' operating models consider the nature and scale of AI adoption within the organisation. The operating model may be centralised, decentralised, or a hybrid of both. Centralised models offer consistent oversight and standards across all AI use cases. Decentralised approaches provide business units and functions greater flexibility and the ability to leverage domain-specific expertise. Hybrid models combine elements of both. Financial institutions may choose to adapt and leverage existing operating models. The operating model is reviewed periodically to maintain effectiveness in supporting the financial institution's strategy, an organisational-wide assessment and management of risks, as well as alignment with risk appetite and tolerance as AI adoption evolves.

Establish cross-functional coordination mechanisms

Financial institutions implement cross-functional coordination mechanisms to leverage on the diverse perspective from business owners, risk managers, technology experts, compliance officers, legal specialists, developers, and other relevant stakeholders. This facilities knowledge sharing so that stakeholders across the financial institution understand the full range of benefits and risks of AI, optimise implementation plans, and clearly delineate responsibilities and accountabilities. Cross-functional coordination mechanisms may include regular fora; standardised communication channels; consistent methodologies for assessing, monitoring and reporting AI risks; and feedback loops that adapt governance and risk frameworks to evolving AI risks.

2.3. AI risk management framework and effective documentation

Sound Practice 3 (Incorporation of AI risks into risk management framework): Financial institutions adopt risk management frameworks that effectively address AI risks and include processes for AI identification documentation as well as materiality and risk assessment.

Consider relevant risks from AI in risk management frameworks

Financial institutions incorporate AI risk considerations into existing risk management frameworks, policies, and procedures as appropriate. This allows AI risks to be identified, assessed, accepted (or approved), documented, monitored, and managed. Section 3 describes AI lifecycle management frameworks relevant to AI adoption in more detail.

Where existing policies, procedures, and frameworks do not sufficiently address AI risks, such as those related to lack of explainability, autonomy, or inadequate or inappropriate human oversight, financial institutions may enhance these policies, procedures, or frameworks, or establish new ones as appropriate.

Financial institutions implement guardrails and controls proportionate to materiality and risks (e.g. streamlined processes for low materiality and low risk use cases, and rigorous processes for high materiality and high risk use cases). Guardrails implemented at the different stages of the AI lifecycle can also vary by use case (see Section 3). They may include staged approvals, comprising supervised experimentation in controlled environments and pilot phases before deployment.

Maintain accurate documentation of AI use cases

To maintain visibility over risks from AI adoption and facilitate effective identification, assessment, monitoring, and management of AI risks in a proportionate manner, financial institutions track and document their AI use throughout the lifecycle.[37] This may leverage existing processes like use case and model inventory, third-party or systems registry, etc. Financial institutions may have more detailed information for higher materiality and higher risk AI use cases.

Documentation captured in the lifecycle can include:

- a description of the use case, such as:
  - (i) its underlying AI model(s) or system(s), and the rationale for its selection;
  - (ii) approved, intended or prohibited applications; and
  - (iii) whether it was developed in-house, uses open-source tools, or was acquired from a third-party AI provider.
- materiality and risk of the AI use case, and mitigants.
- business lines, regions, products, customers, and market participants impacted.
- key assumptions and limitations about the AI use case.
- data sources used during inference and if developed in-house, data used for training.
- allocation of accountability (i.e. key roles and responsibilities) at the financial institution.
- dependencies on other AI use cases, models, systems, or third-party AI providers.
- AI lifecycle stages and status.
- findings from validation, testing, ongoing monitoring, and performance assessment.[38]
- relevant information from third-party AI providers.[39]

---

[37] Proactive record-keeping from inception promotes effective AI risk management.

[38] For example, indicators of whether AI models or systems are functioning properly, the dates when those indicators were last updated, and any outstanding remediation actions.

[39] For example, acceptable use policy; model or system cards; certifications, standards and other assurance; or any customisation or fine-tuning by the financial institution.

Where appropriate, financial institutions deploy enhanced documentation and logging for GenAI and agentic AI use cases to support effective risk management and oversight. For GenAI use cases, this may include prompt versioning[40] and version control to enable rollback to previous configurations, as well as model parameters to monitor performance changes, and chain-of-thought logging to help explain outputs. Where such controls are managed by third-party vendors and rollback or configuration access is limited, financial institutions establish contractual arrangements for vendors to provide timely information and compensating controls.

Case study: Documentation of AI use cases

Some large internationally active banks maintain a centralised, organisation-wide AI inventory supported by dedicated software systems, giving them clear visibility over their AI footprint and the risks arising from it. These systems serve as a repository for comprehensive documentation of individual AI use cases, with records maintained throughout the AI's lifecycle. Key attributes captured may include the AI's purpose, scope of use, risk materiality rating, dependencies, responsible personnel, and approvals obtained, among others. Beyond recording where AI is used, these systems may also support features such as automated tracking of approvals and issues, identification of interdependencies between models, and lifecycle tracking with checkpoints at key risk management stages. Some other banks with more limited AI adoption to date have been observed to rely on spreadsheets for their AI inventories.

Where banks deploy agentic AI systems, additional attributes beyond those captured for standard AI models have been included in their inventories. These include agentic AI-specific identifiers, the tools the agent can access, its components, and guardrails to manage risk. Where banks have adopted a modular approach - with reusable agents deployed across multiple use cases - some banks have tracked risk information at the agent level, including permitted uses and known risks of each agent, rather than solely at the use case level. In more mature organisations, this has taken the form of agent 'certification', where an agent is reviewed and approved for use within defined boundaries - with its certified status and permitted scope recorded as inventory attributes. As governance practices for agentic AI are still rapidly evolving, the attributes captured for such systems are expected to develop further as industry norms and regulatory expectations mature.

2.4. Organisational adaptability

Sound Practice 4 (Organisational adaptability): Financial institutions learn, adapt and adjust their oversight, governance, risk management practices, and capabilities as AI evolves.

Sustain AI expertise and risk-aware culture

Financial institutions develop and maintain AI expertise and embed a risk-aware culture that encourages challenge, escalation, and continuous learning. Organisational adaptability is only possible with adequate capability across the financial institution - from foundational literacy for all staff to specialised expertise for implementers and risk managers, and informed board and senior management oversight.[41] Capability-building efforts specific to AI may include:

[40] The systematic practice of tracking, management, and auditing of AI prompt iterations.

[41] See The MindForge Consortium (2026), pp.109.

- providing foundational AI training for all staff so that they can recognise and respond to inappropriate use, automation errors, automation bias/over-reliance, and unusual model behaviour;
- developing role-specific expertise for developers, validators, risk managers, auditors, and compliance teams to manage AI-specific risks (e.g. data/model drift, instability, unusual AI behaviour, explainability gaps, data vulnerabilities);
- offering targeted oversight training for the board and senior management to enable them to identify and make informed decisions on AI-related risks and opportunities;
- embedding a strong risk-aware culture that encourages detection of emerging risks, early escalation, questioning of model outputs, and consistent application of governance across business lines; and
- integrating insights from incidents, data/model drift events, near-misses, testing outcomes, and validation findings, as well as relevant information from third-party AI providers, into staff training and behavioural expectations, to align capabilities with operational experience.

Maintain external awareness

Financial institutions proactively monitor developments outside the organisation for timely awareness of new technology, risks, and supervisory expectations. External developments can surface early risk signals that internal processes may not yet detect. Such efforts to maintain external awareness specific to AI can include:

- monitoring emerging AI capabilities, model architectures, and risk typologies to detect external shifts that may impact internal risk assessments;
- engaging regularly with peer financial institutions, industry groups, academics, technology providers, and authorities to remain informed about industry developments as well as any relevant regulatory and supervisory initiatives;
- participating in shared-learning fora, collaborative testing, and public-private consortia to identify risks or practices observed across the financial sector; and
- establishing routines to synthesise external intelligence and flag potential areas requiring governance review.

Case study: Learning from the AI ecosystem through public-private sector collaboration

Several authorities have established structured public-private sector collaboration platforms - including industry consortia and regulatory sandboxes - to build shared understanding of AI opportunities and risks. These platforms bring together authorities, financial institutions, and technology partners to develop practical guidance and deepen supervisory insight.

For example, the Monetary Authority of Singapore (MAS) convened Project MindForge in 2023, bringing together financial institutions, industry associations, technology partners, as well as consulting and

technology partners to examine AI risks and develop practical guidance.[42] The consortium produced a whitepaper on the emerging risks and opportunities of GenAI for banks in November 2023, and subsequently an AI Risk Management Toolkit to support financial institutions in implementing AI risk management practices in March 2026. Similarly, the Bank of England (BOE) and Financial Conduct Authority (FCA) established the AI Consortium in May 2025 to gather input from stakeholders on the capabilities, development, deployment, and use of AI in UK financial services.[43] The AI Consortium aims to identify how AI is being used in financial services, discuss emerging benefits and risks, and inform their approach to safe AI adoption.

Other authorities have used regulatory sandboxes to enable financial institutions to pilot AI use cases in a controlled environment. Launched in 2024, the Hong Kong Monetary Authority (HKMA)'s GenAI Sandbox provided participating banks with supervisory guidance, technical support, and access to GPU computing resources, while industry symposia and roundtables facilitated broader knowledge sharing. Building on its success, the GenAI Sandbox has since expanded to cover multiple financial sectors including banking, securities and capital markets, asset and wealth management, insurance, mandatory provident fund and stored valued facilities.[44]

Evolve with insight

Financial institutions update governance, controls, and oversight mechanisms as needed, based on lessons learned from both internal and external insights such as AI incidents, testing outcomes, and emerging risks. These updates can include:

- periodically reviewing governance, lifecycle controls, and oversight roles as risks evolve;
- regularly reassessing internal assumptions, including due to external developments to avoid outdated mental models or blind spots;
- performing root-cause analysis and post-incident reviews for AI-related incidents, recognising that AI incidents may reflect underlying issues that traditional analysis and reviews may not always be designed to capture (e.g. societal biases in training data);
- treating relevant anomalies, drift events, and near-misses as triggers to assess and, where needed refine controls, given AI models' unique and non-linear degradation patterns;
- refreshing policies, standards, and procedures based on learning, and maintain governance as internal experience accumulates;
- implementing governance mechanisms so that they adapt to new forms of AI (e.g. GenAI, foundation and frontier models, agentic systems) that introduce new risks; and
- incorporating lessons learned into capability development and tooling enhancements, strengthening the financial institution's ability to respond to future issues.

