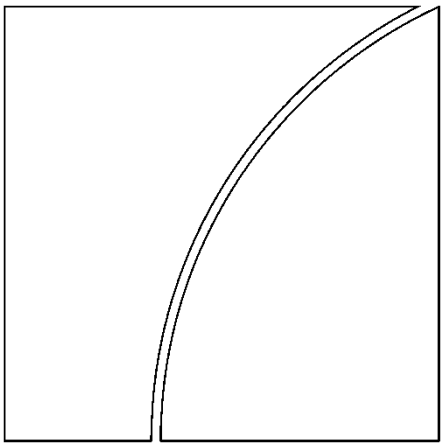
BIS

Consultative Group on Risk Management



Governance of AI adoption in central banks

January 2025

BIS Representative Office for the Americas

This publication is available on the BIS website (www.bis.org).

ISBN 978-92-9259-831-0 (online)

Table of contents

Foreword

Artificial intelligence (AI) presents huge opportunities for central banks. At the same time, its adoption entails complex risk management challenges. The use cases for AI span a broad range of critical functions of a central bank including data analysis, research, economic forecasting, payments, supervision and banknote production. The adoption of AI presents new risks and can amplify existing ones. The potential risks are wide-ranging and include those around data security and confidentiality, risks inherent to AI models (eg "hallucinations") and, importantly, reputational risks. The potential risk exposure for central banks can be significant, owing to the criticality and sensitivity of the data they handle as well as their central role in financial markets.

This report on the governance of AI adoption in central banks provides guidance on the implementation of AI at central banks and proposes a governance and risk management framework. A comprehensive risk management strategy can leverage existing risk management models and processes, in particular the well established three lines of defence model. In incorporating the specific issues around AI and its use cases, risk managers at central banks can make use of the frameworks proposed by a number of international bodies. A good governance framework is key for adopting AI. The report proposes an adaptive governance framework and recommends ten practical actions that central banks may want to undertake as part of their journey in adopting AI.

The report is the outcome of work conducted by Bank for International Settlements (BIS) member central banks in the Americas within the Consultative Group on Risk Management (CGRM), which brings together representatives of the central banks of Brazil, Canada, Chile, Colombia, Mexico, Peru and the United States. The Artificial Intelligence Task Force that prepared this report was co-led by **Alejandro de los Santos** from the Bank of Mexico and **Angela O'Connor** from the BIS. The BIS Americas Office acted as the secretariat.

Claudia Álvarez Toca
Chair of the CGRM
Bank of Mexico

Alexandre Tombini
Chief Representative for the Americas
Bank for International Settlements

Executive summary

The adoption of artificial intelligence (AI) technologies in the financial sector may usher in a transformative era for financial services, offering unprecedented opportunities for innovation, efficiency and customer focus. As with any groundbreaking new technology, AI also introduces complex risk management challenges for banks and central banks. Risks from the use of AI include but are not limited to operational, information security, privacy and cyber security risks; information and communication technology (ICT) risks; third-party risks such as external dependency; and risks inherent in AI models (eg "hallucinations"). The materialisation of these risks can have significant reputational and financial impacts.

The identification and management of such risks can be complex owing to their variety and interactions between different types. The analysis and management of these risks should therefore be approached with a holistic view that incorporates a wide array of expertise to consider the full range of risks and complex interactions.

The Bank for International Settlements Consultative Group on Risk Management (CGRM) set up a task force to provide guidance to central banks on how AI is being used in different functions, and how to organise and govern risk management related to AI adoption. This report provides specific suggestions on how central banks can identify, analyse, report and manage risk associated with the adoption of AI models and tools in their organisations. The suggestions are based on risk scenarios and risk management practices developed in the central banking community and the financial sector at large, such as the three lines of defence model. The report does not focus on the impact of AI on financial stability or the broader economy, which are covered by other forums.

The report argues that central banks need to find a balance between fostering innovation using AI and mitigating the different risks that this technology may generate. Good governance schemes for the adoption of AI in the organisation, with a holistic view beyond technology, might help to achieve such a balance.

The report is organised as follows. Section 1 begins with a brief overview of AI models, highlights use cases from central bank publications and complements the discussion with answers provided by members of the CGRM task force to a questionnaire on AI usage. Some of the benefits and uses identified by the group include the automation of processes, analysis of large data sets and solving complex problems. Central banks adopt AI to enhance efficiency, improve operational robustness and inform decision-making in different areas of the organisation. This includes core functions such as economic forecasting, payments, supervision and banknote production. Central banks are also exploring the use of AI to provide customer and corporate services, for instance by using chatbots to answer enquires from regulated entities or assist their own researchers. These applications demonstrate the potential of AI to address complex challenges and support central banking operations.

Section 2 describes, from a holistic perspective, risks related to the adoption of AI for central banks, covering both new risks raised by AI and existing risks that are amplified by its use. The section introduces a taxonomy into strategic; operational (legal, compliance, process, people and capacity); information security and cyber security; ICT; third party; AI model; environmental, ethical and social; and reputational risks. It also highlights some considerations about generative artificial intelligence (GAI) and the unintended consequences of its adoption, which can result in significant risk exposure for central banks, if the associated risks are not well managed.

Section 3 provides some guidance for the implementation of AI models in central banks. The report suggests a comprehensive risk management strategy that leverages existing models and processes. It suggests updating the three lines of defence model with some considerations specific to the use of AI. The section also suggests a process that central banks can employ to identify and analyse use cases,

initiatives or projects related to AI before its deployment. This process can be based on the frameworks suggested by various standard-setting bodies, especially those related to information security and cyber security. If central banks decide to rely on third-party AI services, tools, components or algorithms, a mature third-party risk management model is essential. The use of GAI poses additional risks, such as the interpretability of the results of such models, possible biases, limitations and robustness. This means that GAI requires more human supervision and checks for legal implications than other technologies, as well as other specific controls based on a holistic and multidisciplinary perspective.

Section 4 proposes a governance and risk management framework for AI based on concepts developed in previous sections. Good AI governance is important not only for complying with national strategies, laws or regulations but also for ensuring the alignment of AI uses with the organisation's strategy and objectives. Effective AI governance boosts efficiency and stimulates innovation while also identifying and mitigating associated risks. The proposal presented in this section aims to balance both the risks and opportunities offered by AI.

As an initial guide, the section presents current industry standards and suggestions for risk management and governance frameworks that reflect key features and concerns already identified as pertaining to the use of AI. These frameworks share common elements that can help central banks to implement AI processes that are ethical, secure, transparent, explainable, reliable, responsible and comply with data privacy. The safe and proper usage of AI across the central bank functions may demand changes to existing risk management and governance frameworks. In this sense, central banks will eventually have to consider a careful review of their current governance structures and risk management practices, while embracing AI models and tools safely and efficiently. An adaptive AI governance framework may prove helpful in this regard. Finally, this section proposes ten actions that have proven useful for central banks on their journey to adopt AI: (i) establish an interdisciplinary AI committee; (ii) define principles for responsible AI use; (iii) establish an AI framework and update existing guidance; (iv) maintain an AI tools inventory; (v) map AI tools and stakeholders; (vi) perform a detailed assessment of risks and controls; (vii) perform regular monitoring; (viii) report anomalies and incidents; (ix) develop and improve workforce skills; and (x) perform ongoing reviews and adaptations to the framework.

1 Use of AI

The concept of artificial intelligence (AI) can be broad and mean different things in different contexts. A number of different sources were considered as part of the process of arriving at a common definition of AI for this report (Appendix 1). The definition adopted here is based on the widely used definition from the Organisation for Economic Co-operation and Development (OECD), according to which:

> "An AI system 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. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."[1]

AI applications rely on different models, which perform operations based on rules, code, knowledge and data (inputs) to learn and generate the required outputs. According to the definition of AI above, the outputs can be predictions, content, recommendations or decisions. Some of these models are

[1] OECD (2024a).

able to learn or act without human involvement (autonomy) and to continue evolving after they are deployed (adaptiveness).[2] Box A describes different AI models.

Box A

AI models

Machine learning (ML) is a subfield of artificial intelligence (AI). This technique gives systems the ability to "learn" from the input data without being explicitly programmed to do so (ISO/IEC (2022)). ML models are widely used to identify patterns that are not otherwise readily evident, mainly through statistical learning algorithms. ML models are capable of improving (Sharma et al (2021)).

Deep learning is a subfield of ML. Deep learning models learn from large and complex inputs, using neural networks with multiple neurons and hidden layers (Amazon AWS (2024)). This architecture allows deep learning models to be more autonomous, by automatically identifying the most relevant features from the data. This allows for deep learning models to be generalised, meaning that they can be used with different types of inputs or in different applications (Sharma et al (2021)).

Generative AI (GAI) models are capable of generating original content in response to user-supplied inputs called prompts. The outputs of these models can be text, images, audio, video or code. General awareness of AI increased with the rise of GAI technologies in 2022, beginning with image generation and later with chat-based AI interfaces.

Natural language processing (NLP) models enable machines to understand natural language (linguistics) and identify meaning and context. At the same time, these models can also generate natural text, such as phrases, sentences or paragraphs with meaning (Khurana et al (2023)).

Large language models (LLMs) are a type of NLP trained on a very large number of parameters or quantity of data using deep learning. Generative LLMs understand how languages work based on very large volumes of text input and can themselves generate text based on this training (Chiarello et al (2024)).

Generative pre-trained transformers (GPTs) are a type of LLM that are able to consider the context and relationships between words in entire sentences. They are capable of executing a wide range of language-related tasks (Yenduri et al (2024)).

1.1 AI benefits for central banks and use cases

Central banks have been exploring AI applications, developing proof of concept solutions and, in some cases, deploying AI applications.

AI can create opportunities by supporting human activities with the following benefits:

- **Automation** of business processes, optimising the use of resources and time, improving the efficiency of repetitive or highly manual tasks and thereby increasing productivity.
- Swift **analysis** of large volumes of data enabling improvements in decision-making.
- **Execution** of processes that require the involvement of many people, allowing employees to undertake other productive tasks.

[2] OECD (2024b).

- **Solving** complex problems through the analysis of large volumes of data and the recognition of patterns that, with other methods, might go unnoticed or require significant time to identify.
- **Innovation** fostered by the use of AI technologies, which can be applied in different business processes and sectors including central banks.

AI use cases relevant to central banks are set out below. These use cases were identified from papers authored within the central banking community and from responses to a questionnaire from Bank for International Settlements Consultative Group on Risk Management (CGRM) member central banks. In the former case, the report includes the corresponding references. The categories highlight the main themes identified during the report's development and are not exhaustive of all potential AI applications within central banks. Appendix 2 provides a summary of these use cases.

Use case 1: Economic analysis, forecasting and policy analysis

The ability of AI to analyse large data sets and capture nonlinear relationships makes it highly effective for economic analysis, forecasting and policy analysis.

- **GDP nowcasting:** central banks and organisations use machine learning algorithms to nowcast GDP growth (Richardson et al (2021) and Dauphin et al (2022)).
- **Inflation forecasting:** machine learning has long been part of some organisations' forecasting procedures (Benford (2024), Buckmann et al (2023), Joseph et al (2022), Chakraborty and Joseph (2017) and Burgess et al (2013)).
- **Textual analysis:** natural language processing (NLP) and large language models (LLMs) have helped to enhance economic analysis and forecasting through textual analysis from newspapers, reports and social media articles (Chen et al (2023), Denes et al (2022), Gascon and Werner (2022) and Kalamara et al (2020)). Some central banks use NLP to analyse survey responses on economic indicators and perform sentiment analysis on official publications.
- **Policy analysis and implementation:** some central banks use AI to develop models that enhance policy processes. These models help to analyse and assess policy implications, investment impacts, and impacts on the financial sector and the macroeconomy.

Use case 2: Payment systems

- **Payment system enhancements:** AI can detect abnormal transactions, strengthening the functioning of payment systems (BIS (2024)). For example, using synthetic transaction data, Phase 1 of the BIS Innovation Hub's Project Aurora showed that machine learning algorithms could detect money laundering networks more effectively than traditional methods (BISIH (2023)). Some central banks employ artificial neural networks to identify anomalies in payment patterns (Rubio et al (2020)).
- **Research on payment systems:** AI can be used to research standards and innovations that can enhance the resilience and robustness of payment systems.

Use case 3: Regulation

- **Regulatory complexity:** some central banks use NLP to calculate complexity measures of prudential and banking regulations (Amadxarif et al (2021)).
- **Impact analysis:** some central banks use AI to analyse the impact of bills and regulations through web scraping, classification algorithms and similarity analysis. This provides timely alerts about legal provisions that may require the attention of the central bank.

Use case 4: Supervision and oversight

- **Supervision activities:** some central banks have used machine learning, NLP and generative artificial intelligence (GAI) to extract insights from supervisory agents' information or to provide insights for supervisors (Benford (2024), McCaul (2024) and ECB (2023)).
- **Monitoring financial institutions:** AI tools help assess and monitor financial institutions, covering aspects like credit risk, exposure and portfolio risks (eg Devys and von Kalckreuth (2022)).

Use case 5: Banknote production and distribution

- **Banknote and coin production:** central banks have used deep learning techniques to enhance the efficiency of printing operations and banknote quality control (eg Kerdsri and Treeratpituk (2022)).
- **Forecasting banknote demand and distribution needs:** AI is used to forecast cash flow and plan production and distribution schedules for banknotes and coins.

Use case 6: Anomaly detection

AI can analyse vast data sets to detect unusual patterns quickly, flag anomalies for investigation and support decision-making. Central banks can use AI for:

- **Data quality enhancements:** in 2022, the BIS published a compendium on the application of data science and machine learning in central banks, including AI use cases for anomaly detection to improve data quality (Araujo et al (2022)). Some other organisations and central banks have employed machine learning to detect anomalies or incorrect reporting (Accornero and Boscariol (2022), Cagala et al (2022), Faria da Costa et al (2022), Haghighi et al (2022) and Jiménez and Serrano (2022)). Some other central banks use AI to detect anomalies in price data and account balances – identifying outliers and unusual changes.
- **Cyber security enhancements:** the BIS Innovation Hub Nordic Centre is working on Project Raven, which proposes a solution for central banks and regulatory authorities to assess cyber security maturity in their financial sectors using AI (BISIH (2024)).

Use case 7: Risk assessment

AI excels at identifying patterns, analysing granular data and capturing nonlinear relationships for risk assessment (Nistor (2023)). Key examples include:

- **Early warning systems:** AI can analyse diverse data sources to detect early signs of potential risks or financial crises (Bluwstein et al (2020)).
- **Stress testing:** AI can generate synthetic data for scenario analysis and enhance stress testing with complex simulations (Petropoulos et al (2019)).
- **Risk event analysis:** NLP is used to categorise risk events, identify trends and extract actionable insights.

Use case 8: Customer and corporate services

- **Research assistance:** LLMs assist with summarising documents, providing coding support, preparing slides, drafting reports and analysing frequently asked questions from the public (Benford (2024)).

- **Public services:** some central banks are using AI to implement chatbots that answer public questions.
- **Corporate services:** some central banks use AI-powered chatbots to assist employees with questions about processes, HR benefits and policies (Benford (2024)).
- **Information technology:** central banks use AI to enhance code development processes and database management.
- **Communications:** some central banks are exploring the use of AI for sentiment analysis on publications to ensure appropriate wording. AI also aids in producing official speeches with a consistent institutional style. Additionally, central banks use AI to support communication processes through document translation, grammar refinement and presentation transcription.

2 Risks and their impact

While there are significant benefits to the use of AI, its use also entails risks that must be identified, analysed and mitigated. The following subsections outline the main risks associated with the adoption of AI, provide examples of how to classify AI risks and propose methods for mitigating these risks.

2.1 Risks associated with the adoption of AI

The use of AI can introduce new risks and vulnerabilities and amplify existing risks. Like other emerging technologies that access and process large volumes of data, AI carries a risk for users. The level of risk will be determined by the access granted to information and data used by AI models and tools.

Below is a proposed classification of AI risks for central banks. Appendix 3 includes a summary table of risks associated with the adoption of AI.

Proposed classification of AI risks for central banks Table 1

Type of risk	Description
Strategic	Absence of a clear strategy for the use of AI can lead to the application of AI in ways that negatively affect the reputation of the organisation. For example, using GAI could raise questions around bias.
Operational	AI adoption could introduce additional complexity in the delivery of internal and external services, exacerbating or amplifying a range of operational risks including: **Legal uncertainty.** AI outputs may not meet legal obligations such as copyright protection. **Compliance.** Lack of explainability or the "black box" nature of some AI models may exacerbate compliance risks due to the difficulty of explaining outcomes or providing evidence on the reasoning behind a decision. **Processes.** Failures or deficiencies in the design, implementation, or ongoing management of processes or controls when incorporating AI tools. **People.** Inadequate digital skills among staff or the lack of a solid understanding of AI models and safe usage of AI tools, which could place central banks at a strategic disadvantage or make them more vulnerable to current threats. A lack of AI knowledge, specialised training and security awareness among central bank staff can lead to greater reliance on AI providers. This dependence on third-party expertise may pose additional operational risks, such as the inappropriate use or mishandling of data by those third parties. **Data governance.** AI models are data intensive and leverage new data sources or introduce new ways to use the existing ones, thereby increasing data governance challenges. In particular, the volume, velocity, variety and quality of data may require intensified efforts for effective data governance.

	Resiliency. Increased use of AI can amplify traditional threats to business continuity such as geopolitical conflicts or wars, civil conflicts, terrorism and other illicit activities including organised crime. This could make it more difficult to prevent or respond to business disruptions. Awareness and prevention around AI-related threats will be critical to foster resiliency.
Information security, privacy and cyber security	**Information security and privacy.** GAI outputs may contain confidential data, leading to security breaches, unauthorised access or privacy violations.[3] It could produce incorrect results (integrity) or suffer disruptions (availability). The use of sensitive information or critical assets may present a risk of data leakage, data corruption, intellectual property rights violations and data privacy violations. Users can send sensitive and proprietary information as part of prompts. These data could be stored by AI tools to continue training the model and may be inadvertently leaked to other users. **Cyber security** **New entry points.** New AI systems and their integration with other central bank systems add entry points, eg through third or nth parties, for cyber attacks expanding the attack surface. Novel attacks include prompt injection, training data poisoning, and model denial of service and model theft. **Data poisoning.** The use of AI models could corrupt the original data, affecting its integrity, reliability and completeness. **Malicious activities.** AI executes tasks as requested by its users, typically without questioning the context, intentions or taking into account ethical considerations. Malicious actors or criminal groups can therefore exploit AI to conduct malicious activities, both within and beyond cyberspace, including scams, fraud, deep fake videos and disseminating misinformation. GAI tools may augment attackers' capabilities, allowing them to carry out more precise and faster attacks such as hacking, malware and phishing. LLMs are already able to find new vulnerabilities in systems and write code to successfully exploit them. AI can facilitate new types of scams and frauds, with attackers and fraudsters deploying more sophisticated techniques to target individuals and organisations, for example, deepfakes (visual and audio content that appear real but were generated with AI) using the image of central bank Governors have been reported lately.
Information and communication technology (ICT)	**ICT system risk.** ICT risks are often originated from the ICT systems themselves, including sources such as software bugs, hardware failures and inadequate system design. Challenges in ensuring compatibility between new technologies and established ICT infrastructure can exacerbate these risks, potentially leading to systems downtime, operational disruptions and reduced reliability, impacting the overall efficiency and effectiveness of financial operations. **ICT infrastructure and maintenance.** Infrastructure failure, compromised information attributes or deficient backups, which can impact business continuity, operations and resilience. **Functionality and permissions.** AI systems with excessive functionality or permissions may make decisions and undertake actions with unintended consequences. For example, an AI tool for reading documents may inadvertently be given permissions to delete documents as well. **Lack of governance.** During the implementation and maintenance of AI systems throughout their life cycle, a lack of governance might lead to inappropriate usage and introduce new vulnerabilities. AI systems might be unavailable due to non-malicious causes (eg incorrect patch or human error). Incompatibility issues between AI systems and legacy systems may use up significant resources.
Third party	Dependence on external AI models or tools developed by third-party providers have several implications for central banks: **Third-party privacy and information security risks.** If a vendor/provider has access to personal or sensitive information, there are additional risks that must be considered, such as intentional or unintentional misuse of data. Also, third-party vendors that store or transmit data through unregulated systems and networks, or process and host data in certain countries, can expose the organisation to compliance risk. **Third-party concentration risk.** Dependency on a single or a few suppliers can create significant risks. Any issues on the side of the supplier could affect the operations of organisations that depend on it. Additionally, the lack of alternatives limits negotiating power and exposes the organisation to security and compliance risks if the supplier fails to meet required standards. **Third-party cyber security risk.** One of the most relevant risks is the impact on central bank processes caused by a failure or cyber attack on these providers, which limits the response capacity of the central bank in the event of an interruption or failure.

[3] The General Data Protection Regulation (GDPR) is applicable to any company that processes the data of European Union (EU) citizens or the personal data of subjects who are in the EU or the European Economic Area.

	Inflexibility. Dependency on third-party providers means that central banks cannot modify or adjust the technologies according to the evolution of their business needs.
Model risk	**Interpretation and reliability** **Imprecise results.** The accuracy of AI models is highly dependent on the quality of data. Poor data quality[4] and poor prompting lead to poor analysis or predictions, biases and errors. In many cases, the data sources used to train AI models to generate knowledge or make decisions may be of low quality or cover general topics. Therefore, certain AI models may struggle to respond to a question or execute activities on very specific, rare or exceptional topics. There are other causes of inaccuracy in the results related to the models and data, such as overfitting,[5] data sparsity and biases. **Reasoning errors and "hallucinations".** AI tools, especially GAI, may be susceptible to reasoning errors and "hallucinations", ie perceiving patterns or objects that do not exist, creating meaningless or inaccurate results. Likewise, handling large quantities of data (big data) by AI could result in apophenia, ie "seeing patterns where they do not actually exist, simply because enormous amounts of data can offer connections that radiate in all directions" (Boyd and Crawford (2012)). Similarly, overfitting can lead to establishing spurious relationships. **Repeatability issues.** AI tools, especially GAI may not produce consistent responses over time, even when given similar inputs. **Overconfidence in generated output.** AI models and tools can generate what appears to be authoritative output, potentially weakening due diligence efforts to confirm validity. This increases the exposure to error risk in decision-making. **Transparency and accountability** **Lack of transparency.** Absence of public and detailed knowledge about the programming of AI models can lead to governance and accountability issues for those who design or operate them. This includes a lack of understanding or inappropriate controls of databases and information used to "train" the algorithms of AI applications. **Self-modification of algorithms.** GAI and ML algorithms can be programmed to automatically incorporate different or additional data beyond their initial training data. These algorithms can initiate changes to themselves, leading to the identification of new patterns or correlations with other data. However, this "evolutionary nature" can also produce outcomes not intended by their designers.
Environmental, ethical and social	**Energy consumption and carbon emissions.** The use of AI models, particularly GAI and LLMs, requires large energy consuming technologies and demand significant computational resources, consuming considerable energy and impacting the carbon footprint. Several factors affecting AI models' carbon emissions. These factors include the volume of training data, the cost of running the model, the location of the training server and its energy grid, the complexity of the training procedure and the hardware used. **Limitations in understanding ethical and social context.** When using AI, individuals interact with computer systems whose "understanding of reality" is not comparable to that of a human. Therefore, AI tools may sometimes provide inappropriate responses from an ethical or social point of view. Biases in the data used to train a model can also result in unethical, discriminating or stigmatising outcomes.[6]
Reputational	Reputational risks may be a consequence of the realisation of any of the above risks. AI-related controversies can attract intense media scrutiny, amplifying negative impact on reputation.

Source: authors' elaboration.

[4] Poor data quality includes missing historical, unlabelled, biased, incomplete, noisy, untimely, inaccurate or inadaptable data.

[5] Overfitting occurs when too many variables are included in the algorithm and it fits its training data too closely or even exactly, resulting in a model that cannot make accurate predictions from any other data.

[6] For instance, according to the literature on the ethical impact of data science, privacy is a concept that evolves alongside technological and social changes. In this regard, people currently perceive privacy differently in public spaces, such as parks and streets, compared with socially and legally designated private spaces, such as homes. Training technologies like ML, AI and GAI with data and information from multiple perspectives or "dimensions" can reduce this risk. See Mulligan et al (2016).

2.2 Broader considerations and impacts

Differentiating between risks and other impacts is important when adopting AI models and tools. Risks are identifiable challenges that can be anticipated and managed. In contrast, other considerations refer to unexpected outcomes that may emerge due to the complex and dynamic nature of AI technologies. Therefore, a comprehensive approach to AI adoption in central banking must encompass a thorough assessment of both risks, as well as detect and incorporate other considerations that are a natural part of using AI.

Examples of such other considerations include:

- **Unplanned uses –** GAI can help identify new issues in central banks' processes that need to be explored.
- **Job impact –** not all employees benefit from AI in their roles, nor is AI necessarily a solution to addressing all business challenges.
- **New activities –** staff responsible for information, processes and technological tools may need to interact given the new activities introduced by AI.
- **Complexity challenges –** the complexity of training or deploying AI models can be underestimated, as can the effort needed to integrate new tools with legacy systems, especially when the information categories of the systems differ. This can lead to delays rather than potential productivity increases. Additionally, some processes will require specially trained or personalised models, which may require significant resources for maintenance. Acquiring and retaining more specialised skills can often be more difficult than continuing existing processes that rely on analysts.
- **Project management –** AI tools are currently in constant evolution, which has made it difficult to precisely programme their implementation and the pressure to rapidly adopt AI tools may lead to rushed implementations, increasing the probability of errors or vulnerabilities.
- **Impact on technological infrastructure –** the widespread use of AI tools within central banks could affect their technological infrastructure performance. AI tools, particularly GAI, require a vast amount of energy and a lot of computing and storage resources for training and responding to user queries.
- **Dependency on AI –** if AI models produce accurate and useful outcomes, they could quickly become key elements of central banks' processes, generating a new dependency on this type of tool and introducing new costs.

Box B

Examples of malicious use and failures

Below are some examples illustrating the risks of various artificial intelligence (AI) applications:

GPT – Derner and Batistič (2023) identified six main risks when using this technology: (i) generation of fraudulent services; (ii) collection of harmful information; (iii) disclosure of private data; (iv) generation of malicious text; (v) generation of malicious code; and (vi) production of offensive content. Computer systems are often designed without explicit ethical standards and lack parameters to make value judgments.

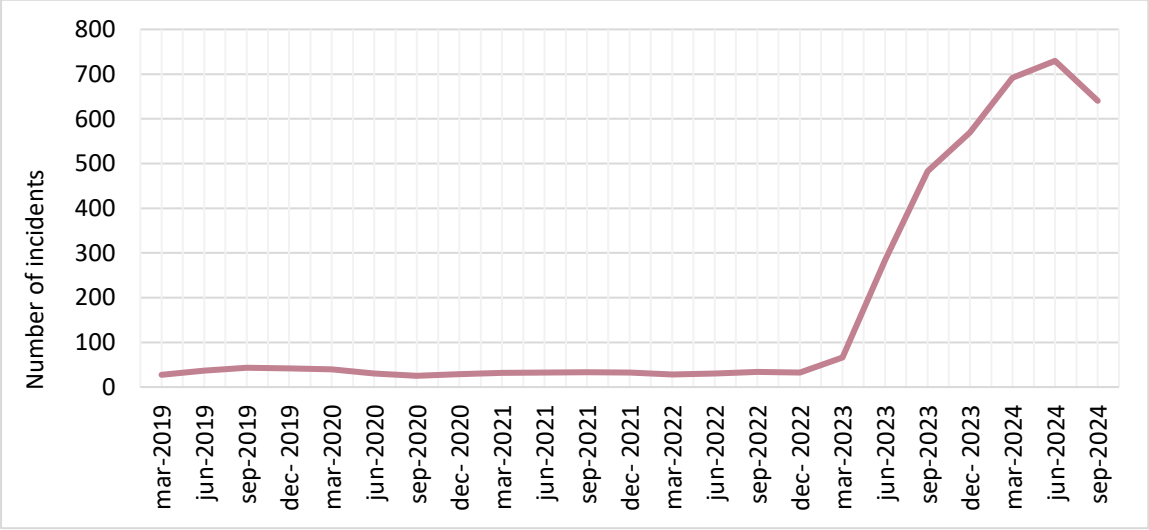
Machine learning (ML) – as ML systems base decisions on algorithms, probabilities and data, the future actions suggested by ML tools may lead to errors with impacts that are challenging to quantify (Babic et al (2021)). Due to their evolutionary nature, ML systems can create discrepancies between the data they are initially trained on and the data they subsequently process. The complexity of ML systems makes it challenging to identify the error or the logical reason behind an unforeseen result when it occurs.