---

42 See MAS website.

43 See BoE website.

44 See HKMA (2026).

Case study: How an adaptive AI strategy delivered industry leadership

A G-SIB's AI approach has evolved through distinct phases, adapting governance, operating models, and controls as AI moved from experimental use to core business infrastructure.

The bank started by investing in internal research capabilities, specialised technical talent, and proprietary data infrastructure. This research-first approach enabled the bank to develop institution-specific AI methods suited to complex financial decision making, while also building a substantial portfolio of AI-related intellectual property. After about a year, the bank shifted from research to implementation in areas where AI seemed to be most fruitful and governable. These early deployments demonstrated that advanced AI systems could operate reliably and deliver measurable outcomes without undermining risk controls. In parallel, the bank invested in enterprise grade data and computing infrastructure to enable secure development and deployment at scale. AI capabilities were progressively embedded into internal tools used by tens of thousands of employees, including productivity, research, and software development functions. As AI adoption expanded over a couple years, the bank identified the limitations of decentralised development and adapted its organisational approach, by creating a centralised AI group reporting directly to the Chief Executive Officer. This group now functions as an enterprise accelerator, responsible for prioritising use cases, coordinating deployment across business lines, and ensuring consistent application of risk, security, and compliance standards.

This structural change marked a transition from project-based experimentation to industrial scale AI delivery. The new approach embeds privacy, security, fairness, accountability, and transparency requirements across the AI development and deployment stages - from design and training to monitoring and audit.

3. AI lifecycle management

This section focuses on sound practices related to the management and mitigation of AI risks throughout the stages of AI development and deployment or 'AI lifecycle', which is explained in Section 3.1. Sound practices 5-10 relate to certain stages of AI development and deployment, while sound practices 11 and 12 relate to cyber and ICT risk management as well as third-party risk management that apply across all stages.

3.1. Stages of AI development and deployment

Regardless of the term used, it is important to note there is no standard, one-size-fits-all stages or AI lifecycle. Different AI use cases may involve different stages of development and deployment of varying length.

In this report, the following illustrative AI lifecycle with seven distinct stages is provided to help financial institutions manage AI risks and adopt the relevant sound practices.[45] The seven stages of AI development and deployment in this report include:

[45] Financial institutions may develop their own AI lifecycle management processes with their choice of stages or adopt the lifecycle stages of different regulation, guidance and standards. For example, NIST AI Risk Management Framework uses a four-stage lifecycle (plan, design, develop, and deploy). For details, see NIST (2023).

- **Inception**: Identifying the needs, goals, and feasibility of the AI use case. This stage involves defining the context, potential use case(s), materiality, and risk assessment.
- **Design and development**: Defining the system architecture, data flows, and selecting AI model or system. Key activities include data governance, component selection, and implementing initial safety controls. If applicable, this stage may also include training an AI model, or customising or fine-tuning of third-party or open source AI models or systems.
- **Verification and validation**: Testing and confirming that the AI use case meets specified requirements (verification) and performs as intended in real-world scenarios (validation).
- **Deployment**: Releasing the AI model or system into its operational environment, which includes setting up human oversight and monitoring.
- **Operation and monitoring**: Running the AI model or system, logging activity, and monitoring for performance, safety, and security risks (including data drift).
- **Re-evaluation**: Periodically assessing whether the AI use case continues to meet its objectives under changing conditions, which may lead to retraining or updating.
- **Retirement**: Decommissioning the AI model or system as well as addressing long-term data and access risks, such as archiving or data disposal.

3.2. Materiality and risk assessment

Sound Practice 5 (Materiality and risk assessment): Financial institutions implement an effective and systematic approach to assess the materiality and risk of AI use cases at the inception stage and thereafter.

In assessing the materiality and risk of their AI use cases, identifying the relevance of these use cases to their 'critical operations',[46] and assessing the criticality and risks of their third-party service relationships involved in the AI use cases,[47] financial institutions can build on the work of the SSBs as well as financial authorities' respective regulations or guidance on AI risk management, model risk management, [48] operational resilience, [49] and third-party risk management.[50]

---

46 See FSB (2024a) and BCBS (2021a), pp.3. Some jurisdictions use different terminology, such as 'important business services' or 'critical or important functions'.

47 Assessing the criticality and risks of their third-party service relationships means assessing "third-party services whose failure or disruption could significantly impair their viability, critical operations, or its ability to meet key legal and regulatory obligations" (see FSB (2023b)). Some jurisdictions use different terminology, such as 'material third-party arrangements'.

48 For example, see BOE/PRA (2023), ECB (2025a), JFSA (2021), OSFI (2025b), and FRB et al (2026).

49 For example, see BOE/PRA et al (2021).

50 For example, see BOE/PRA (2021) and FRB et al (2023).

Financial institutions can develop a specific framework for assessing the materiality and risk of their AI use cases, or adapt and enhance relevant frameworks[51] to take account of AI-specific considerations and risks. Criteria used in existing frameworks can be directly relevant to assessing the materiality and risks of AI use cases (e.g. the complexity of the underlying AI model or system, criticality to operations,[52] or the nature of the data or information that the financial institution shares with third-party service providers). Regardless of the approach used, it is important to capture all relevant dimensions of materiality and risk systematically.

The materiality and risk profile of an AI use case are interrelated but are different concepts. In practice, financial institutions may assess them within a single process. Both materiality and risk are relevant in deciding the appropriate and proportionate level of risk management and risk mitigation for an AI use case, including guardrails and level of human oversight.

Assess materiality and risks of AI use cases at the inception stage and thereafter

The materiality of an AI use case refers to its importance to the financial institution's viability, critical operations, and its ability to meet key legal and regulatory obligations (including towards customers). The materiality assessment of an AI use case will typically consider the criticality of the operations or business lines that an AI use case will support, and the relative importance of the AI use case to those business lines or operations. On the latter, an AI use case may support critical operations or core business lines in various ways, ranging from narrow administrative tasks (e.g. automated document sorting) to being essential to their effective functioning.

The materiality of an AI use case may change over time. Financial institutions often deploy AI use cases in non-critical operations and business lines initially, and scale to critical operations progressively as they gain comfort with use. Therefore, in addition to assessing the materiality of a proposed AI use case at inception, financial institutions reassess it at regular intervals and upon certain events or milestones (e.g. a decision to scale up the use of the AI use case to critical operations).

The risks of an AI use case refer to the attributes that can cause harm to the financial institution (including but not limited to financial, legal and regulatory, and operational risks) or its customers. An AI use case can pose high risks even if it is not material. For example, an AI agent performing a low materiality task may still have access to ICT assets or a significant amount of sensitive data. Consequently, the agent could cause significant harm to the financial institution or its customers if it accidentally or intentionally acts outside defined parameters.

Financial institutions systemically assess risks that an AI use case may create or amplify, including as a result of:

- the AI's autonomy, capabilities, complexity, the extent of access to systems and sensitive data, and scope of action;[53]

---

51 Existing frameworks include model risk management, operational resilience and/or third-party risk.

52 See FSB (2024a) and BCBS (2021a), pp.3.

53 This may include the financial institution's evolving understanding of what the AI is capable of doing and allowed to do. There are examples of AI agents, models, or systems being capable of performing actions they were not initially designed or expected to do.

- cyber/ICT vulnerabilities, data dependencies, explainability, reliability, substitutability, susceptibility to performance degradation, and/or third-party/supply chain dependencies;
- the feasibility of human oversight and level of reliance on the AI use case's outputs; and
- relevant legal and regulatory requirements.

For each AI use case, financial institutions assess both inherent risk (i.e. the risk posed before relevant risk management and mitigation is applied) and residual risk (i.e. the leftover risk after risk management and mitigation is put in place).

As with materiality assessments, financial institutions take a dynamic approach to the risk assessment of AI use cases. Given rapid advances in technology and AI adoption, the AI risk landscape can change quickly. In addition to assessing the risks of a potential AI use case at inception, financial institutions generally repeat risk assessments periodically or upon the occurrence of certain events (e.g. after detecting incidents or performance degradation). Such reassessments can lead to risks being given a higher or lower tiering.[54]

3.3. Selection

Sound Practice 6 (Selection): When selecting AI models or systems, financial institutions consider business objectives, operational and technical needs, as well as the materiality and risks of AI use cases.

Financial institutions have the ultimate choice of whether to adopt AI for specific use cases, taking into account alignment to their business strategy and their assessment of potential benefits and risks.

If a financial institution chooses to adopt AI for a given use case, during the selection stage it decides:

- whether to develop AI in-house, procure it from third-party AI providers, use open-source tools, or use a combination of these; and
- the appropriate AI model or system (such as traditional, GenAI or agentic) for the use case.

In making these decisions, financial institutions consider factors such as:

- materiality and risks;
- customer impact;
- intended use;

[54] For example, risks relating to the complexity of an AI model or system can reduce over time as both technology and its adoption mature, and these risks become better understood and managed or mitigated.

- whether the AI use case will apply across the financial institution, or to a specific business line or function;
- infrastructure and technical requirements;
- data needs and availability;
- readiness of the financial institution, including expertise and skills;
- legal and regulatory requirements;
- available resources;
- costs; and
- if applicable, third-party dependencies and risks.

Certain considerations may involve trade-offs. For example, high performing AI models or systems can be more complex, computationally intensive, and have less explainability or transparency than simpler AI models or systems with lower performance. The most complex AI models or systems may not invariably be the most appropriate. In some instances, a simpler, less costly, more explainable model or system may still deliver the financial institution's business needs.

Financial institutions' assessment of costs includes the commercial, computing, energy, financial and staff costs of different AI models or systems, as well as costs associated with business continuity and contingency planning with third-party AI providers. Financial institutions consider ways to improve cost efficiency. The selection process is informed by research into the capabilities of the different AI models, and inputs from business users, control functions, subject matter experts, technical experts, and industry practices.

Once they have selected an AI model or system, financial institutions implement appropriate, documented change management processes throughout the AI lifecycle, in particular, for high materiality or risk use cases.

3.4. Data governance

Sound Practice 7 (Data governance): Financial institutions establish appropriate data governance to maintain data that is fit-for purpose for training, testing, and using AI (i.e. accurate, complete, consistent, reliable, secure).

Effective data governance is essential for effective AI risk management, and for promoting high-quality data to underpin effective and reliable AI outputs.

Data governance and data quality are key regardless of whether a financial institution develops an AI model or system in-house using its own data, or relies on an AI model, system and/or data from third parties. However, in the latter case, there may contractual, legal, or practical limitations on the financial institution's ability to access or process third-party data. In such cases, financial institutions consider alternative or compensating ways to deliver the overall outcomes of

ensuring sound data governance and high-quality data, including by assessing the third-party service providers' data governance practices, processes, and data quality controls.