Generative artificial intelligence (GAI) – according to interviews with cyber security experts, GAI may enhance organisations' cyber defence capabilities worldwide but can also be used by attackers to develop more potent cyber weapons that will be harder to defend against (Moody's (2023)). Cyber security departments may lack sufficient data to feed their GAI-based cyber defence systems as most organisations are reluctant to share technical details of cyber attacks with other companies.

According to OECD (2023), risks or incidents resulting from the use of GAI grew exponentially between December 2022 and September 2024, reaching an unprecedented peak of 730 incidents in June 2024, as illustrated in Graph B.1.

Increase in the number of registered risks or incidents resulting from the use of GAI — Graph B.1

Count



Source: adapted from OECD (2023).

Considering the above, as well as the increasingly widespread use of AI worldwide, some countries are evaluating the need to regulate its usage. For example, in the European Union, the Artificial Intelligence Act came into force on 1 August 2024 and will be fully applicable from 1 August 2026. Appendix 4 of this report provides an overview of current AI frameworks (European Parliament (2024a)).

3 Risk management for AI

3.1 Risk management strategy

When defining an AI risk management strategy, the most important consideration is that the implementation of AI models and tools is supported by a comprehensive risk management model. This strategy should: **define the AI risk profile; identify, evaluate and select AI projects; leverage and adapt governance management models that are already in place; and protect information through the full life cycle**.

1. **Define the AI risk profile**. Determine the central bank's appetite for AI-related risks, taking into account the intended strategic goals and outcomes for AI use. This can then guide risk management decisions on AI development, management and use. This also helps to comply with regulations and align usage with the central bank's resilience objectives (CGRM (2023)). By its nature the output from GAI is not precise so it may be necessary to accept certain risks in favour of innovation with AI. Central banks should carefully consider whether the output of AI or the means to manage the risks around AI are aligned with its risk appetite for innovation.

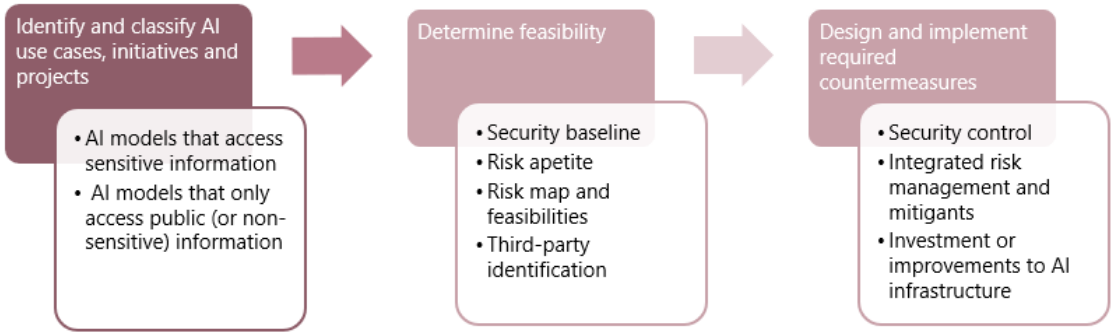
   At the initial stages of AI adoption, it is often safer to use AI only in internal processes that are not critical for central banking functions until the risks associated with such technology are properly assessed and controlled.

2. **Establish a process to identify, evaluate and select feasible AI projects**. Establish a multidisciplinary team responsible for selecting feasible AI projects that align with the strategic and innovation objectives of the central bank. The team's main objective is to evaluate the suitability of adopting certain AI models and tools, and verify whether the proposed models and tools align with the central bank's risk appetite. This process should take place before other risk management activities.

   Such a process starts with the identification and analysis of AI use cases, initiatives and projects, their classification by risk level depending on the sensitivity of the information used to train them and the information accessed and generated by the model (Graph 1). The second step is to determine the appropriate controls, striking a balance between risks and controls that is determined by the risk profile of the institution. One factor to consider here is risk introduced by the use of third-party information or solutions. Once it has been decided that a use case, initiative or project is feasible to implement, the new controls should be designed, implemented and tested, and the technological infrastructure adapted if needed. Finally, the solution should be integrated into the risk mitigation mechanisms that are already in place and must be monitored during its operation.

   These approaches allow for more efficient allocation of resources and greater precision in terms of managing vulnerabilities. Box C explains an approach that central banks could follow to start addressing AI risks according to AI models' intended use.

AI project evaluation process diagram Graph 1



Source: authors' elaboration.

Box C

AI risks based on use cases, initiatives and projects

Classifying artificial intelligence (AI) risks based on use cases, initiatives and projects allows more flexibility to focus on the level of risk of each use case (Amazon AWS (2023)). Central banks may employ third-party AI services or applications for non-sensitive daily use. Such arrangements warrant specific considerations for each use case, initiative or project in particular.

Risks associated with the daily use of AI tools

Buying or using open AI models for daily use, often using non-sensitive information, includes the following scenarios:

- **Consumer application**. The organisation consumes a public third-party AI service, either at no cost or paid. The organisation does not own or know the training data or the model. Usage is through application programming interfaces (APIs) or direct interaction with the application, according to the terms of service of the provider.

  Enterprise application. The organisation uses a third-party enterprise application that has AI features embedded within, and a business relationship is established between the organisation and the vendor.

The risks associated with the daily use of AI tools, such as consumer or enterprise applications, are generally lower compared with the risks tied to newly developed AI projects. Daily use scenarios involve established third-party services in which the organisation interacts with public or enterprise applications under defined terms of service. These scenarios, while not free from risk, benefit from the inherent stability and predictability of using pre-existing AI models and services. Organisations must remain vigilant about compliance and data privacy, but the risks are typically well managed through robust vendor relationships and standardised security protocols.

3. **Leverage and adapt governance management models already in place**. Using and adapting existing governance management models is important for both completeness and consistency. For example, the three lines of defence model, which separates management and control responsibilities into three distinct layers, should be adapted and used to clarify roles and responsibilities in managing AI risks.

   - **First line roles:** these are directly aligned with delivering products and/or services to internal and external stakeholders, including support functions. They own and manage related risks with AI outputs.

- **Second line roles:** these are responsible for the regulation and oversight of AI standards, supporting the first line in identifying relevant risks and challenging their risk assessments and control effectiveness.
- **Third line roles:** these provide independent and objective assurance and advice on the adequacy and effectiveness of governance and risk management.

The use of the three lines of defence model has several advantages in managing AI risks. The clear structure of roles and responsibilities facilitates the integration of AI risk management, though it may be necessary to consider specific aspects of each line's function for risks unique to AI. This model enhances the ability to quickly identify and mitigate AI risks and strengthens resilience against potential incidents. In addition to the three lines of defence model, central banks have developed robust risk management frameworks and processes to identify, monitor and mitigate many kinds of risks. Risks associated with AI should be included in these existing mechanisms, provided they are updated to capture risks that are unique or more pronounced with AI use (see Box A).

The table below contains some AI considerations for the three lines of defence (3LoD) model.

AI considerations for the 3LoD model

Table 2

Function	Considerations
First line	– **Identify AI use cases, initiatives and projects** that are aligned with the central bank's strategy or add innovation value. Identify use cases that are valid and reliable with clear benefits. – **Identify and assess risks.** Conduct risk assessments (identification, analysis and evaluation) that consider AI-specific threats and vulnerabilities, and determine whether use cases match the risk appetite. – **AI-specific technical controls**. Implement new controls, based on strengths, weaknesses, opportunities and threats (SWOT) analysis, such as AI model robustness testing, training data and outcome validation, and bias detection techniques. – **Monitoring mechanisms**. Set up continuous monitoring mechanisms to detect performance and security issues. Monitor data drift in AI models where changes to the statistical distribution of the underlying data can potentially cause a decline in model performance. – **Deliver specialised training**. Ensure technical and risk management skills are regularly refreshed.
Second line	– **Alignment of AI use with risk appetite and risk profile**. Support the first line to determine if the use case, initiative or project is aligned with the central bank's risk appetite and profile. – **Risk methodology and prioritisation.** Develop the institutional risk management methodology, coordinate the execution of risk assessments made by the first line and prioritise risks based on the AI risk profile. – **AI-specific policies and guidelines**. Update existing policies or develop new policies and guidelines to address AI-specific risks and promote the adoption of AI while ensuring ethical and secure use. Update data governance policies if necessary. – **Compliance and ethics**. Supervise compliance with regulations and adherence to ethical principles. Analyse legal framework according to the finance sector and central bank contexts. – **Specialised training and awareness**. Design and provide AI-specific risk training materials and instil awareness.
Third line	– **Technical and ethical audits**. Perform audits that evaluate the security controls as well as the fairness, transparency and ethics of AI models. – **Continuous review**. Provide ongoing reviews and recommendations for improving AI controls and policies, adapting to fast-evolving technology and associated risks.

Source: authors' elaboration.

When applying a risk management framework to the adoption of AI models and tools, it is crucial to define specific objectives based on principles. For example, a key principle is that central banks should implement a robust ICT risk management programme in alignment with their operational risk management framework.[7] Existing frameworks, based on standards like ISO 31000[8], can be very useful regarding promoting strong data governance and operational resiliency (BCBS (2021b)). Key elements include:

- governance (discussed in more detail in Section 5);
- operational-digital (integrated) risk management;
- business continuity plan;
- mapping interconnections, interdependencies and dependency management (as part of third-party risk management);
- incident management;
- ICT and information security risk management; and
- cyber security.

4. **Protect information through the full life cycle**. Due to the large volume of data that AI models require to operate and the opaqueness associated with AI models, one of the main challenges associated with AI risk management is protecting information, in terms of confidentiality, integrity, availability and privacy. A resilience perspective ensures that organisations not only prevent cyber attacks and protect ICT infrastructure, but also have the capacity to quickly recover and maintain essential operations during and after cyber incidents. As a result, protecting information throughout the AI life cycle is of the utmost importance.

3.2 Information security, privacy and cyber security risks

Information security and cyber security programs that have been developed based on standards such as ISO/IEC 27001:2022 – *Information security management*[9] or the National Institute of Standards and Technology (NIST) *Cybersecurity framework*[10] will continue to be adequate to manage the risks posed by AI models. These frameworks are flexible and scalable, so they can be adapted to different industries and technologies. Nevertheless, they need to be modified to capture the specific AI-related risks described in Section 2. A key aspect of AI models is that systems or tools that support these models should be valid, safe and reliable. In this way, cyber security is crucial for constructing reliable AI systems; by directly contributing to the confidentiality of sensitive information, mitigation of integrity and availability risks, as well as ensuring adequate access management.

Beyond preventing attacks and mitigating vulnerabilities, several standards focus on data governance and ensure that the results from AI are explainable and interpretable:

- NIST *Artificial intelligence risk management framework* (NIST AI RMF);[11]

[7] See principle 10 of BCBS (2021a).

[8] ISO (2018).

[9] ISO and IEC (2022a).

[10] NIST (2024).

[11] NIST (2023).

- ISO/IEC 23894:2023 – *Information technology – artificial intelligence – guidance on risk management*;[12] and
- ISO/IEC 38507:2022 – *Information technology – governance of IT – governance implications of the use of artificial intelligence by organisations*.[13]

These publications promote an integrated AI risk management framework into general risk models already in place. NIST AI RMF is a specialised technical approach to managing AI risks and complements the widely used general risk management model NIST *Risk management framework* (NIST RMF)[14]. These security frameworks might be considered part of the corporate mitigants, within the information security corporate definitions and risk framework to ensure the adequate treatment of information and data, as well as information technology (IT) assets, when adopting AI models. For example, if an AI model needs access to a database, it is necessary to:

– know the classification and sensitivity level of the data stored in the database in terms of confidentiality, integrity and availability;

– establish the control baseline for that particular classification and sensitivity;

– identify who or which systems are allowed to access those data and manage access; and

– monitor threats and report incidents related to information security and technological assets involved in the AI model.

If central banks decide to use or rely on third-party AI services, tools, components or algorithms, **a third-party risk management** model must be in place. This model begins by identifying external participants, assessing risks, determining the central bank risk profile and level of acceptance of external risks and their potential impact on critical processes. Central banks should perform due diligence to select third parties and negotiate contracts that define the rights and responsibilities of all parties, considering specific AI model risks.

When evaluating third-party data sets used by AI systems, it is important to have a clear understanding of data quality, training data sources, data ownership and traceability. Regarding AI model attributes, it is necessary to clarify the type of model, learning method, biases that may be present, autonomy level and how much human oversight employed. In this way, a central bank should align the identified third-party AI risks with the current corporate risk evaluation, before taking decisions on legal agreements, to ensure that the risk appetite is adequately represented by the conditions and duties of third-party service providers.

3.3 Specific actions to mitigate GAI risks

The complexity of GAI models requires special attention on trying to understand the outputs of those models, including potential biases, limitations and robustness, as well as considering the greater human supervision they will require. Some specific controls and practices examples to address GAI risks are listed below.

[12] ISO and IEC (2023a).

[13] ISO and IEC (2022b).

[14] NIST (2018).

Examples of controls to address GAI risks Table 3

Risk category	Approach	Specific controls or practices
Strategic	Define key functions	Define organisational functions related to GAI adoption that include creating a risk management culture for all phases of the GAI systems life cycle, assessing and monitoring GAI risks, and responding to and treating GAI risks.
	Enhanced governance structures	Establish a dedicated committee to identify and assess use cases, complementing existing governance structures. The key is to identify and carefully select cases in which AI generates a productivity impact and those where it does not. It is necessary to maintain an inventory of use cases to avoid project duplication and encourage innovative uses.
		Establish formal approval processes for use of GAI tools.
		Clear allocation of roles and responsibilities between model owner/developer/user and model validators.
	Enhanced policies and guidelines	Establish specific GAI requirements in existing policies or if necessary in an AI-dedicated policy that: define permitted and prohibited practices, and encourage its usage while balancing trustworthy AI principles or characteristics.[15] Examples of prohibited practices are: (i) publishing AI generated documents without proper review due to urgency; (ii) using GAI tools in critical processes before these tools are well calibrated; (iii) using sensitive information to train GAI models; and (iv) letting GAI tools modify their algorithms by themselves. If an AI-specific policy or guideline is created, it must be aligned to existing policies and governance mechanisms without duplicating work. Protect data used to train GAI models according to existing data policies.
Operational	Legal certainty	Review applicable end-user licence agreement with support from the legal team.
	Compliance	Align GAI usage to regulatory requirements and include it in the monitoring and reporting processes. Regulatory requirements must be considered when sensitive information assets or personally identifiable information (PII) are used as inputs of AI models. Particular attention is needed when AI solutions are deployed on public clouds, or third parties without proper contracts in which clear legal duties and responsibilities are established in case of any negative impact on the attributes of sensitive information.
	Business continuity	Review and, if necessary, update business continuity plans.
	People	Involve people with experience in various areas to define GAI usage guidelines to foster acceptable use of tools.

[15] NIST (2023) defines the following key characteristics of trustworthy AI systems: valid and reliable, safe, secure and resilient, explainable and interpretable, privacy enhanced, fair (with harmful bias managed), and accountable and transparent.

Information security, privacy and cyber security	**Information security and privacy**	
	Data classification	Determine what information can be used in AI models based on their classification.
		Classify AI models according to the most sensitive data used for training.
	Cyber security	
	Limitation on the use of AI tools	Identify GAI solutions that staff are using without the knowledge of the IT and cyber security units (shadow AI) and propose how these use cases should be addressed in the organisation. This identification could be done through network traffic analysis, considering the main AI public platforms.
		Limiting use to AI tools which are installed locally to avoid sharing data with external parties.
		Blocking access to public online models except for non-risk activities.
		Use separate networks or sandboxes for AI tools.
		Limit access to relevant data and/or limit the data that can be used for creating inputs (prompts) to non-sensitive data.
	Data protection	Filter outputs at the application level to remove sensitive or inappropriate information.
	System security	Perform security checks to open source or third-party models according to cyber controls.
		Identify and mitigate specific GAI and LLM systems vulnerabilities OWASP (2024).
		Consider specific GAI threats in application threat models, ie prompt injection, insecure plugin design, supply chain vulnerabilities and training data poisoning.
		Make sure AI systems follow both AI and security standards.
Information and communication technology (ICT)	Capacity planning	Analyse current and future resource needs, including processing power, data storage and bandwidth. Ensure that the infrastructure can handle AI workloads. Deploy or utilise elastic architectures, based on containers or virtual machines. Continuously monitor and optimise resource usage for maximum efficiency and performance.
Third party	Terms and conditions	Understand the provider's terms of service and privacy policy, and review them periodically as they can change without notice. Understand the service's data flow, particularly if it stores information, uses it to improve the model or shares it with fourth parties. Understand ownership of prompts and responses.
Reputational	Explainability	Understand the data used to train the model and the quality of the information. Only use explicable AI outcomes.

AI models	**Interpretation and reliability**	
	Human supervision	Even though AI models were trained with large amounts of data, their results must be considered as drafts that must be reviewed and discussed. These reviews must be carried out by humans and in many cases are new process activities that should be done by experts. Notably, output can seem accurate even when it is not. This requires close attention. Establish procedures for output validation that consider the review of a human expert on the subject.
	Employee education	Training people to ensure the risks and limitations of AI models are known.
		Learn how to create adequate inputs (prompts) to increase the probability of obtaining the expected results.
	Transparency and accountability	
	Transparent usage	Disclose AI usage. Appropriately denote, label or identify work and processes that use GAI tools.
	Enhanced validation	Requirement for more validation and authorisation even for lower risk use cases.
	Environmental	
	Efficient energy use	Optimise the life cycle of AI systems.
	Ethical and social	
	Internal rules and manuals	Elaborate internal rules or manuals that set out the responsible use of AI tools.

Source: authors' elaboration.

4 Governance

The inherent complexity and uncertainty associated with AI systems makes their adoption challenging. Effective governance[16] mechanisms can help balance the risks and rewards of AI adoption in a consistent risk-based manner. This is important because AI may affect business processes and decision-making across the institution. AI governance is important not only for complying with national and international strategies, laws or regulations (including multilateral agreements between countries), but also for ensuring the alignment of any AI initiative with the organisation's strategy. Effective governance mechanisms should support efficiency and innovation in the organisation while effectively identifying and balancing associated risks. That said, AI governance does not have to start from scratch but can be built on policies, procedures and management tools that are already in place.

[16] AI governance refers to the set of policies, rules, frameworks, principles and/or standards that guide organisations in their adoption or development of secure, responsible and ethical use of AI.

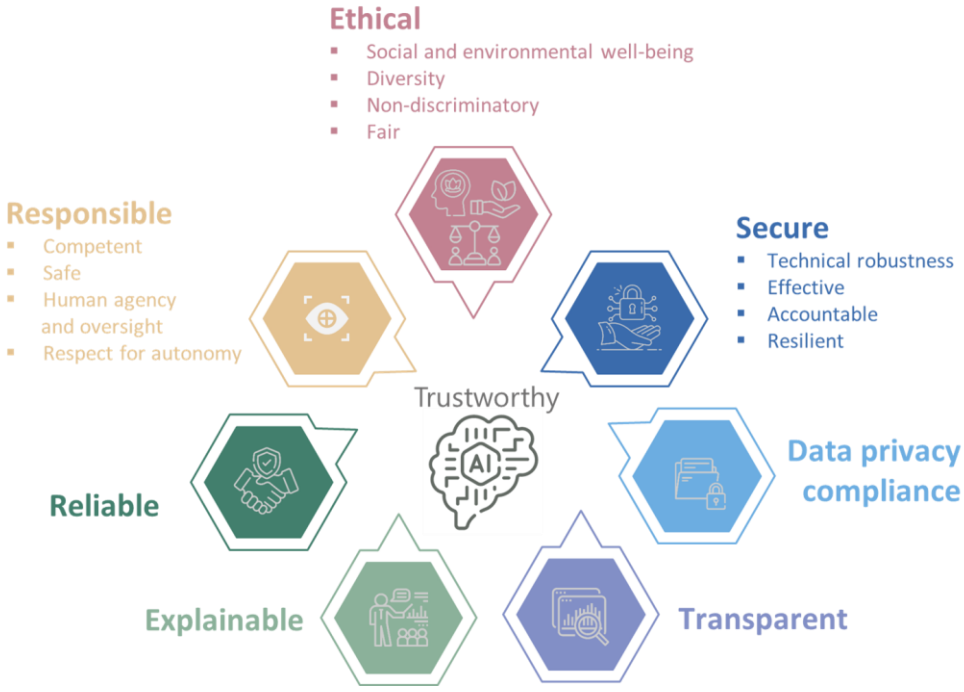
4.1 Current industry frameworks

Several institutions worldwide[17] have developed AI frameworks for organisations to exercise AI governance (Appendix 4). While guidelines can be a good starting point for setting up governance practices, these need to be commensurate with each organisation's operational environment and risk appetite. In the case of central banks, these include the generally low risk appetite and the importance of fulfilling its mandate and maintaining transparency.

For AI to be trustworthy, the majority of industry guidance considers balancing each of the principles listed below:

- **secure –** AI systems should be robust, secure, effective and resilient;
- **data privacy compliant –** AI users should ensure data privacy internal governance is maintained;
- **explainable and transparent –** AI users should ensure that decisions taken with the support of AI are understandable and clear;
- **reliable –** AI systems should consistently perform as expected;
- **ethical –** AI users should ensure that the results of AI systems help promote social and environmental well-being, ensure diversity, and are non-discriminatory and fair;
- **responsible –** AI users must ensure that AI systems assist the human decision-making process, allowing people to oversee the system and override its decisions; and
- **accountable –** users should always know and follow the internal governance of AI usage.

Common AI principles of different frameworks — Graph 2



Source: authors' elaboration.

[17] Eg the International Organization for Standardization (ISO), the Organisation for Economic Co-operation and Development (OECD) and the National Institute of Standards and Technology (NIST) of the US Department of Commerce.

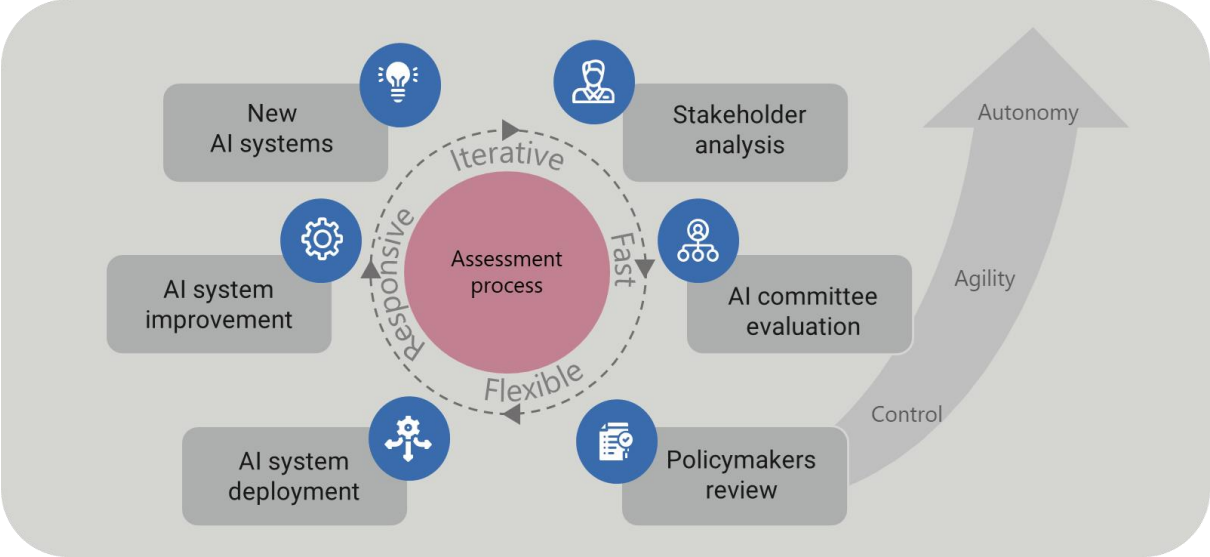
4.1.1 Adaptive governance frameworks

To deal with the fast pace of change in the field, AI governance needs to be dynamic and adaptative. A governance framework is dynamic if it is able to keep up with developments in the field and is flexible and responsive. A dynamic and adaptive framework can be based on the following principles:

1. Control – the organisation defines guard rails to comply with policies, requirements and regulations.
2. Agility – delegating roles and teams with authority to make distributed and/or mandated decisions.
3. Autonomy – people in the organisation can make autonomous governance decisions based, for instance, on real-time data and in line with the enterprise's objectives and goals.

Graph 3 sketches an adaptive AI governance framework, in which AI projects are evaluated on their ethical, data security, transparency, accountability, non-discrimination, regulatory and compliance implications and can only be implemented once these points have been analysed and addressed.

Adaptive AI governance framework — Graph 3



Sources: Grasso (2021); Gartner (2022).

4.2 Proposed actions for AI governance at central banks

The CGRM has identified the following actions as useful for governance of the adoption of AI in central banks:

1. Establish an interdisciplinary AI committee

Prior to establishing specific AI governance, central banks should establish a dedicated AI committee, which will serve as an oversight body to help guide the implementation of governance requirements. This oversight body should be sufficiently interdisciplinary, given the

far-reaching nature of AI risk, and at a sufficiently senior level to ensure organisational buy-in. While this could mean establishing a specific committee to oversee AI use, it could also be an additional responsibility of an existing committee.

2. Define principles for responsible AI use

The committee overseeing AI use should establish a set of principles that outlines the organisation's philosophy and approach to AI use. These principles should be sufficiently clear to capture the central bank's overall posture towards AI use and consistently guide governance decisions. The information from standard-setting bodies (such as ISO) or other organisations (such as OECD) can help with defining these principles.

3. Establish an AI framework and update existing guidance

Set up a robust AI governance framework in line with the central bank's strategy, objectives, values and established AI principles. This framework should be used when assessing AI initiatives proposed by different parts of the bank and when managing projects.

4. Maintain an AI tools inventory

An inventory ensures that all AI systems within an organisation are known and can be properly managed, monitored and aligned with the central bank's AI governance framework.

5. Map AI tools and stakeholders

Once the principles and inventory are defined, there is a need to establish an understanding of which processes might benefit from AI usage and the corresponding stakeholders (current users, potential users, data providers and third parties etc). This mapping would constitute the use cases that are relevant for the central bank. Incorporating stakeholders' participation, both internal and external, is crucial to ensure that AI systems meet regulatory and ethical requirements.

6. Perform a detailed assessment of risks and controls

The assessment of risks and controls should not only involve ICT technical aspects but also other relevant functions such as cyber security, legal, information security operational risk and external third parties. It should identify controls that need to be in place before testing starts or before an AI tool goes live.