AI-specific considerations in data governance, which can apply to all types of data usage, include:

- the role of data in shaping AI design choices and outputs of AI, depending on the specific AI technology and use cases;
- the variety of data that are used in different AI applications, which may include multimodal (audio, text, image, video), unconventional (social media feeds), and synthetic data,
- the forms that data used by AI can take (structured, semi-structured, unstructured);
- the speed at which some data used by AI are generated and processed (velocity), which can be real-time or near real-time (e.g. for financial trading data and social media feeds);
- the volume of data used by, in particular, GenAI; and
- the ability of data to influence the behaviour (as well as outputs) of AI agents which creates additional risks around how agents ingest, interpret, memorise, use, and re-use data.

Establish effective data governance frameworks

Financial institutions adopt a centralised data governance framework, decentralised but consistent frameworks, or a hybrid framework. This framework or frameworks may be pre-existing and cover the financial institution's approach to data in general, provided they cover data used for AI purposes and address AI-specific considerations.

The data governance framework also addresses data issues in third-party AI models, systems, or applications that the financial institution deploys, including where the financial institution provides prompts, internal documents, or other context to a third-party tool, or where the tool accesses the financial institution's data for its outputs (e.g. an LLM that accesses confidential or sensitive data for Retrieval-Augmented Generation (RAG)).

An effective data governance framework addresses the following AI considerations:

- alignment with the financial institution's risk appetite and strategy;
- clear accountability in different scenarios, including where the financial institution uses its own data, public data, and/or data from third-parties;
- basic data literacy for all staff and additional, targeted training for staff working with data;
- resources needed for AI data processing, including computation and storage;

- procedures for the classification and labelling of data, including metadata;
- guidelines for processing unconventional data, and data in different formats;
- data security including controls for (i) authorised access and data recovery, and (ii) confidential and personal data (including privacy enhancing techniques); and
- an adaptable but consistent approach for assessing data quality and performance.

Effectively manage different stages of AI data use

Managing the different stages of AI data use is key to sound data governance. Effective management of these stages can vary by use case, but typically includes:

- an initial assessment of the availability, quality, quantity, and suitability of data needed for the AI use case, which may in turn influence the AI design choice. This stage includes considering whether to use internal data, public data, and/or data from third parties.
- a data collection stage that tracks and documents the authenticity and origin of data (data provenance). This stage can include due diligence on public and/or third-party data.[55]
- a data classification stage that documents the materiality, risk, sensitivity, intended and prohibited uses, and other characteristics of data.
- a data preparation stage which may involve aggregating, annotating, cleaning, enriching, labelling, or updating data. This is an essential stage of the data management lifecycle and the most resource-intensive.[56]
- an assessment of data quality after the data preparation stage and before deployment, which identifies gaps and limitations in the dataset as well as mitigating or compensating measures.
- post-deployment monitoring to identify and remediate problems that may be caused by data issues, such as data drift.

Documenting key steps throughout these stages is integral for tracking and understanding how data is enhanced, enriched, transformed, and used as it flows through the financial institution (known as 'data lineage'), which can enhance data quality, facilitate auditability, and aid the remediation of issues and risks, including root cause analysis.

---

55 For example, checking data generated by a GenAI model for accuracy and hallucinations before using that data to train another AI model or system.

56 This stage may include embedding 'instructions' in the metadata of data used by AI setting out the data's sensitivity, applicable legal requirements, permitted and prohibited uses etc. These instructions can act as guardrails on the actions the AI model, system or agent takes to fulfil its tasks (policy aware' data').

3.5. Explainability and transparency

Sound Practice 8 (Explainability and transparency): Financial institutions understand differences in the explainability of various types of AI. If appropriate and feasible, financial institutions adopt more explainable AI, or consider compensating controls. Financial institutions also provide appropriate transparency tailored to different stakeholders.

In this report, explainability refers to the ability of an AI model or system to provide clear and interpretable outputs or decisions, specifically the extent to which humans can understand how an AI model or system generates outputs from inputs, including: (i) the factors, logic, or processes that contribute to outputs; (ii) the relative importance of input variables or features; and (iii) how methodological relationships vary across context or use cases. Explainability stems from characteristics in AI models or systems. It is not binary but a continuum.[57]

Transparency refers to the extent to which appropriate information about an AI model or system is made available to relevant stakeholders, including the financial institution's board and senior management, counterparties and customers interacting with the AI model or system, and authorities.

Explainability and transparency are complementary but not synonymous. Even if appropriate information about an AI model or system is available, it may still have inherent explainability limitations and neither its developers not its deployers may understand how it specifically translates inputs into outputs.

Explainability

Financial institutions identify, understand, and (where appropriate) communicate the degree of explainability of different AI models and systems for specific AI use cases. Financial institutions recognise that: (i) the importance of explainability may vary depending on the level of autonomy, materiality, and risk of different AI models and systems; and (ii) explainability is not an end in itself but enables accountability, the assessment of conceptual soundness, and compliance with legal and regulatory requirements.

Where justified, financial institutions consider trade-offs between explainability and performance. Limited explainability might be acceptable in some use cases. For example, in certain fraud detection use cases the superior predictive power of less explainable forms of AI may warrant their use. Conversely, AI that is heavily relied upon for credit decisioning, insurance underwriting, or other activities with potentially high impact on customers may warrant greater explainability.

There is a range of choices available to financial institutions to address explainability issues. These include building explainability into AI model or system development, using explainability methods, and applying compensating controls where explainability is limited. The appropriate choice(s) will depend on the AI use case.

---

[57] A related concept to 'explainability' is 'interpretability', which is the ease or difficulty of predicting what an AI model or system will do, i.e. the degree to which the cause of a decision can be understood.

Build explainability into model development

During design and development financial institutions may create or select more explainable forms of AI if they do not perform significantly differently compared to more complex, less explainable forms. Financial institutions acknowledge that, all else equal, lower explainability increases the uncertainty about the AI model or system.

Apply explainability methods

Financial institutions apply explainability methods usually after the AI model or system generates its outputs ('post-hoc explainability').[58] These methods are generally applied when explainability challenges arise in model development. Financial institutions understand that: explainability methods vary in applicability, maturity, and scope; have limitations that may introduce additional uncertainty; have prerequisite conditions to enable their functionality; and can involve trade-offs.[59] Consequently, financial institutions:

- use explainability methods as part of a broader suite of AI risk management and mitigation tools, rather than as standalone solutions.
- avoid over-reliance on a single method. If appropriate, financial institutions combine two or more explainability methods (known as 'triangulation').
- recognise that explanations may be susceptible to assumptions, data characteristics or feature correlations, and some explainability methods may themselves rely on additional models.[60]

Explainability methods can provide insights into the behaviour of AI models or systems at the global or local level. Both can be relevant and useful for understanding AI model or system behaviour, depending on the AI use case, its materiality and risk, and the type of AI being used. For example, agentic AI systems that execute sequential decisions may warrant additional or different approaches to explainability that look not just at the agent's final output but intermediate decisions and steps.[61] In this manner, employing methods such as logs for AI agent activities can help humans track the agent's actions in real time or evaluate the actions after the fact (or both).

- 'Global explainability' refers to how an AI model or system behaves overall. For example, in a ML credit scoring model, it refers to the key drivers of model output across all relevant borrowers. As another example, in an agentic system, global explainability would characterise typical execution patterns and common decision pathways.
- 'Local explainability' explains how an AI model or system delivered a given, single output. In the same ML credit scoring model, it refers to the key drivers of model output for an individual borrower. For the agentic system, local explainability would trace the

[58] Examples of explainability methods include partial dependence plots and feature importance analysis.

[59] See BIS-FSI (2025b).

[60] See BOE (2025).

[61] Financial institutions deploying multi-step agentic AI systems are developing explainability approaches that can trace and evaluate the full reasoning path, not the individual outputs.

specific sequence of reasoning steps, tool invocations, and intermediate decisions that led to the final, specific outcome.

Case study: Project Noor

Project Noor is an initiative of the BIS Innovation Hub that seeks to equip financial supervisors with independent, practical tools to evaluate and interpret the inner workings of AI models used by banks and other financial institutions.

Led by the BIS Innovation Hub Hong Kong Centre in collaboration with the HKMA, the UK FCA, and Saudi Central Bank, Project Noor ('light' in Arabic) will prototype the latest Explainable AI (XAI) techniques in a controlled setting. XAI converts complex model logic into plain language and intuitive visuals, making it easier to see which factors influenced a decision and how sensitive that decision is to change, all while preserving privacy.

By combining explainable AI methods with risk analytics, the project aims to deliver a prototype through which supervisors can verify model transparency and test robustness. Project Noor aims to shed light into AI black box models using the latest XAI techniques in the context of financial supervision in a controlled setting.[62]

Deploy compensating controls

Where explainability is constrained due to the complexity of AI models or systems, limited accessibility to information about the AI model or system, or third-party dependencies, financial institutions deploy compensating controls including:[63]

- guardrails, such as stricter limits on permissible use cases;
- benchmarking and/or the use of challenger models or systems;[64]
- sensitivity and stress testing;
- enhanced data governance;
- more extensive validation and testing; and
- increased human oversight, such as heightened scrutiny of outputs following deployment.

Compensating controls are particularly important in high materiality and high-risk AI use cases, where they can help financial institutions accept limited explainability while mitigating potential risks.

If compensating controls cannot mitigate the risk stemming from the limited explainability of an AI model or system in line with the financial institution's risk appetite and tolerance, the financial

62 See BIS website.

63 See IAIS (2025) pp.27 and ECB Banking Supervision (2025).

64 A new, alternative, or retrained model or system that is tested against an existing, deployed model (known as the 'champion' model) to see if it can perform better.

institution considers alternative solutions, such as deploying a more explainable AI or non-AI model or system, even if performance is inferior.

Transparency

Transparency is key for ensuring that both internal and external stakeholders are appropriately informed about the AI models or systems that affect them. For internal stakeholders, including those deploying an AI model or system developed by a third-party, transparency enables them to assess whether it continues to be fit-for-purpose and is performing as intended. For external stakeholders such as customers, transparency helps them understand the role AI plays in the products or services they use, and how AI may have influenced outcomes that are relevant to them.

Financial institutions make available appropriate information relating to their use of AI to internal and external stakeholders, which may include: the intended, permitted and prohibited uses of an AI model or system; its capability and its limitations; training data if available; and its development process. The optimal level of transparency may vary depending on the intended recipients of the information. Additionally, the degree of transparency may vary across different stages of model development and deployment, and whether the model or system is developed internally or procured from a third-party AI provider.