7. Perform regular monitoring

The second line of defence should develop compliance monitoring and reporting to the AI committee to ensure that systems, operational practices, information treatment and ethical implications are in line with defined guidelines. Effective compliance monitoring for AI governance includes:

a. The deployment of tools that continuously monitor AI systems for data drift, performance, fairness, bias and adherence to defined parameters. They should also ensure that data used for AI training and operations are handled in compliance with privacy laws and security standards.

b. Implementing processes to ensure the quality and integrity of data used by AI systems.

c. Developing a policy to respond to incidents involving AI systems, including their investigation, mitigation and reporting. Conducting root cause analysis for any compliance breaches or performance issues to prevent future occurrences.

d. Considering human oversight for compliance monitoring as another component to validate the accomplishment of the defined guidelines.

8. Report anomalies and incidents

Once AI tools have been implemented, there should be a process for reporting incidents involving AI to the overseeing AI committee to capture residual risks that require further reduction or controls that require further improvement.

9. Develop and improve workforce skills

The proper functioning of central banks largely depends on the specialisation and expertise of their staff. AI applications reduce intensive manual processing, allowing staff to dedicate a larger share of their time to more productive and innovative activities. To build AI knowledge, central banks should define basic and specialist training and awareness programmes on AI usage, governance and compliance.

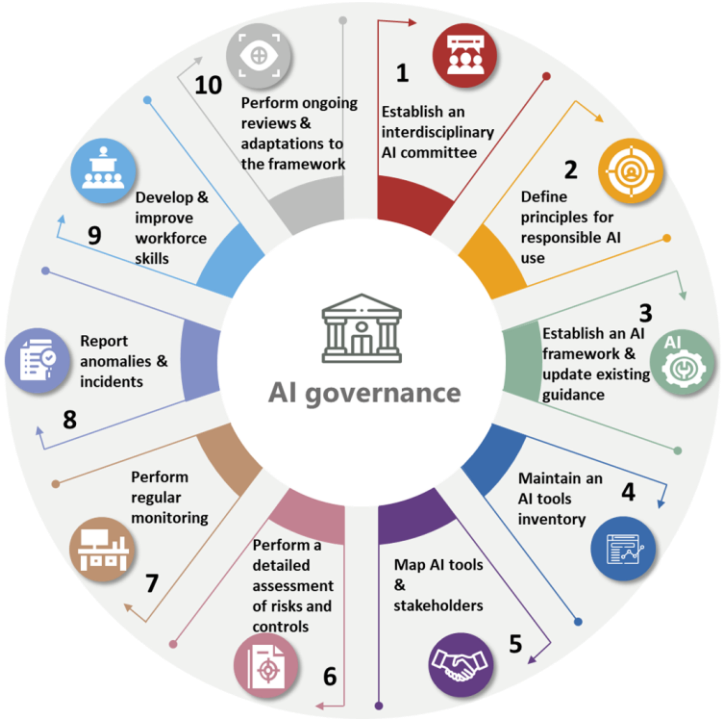
10. Perform ongoing reviews and adaptations to the framework

Given the fast pace of change in AI, governance frameworks need to be reviewed regularly to ensure validity and appropriateness on an ongoing basis. Central banks need to regularly review the outcomes of AI systems to ensure they align with their organisational goals and risk management measures. They should also consider the outcomes of risk and control assessments performed, as well as incident reports involving any AI tools.

Together, these actions could form an AI governance system for central banks. These actions should all be aligned to the central bank's objectives and do not need to be created separately from existing forms of technology governance.

General proposal for governance and risk management associated with the use of AI models by central banks

Graph 4



Source: authors' elaboration.

5 Conclusions

Central banks are increasingly adopting AI to improve data quality, enhance operations and support decision making. AI provides powerful tools for addressing complex challenges in areas such as data analysis, risk assessment, forecasting, operations, and customer and corporate services.

But the use of AI tools can also pose new risks and have the potential to exacerbate existing risks in central banks. Such risks are best managed holistically at the design, implementation and operations phases. Information security risks deserve special mention, given the criticality and sensitivity of the information handled by central banks. In general, central banks could start adopting AI in non-critical processes where risks can be better controlled.

A safe AI adoption could cover the following domains: (i) governance; (ii) legal and compliance; (iii) information security and privacy; (iv) cyber security; (v) third-party risk management; (vi) business continuity; and (viii) other operational risks associated with the level of digitalisation and exposure of the organisation.

Instead of reinventing the wheel, such a framework can build on existing risk management frameworks and governance schemes. The task force did not identify any entirely new approaches to managing AI risks, but recommends adapting existing mechanisms to the specific risks posed by AI tools. Central banks can use existing risk management mechanisms such as the three lines of defence model with specific adjustments.

We recommend that the AI risk profile is discussed and defined by upper management before adopting AI models as this would help to allocate resources and set priorities. Benefits need to be weighed against risks posed to information and core functions and be aligned with the risk appetite and security capabilities of the organisation.

Given the transformative potential of AI technology, both in terms of its business impact and possible externalities to society at large, developing a governance framework for AI adoption is of high importance for central banks interested in the use of AI. This involves revising policies that cover various procedures for the organisation's governance and operations, such as systems and risk management, compliance and data maintenance, and transparency and communication with internal and external stakeholders.

Good practices laid out in international standards can serve as a starting point. These practices include:

- **systems and risk management –** updating systems to integrate AI while ensuring robust risk management practices;
- **compliance and data maintenance –** ensuring AI systems comply with existing regulations and maintain high standards of data integrity and privacy; and
- **transparency and communication –** enhancing transparency in AI decision-making processes and effectively communicating these processes to internal and external audiences.

By leveraging these international standards and best practices, central banks can effectively incorporate AI into their governance frameworks – ensuring the secure, responsible and ethical use of AI technology.

References

Accornero, M and G Boscariol (2022): "Machine learning for anomaly detection in datasets with categorical variables and skewed distributions" in "Machine learning in central banking", *IFC Bulletin*, no 57, November, presentation given at an Irving Fisher Committee on Central Bank Statistics and Bank of Italy workshop, 19–22 October 2021, www.bis.org/ifc/publ/ifcb57_05.pdf.

Amadxarif, Z, J Brookes, N Garbarino, R Patel and E Walczak (2021): "The language of rules: textual complexity in banking reforms" *Bank of England Staff Working Paper*, no 834, November, www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/the-language-of-rules-textual-complexity-in-banking-reforms.pdf.

Amazon AWS (2023): "Securing generative AI: an introduction to the generative AI security scoping matrix", *AWS Security Blog*, 19 October, aws.amazon.com/es/blogs/security/securing-generative-ai-an-introduction-to-the-generative-ai-security-scoping-matrix/.

——— (2024): "What's the difference between deep learning and neural networks?", aws.amazon.com/compare/the-difference-between-deep-learning-and-neural-networks/?nc2=h_mo-lang.

Araujo, D, G Bruno, J Marcucci, R Schmidt and B Tissot (2022): "Machine learning applications in central banking: an overview" in "Machine learning in central banking", *IFC Bulletin*, no 57, November, www.bis.org/ifc/publ/ifcb57_01_rh.pdf.

Babic, B, I Cohen, T Evgeniou and S Gerke (2021): "When machine learning goes off the rails", *Harvard Business Review*, January–February 2021, hbr.org/2021/01/when-machine-learning-goes-off-the-rails.

Baker-Brunnbauer, J (2023): *Trustworthy artificial intelligence implementation: introduction to the TAII framework*, Springer, link.springer.com/book/10.1007/978-3-031-18275-4.

Bank for International Settlements (BIS) (2024): "Artificial intelligence and the economy: implications for central banks", *Annual Economic Report 2024*, June, Chapter III, www.bis.org/publ/arpdf/ar2024e3.pdf.

Bank for International Settlements Innovation Hub (BISIH) (2023): *Project Aurora: the power of data, technology and collaboration to combat money laundering across institutions and borders*, May, www.bis.org/publ/othp66.htm.

——— (2024): *Project Raven: using AI to assess financial system's cyber security and resilience*, April, www.bis.org/about/bisih/topics/cyber_security/raven.htm.

Basel Committee on Banking Supervision (BCBS) (2021a): *Revisions to the principles for the sound management of operational risk*, March, www.bis.org/bcbs/publ/d515.htm.

——— (2021b): *Principles for operational resilience*, March, www.bis.org/bcbs/publ/d516.htm.

Benford, J (2024): "Trusted AI: ethical, safe and effective application of artificial intelligence at the Bank of England", speech given at the Central Bank AI Conference, Bank of England, September, www.bankofengland.co.uk/speech/2024/september/james-benford-speech-at-the-central-bank-ai-inaugural-conference.

Bluwstein, K, M Buckmann, A Joseph, M Kang, S Kapadia and Ö Simsek (2020): "Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach", *Bank of England Staff Working Paper*, no 848, January, www.bankofengland.co.uk/-/media/boe/files/working-paper/2020/credit-growth-the-yield-curve-and-financial-crisis-prediction-evidence-from-a-machine-learning.pdf.

Boyd, D and K Crawford (2012): "Critical questions for big data", *Information, Communication & Society*, vol 15, no 5, pp 662–79, doi.org/10.1080/1369118X.2012.678878.

Buckmann, M, G Potjagailo and P Schnattinger (2023): "Dissecting UK service inflation via a neural network Phillips curve," *Bank Underground*, 10 July, bankunderground.co.uk/2023/07/10/dissecting-uk-service-inflation-via-a-neural-network-phillips-curve/.

Burgess, S, E Fernandez-Corugedo, C Groth, R Harrison, F Monti, K Theodoridis and M Waldron (2013): "The Bank of England's forecasting platform: COMPASS, MAPS, EASE and the suite of models", *Bank of England Staff Working Paper*, no 471, May, www.bankofengland.co.uk/-/media/boe/files/working-paper/2013/the-boes-forecasting-platform-compass-maps-ease-and-the-suite-of-models.pdf.

Cagala, T, J Hees, D Herurkar, M Meier, N-T Nguyen, T Sattarov, K Troutman and P Weber (2022): "Unsupervised outlier detection in official statistics" in "Machine learning in central banking", *IFC Bulletin,* no 57, November, presentation given at an Irving Fisher Committee on Central Bank Statistics and Bank of Italy workshop, 19–22 October 2021, www.bis.org/ifc/publ/ifcb57_09.pdf.

Chakraborty, C and A Joseph (2017): "Machine learning at central banks", *Bank of England Staff Working Paper*, no 674, www.bankofengland.co.uk/-/media/boe/files/working-paper/2017/machine-learning-at-central-banks.pdf.

Chen, M, M DeHaven, I Kitschelt, S Lee and M Sicilian (2023): "Identifying financial crises using machine learning on textual data", Board of Governors of the Federal Reserve System, *International Finance Discussion Papers*, no 1374, doi.org/10.17016/IFDP.2023.1374.

Chiarello, F, V Giordano, I Spada, S Barandoni and G Fantoni (2024): "Future applications of generative large language models: a data-driven case study on ChatGPT", *Technovation*, vol 133, doi.org/10.1016/j.technovation.2024.103002.

Consultancy.eu (2023): "European AI Act: Implications for the financial services industry", 19 October, www.consultancy.eu/news/9392/european-ai-act-implications-for-the-financial-services-industry.

Consultative Group on Risk Management (CGRM) (2023): *Central bank digital currency (CBDC) information security and operational risks to central banks*, November, www.bis.org/publ/othp81.pdf.

Dauphin, J-F, K Dybczak, M Maneely, M Sanjani, N Suphaphiphat, Y Wang and H Zhang (2022): "Nowcasting GDP: a scalable approach using DFM, machine learning and novel data, applied to European economies", *IMF Working Paper*, no 52, March, www.imf.org/en/Publications/WP/Issues/2022/03/11/Nowcasting-GDP-A-Scalable-Approach-Using-DFM-Machine-Learning-and-Novel-Data-Applied-to-513703.

Denes, J, A Lestrade and L Richardet (2022): "Using Twitter data to measure inflation perception" in "Machine learning in central banking", *IFC Bulletin*, no 57, November, presentation given at an Irving Fisher Committee on Central Bank Statistics and Bank of Italy workshop, 19–22 October 2021, www.bis.org/ifc/publ/ifcb57_13.pdf.

Derner, E and K Batistič (2023): "Beyond the safeguards: exploring the security risks of ChatGPT", *arXiv preprint arXiv:2305.08005*, arxiv.org/pdf/2305.08005

Devys, E and U von Kalckreuth (2022): "The use of AI for company data gathering – finding and monitoring fintechs in Germany and France" in "Machine learning in central banking", *IFC Bulletin*, no 57, November, presentation given at an Irving Fisher Committee on Central Bank Statistics and Bank of Italy workshop, 19–22 October 2021, www.bis.org/ifc/publ/ifcb57_28.pdf.

European Central Bank (ECB) (2023): "Suptech: thriving in the digital age", *Supervision Newsletter*, 15 November, www.bankingsupervision.europa.eu/press/supervisory-newsletters/newsletter/2023/html/ssm.nl231115_2.en.html.

European Parliament (2024a): "EU AI Act: first regulation on artificial intelligence", 18 June, www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

European Parliament (2024b): "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)", *Official Journal of the European Union*, 7 July, eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Faria da Costa, A, F Fonseca and S Maurício (2022): "Novel methodologies for data quality management anomaly detection in the Portuguese central credit register" in "Machine learning in central banking", *IFC Bulletin*, no 57, November, presentation given at an Irving Fisher Committee on Central Bank Statistics and Bank of Italy workshop, 19–22 October 2021, www.bis.org/ifc/publ/ifcb57_29.pdf.

Future of Life Institute (2024): "The AI Act Explorer", *EU Artificial Intelligence Act*, artificialintelligenceact.eu/ai-act-explorer/.

Gartner (2022): "Choose adaptive data governance over one-size-fits-all for greater flexibility", 11 April, www.gartner.com/en/articles/choose-adaptive-data-governance-over-one-size-fits-all-for-greater-flexibility.