Financial institutions tailor the information they provide to different stakeholders to support effective challenge and accountability, oversight and use of AI, while maintaining proportionality and avoiding undue complexity. Appropriate transparency involves providing information about how AI models or systems produce their outputs (subject to any explainability limitations referred to above) in ways that are understandable to the intended audience. Information provided to customers, authorities, auditors, risk managers, the board and senior management differ, reflecting differences in needs, expertise, responsibilities, and decision making of these different stakeholders.

Transparency to customers and end-users[65]

Financial institutions consider the information needed for customers to understand the nature and limitations of financial products or services where AI plays a material role. Appropriate transparency does not necessarily entail financial institutions disclosing to end-users that they are using AI in all circumstances. However, it may include informing end-users: (i) that they are interacting with AI (e.g. through chatbots); (ii) if AI might materially influence decisions that could affect them; (iii) if there are limitations associated with the AI model or system; and/or (iv) of their rights to appropriate explanations, contestability, or redress mechanisms (if applicable). Financial institutions may also consider offering “opt-out” options or channels that allow customers to request a review of AI-generated outputs or human assistance.

[65] This section can also apply to investors.

3.6. Performance management

Sound Practice 9 (Performance management): Financial institutions evaluate the performance of AI use cases proportionately to their materiality and risk, including through performance assessments, testing, and ongoing monitoring**.**

Assess dimensions of AI performance

Financial institutions evaluate AI performance across different dimensions, proportionately, and over time (including via ongoing monitoring), and adapt their performance assessment to the objectives and anticipated outcomes of different AI use cases. They adjust the frequency, intensity, scope, and types of performance assessment based on factors including:

- the materiality and risk of the AI use case.[66] Material or high risk AI use cases may justify higher performance requirements and a lower tolerance for errors.
- the AI model's or system's attributes, including complexity and scope of action. More complex models or systems may be more difficult to evaluate, while those with a broader scope of action may have greater downstream consequences if issues arise.
- the AI use case's prior performance. A model's or system's performance track record may signal where closer scrutiny is needed and helps financial institutions calibrate their assessment benchmarks.
- any recent changes to the AI model or system, or its data. [67]
- any significant changes to relevant real-world conditions.

Several types of performance assessments used for non-AI models or systems and other analytical approaches can help financial institutions assess AI models or systems.[68] Several key types are examined below (e.g. accuracy, benchmarking, range of output, robustness).

Assessing the performance of traditional AI models or systems generally involves comparing a quantitative output with predetermined targets or thresholds (e.g. predicted versus actual quantitative values). It is therefore conceptually straightforward and easier to automate.

GenAI models or systems produce a range of potentially acceptable outputs but not necessarily a single correct output. Assessing GenAI performance is harder to automate and may involve more human judgment and qualitative metrics such as the coherence or relevance of an output.

Assessing the performance of agentic AI is a live area of research, which can be even more complex as it needs to consider not only whether the AI agent achieved its overall objective but also whether the actions it took to do so were appropriate.

---

66 MAS (2025b) pp.23.

67 See IAIS (2025) pp.25.

68 See BOE/PRA (2023) in particular Principles 3.3 and 4.4-4.5, Also see ECB (2025a), JFSA (2021), OSFI (2025), and FRB et al (2026).

Financial institutions assess dimensions of performance (individually and collectively) with the specific weighting of different dimensions and their corresponding performance measures determined by the AI use case. Dimensions of performance often include:

- *accuracy*: how close the output aligns with target or realised outcomes. Accuracy can be measured in different ways and along various horizons. Some AI models or systems may perform better in the short-term and others in the medium or long-term.
- *range of output* (also known as a 'confidence band' or 'distribution of outcomes'): whether output is centrally clustered or more widely dispersed, with a narrower range generally preferred. This dimension of performance can indicate how much uncertainty is associated with an output and whether a given point estimate is representative.
- *robustness*: how well the AI performs in new conditions or circumstances (also known as the ability to generalise to new data). Some forms of AI may experience natural degradation over time.
- *stability*: how reliable an AI model's or system's outputs are over time, including whether the same task performed or query asked at different times produces different outputs.[69]
- *sensitivity* or *stress testing*, during which extreme inputs or assumptions are fed into the AI model or system to determine the effects on outputs. This type of testing can help identify limitations and uncertainties of AI, including the conditions under which performance deteriorates or fails.

Financial institutions specify thresholds across all relevant dimensions of performance to indicate which results constitute acceptable performance, warrant further investigation, or indicate significant issues. Where possible, establishing clear performance thresholds *ante hoc* (i.e. before outputs are generated) is generally preferable to reviewing outputs *post hoc* and retrospectively deciding what constitutes adequate performance. As noted above, this might be harder for certain types of AI, such as agentic AI. Financial institutions generally review performance thresholds periodically to confirm they are still appropriate.

Carry out developmental testing

Financial institutions carry out testing during the design and development stage ('developmental testing') to understand early on whether an AI model or system needs adjustments or additional calibration. Developmental testing typically covers the dimensions of performance noted above, where feasible. Financial institutions may also use this testing to decide whether a given AI use case (or its underlying model or system) is worth pursuing further.

Financial institutions' developmental testing generally considers how well an AI model or system performs, according to its intended use. This may include checking whether there is:

---

[69] One helpful way to consider stability is the potential 'capability-reliability' gap, i.e. if an AI model or system is capable of executing operations but not consistently enough to be reliable.

- *overfitting*, occurring when the AI model or system is narrowly focused on its training data, performing well for that training data but poorly on new data, contributing to a lack of robustness and inability to generalise;[70] and

- *unacceptable bias*, including statistical biases such as trading platforms favouring large-cap stocks or stocks in certain sectors. Biases can also lead to legal or regulatory breaches or discrimination against certain groups or parties. Financial institutions test for unacceptable bias in various parts of an AI model or system, including its training data, its design choice, or how its output is used.

It may be challenging to identify and detect unacceptable bias, especially if a financial institution does not have access to the underlying dataset, or if the AI model or system has explainability limitations.

Developmental testing of GenAI and agentic AI models or systems can give rise to additional challenges and considerations. For instance, small linguistic changes to the input (prompt) of a LLM can significantly change the output, which increases unpredictability. Moreover, in certain instances, there may not be a 'ground truth' (i.e. verified, true data used for training, and testing AI performance), meaning the assessment may not be as straightforward, and financial institutions may include subject-matter experts as evaluators. Including a multi-disciplinary set of evaluators can help assess all dimensions of performance.

Case study: Developmental testing

A large internationally active bank began developing a ML model in its trading business with the expectation that it would perform better than the bank's current non-AI model. The bank developed an initial version of its ML model, conducting various tests to assess performance. Testing showed that the new ML model did not consistently outperform the existing non-AI model, including for stability. In addition, the training dataset was found to be not representative of the production data to be used with the ML model over time. Other aspects of the ML model did not fully align with the conditions in which it would be deployed. Given these substantive adverse outcomes, as well as the higher complexity of the ML model that made it more challenging to assess, the bank decided to postpone further deployment of the ML model and continue to use its existing non-AI model.

Conduct ongoing monitoring and testing

After confirming through developmental testing that the AI model or system can be responsibly deployed, financial institutions undertake ongoing monitoring and testing to: (i) confirm that the AI model or system performs well over time and relative to a range of realised outcomes; and (ii) support early identification of material deviations in results from the intended design and timely responses by the financial institution. In certain cases, such as agentic AI, monitoring may involve regular, routinised assessment of specific actions or consequences - and not just a review of output.

[70] There is evidence that AI models or systems can have a higher propensity for overfitting, compared to non-AI models and other quantitative approaches.

Financial institutions align the frequency, horizon, intensity, and type of ongoing monitoring and testing to the characteristics, materiality, and risks of different AI use cases. Key steps in ongoing monitoring and testing may include:

- testing over time using new data, or data not used in training;
- repeating tests conducted prior to deployment;
- monitoring degradation in performance;
- checking for data or model drift;
- periodically confirming that established performance metrics continue to be appropriate to the use case; and
- assigning staff to try to 'break' the AI model or system, or identify vulnerabilities, weaknesses, and uncertainties ('red-teaming').

Financial institutions also use benchmarking where appropriate to evaluate AI performance by comparing an AI model's or system's output with output from other relevant models or systems, or other relevant information. Benchmarking is especially helpful when there are insufficient realised outcomes against which to test the AI model's or system's output, or when attempting to triangulate accurate outputs. Benchmarking is most helpful when the alternative output has been produced by a model or system that takes a different perspective, such as using different data sources or methodologies.

As the number of AI use cases (in particular high materiality and risk use cases) increases, financial institutions assess the resource implications, scalability, and viability of different types of performance assessment and monitoring. Where appropriate, financial institutions explore ways to augment performance assessments potentially through automation.

Financial institutions manage the above challenges with dynamic updating of AI models or systems that draw in new data, alter inputs used, change feature relationships, or undergo other modifications. In doing so, financial institutions maintain comprehensive, objective, and appropriate testing, including developing new tests and conducting additional bias checks as needed. Financial institutions also conduct more frequent and intensive ongoing monitoring, and alert relevant stakeholders about updates to AI models or systems.

Financial institutions respond appropriately to performance assessment results based on the AI model or system, materiality, risk level, and nature of findings. For example:

- poor results can indicate that the AI model or system needs updating, redeveloping, or retiring.
- results that indicate only mild concern might warrant further testing and closer monitoring.
- in certain cases, compensating controls, such as an overlay or other adjustment to outputs, might provide an appropriate response. Compensating controls can also

include more intensive human oversight or limiting the use of the AI model or system to lower-risk areas until concerns abate.

- in many cases compensating controls may be just temporary measures until more substantial steps can be taken to address underlying issues.

3.7. Human oversight

Sound Practice 10 (Human oversight): Financial institutions implement appropriate and effective human oversight relevant to the materiality, risk, autonomy, complexity, and explainability of different AI use cases.

Effective human oversight has a tangible positive impact and supports responsible AI adoption (i.e. by improving decision-making, preventing harm, or providing redress). Effective oversight is meaningful, i.e. it involves humans having sufficient ability, authority, and incentive to intervene, as opposed to oversight that is nominal or driven by 'tick-the-box' compliance.