Gascon, C and D Werner (2022): "Does the Beige Book reflect US employment and inflation trends?", *Economic Synopses*, Federal Reserve Bank of St. Louis, doi.org/10.20955/es.2022.13.

Grasso, C (2021): "Governance: from research trend to enterprise standard", *Dataiku blog*, 2 August, blog.dataiku.com/governance-from-research-trend-to-enterprise-standard.

Haghighi, M, C Jones and J Younker (2022): "Machine learning for anomaly detection in financial regulatory data" in "Machine learning in central banking", *IFC Bulletin*, no 57, November, presentation given at an Irving Fisher Committee on Central Bank Statistics and Bank of Italy workshop, 19–22 October 2021, www.bis.org/ifc/publ/ifcb57_24.pdf.

Hirsch, T, K Merced, S Narayanan, Z E Imel and D C Atkins (2017): "Designing contestability: Interaction design, machine learning, and mental health", Proceedings of the 2017 Conference on Designing Interactive Systems, pp 95–99.

Info-communications Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC) (2020): *Model artificial intelligence governance framework*, second edition, www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf.

Institute of Electrical and Electronics Engineers (IEEE) (2023) "General Principles of Ethically Aligned Design", *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*, standards.ieee.org/wp-content/uploads/import/documents/other/ead1e_general_principles.pdf.

International Organization for Standardization (ISO) (2018): "Risk management – guidelines", *ISO 31000:2018*, www.iso.org/standard/65694.html.

International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) (2017): "Information technology – cloud computing – interoperability and portability", *ISO/IEC 19941:2017*, www.iso.org/standard/66639.html.

——— (2022a): "Information security, cybersecurity and privacy protection – information security management systems – requirements", *ISO/IEC 27001:2022*, www.iso.org/standard/27001.

——— (2022b): "Information technology – governance of IT – governance implications of the use of artificial intelligence by organizations", *ISO/IEC 38507:2022*, www.iso.org/standard/56641.html.

——— (2022c): "Trustworthiness – vocabulary", *ISO/IEC TS 5723:2022*, www.iso.org/standard/81608.html.

——— (2023a): "Information technology – artificial intelligence – guidance on risk management", *ISO/IEC 23894:2023*, www.iso.org/standard/77304.html.

——— (2023b): "Information technology – artificial intelligence – management system", *ISO/IEC 42001:2023*, www.iso.org/standard/81230.html.

Jiménez, P and T Serrano (2022): "An artificial intelligence application for accounting data cleansing" in "Machine learning in central banking", *IFC Bulletin*, no 57, November, presentation given at an Irving Fisher Committee on Central Bank Statistics and Bank of Italy workshop, 19–22 October 2021, www.bis.org/ifc/publ/ifcb57_23.pdf.

Joseph, A, G Potjagailo, E Kalamara, C Chakraborty and G Kapetanios (2022): "Forecasting UK inflation bottom up", *Bank of England Staff Working Paper*, no 915, March, www.bankofengland.co.uk/-/media/boe/files/working-paper/2021/forecasting-uk-inflation-bottom-up.pdf.

Kalamara, E, A Turrell, C Redl, G Kapetanios and S Kapadia (2020): "Making text count: economic forecasting using newspaper text", *Bank of England Staff Working Paper*, no 865, May, www.bankofengland.co.uk/working-paper/2020/making-text-count-economic-forecasting-using-newspaper-text.

Kerdsri, J and P Treeratpituk (2022): "Using deep learning technique to automate banknote defect classification" in "Machine learning in central banking", *IFC Bulletin*, no 57, November, presentation given at an Irving Fisher Committee on Central Bank Statistics and Bank of Italy workshop, 19–22 October 2021, www.bis.org/ifc/publ/ifcb57_34.pdf.

Khurana, D, A Koli, K Khatter and S Singh (2023): "Natural language processing: state of the art, current trends and challenges", *Multimedia Tools and Applications*, vol 82, pp 3713–44, doi.org/10.1007/s11042-022-13428-4.

McCaul, E (2024): "From data to decisions: AI and supervision," February, www.bankingsupervision.europa.eu/press/interviews/date/2024/html/ssm.in240226~c6f7fc9251.en.html.

Moody's Investor Service (Moody's) (2023): "Generative AI may increase cyber risk rather than lessen it", Cybersecurity – Cross region, 18 September.

Mulligan, D, C Koopman and N Doty (2016): "Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy", *Philosophical Transactions of the Royal Society A*, vol 374, no 2083, December, dx.doi.org/10.1098/rsta.2016.0118.

National Institute of Standards and Technology (NIST) (2018): *Risk management framework for information systems and organizations*, nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-37r2.pdf.

National Institute of Standards and Technology (NIST) (2023): *Artificial intelligence risk management framework (AI RMF 1.0)*, nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

National Institute of Standards and Technology (NIST) (2024): *The NIST Cybersecurity Framework (CSF) 2.0*, nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf.

Nistor, M (2023): "Artificial intelligence applications in finance: considering the benefits and risks", *The Teller Window*, Federal Reserve Bank of New York, December, tellerwindow.newyorkfed.org/2023/12/14/ai-applications-in-finance-considering-the-benefits-and-risks/.

Njoroge, L (2024): "Role of artificial intelligence (AI) in central banking: implications for COMESA member central banks", *COMESA Monetary Institute Special Report*, March, www.comesa.int/wp-content/uploads/2020/09/Special-Report-AI-and-Big-Data-Implication-for-Central-Banking-in-COMESA-region.pdf.

Oh, P (2024): "Navigating the AI Maze: How ISO Standards Guide Responsible Development and Deployment", *Medium*, 21 January, medium.com/@patrick-oh-sglion65/navigating-the-ai-maze-09ff5909ae78.

Organisation for Economic Co-operation and Development (OECD) (2022): "OECD framework for the classification of AI systems", *OECD Digital Economy Papers*, February, doi.org/10.1787/cb6d9eca-en.

——— (2023): *G7 Hiroshima process on generative artificial intelligence (AI): towards a G7 common understanding on generative AI*, doi.org/10.1787/bf3c0c60-en.

——— (2024a): "Recommendation of the Council on Artificial Intelligence", *OECD Legal Instruments*, legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

——— (2024b): "What is AI? Can you make a clear distinction between AI and non-AI systems?", *The AI Wonk*, March, oecd.ai/en/wonk/definition.

Open Web Application Security Project (OWASP) (2024): "OWASP Top 10 for Large Language Model Applications", *The OWASP foundation*, owasp.org/www-project-top-10-for-large-language-model-applications/

Petropoulos, A, V Siakoulis, K Panousis, L Papadoulas and S Chatzis (2019): "A deep learning approach for dynamic balance sheet stress testing", in D Magazzeni, S Kumar et al (eds), *ICAIF '22: proceedings of the third ACM international conference on AI in finance*, Association for Computing Machinery, pp 53–61, doi.org/10.1145/3533271.3561656.

Pinsent and Masons (2024): "A guide to high-risk AI systems under the EU AI Act", *Out-law guide*, 13 February, www.pinsentmasons.com/out-law/guides/guide-to-high-risk-ai-systems-under-the-eu-ai-act.

Richardson, A, T van Florenstein Mulder and T Vehbi (2021): "Nowcasting GDP using machine learning algorithms: a real-time assessment", *International Journal of Forecasting*, vol 37, no 2, pp 941–48, doi.org/10.1016/j.ijforecast.2020.10.005.

Rubio, J, P Barucca, G Gage, J Arroyo and R Morales-Resendiz (2020): "Classifying payment patterns with artificial neural networks: an autoencoder approach", *Latin American Journal of Central Banking*, vol 1, nos 1–4, doi.org/10.1016/j.latcb.2020.100013.

Sharma, N, R Sharma and N Jindal (2021): "Machine learning and deep learning applications – a vision", *Global Transitions Proceedings*, vol 2, no 1, June, pp 24–28, doi.org/10.1016/j.gltp.2021.01.004.

Université de Montréal (2018): "Montréal declaration for a responsible development of artificial intelligence", *Montréal Declaration Responsible AI*, declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl-IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf.

Yenduri, G, M Ramalingam, G Chemmalar Selvi et al (2024): "GPT (generative pre-trained transformer) – a comprehensive review on enabling technologies, potential applications, emerging challenges and future directions", *IEEE Access*, April, ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10500411.

Appendix 1: Artificial intelligence definitions

Artificial intelligence risk management framework (AI RMF 1.0) **(NIST (2023))**

"An AI system is an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy."

Recommendation of the Council on Artificial Intelligence **(OECD (2024a))**

"An AI system 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. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."

Trustworthy artificial intelligence implementation: introduction to the TAII framework **(Baker-Brunnbauer (2023))**

"AI systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best actions to take to achieve the given goal."

Model artificial intelligence governance framework **(IMDA and PDPC (2020))**

"AI refers to a set of technologies that seek to simulate human traits such as knowledge, reasoning, problem solving, perception, learning and planning, and, depending on the AI model, produce an output or decision (such as a prediction, recommendation, and/or classification)."

Role of artificial intelligence (AI) in central banking: implications for COMESA member central banks **(Njoroge (2024))**

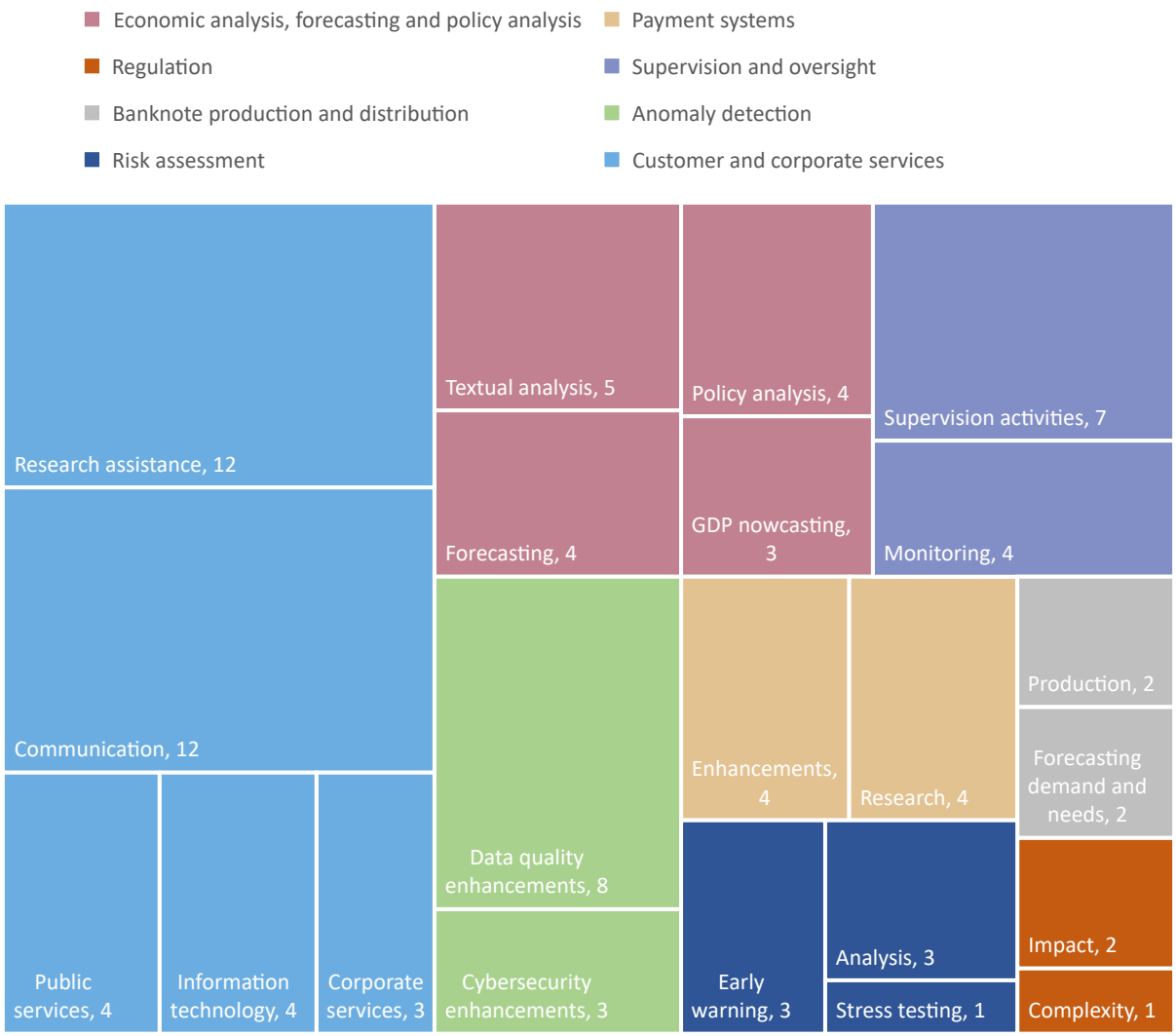
"Artificial intelligence (AI) uses computing to create intelligence artificially and is described as the ability of machines to imitate human intelligence. It entails a collection of tools that learn with given data and understand patterns and interactions between series and values."

Appendix 2: Summary of AI use cases described in Section 1.1

The use cases described in Section 1.1 were identified from both public sources and from responses to a questionnaire from members of the Consultative Group on Risk Management (CGRM) task force. Graph 2.1 below displays the distribution of the number of use cases.

AI use cases among CGRM task force members[1] Graph 2.1

Number of responses



[1] This graph includes more use case categories than those described in Section 1.1.

Source: authors' elaborations.