Design and develop appropriate and meaningful human oversight of AI

Financial institutions design and develop AI to enable effective and meaningful human oversight (as far as technically possible) which includes:

- enabling staff to assess the outputs of the AI model or system, and understand its capabilities and limitations;
- preventing automation bias (i.e. overreliance)[71] by creating incentive alignment for thorough, thoughtful review, such as by incorporating performance metrics that reward quality of oversight decisions;
- creating feedback mechanisms for the AI model or system to learn and improve from oversight decisions, including tracking when humans override AI recommendations, analysing patterns in these interventions, and using this information to refine both the AI system and the oversight processes; and
- assessing whether human overseers are providing effective oversight.

When obtaining an AI model or system from an open-source repository or third-party AI provider, financial institutions consider the following as part of their due diligence:

- the extent to which the AI model or system has been designed to enable effective and meaningful human oversight by those deploying it; and

[71] For example, an AI model or system may be designed to seek human feedback in specific scenarios or provide confidence scores for outputs.

- (if applicable) the extent to which the third-party AI provider will provide human oversight. For example, in the case of a GenAI model, the extent of human oversight when the third-party AI provider deploys model updates and new versions.

Select appropriate forms of human oversight

Financial institutions select the most appropriate form(s) of human oversight taking into account the materiality, risk, and other relevant factors, including technical feasibility, proportionality, resource effectiveness, as well as legal, regulatory and supervisory considerations. Common forms of human oversight may include but are not limited to the following:

- *'Human-in-the-loop'*: human approval of all decisions made by AI, which might be appropriate in situations where accuracy is more important than speed of decision-making (e.g. approval of a health insurance claim).

- *'AI-in-the-loop'*: integrates AI into human oversight to augment performance monitoring rather than merely using humans to oversee AI. It uses AI as a supportive layer for decision-making and task automation while keeping humans in control (e.g. humans monitor and respond to AI-enabled automated alerts).[72] AI-in-the-loop may become warranted as financial institutions scale the number of AI use cases.

- *'Human-on-the-loop'*: human intervention periodically or only where necessary. For example, when the AI's outputs have a low confidence score or an override is necessary (e.g. AI chatbot autonomously handles routine customer inquiries, with human agents alerted to intervene only when sentiment analysis detects frustration or the request volume exceeds the AI's capability).

- *'Human-in-command'*: high-level human oversight, including deciding the extent of autonomy, setting guardrails, and managing its overall impact. This form of human oversight is aimed at AI with high levels of autonomy, such as agentic AI.

- *'Kill switch'*: mechanisms enabling human intervention to halt or constrain AI operations, ranging from complete shutdown (which can be used in disabling algorithmic trading) to dynamic or graduated degradation where systems transition from autonomous to human operation.

- *'Contestability'*: mechanisms for human appeal and redress of AI outputs (e.g. allowing customers to appeal AI-enabled loan denials if they suspect unacceptable bias).

Set clear responsibility for human oversight throughout the AI lifecycle

The specific individuals responsible for human oversight and their related functions may change during the lifecycle of an AI use case. Financial institutions:

---

[72] For example, agentic AI systems may automatically escalate to humans when accuracy metrics are beached or when requested by users, maintaining service while addressing performance issues.

- allocate responsibility for monitoring an AI use case after deployment to staff who were not part of design and development to promote objectivity;
- document who the responsible staff are at different stages of the lifecycle, for example, by maintaining responsibility matrices; and
- tailor the training of staff to the lifecycle stage(s) they will be responsible for.

Additional considerations for GenAI and agentic AI

Some steps that financial institutions take for human oversight of GenAI and agentic AI include:[73]

- requiring a human to intervene when an LLM starts generating offensive, illegal, or inappropriate output.
- having humans periodically review and spot-check GenAI outputs to confirm consistency and stability over time.
- as a form of testing, having humans try to coerce an LLM or an AI agent to act outside its defined parameters ("jailbreak").
- assigning and documenting individual identifiers to AI agents to facilitate monitoring and identity management;
- placing boundaries on AI agents, such as defining actions that are prohibited or that require human approval at appropriate checkpoints (i.e. high risk actions or steps outside of the agent's initial human-defined scope).
- testing the behaviour of AI agents, including through 'red-teaming'.[74] This can include testing how agents interact with other agents both pre and post deployment.
- defining and, if appropriate, limiting AI agents' abilities to interact with their external environment (including APIs, data, ICT systems, and other agents) until financial institutions are confident that AI agents can do so safely.
- placing controls on AI agents executing financial transactions especially with customers' funds; including human approval or dual authorisation for transactions above a certain value; restricting direct agent access to payment systems (requiring human intermediation for execution); and maintaining audit trails of agent transactions.
- analysing human oversight patterns to verify that human oversight is effective. For example, investigating cases where humans either consistently approve agent recommendations without modification (rubber-stamping) or override agents in ways that suggest pervasive misalignment.

---

73 Various organisations are actively developing best practices for agentic AI. See for example NIST (2026).

74 Red teaming is a structured testing effort to find flaws and vulnerabilities in an AI model or system, often in a controlled environment and in collaboration with developers of AI.

- defining an appropriate frequency and operating model for human oversight based on the use case (see above) and triggers for transitioning between different forms of human oversight based on performance metrics or user requests. In some cases, monitoring may need to be continuous or performed by AI. However, even when active monitoring is primarily undertaken by machines, financial institutions and individuals retain ultimate accountability.

- monitoring and documenting the process of AI agents (including key intermediate steps), with particular attention to tool access patterns that may indicate scope creep or misuse (e.g. agent querying databases beyond those necessary for the task, or accessing customer records more broadly than required). This may include: the reasoning at each autonomous decision point; the APIs' data and other tools accessed by the agent; and the final outcome (including the agent's interpretation of success). This can improve staff's understanding of how an agent fulfilled its objectives (short of full explainability) and facilitate certain forms of human oversight (e.g. contestability, kill-switches, or overrides).

- as the use of AI agents across the financial institution increases, adapting human resources controls and processes to AI agents in a way that treats them as synthetic employees.

3.8. Cyber and ICT risk management

Sound Practice 11 (Cyber and ICT risk management): Financial institutions manage AI-related cyber and ICT risks, including, where appropriate, by incorporating AI cyber and ICT risk scenarios into tests and exercises, sharing relevant information, and using AI tools in cyber and ICT risk management.

Adapt or enhance cyber security and ICT risk management to address AI risks

To effectively monitor, manage, and mitigate cyber and ICT risks relating to AI (hereafter AI cyber and ICT risks), financial institutions adapt and enhance, as needed, existing practices on cyber and ICT risk management. There are extensive regulations, guidance, supervisory expectations, information bulletin, and practices on operational resilience, cyber and ICT risk management, and third-party risk management by the FSB,[75] SSBs,[76] G7,[77] national authorities,[78] national cybersecurity agencies, and other organisations active in the field of cyber and ICT risk management[79] that financial institutions can adapt. These measures can even be effective risk mitigants against the capabilities of advanced AI models, and include:

---

75 FSB (2020), FSB (2023a), FSB (2023b), and FSB (2025a).

76 For example, see BCBS (2021a); BCBS (2021b), BCBS (2025), CPMI-IOSCO (2016), and IAIS (2026).

77 For example, see G7 Cyber Expert Group (2020).

78 For cyber and ICT risk management in the use of AI at financial institutions, see for example BaFin (2026), APRA (2026), OSFI (2026), and BOE et al (2026).

79 See UK Department for Science, Innovation & Technology (2025a) and UK Department for Science, Innovation & Technology (2025b), Also see ANSSI (2025) and NIST (2025).

- cyber and ICT governance, and cyber hygiene.
- reducing the potential attack surface by implementing strong network security. This may include:
  - documenting which systems are connected to the Internet and whether these connections are necessary for functionality; and
  - segmenting and micro-segmenting the most valuable ICT assets.
- implementing measures to monitor, prevent, and remedy the use of 'shadow AI'.
- applying secure development practices to AI.
- authentication and 'identity and access management' (IAM) including 'dynamic IAM' for AI agents to address the limitations of IAM for human users,[80] and phishing-resistant multi-factor authentication (MFA) to mitigate some of the risks of advanced AI models.
- adopting a multi-layered cyber defence strategy ('defence-in-depth') to AI so that, if one or more security layers were breached, the remaining layers are still effective in mitigating the risks. This is particularly important for advanced forms of AI, which can sometimes compromise certain layers, such as detection.
- applying secure-by-design principles to AI design, including performing threat modelling and incorporating security considerations during design, as well as implementing secure coding and code reviews during development.
- applying the principle of 'least privilege' to AI agents and their sub-agents (giving them the minimum permissions and access controls necessary to accomplish their tasks).[81]
- implementing security patches and updates promptly but securely. Some advanced forms of AI are particularly adept at discovering vulnerabilities faster and at greater volume which may accelerate the pace and volume of patching in future. Financial institutions assess if they can keep pace with compressed exploitation windows driven by AI-driven reconnaissance and exploitation, or if changes are necessary. Where patches are not yet available or cannot be applied immediately, financial institutions consider compensating controls such as virtual patching[82] to reduce exposure while permanent remediation is planned and implemented.
- enhancing monitoring and logging of anomalies and cyber incidents with controls on data provenance, cross-validation of datasets, and model behaviour tracking.

---

80 Dynamic IAM (IAM) grants, modifies, or revokes user permissions in real-time based on context, such as user behaviour, device security posture and location. It uses Attribute-Based Access Control and AI to allow for continuous verification. It can minimise unauthorised access risks by adjusting access dynamically rather than manually or statically

81 For example, see Cloud Security Alliance (2025).

82 A security policy enforcement layer which does not fix a known vulnerability, but mitigates its exploitation, for example through monitoring of exploitation attempts and intercepts such attempts.

Understand and monitor AI cyber and ICT risks

Financial institutions use specialist sources of information to identify common, emerging, and new AI cyber and ICT risks. These sources include:

- OWASP Machine Learning (ML) Security Top 10: an up-to-date list of the top 10 security issues of ML systems.[83] OWASP also publishes a top 10 list of security issues for agentic and LLM applications respectively.[84]
- MITRE ATLAS: a knowledge base of adversary tactics and techniques tailored to ML systems and covering the entire AI lifecycle.[85]
- NIST AI 100-2 E2025, Adversarial Machine Learning: a taxonomy and terminology of Attacks and Mitigation.[86]

Use AI-powered tools to manage AI cyber and ICT risks

Financial institutions consider adopting AI-powered cyber and ICT risk management tools, including:

- AI-powered firewalls to help protect financial institutions against data exfiltration, harmful content, and prompt injection attacks;
- active liveness detection tools to detect deepfakes;[87]
- AI-powered behavioural analysis to detect anomalous behaviour, which can help detect distillation, pre-positioning, or prompt injection attacks;
- advanced AI models to identify and remediate vulnerabilities;
- threat modelling and vulnerability identification, using adversary tactics and techniques against AI-enabled systems based on threat intelligence;
- data loss prevention tools that use contextual understanding and behavioural monitoring to protect data against attacks such as privileged escalation;
- AI-augmented incident response, including automating incident response playbooks or speeding up forensic examinations and threat hunting to support humans in security operations centres; and
- using advanced AI models or agentic AI for autonomous cyber and ICT risk management, including automating aspects of vulnerability management (e.g. software updating and patching), security testing including source code reviews, penetration

---

83 OWASP Machine Learning (ML) Security Top 10.

84 OWASP Top 10 for Agentic Applications for 2026; OWASP Top 10 for Large Language Model Applications .

85 https://atlas.mitre.org/.