Appendix 3: Risks associated with AI

Risks associated with AI

Table 3.1

Type of risk		Description	Applies to new technologies	Applies to AI
	**Strategic**	Lack of strategy and governance in AI, which may impede central banks' ability to achieve their objectives.	✔	✔
	**Operational**	Loss resulting from inadequate or failed internal processes, people, systems or from external events. Includes legal uncertainties, compliance risks due to the black box nature of AI models, deficiencies in processes, risks associated with personnel's digital capabilities, third-party dependency and capacity issues related to the data intensive nature of AI models.	✔	✔
	**Information security, privacy and cyber security**	Safeguard sensitive assets and personally identifiable information in systems, particularly in public cloud infrastructure, while ensuring regulatory compliance. The misuse or mishandling of sensitive data or personal information by AI models, exposure of confidential information due to inadequate cyber security controls, and vulnerabilities such as model extraction and data poisoning, all of which that can result in significant legal and reputational consequences.	✔	✔
	**Information and communication technology (ICT)**	Potential failures or misconfigurations in ICT infrastructure that could impact the performance, availability or integration of systems, compromising sensitive data. This includes risks arising from software bugs, hardware failures, inadequate system design and challenges in compatibility with new technologies, as well as insufficient safeguards for critical processes during disruptions, which can affect business continuity and operational resilience.	✔	✔
	**Third party**	Incidents originating from dependence on external AI models or tools developed by a third party outside the organisation including privacy breaches, operational disruptions, compliance failures and cyber security threats.	✔	✔
	**Reputational**	Reputational damage or adverse publicity because of errors, data leaks or lack of transparency, particularly due to inadequate control of complex AI models. Such incidents can lead to public scrutiny and diminished trust in central banks.	✔	✔
	**AI models**	Risks associated with AI models include issues of interpretation and reliability from imprecise results and black box models, as well as challenges in transparency and accountability due to algorithm self-modification. Environmental concerns arise from the high carbon footprint of energy intensive AI models, while ethical issues stem from potential biases and inappropriate outcomes.		✔

Source: authors' elaboration.

Appendix 4: Review of current AI frameworks

The European Union's Artificial Intelligence Act[18]

The Artificial Intelligence Act (AI Act) 2024[19] mandates that providers follow six ethical principles. This approach promotes innovation while ensuring responsible and safe AI development.

Values-based principles — Table 4.1

1. Social and environmental well-being	AI system creators should design these systems to promote sustainable and inclusive growth, social progress and environmental well-being. Providers must consider the potential societal and environmental impacts of AI systems to ensure they contribute positively to these areas.
2. Diversity, non-discrimination and fairness	Developers and providers of AI systems should create these systems to avoid discrimination and bias, while promoting diversity. Providers must rigorously examine data sources for bias and implement appropriate measures to mitigate any potential biases.
3. Transparency	AI systems must be transparent. Providers should offer clear information regarding the system's capabilities, limitations and the data sources used for training.
4. Technical robustness and safety	Providers and developers should design AI systems to be reliable, predictable and safe for use. AI providers must ensure their systems comply with established quality management standards.
5. Human agency and oversight	AI systems should assist humans in decision-making, allowing humans the ability to override system-generated decisions.
6. Privacy and data governance	AI system providers and developers should design AI systems with a focus on data privacy and protection. The data sets used for training should be subject to stringent governance.

Source: adapted from European Parliament (2024b).

The AI Act classifies models according to its risks, categorising them into four different levels (Graph 4.1):[20]

The **unacceptable risk** category establishes that systems considered a clear threat to the safety, livelihoods and rights of people will be banned. This includes social scoring by governments, devices using voice assistance that encourage dangerous behaviour and manipulative AI.

The second category, **high risk**, refers to any AI system which: (i) constitutes a certain type of product (eg medical devices, industrial machinery, toys, aircraft or cars); (ii) is a safety component of a certain type of product (eg components for rail infrastructure, lifts or appliances burning gaseous fuels); or (iii) AI systems that are high risk by their nature eg biometrics, critical infrastructure, exploiting vulnerabilities, education, employment, access to essential services both public and private, law enforcement, immigration, the administration of justice and democratic processes.

[18] Future of Life Institute (2024) and European Parliament (2024a).

[19] European Parliament (2024b).

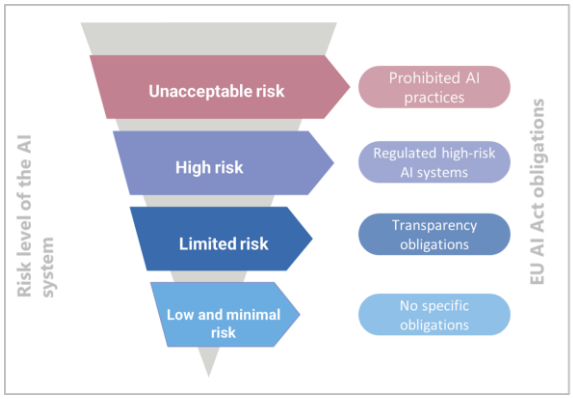
[20] Pinsent and Masons (2024) and Future of Life Institute (2024).

High-risk AI providers have specific obligations including establishing a risk management system to operate throughout the AI system's life cycle. Such providers must also conduct data governance, ensure adequate training is carried out, and validate and test that data sets are free of errors. Another obligation is to draw up technical documentation to demonstrate and assess compliance, and to design high-risk AI systems to allow deployers to implement human oversight and achieve appropriate levels of accuracy, robustness and cyber security. Finally, they must consider establishing a quality management system to ensure compliance.

For AI systems considered as **limited risk**, the AI Act defines transparency obligations for providers and deployers of AI systems. This includes ensuring that such systems are designed to inform people that they are interacting with an AI system. Another obligation is to ensure that the outputs of AI systems generating synthetic content (eg audio, image, video or text), are marked and detectable as artificially generated. Finally, there is an obligation to ensure that deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake shall disclose that the content has been artificially generated or manipulated. In contrast, AI systems catalogued as **low and minimal risk** do not have any specific obligations.

Moreover, the AI Act, specifies other obligations regarding the AI models' technical documentation, including training and testing processes and the results of evaluations. There is also an obligation to make information and documentation available to providers of AI systems who intend to integrate the general purpose AI model in their AI systems. Additional obligations include having a policy in place regarding copyright law and making a detailed summary about the content used for training the general purpose AI model publicly available.

European Union AI Act risk model — Graph 4.1



Source: adapted from Consultancy.eu (2023).

The Institute of Electrical and Electronics Engineers (IEEE) Global Initiative on Ethics of Autonomous and Intelligent Systems[21]

This framework consists of eight general principles applicable to the creation and operation of all types of autonomous and intelligent systems (AIS), regardless of whether they are physical robots or software systems in real, virtual or mixed-reality environments.

[21] IEEE (2023).

Values-based principles

Table 4.2

1. Human rights	Systems should be created and operated to respect, promote and protect internationally recognised human rights.
2. Well-being	Systems should adopt increased human well-being as a primary success criterion for development.
3. Data agency	Creators should empower individuals with the ability to access and securely share their data, to maintain people's control over their identity.
4. Effectiveness	Developers and operators should provide evidence of the effectiveness and suitability for purpose of the AI system.
5. Transparency	The basis of a particular AI system decision should always be discoverable.
6. Accountability	Systems should be developed and operated to provide an unambiguous rationale for all decisions made.
7. Awareness of misuse	Systems should guard against all potential misuses and risks during operation.
8. Competence	Systems should specify competences and operators should have the knowledge and skills required for secure and effective operation.

Source: adapted from IEEE (2023).

The Montreal Declaration for Responsible Development of Artificial Intelligence

The Montreal Declaration for Responsible Development of Artificial Intelligence **is based on 10 principles**. It covers areas similar to the two frameworks above and additionally emphasises **democratic participation, respect for autonomy** and **prudence during development**.

Values-based principles

Table 4.3

1. Well-being	AI systems must permit the growth of well-being for individuals such as living conditions, health, and the exercise of their mental and physical capacities.
2. Respect for autonomy	AI systems must be developed and used while respecting people's autonomy, and with the goal of increasing people's control over their lives and their surroundings.
3. Protection of privacy and intimacy	Privacy and intimacy must be protected from artificial intelligence system (AIS) intrusion, and data acquisition and archiving systems.
4. Solidarity	The development of AIS must be compatible with maintaining the bonds of solidarity among people and generations.
5. Democratic participation	AIS must meet intelligibility, justifiability and accessibility criteria, and must be subject to democratic scrutiny, debate and control.
6. Equity	The development and use of AIS must contribute to the creation of a just and equitable society.

7. Diversity inclusion	The development and use of AIS must be compatible with maintaining social and cultural diversity and must not restrict the scope of lifestyle choices or personal experiences.
8. Caution	Every person involved in AI development must exercise caution by anticipating, as far as possible, the adverse consequences of AIS use and by taking appropriate measures to avoid them.
9. Responsibility	The development and use of AIS must not contribute to lessening the responsibility of human beings when decisions must be made.
10. Sustainable development	The development and use of AIS must be carried out to ensure strong environmental sustainability of the planet.

Source: Université de Montréal (2018).

These principles have resulted in eight recommendations that provide guidelines for achieving digital transition within the ethical framework of the declaration. These include implementing audits and certifications, independent controlling organisations, ethics education for developing stakeholders, empowerment of the user and ecological sustainability.

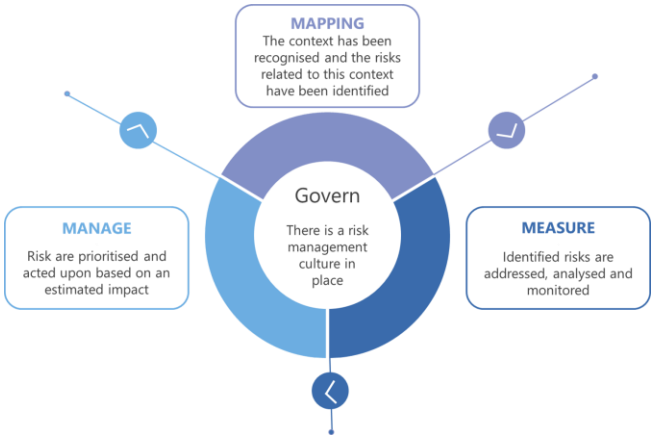
NIST *AI risk management framework* (AI RMF)

The US Department of Commerce's National Institute of Standards and Technology (NIST) has published this framework to help organisations to manage the risks of AI. The framework proposes two lines of action.

The first line of action is related to establishing specific processes to identify, assess and address risks associated with the development and use of a specific AI, aiming to enhance its safe use and reduce the likelihood of incidents related to its use.

The management process referred to in the first line of action is composed of four functions: govern, map, measure and manage (Graph 4.2).

Process functions to manage risks associated with AI — Graph 4.2



Source: adapted from NIST (2023).

The govern function is integral to the other three, providing the necessary framework to implement, follow up and control the entire risk management process associated with AI, while aligning it with the organisation's principles and policies. The map function provides context to identify risks associated with AI in terms of various factors related to its use, for instance stakeholders, values and interests. This function is the basis for the other two functions. The measure function addresses, analyses and monitors the AI risks and their impacts previously identified in the map function, and reports to the govern function by using different qualitative and quantitative tools and methodologies. Finally, the manage function's aim is to prioritise risks and to take action based on their estimated impact. This function includes developing plans to respond, recover and communicate about incidents or events in the AI development life cycle and use. Regarding the second line of action, NIST refers to implementing general processes to validate that AI systems are equipped with features that ensure their safe use and reliability. It also suggests that organisations should ensure that AI is used in accordance with principles that facilitate its governance, such as the ones described in the following table.

Characteristics of trustworthy AI — Table 4.4

1. Safe	The AI subject to evaluation must not pose risks to human life, health, property or the environment.[22]
2. Secure and resilient	The AI under evaluation and the ecosystems in which it operates must be resilient and able to tolerate adverse and unexpected events in its environment, maintain its functions and structure despite internal and external challenges and, if necessary, degrade safely.
3. Explainable and interpretable	AI must provide comprehensive information about its functionality and reliability, including its results, allowing humans to understand its behaviour, an attribute that is also referred to as contestability in specialised literature.[23] Similarly, AI must be implemented with transparency and accountability mechanisms regarding its operation.
4. Privacy enhanced	AI must safeguard human autonomy, identity and dignity, as well as protect information from intrusion. It should limit its observation to strictly necessary parties, and consider the rights of those who operate, monitor or use it to consent to the disclosure or control of their identities**.**
5. Fair – with harmful bias management	AI should minimise bias, and thus the damage caused to individuals, groups, communities, organisations and society as a whole. It is therefore important to consider multiple human and social values, such as transparency and fairness in this process. Methodologies have been developed to examine the role of human values in technological contexts, in line with the theoretical currents known as "values in design" and "value-sensitive design".
6. Valid and reliable	The AI under evaluation must provide objective evidence of compliance with specific use or application requirements. Additionally, it must function as intended, without failures, during a given time interval and under given conditions.
7. Accountable and transparent	Accountability presupposes transparency and both are necessary for trustworthy AI. Transparency provides and promotes access to information, increasing confidence in the AI system. Having organisational best practices and governance to address risks can contribute to more accountable AI systems.

Source: NIST (2023).

[22] ISO and IEC (2022c).

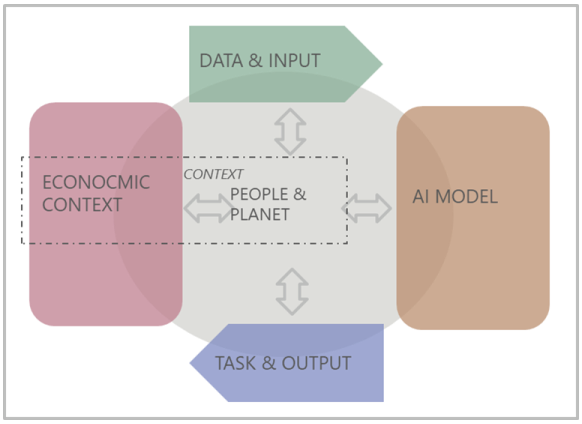
[23] Hirsch et al (2017)

OECD framework for the classification of AI systems[24]

This framework is a reference for "*policy makers, regulators, legislators and others to characterise AI systems for specific projects and contexts. The framework links AI system characteristics with the OECD AI Principles …, the first set of AI standards that governments pledged to incorporate into policy making and promote the innovative and trustworthy use of AI*."