86 NIST (2025).

87 MAS (2025a).

testing, and adversarial simulation. Cyber security AI agents could even be used to combat agentic AI cyber-attacks.[88]

Incorporate AI cyber and ICT risk scenarios in tests and exercises

Financial institutions incorporate AI cyber and ICT risks in their scenario testing, other forms of cyber and ICT testing (e.g. vulnerability scans) and exercises, especially for high materiality or risk use cases.[89] Scenario testing allows financial institutions to assess and strengthen their ability to respond and recover from disruption, with a view to ensuring that their critical operations can remain within agreed tolerance for disruption in severe but plausible scenarios.

Case study: UK Cross-Market Operational Resilience Group

Some authorities and industry groups maintain scenario libraries to assist financial institutions in their scenario testing. Some have started including AI and cyber ICT scenarios in these libraries.

The Cross-Market Operational Resilience Group (CMORG), a public-private collaboration forum in the UK, maintains a Dynamic Scenario Library (DSL) which provides a structured, sector-wide catalogue of severe but plausible scenarios to support operational resilience planning and testing across the UK financial sector.[90]

The 2026 edition of the DSL introduced AI and cyber ICT scenarios. One scenario involved a threat actor using GenAI and agentic AI to bypass a financial institutions' internal controls and create staff accounts, which it then utilises to conduct criminal activities at scale (e.g. data theft, unauthorised transactions and fraud).

Exercises, such as tabletop, sector-wide, and simulation exercises, are another valuable way to assess the ability of financial institutions and the financial sector as a whole (including third-party service providers) to coordinate and respond collectively to incidents.[91]

Financial institutions incorporate AI cyber and ICT risks in other types of testing. For example, 'red teaming' can test not just the performance of GenAI but also the vulnerability of AI models or systems against adversarial attacks such as prompt injection and model inversion, and the effectiveness of associated controls.

Controlled environments are also used to test the capabilities and risks of emerging forms of AI.[92] For example, they could help monitor and understand the behaviour of AI agents before deployment into production.

There might be potential to incorporate AI scenarios and tools in threat-led penetration testing (TLPT) in future. However, further work is needed to understand the ethical, legal, and practical implications of doing so.

---

88 There are currently fewer AI-powered tools for incident recovery, which may imply that this might be an area for future development.

89 IAIS (2026) sets out the important role for scenario analysis in maintaining operational resilience.

90 See CMORG website.

91 See FSB (2023b) pp.40 and G7 Cyber Experts Group (2020).

92 See, for example, AI Lab | FCA.

Collaborate and share information

Financial institutions establish or use mechanisms to securely share information on AI incidents, near-misses, deteriorating or unusual performance, risks, lessons learnt, threat intelligence and good practices. Such mechanisms include Information Sharing and Analysis Centres (ISACs).[93] Over time, proactive information-sharing can help to highlight additional sound and bad practices.

3.9. Third-party risk management

Sound Practice 12 (Third-party AI risk management): Financial institutions appropriately manage risks from AI third-party use with a focus on performance, transparency, data quality, supply chain and concentration risks, and business continuity.

Leverage and adapt existing regulation and guidance on third-party risk management

Financial institutions leverage existing technology-neutral publications on third-party risk management issued by the FSB,[94] SSBs,[95] international bodies and authorities in several jurisdictions[96] as a foundation for managing third-party AI risks, adapting them as appropriate to consider AI-specific risks. In particular, financial institutions:

- conduct appropriate due diligence on prospective third-party AI providers;
- select an appropriate third-party AI provider considering business needs, risk considerations as well as legal and regulatory obligations;
- include terms and conditions in contracts with third-party AI providers that enable: effective oversight, including monitoring;[97] access, audit, and information rights; business continuity and contingency planning; provisions on data access, quality and security, including responsibility for updating data and information on jurisdictions where it will be processed; and termination;[98] and
- appropriately oversee the provision of third-party AI services. Financial institutions may (where appropriate) use independent auditors and pooled audits.

---

93 See FS-ISAC | Advancing cybersecurity for the global financial system.

94 See FSB (2023b).

95 See, for example, BCBS (2025).

96 For example, see BOE/PRA (2021) and FRB et al (2023).

97 This may include monitoring the embedding of AI by third-party service providers of non-AI critical services into these services.

98 See IAIS (2026).

Use AI transparency and assurance tools

Financial institutions enhance third-party AI transparency through various channels, including but not limited to legal and regulatory disclosures, AI model and system cards, as well as certifications and standards.

Financial institutions leverage disclosure obligations, where available. For example, in some jurisdictions, AI developers are required to publish or make available to organisations deploying their AI models, systems or applications information on: intended, prohibited, and restricted uses; data; capabilities, characteristics and limitations (including on explainability); feasibility of human oversight; and cyber and ICT security. In some cases, third-party AI providers may also be subject to incident reporting obligations.

Financial institutions' contractual arrangements with third-party AI providers provide for the use of AI transparency and assurance tools. For example, AI model and system cards contain information about AI models' or systems' data, functionality, use cases, potential biases, errors, and limitations.[99]

AI certifications and standards provide assurance on the governance of third-party AI providers. A third-party audit is required to obtain certain AI certifications such as ISO/IEC42001, which provides a layer of independent validation.[100]

The AI transparency tools referred to above can help balance financial institutions' need for appropriate information from third-party AI providers, with the need of third-party AI providers to retain some commercially sensitive proprietary information.

A current limitation of AI transparency tools is that they are generally not tailored to the financial sector and may not cater to financial institution-specific risks and regulatory requirements. For example, model or system cards may provide information of how an AI model, system, or agent was built and its capabilities, but may not provide guidance on how financial institutions can integrate that model, system or agent safely into their environment (including constraints and guardrails).

Financial institutions could collaborate with authorities and third-party AI providers to develop targeted AI transparency and assurance tools (e.g. financial sector-specific model or system cards). In this regard, there are examples of these tools in some jurisdictions, such as a shared AI responsibility model.[101]

Consider compensating controls

If a financial institution cannot obtain sufficient assurance and information from third-party AI providers, it implements compensating controls. Although sub-optimal, these controls provide risk mitigation and may include: additional pre-deployment testing; limiting the initial use of the third-party AI to low materiality and risk use cases; and more intense human oversight (at least

---

[99] For example, see OECD Model Card Regulatory Check.

[100] See ISO/IEC 42001:2023 - AI management systems.

[101] For example, CMORG AI Shared Responsibility Model.

temporarily) such as enhanced monitoring, benchmarking, or stricter approval for certain actions. Contractual tools such as warranties and indemnities from third-party AI providers can also help compensate for limited transparency. Where there are no satisfactory compensating controls available to the financial institution that can bring relevant risks within agreed appetite, it may decide not to use the third-party AI provider's model or system.

Monitor and manage concentration risks

Financial institutions monitor and manage concentration risks at the individual financial institution (including group level). In particular, they identify and document their exposures to third-party AI providers and the AI supply chain, using AI inventories, third-party registers etc., and seek assurances on how third-party AI providers manage supply chain risks.[102]

Financial institutions using third-party AI for high materiality and high risk use cases incorporate concentration risk considerations into their business continuity planning and exit strategies. In particular, financial institutions develop robust exit strategies which may include alternative third-party AI providers or the ability to continue providing a given AI-enabled service manually in the event of disruption or termination. Ensuring data access and portability can enable substitutability while minimising the risk of disruption for financial institutions.

Financial institutions also explore the use of joint scenario tests and sector exercises with third-party AI providers as a means to mitigate potential sector-wide risks.[103]

Case study: Financial institutions' management of third-party AI risks

Some internationally active banks have uplifted their procurement and third-party risk management frameworks to address AI-specific risks. Common challenges these banks seek to address include information gaps (financial institutions often have limited visibility into third-party AI systems, such as training data, model details, or how their data is used by the vendor) and accountability gaps (financial institutions remain accountable for AI outcomes even when the AI was not built in-house).

These banks have developed internal frameworks for categorising common third-party AI deployment patterns, including: onboarding with customisation (e.g. fine-tuning); onboarding without customisation (including SaaS); and embedded AI within a wider product. They have recognised that "open source" AI spans a spectrum, from fully open-source models to open-weights models and models with restricted-use licensing, with different risk and due diligence implications for each.

A key principle adopted is to subject third-party AI to the same lifecycle governance as in-house AI, with additional steps where vendor visibility is limited. The practices include:

- requesting standardised vendor disclosures (often via 'AI Cards") proportionate to risk;
- deciding how to respond when disclosures are incomplete, using mitigants such as indemnities, external attestations, or compensatory testing;
- conducting AI-specific due diligence during procurement and on an ongoing basis;
- monitoring for unapproved new AI features via vendor release notes;

[102] For example, by asking a third-party AI provider how it has tested its ability to deal with severe shortages of essential hardware or energy.

[103] In some jurisdictions third-party service providers designated as critical service third-party providers by authorities are subject `to direct oversight by authorities. This can include engagement in joint incident management exercises with financial institutions.

- scheduling periodic reassessments and including contract notification clauses;
- updating contracts to cover AI-specific issues such as intellectual property, data retention, cybersecurity, and change notification; and
- ensuring appropriate legal, technical, and risk expertise is involved throughout.

Annex 1: List of Sound Practices for financial institutions' responsible AI adoption

Sound Practice 1 (Strategic direction and oversight): The board and senior management align AI adoption and governance with the financial institution's business model, risk appetite, and strategy.

Sound Practice 2 (Governance and accountability): Financial institutions define clear roles and responsibilities, and maintain an appropriate governance framework to enable responsible AI adoption.

Sound Practice 3 (Incorporation of AI risks into risk management framework): Financial institutions adopt risk management frameworks that effectively address AI risks, and include processes for AI identification, documentation, as well as materiality and risk assessment.

Sound Practice 4 (Organisational adaptability): Financial institutions learn, adapt and adjust their oversight, governance, risk management practices, and capabilities as AI evolves.

Sound Practice 5 (Materiality and risk assessment): Financial institutions implement an effective and systematic approach to assess the materiality and risk of AI use cases at the inception stage and thereafter.

Sound Practice 6 (Selection): When selecting AI models or systems, financial institutions consider business objectives, operational and technical needs, as well as the materiality and risks of AI use cases.

Sound Practice 7 (Data governance): Financial institutions establish appropriate data governance to maintain data that is fit-for-purpose for training, testing, and using AI (i.e. accurate, complete, consistent, reliable, secure).

Sound Practice 8 (Explainability and transparency): Financial institutions understand differences in the explainability of various types of AI. If appropriate and feasible, financial institutions adopt more explainable AI, or consider compensating controls. Financial institutions also provide appropriate transparency tailored to different stakeholders.

Sound Practice 9 (Performance management): Financial institutions evaluate the performance of AI use cases proportionately to their materiality and risk, including through performance assessments, testing, and ongoing monitoring**.**

Sound Practice 10 (Human oversight): Financial institutions implement appropriate and effective human oversight relevant to the materiality, risk, autonomy, complexity, and explainability of different AI use cases.

Sound Practice 11 (Cyber and ICT risk management): Financial institutions manage AI-related cyber and ICT risks, including, where appropriate, by incorporating AI cyber and ICT risk scenarios into tests and exercises, sharing relevant information, and using AI tools in cyber and ICT risk management.

Sound Practice 12 (Third-party AI risk management): Financial institutions appropriately manage risks from AI third-party use with a focus on performance, transparency, data quality, supply chain and concentration risks, and business continuity.

Annex 2: Examples of explainability for different AI approaches

Traditional AI (machine-learning (ML))

A financial institution engaging in retail lending is looking to employ a traditional ML approach to augment earnings and reduce losses. In initial testing, that ML approach has shown to exhibit somewhat better performance, across several measures, than the non-AI model used to date. However, the ML approach, which uses hundreds of features and involves complex interaction terms among them, presents considerable explainability challenges. These include a lack of understanding of how the ML translates inputs into outputs, which features are included in processing runs, and how features are weighted. This lack of explainability creates some questions about the conceptual soundness of the ML approach. Furthermore, given that the ML is issued for retail lending, there are concerns about whether the ML approach might violate relevant consumer laws or regulations.

In response, the financial institution has chosen to include its consumer compliance experts into the process before the ML is deployed, to review both the input data and the output to monitor for potential discrimination or other bias against consumers. Those consumer experts identify some features that while not problematic themselves, have shown to be correlated with protected features (i.e. ones not to be used with consumers) - which are then dropped from the dataset. In addition, given the lack of ability to assess conceptual soundness in full, the financial institution places higher standards for performance and plans to conduct close and frequent monitoring of the ML approach over time. The financial institution also deploys some post hoc explainers to better understand how the ML processes inputs to produce outputs (while recognising the limitations of such explainers, for example if features are correlated with one another). Finally, the financial institution intends to use the ML approach in tandem with the previous non-AI model to benchmark the two sets of output and ‘triangulate’ around the most accurate outcomes.

GenAI

A large insurance company has decided to deploy an LLM-based AI system in its claims management unit to help claims adjustors with initial processing of customer claim submissions and other tasks. The LLM is a vendor-developed foundation model, and the vendor itself has admitted that it does not fully understand how the LLM arrives at outputs for a particular input. The insurance company has some concerns about the potential for hallucinations from the LLM and the inability to determine how they arise and how to minimise them. In response, the insurance company decided to employ retrieval augmented generation (RAG) based on its existing corpus of insurance information to help provide grounding information from which the LLM draws and thus help narrow the range of outputs. In addition, the insurance company employs a reasoning feature with the LLM, to prompt it to demonstrate the steps by which it arrived at its output. However, the insurance company is also mindful of recent research showing that some LLM reasoning may not be accurate, i.e. the LLM may be describing one set of steps to the operator/user but may actually be employing a different set of steps in its operation. Given this level of uncertainty, the insurance company plans to deploy the LLM-based AI system in a controlled manner, with only partial deployment among some claims adjustors and only with qualified human oversight of the output before it is used. The insurance company does not allow any LLM output to be shared directly with customers. Additional implementation of the LLM will

only occur after further testing, including stability tests to confirm the LLM operates within a certain range. Furthermore, operators/users are given specific instructions for prompts, to help narrow the range of potential output and minimise hallucinations. Employees using the LLM are given proper training for its use, including clear indications of the LLM's limitations and parameters for its use.

Agentic AI

A financial institution has decided to start deploying an AI agent for some operational tasks. It is pairing the agent with an LLM to provide information about steps to take. While the agent's supposed range of tasks involve a discrete, concise set of steps, the pairing with the LLM introduces a stochastic element to the agent's activities, given the probabilistic nature of the LLM. This means that there could be some uncertainty about the steps to be taken by the agent should the LLM hallucinate or otherwise generate erroneous or problematic output - which could, in turn, cause the agent to duplicate tasks, conduct them in the wrong order, or attempt to operate outside its designated range of activity. Controls deployed for use of the agent include prohibiting its use in any financial transactions until further testing is conducted, including some intentional sensitivity tests by validation/reviewers to try to convince the agent to act beyond its established parameters. Finally, the financial institution conducts monitoring of the agent on a sampled basis, reviewing selected activities and tasks on a regular basis to confirm performance.

Annex 3: Examples of AI performance testing

Example A

A financial institution has developed a new ML fraud detection and prevention model for its credit cards business. Initial tests demonstrate that this ML model outperforms the previous non-AI model used, including to reduce both false positive and false negatives. The financial institution continues further testing, including sensitivity tests, to determine the limits of the model and where it may not function as well. The financial institution also continues to collect information about realised outcomes to expand testing out of sample - especially to check whether the model may be overfit. Furthermore, the financial institution conducts ongoing testing for bias, including potential discrimination against certain consumers - especially since the ML model is quite complex and has some explainability challenges. The ongoing monitoring plan also includes checks for changes in fraud behaviour and activities to detect whether the established relationships in the ML model continue to apply.

Example B

A financial institution has deployed an LLM-based AI system to help with document retrieval and summarisation, including to evaluate prospectuses and draft outlines of public financial statements. The financial institution recognises that evaluating the blocks of text produced by the LLM can be challenging, especially since there is not a 'ground truth' against which to compare the output. The financial institution tasks certain employees to review the output of the LLM to determine the faithfulness to the underlying documents and check for hallucinations. Subject-matter experts are also tasked with evaluating the LLM's facility with more technical topics. Further testing includes checks for stability of output, such as running the same query multiple times and several hours/days apart to confirm that output is similar over time. Based on such testing, the financial institution determines that the LLM could continue to be used, but that a human has to review all output for accuracy before it is distributed within the financial institution. Also, staff using the LLM are given specific instructions to submit prompts and queries within certain parameters, to help narrow the range of potential output. Additional testing by subject-matter experts is conducted on a sampled basis.

Example C

As a first step to experiment with AI agents, a financial institution has started to deploy a few for simple operational tasks. The financial institution conducts testing on the full range of tasks the agent is permitted to execute, running repeated tests at different times to try to identify deviation in performance. In addition, some staff reviewing the AI agent make attempts to force it to take actions that are not permitted, for example to see if it would try to access financial accounts. As the financial institution works with the agents, it collects information on performance, including from users. Use of the agents is limited to certain areas that do not involve financial accounts, areas requiring legal or regulatory compliance, or interactions with customers. Furthermore, one could direct the agent to follow regular checklists to confirm operational adequacy - if the agent does not follow the checklist or is unable to execute its steps, staff investigate further. Staff continue to monitor agent performance on a sampled basis.

Glossary

This glossary provides a (non-exhaustive) list of terms and definitions used in this report. Many of these terms are commonly used in the field of AI and have been compiled based on their general understanding and usage within the community. The definitions draw on the previous FSB reports on AI[104] and other relevant works,[105] and also on the extensive work that has previously been done or is underway by other groups in developing lexicons and glossaries related to financial institutions' AI adoption.

Agentic AI: Systems designed to autonomously perform complex and extended tasks, often making decisions and taking actions with limited human oversight. These systems are distinct from GenAI but may incorporate generative capabilities to enhance content generation and decision making.

Artificial Intelligence (AI): AI is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

Criticality: The importance of a service, system, or provider to the operations of a financial institution.

Data drift: When data used to train, validate, test and run an AI model or system ceases to be appropriate. For example, because it no longer applies to current circumstances or when the relationships among features established by the AI methodology no longer hold.

Data/model poisoning: An attack where malicious actors manipulate training data or model parameters to introduce errors or biases into an AI model's outputs. This can compromise the integrity, reliability, and security of AI systems.

Deep learning: A form of machine learning that uses algorithms that work in 'layers', inspired by the structure and function of the brain. Deep learning algorithms can be used for supervised, unsupervised, or reinforcement learning.

Disinformation: The deliberate creation and dissemination of false or misleading information, with the intent to deceive or manipulate public opinion, market behaviour, or decision-making.

Explainability: The ability of an AI model or system to provide clear and interpretable outputs or decisions, specifically the extent to which humans can understand how an AI model or system generates outputs from inputs. This includes: (i) the factors, logic, or processes that contribute to outputs; (ii) the relative importance of input variables or features; and (iii) how methodological relationships vary across context or use cases. Explainability stems from characteristics in AI models or systems. It is not binary but a continuum.

Foundation models: An umbrella term referring to a diversity of models that are usually trained by applying deep learning to massive quantities of data, such as text and images. Because the

[104] FSB (2024b) and FSB (2025b).

[105] For example, see FSB (2023a) and FSB (2023b).

expertise, time, and computing power involved in training foundation models from scratch are typically prohibitive for most non-specialist firms, these models are usually pre-trained and shared with end-users for further use and refinement.

Generative AI (GenAI): AI that generates new content, such as text, images, and videos, often based on user prompts. GenAI is usually powered by foundation models, such as LLMs.

Large Language Models (LLMs): A type of foundation model that is trained on and designed to perform tasks with natural language. Key tasks LLMs perform include text generation, document classification, summarisation, question-and-answer, and sentiment analysis, among other tasks.

LLM hallucinations: Occur when a large language model (LLM) generates seemingly confident but inaccurate outputs, or fabricates nonsensical outputs in response to user inputs.