This framework classifies AI systems and applications along five dimensions.

Key high-level dimensions of the OECD framework — Graph 4.3



Source: authors' elaboration.

The dimension **people and planet** considers different criteria such as users of the system, impacted stakeholders, human rights and democratic values, well-being, society and the environment including sustainability and stewardship.

Economic context refers to industrial sectors. It describes the business function and business model, the level of criticality regarding the extent to which a disruption impacts systems functions or essential services, and is also related to scalability and maturity (breadth of deployment).

The third criterion, **data and input**, refers to the detection and collection of data either by humans, automated sensors or both, the provenance of the data and input, as well as their nature (dynamic, static, updated or real time). Further, it covers proprietary, public or personal rights and the identifiability of personal data. ISO/IEC 19441 (2017)[25] distinguishes various categories – or "states" – of data identifiability.

The AI model dimension refers to model characteristics like AI model type, model building either from machine or human knowledge, model evolution, machine learning, central or federated machine learning and model inference about transparency and explainability.

Finally, the **task and outputs** criterion describes the tasks that the system performs, the level of autonomy and the role that humans play. It also considers core application areas such as human language, computer vision, automation/optimisation or robotics.

[24] OECD (2022).

[25] ISO and IEC (2017)

Each of the framework's dimensions have distinct properties and attributes that are relevant to assessing policy considerations associated with a particular AI system. The 10 OECD AI Principles, adopted in 2019[26], help to structure the analysis of policy considerations associated with each dimension and sub-dimension.

Values-based principles

Table 4.5

1.People and planet	Such as non-discrimination and equality, freedom, dignity, autonomy of individuals, privacy and data protection, diversity, fairness, social justice and internationally recognised labour rights. This **also includes addressing misinformation** and disinformation amplified by AI, while respecting freedom of expression and other rights and freedoms protected by applicable international law. **Addressing risks** arising from uses outside intended purpose.
2. Human rights, privacy and fairness	
3.Transparency and explainability	Commit to transparency and responsible disclosure regarding AI systems. Provide meaningful information appropriate to the context and consistent with state of the art technology to foster a general understanding of AI systems, including their capabilities and limitations; and to make stakeholders aware of their interactions with AI systems, including in the workplace.
4. Robustness, security and safety	AI systems should be robust, secure and safe throughout their entire life cycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety and/or security risks.
5. Accountability	AI actors should be accountable for the proper functioning of AI systems and for respect of the above principles, based on their roles, the context, and consistent with state of the art technology. **Ensure traceability**, **apply a systematic risk management approach to each phase of the AI system life cycle** including cooperation between different AI actors, suppliers of AI knowledge and AI resources, AI system users and other stakeholders.
Recommendations for AI policies	
6. Investment in research and development	OECD recommends increasing public and private investment to drive innovation and economic growth. Encourages stronger collaboration between industry, universities and research institutions to enhance knowledge. Also suggests prioritising strategic fields like health, green technologies and digital innovation for sustainable development.
7. Data, compute and technologies	Recommends creating an environment that ensures the responsible use of data and technologies. Highlights the need for investment in technology, the need for robust cyber security measures, and stresses the importance of education and training programmes, among others.
8. Enabling policy and regulatory environment	Recommends creating flexible and adaptive regulatory frameworks that can adapt rapidly to disruptive technologies, advocates for intellectual property rights and recommends policies that promote fair competition and prevent monopolistic practices.

[26] OECD (2022).

9. Jobs, automation and skills	Emphasises the need for continuous upskilling programmes to prepare the workforce for the use of AI technologies, also recommends supporting workers transitioning between jobs or sectors particularly for those affected by automation. Encourage the creation of new job opportunities in emerging sectors, such as health and green energy.
10. International cooperation	Encourages international cooperation on data standards and practices to facilitate cross-border data flows and fosters global innovation.

Source: OECD (2022).

ISO/IEC 42001 – *Information technology – artificial intelligence – management system*[27]

This international standard provides the requirements and guidelines for establishing, implementing, maintaining and continually improving an AI management system within the context of an organisation. The document is intended to be helpful to any organisation regardless of the nature of their activities.

This standard consists of seven key components.

ISO/IEC 42001 key components Table 4.6

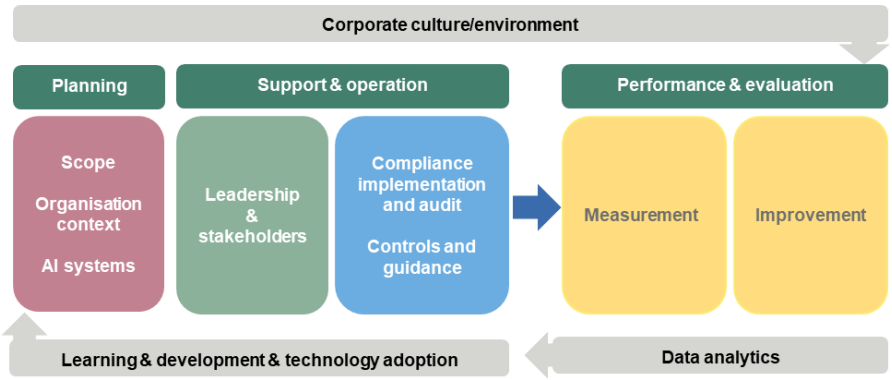
1. Context of the organisation	Organisations should be able to understand the internal and external factors that can impact their AI systems, including regulatory and contractual obligations. This involves identifying the needs and expectations of all stakeholders.
2. Leadership and commitment	The board must establish a privacy policy, assign roles and responsibilities, and ensure the integration of AI systems requirements into the organisation's current processes.
3. Planning	This involves identifying risks and opportunities related to the processing of personal data, setting objectives for the AI systems, and planning actions to address these risks and opportunities. This also includes compliance with applicable laws and regulations.
4. Support	This is related to procuring the necessary resources for the AI systems, including personnel with appropriate skills and competencies. This also involves ensuring awareness and communication about privacy policies and procedures, as well as maintaining documentation.
5. Operation	The standard outlines the processes and controls necessary to meet the requirements of the AI systems. This includes data protection by design and by default, conducting data protection impact assessments, and managing data breaches and incidents.
6. Performance evaluation	Organisations must monitor, measure, analyse and evaluate the performance of their AI systems. This implies conducting internal audits, management reviews and measuring the effectiveness of privacy controls and procedures.
7. Improvement	The standard highlights the need for continual improvement of the AI systems. Some activities are related to identifying discords and taking corrective actions, as well as proactively seeking ways to enhance privacy management practices.

Source: adapted from ISO and IEC (2023b).

[27] ISO and IEC (2023b).

ISO/IEC 42001 offers several benefits for organisations like helping them with regulatory compliance, and providing a structured approach to risk management to identify and mitigate risks related to AI systems and personal data. It also offers enhanced customer trust while providing operational efficiency and helping organisations to establish clear protocols for incident management.

ISO 42001 artificial intelligence management system — Graph 4.4



Sources: Oh (2024); ISO and IEC (2023b); OECD (2022).

ISO/IEC 23894 – *Information technology – artificial intelligence – guidance on risk management*[28]

In close connection with the approach proposed by ISO/IEC 31000 – *Risk management*, ISO/IEC 23894 integrates specific considerations related to the adoption of AI systems by the organisation into the risk management framework. With a focus on identification, assessment and mitigation of risks emerging from AI adoption, this standard provides recommendations to the organisation to ensure that AI systems are properly controlled and safe. Transparency of the risk management process, promotion of ethical AI use and continuous monitoring of AI risks are among the practices suggested by this standard, whose main stakeholders may include technical professionals such as IT managers, AI developers and risk managers.

In summary, the table below highlights the procedures required to adapt current risk management frameworks to cover the adoption of AI systems.

[28] ISO and IEC (2023a).

ISO/IEC 23894 key components Table 4.7

1. Principles of AI risk and impact assessment	Considering the benefits of an integrated risk management framework, this standard describes several principles that shall guide AI risk and impact assessments in organisations, such as dynamism, use of best available information, inclusivity, continuous improvement, and human and cultural factors.
2. Risk and AI systems management framework	The standard provides guidance on adapting the current risk management structure to cover AI systems, like ensuring risk integration to the business processes, maintaining the commitment of leadership towards the responsible adoption of AI, properly mapping external and internal contexts, defining roles, accountability and resource allocation.
3. Ethical impact assessment	The standard highlights potential impacts on individuals from AI implementation involving AI bias, personal data usage and even impacts on fundamental rights and physical security.
4. Social impact assessment	Populations may be affected in terms of their social and cultural values by AI development and usage, so it is important that organisations consider this kind of impact in their risk assessments.
5. Data governance and privacy risk assessment	Given the nature and amount of data required by AI systems, organisations should assess and mitigate risks related to data governance and privacy, particularly with the use of large data sets for training AI systems.

Source: adapted from ISO and IEC (2023a).

ISO/IEC 38507 – *Information technology – governance of IT – governance implications of the use of artificial intelligence by organisations*[29]

On the other hand, ISO/IEC 38507 discusses the consequences for an already existing governance structure as the institution considers the adoption of AI in its processes. The huge differentiation between traditional and AI systems and the recognition of their vast implications on the emerging relations between human and AI systems are the backdrop for the practices discussed in this standard, which provides guidance to the governing body for efficient, effective and acceptable adoption of AI within the organisation.

The document highlights the importance for the organisation to keep current governance pillars, like human supervision and accountability over automated decision-making processes, while being compliant with current, and potentially new, internal and external obligations (like ethically and/or legally defined usage of data, which covers privacy and data protection – even for training models). In order to ensure adherence to best practices in this new reality, the organisation shall consider building a clear and detailed map of AI systems (AI ecosystem) operating across the institution, which will provide the governing body with necessary information regarding data usage, model transparency and outcome explicability.

Finally, keeping human supervision and accountability over AI-aided decisions in all executive levels may require revision of internal commands that define the governance. Review of internal policies, updates to the risk management framework and impacts on the organisation's culture are expected in order to align safe AI usage to the organisation's strategic objectives.

This detailed standard has its main elements described in the following table.

[29] ISO and IEC (2022b).

ISO/IEC 38507 key elements

Table 4.8

1. Keeping governance	Given the possibility of a great impact on the organisation's activities, this standard provides guidance on assessing the adequacy of the current governance structure in the light of AI adoption.
2. Responsibilities of the governing body	Roles and responsibilities of the governing body (eg board of directors) in overseeing AI initiatives are highlighted in this standard. It emphasises the need to develop/review governance structures, decision-making authority and accountability frameworks in the face of greater technological dependency and the need for greater transparency and explainability brought by AI systems.
3. AI strategy and investment	The standard provides guidelines on how the governing body should define the organisation's AI strategy, ensuring that AI investments align with long-term goals, provide value and are in line with its risk appetite. It also suggests attention to new needs related to the technology, like improvement compliance supervision and control, assessment of the impact of usage of AI across the organisation, and special care with legal requirements and the consequences of deploying AI.
4. Policies review and decision-making governance	Governing body shall ensure that proper policies, responsibility chain and human supervision are in place for the controlled use of AI, since automated decision-making delivered by AI systems does not change the responsibility of the governing body.
5. Data governance	The standard recommends robust data governance for responsible and effective AI usage. Data collection, treatment and storage processes shall be enhanced, in order to ensure quality in data processing and output. Bias analysis shall also be performed.
6. Culture and values	As an AI system does not understand context (moral values and common sense etc) like humans, it is important that the governing body is explicit about the organisation's culture and values and is able to monitor, and when necessary correct, AI's behaviour.
7. Compliance	The governing body should seek assurances that management configures and maintains any AI system used by the organisation to meet compliance obligations and avoid compliance violations, such as pricing mechanisms that violate antitrust legal requirements or the use of data for training that violates civil rights or is discriminatory. Compliance management shall also be enhanced to cover new needs, like the sophistication of AI systems (new controls may be implemented) and human reassessment of decisions made by AI systems.
8. Risk management	As AI adoption provides risks and opportunities, it is paramount that the current risk management process review examines whether the risks involved are fully understood and managed, especially in decision-making processes, data usage, culture and values development, and compliance. If so, the governing body must be aware of the acceptability of those risks with regard to its stated risk appetite.
9. Objectives	The standard reminds the reader that the organisation's objectives and assets shall be carefully considered in an AI adoption scenario, like accountability, duty of care, physical safety, security and privacy, transparency and data itself (whose protection and integrity may be considered an organisational objective).

10. Sources of risks	The governing body shall be aware of risks that depend on the nature and domain where an AI system is deployed, as well as on the maturity of technologies used by the organisation. Other risks, already mapped on current IT processes, may emerge with new and unknown intensity, like risks related to data sources, poorly specified systems, value chain, undesired bias, lack of explicability, lack of experience in AI and cyber threats. Impacts on human professional autonomy, contractual and environmental issues may also be expected.

Source: ISO and IEC (2022b).

To sum up, these three ISO/IEC standards can be seen as complementary, each addressing a different layer of responsibility and scope within the life cycle of AI adoption, from operational management and risk assessment to high-level governance and strategic oversight.

Appendix 5: Members of the Artificial Intelligence Task Force

Co-chairs

Bank of Mexico	Alejandro De Los Santos
Bank for International Settlements, Basel	Angela O'Connor

Members of the task force

Central Bank of Brazil	Germano Machado
Bank of Canada	Stefan Smith
	Maryam Haghighi
Central Bank of Chile	María Jesus Orellana A
Central Bank of Colombia	María Carolina Trespalacios
Federal Reserve Bank of New York	Li He
Bank of Mexico	Alberto Mendoza
	Aldo Hernández
	Julieta Carmona
	Gloria Guadarrama
	Alfonso Murillo
Bank for International Settlements, Basel	Steffen Grosse
	Graham Cameron

Secretariat

Bank for International Settlements, Representative Office for the Americas	Christian Upper
	José Aurazo
	Torsten Ehlers