Machine learning: A method of designing a sequence of actions, known as algorithms, to solve a problem, which optimise automatically through experience and with limited or no human intervention.

Model drift: When the relationships between model inputs and outputs change. For example, because the assumed linkage between an independent (or input) variable and a dependent (or output) variable no longer holds.

Model risk: The potential for adverse consequences arising from decisions based on incorrect or misused models.

Open-source models: AI models where the full training code, and in some cases the training data or its composition, is made publicly available for use and modification. These models offer customisability and reduce vendor lock-in but may introduce additional risks, such as security vulnerabilities, data privacy, or data quality concerns.

Operational resilience: The ability of a financial institution to deliver its critical operations despite disruptions. This includes the capacity to prepare for, withstand, respond to, recover from, and adapt to adverse events, ensuring the continuity of critical services.

Reinforcement learning: A subset of machine learning which falls in between supervised and unsupervised learning. The algorithm is fed an unlabelled set of data, chooses an action for each data point, and receives feedback (perhaps from a human) that helps the algorithm learn. For instance, reinforcement learning can be used in robotics, game theory, and self-driving cars.

Retrieval-Augmented Generation (RAG): An AI technique that improves LLM accuracy by retrieving data from a selected corpus of sources, such as internal company databases or live data feeds, to inform its outputs.

Supervised learning: The algorithm is fed a set of 'training' data that contains labels on all of the observations. For instance, a data set of transactions may contain labels on data points identifying those that are fraudulent and those that are not fraudulent. The algorithm will 'learn' a general rule of classification that it will use to predict the labels for observations when deployed on a data set.

Supply chain: The network of entities that provide infrastructure, physical goods, services and other inputs directly or indirectly utilised for the delivery of a service to a financial institution. For the purposes of this report, the scope of supply chain is limited to the services under a third-party service relationship.

Traditional AI: A suite of computational techniques, such as longstanding machine learning methods, that pre-date recent advances, such as GenAI.

Unsupervised learning: The algorithm is asked to detect patterns in the data by identifying clusters of observations that depend on similar underlying characteristics. For example, an unsupervised machine learning algorithm could be set up to look for securities that have characteristics similar to an illiquid security that is hard to price. If it finds an appropriate cluster for the illiquid security, pricing of other securities in the cluster can be used to help price the illiquid security.

Vertical integration: The consolidation of multiple stages of a supply chain or production process within a single organisation. In the context of AI, vertical integration occurs when one company controls various layers of the AI supply chain, such as hardware, cloud infrastructure, and AI models, offering end-to-end services.

References

Australian Prudential Regulation Authority (APRA) (2026), *APRA Letter to Industry on Artificial Intelligence (AI)*, April

BIS Financial Stability Institute (BIS-FSI) (2025a), *Financial stability implications of artificial intelligence – Executive Summary*, June

BIS Financial Stability Institute (BIS-FSI) (2025b), Managing explanations: how regulators can address AI explainability, *FSI Occasional Paper*, No. 24, September

BIS (2025), *The use of artificial intelligence for policy purposes*, October

Bank of England (BOE)/ Prudential Regulation Authority (PRA) (2021), *SS2/21 Outsourcing and third party risk management*, March

BOE/PRA (2023), *SS1/23 – Model risk management principles for banks*, May

BOE (2025), *Model Risk Management AI and ML Roundtable - Current thinking and risks identified in artificial intelligence and machine learning (AI and ML) adoption*, October

BOE, Financial Conduct Authority (FCA) and HM Treasury (HMT) (2026), *BoE, FCA and HM Treasury joint statement on Frontier AI models and cyber resilience*, May

Basel Committee on Banking Supervision (BCBS) (2021a), *Principles for operational resilience*, March

BCBS (2021b), *Principles for the Sound Management of Operational Risk*, March

BCBS (2024), *Digitalisation of Finance*, May

BCBS (2025), *Principles for the sound management of third-party risk*, December

Board of Governors of the Federal Reserve System (FRB), Consumer Financial Protection Bureau, Federal Deposit Insurance Corporation (FDIC), National Credit Union Administration, and Office of the Comptroller of the Currency (OCC) (2019), *Interagency Statement on the Use of Alternative Data in Credit Underwriting*, December

FRB, FDIC, and OCC (2023), *Interagency Guidance on Third-Party Relationships*, SR 23-4, June

FRB, OCC, and FDIC (2026), *Revised Guidance on Model Risk Management*, SR26-2, April

Bowman, M W (2024), *Artificial Intelligence in the Financial System*, Remarks at the 27th Annual Symposium on Building the Financial System of the 21st Century: An Agenda for Japan and the United States, November

Breeden, S (2024), *Engaging with the machine: AI and financial stability*, Speech given at the HKMA-BIS Joint Conference on Opportunities and Challenges of Emerging Technologies in the Financial Ecosystem

Cloud Security Alliance (2025), *Agentic AI Identity Management Approach*, November

Committee on Payments and Market Infrastructures (CPMI) and Board of the International Organization of Securities Commissions (IOSCO) (2016), *Guidance on cyber resilience for financial market infrastructures*, June

European Central Bank (ECB) (2024), The rise of artificial intelligence: benefits and risks for financial stability, *Financial Stability Review*, May

ECB Banking Supervision (2025), *ECB guide to internal models*, July

Federal Financial Supervisory Authority (BaFin) (2026), *Guidance on ICT Risks in the Use of AI at Financial Entities*, January

Financial Markets Standards Board (FMSB) (2026), AI in Trading: A practitioners' view of the current landscape, *Spotlight Review*, February

Financial Services Agency (JFSA) (2021), *Principles for Model Risk Management*, November

FSB (2013), *Principles for an Effective Risk Appetite Framework*, November

FSB (2017), *Artificial intelligence and machine learning in financial services*, November

FSB (2020), *Effective Practices for Cyber Incident Response and Recovery*, October

FSB (2023a), *Cyber Lexicon: Updated in 2023*, April

FSB (2023b), *Final Report on Enhancing Third-party Risk Management and Oversight – A Toolkit for Financial Institutions and Financial Authorities*, December

FSB (2024a), *Guidance on Arrangements to Support Operational Continuity in Resolution - Revised version*, March

FSB (2024b), *The Financial Stability Implications of Artificial Intelligence*, November

FSB (2025a), *Format for Incident Reporting Exchange (FIRE), Final report*, April

FSB (2025b), *Monitoring AI Adoption and Related Vulnerabilities in the Financial Sector*, October

G7 Cyber Expert Group (2020), *G-7 Fundamental Elements of Cyber Exercise Programmes*, December

G7 Cyber Expert Group (2025), *G7 Cyber Expert Group statement on Artificial Intelligence and Cybersecurity*, September

Hlophe, N and L Mabetha (2025), *Artificial Intelligence in the South African Financial Sector*, Financial Sector Conduct Authority and Prudential Authority, November

Hong Kong Monetary Authority (HKMA) (2024a), *Research Paper on Generative Artificial Intelligence in the Financial Services Sector,* September

HKMA (2024b), *HKMA announces inaugural cohort of GenA.I. Sandbox*, Press Release, December

HKMA (2025), *Responsible Innovation with GenA.I. in the Banking Industry – Practical Insights from the Generative Artificial Intelligence Sandbox,* October

HKMA (2026), *Regulators launch GenA.I. Sandbox++ to foster A.I. innovation across financial services*, Press Release, May.

Holphe, N and L Mabetha (2025), *Artificial Intelligence in the South African Financial Sector*, Financial Sector Conduct Authority and South African Reserve Bank/Prudential Authority, November

International Association of Insurance Supervisors (IAIS) (2025) *Application Paper on the supervision of artificial intelligence*, July

IAIS (2026), *Application Paper on operational resilience objectives and toolkit - draft*, February

International Monetary Fund (IMF) (2024), Advances in Artificial Intelligence: Implications for Capital Market Activities, *Global Financial Stability Report*, Chapter 3, October

IOSCO (2025) *Artificial Intelligence in Capital Markets: Use Cases, Risks and Challenges – Consultation Report*, March

IOSCO (2026), *Supervisory Toolkit for AI Use in Capital Markets – Final Report*, May

International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) (2023) *ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system,* Edition 1

Lim, X-L, P Singh and R Stobo (2025) Regulatory Considerations Regarding Accelerated Use of AI in Securities Markets, *IMF Technical Notes & Manuals*, Volume 2025, Issue 016

Machado, P (2026), *Technology is neutral, governance is not: AI adoption in the banking sector*, Speech at KPMG RiskTech Conference, February

Monetary Authority of Singapore (MAS) (2024), *Artificial Intelligence (AI) Model Risk Management*, December

MAS (2025a), Cyber Risks Associated with Deepfakes, *Information Paper*, September

MAS (2025b), *Guidelines for Artificial Intelligence (AI) Risk Management*, November

National Cyber Security Agency for France (ANSSI) (2025), *Building trust in AI through a cyber risk-based approach*, February

National Institute of Standards and Technology (NIST) (2023), *AI Risk Management Framework (AI RMF 1.0)*, January

NIST (2025), *Adversarial Machine Learning - A Taxonomy and Terminology of Attacks and Mitigations*, March

NIST (2026), *Announcing the "AI Agent Standards Initiative" for Interoperable and Secure Innovation*, News, February

Office of the Superintendent of Financial Institutions (OSFI) (2022), *Technology and Cyber Risk Management*, Guideline, July

OSFI (2023), *Financial Industry Forum on Artificial Intelligence (FIFAI), A Canadian Perspective on Responsible AI*, April

OSFI (2025a), *FIFAI II: A Collaborative Approach to AI Threats, Opportunities, and Best Practices, Workshop 1 - Security and Cybersecurity*, July

OSFI (2025b), *Guideline E-23 – Model Risk Management (2027)*, September

OSFI (2026), *Frontier Artificial Intelligence: Implications for Technology, Cyber Security, and Operational Resilience*, Technology Risk Bulletin, April

Organisation for Economic Co-operation and Development (OECD) (2024), *Regulatory Approaches to Artificial Intelligence in Finance*, September

OECD (2026), *Supervision of artificial intelligence in finance*, January

Swiss Financial Market Supervisory Authority (FINMA) (2024), *Governance and risk management when using artificial intelligence*, FINMA Guidance 08/2024, December

The MindForge Consortium (2026), *AI Risk Management: Operationalisation Handbook*, January

UK Department for Science, Innovation & Technology (2025a), *Code of Practice for the Cyber Security of AI*, Guidance, January

UK Department for Science, Innovation & Technology (2025b), *Implementation Guide for the AI Cyber Security Code of Practice*, January