ARTIFICIAL INTELLIGENCE IN THE FINANCIAL SYSTEM: IMPLICATIONS AND PROGRESS FROM A CENTRAL BANK PERSPECTIVE

Iván Balsategui, Sergio Gorjón and José Manuel Marqués

BANCO DE ESPAÑA

https://doi.org/10.53479/38942

The authors belong to the Financial Innovation and Market Infrastructures Department. They are grateful for the feedback from José Manuel Carbó Martínez, Andrés Alonso-Robisco and an anonymous referee. Contact form for comments.

This article is the sole responsibility of the authors and does not necessarily reflect the opinion of the Banco de España or of the Eurosystem.

ARTIFICIAL INTELLIGENCE IN THE FINANCIAL SYSTEM: IMPLICATIONS AND PROGRESS FROM A CENTRAL BANK PERSPECTIVE

Abstract

The adoption of the Artificial Intelligence Act by the European Union, together with the emergence of large language models (LLMs) based on foundation models, or, more generally, of generative artificial intelligence (GenAI), has attracted renewed interest, within and beyond the financial sector, in the opportunities and limitations of this technology as a driver of change in society. This article aims to provide the context for recent developments and put them into perspective, identify possible road maps within the financial system and also set out what could hold back or spur progress in the medium and long term.

Keywords: generative artificial intelligence, foundation models.

1 Introduction

Artificial intelligence (AI) essentially consists of developing systems that perform tasks typically requiring human skills. Although it has a long history, AI progress is now driven by the recent advances in computing power and a major shift in approach. Specifically, this technology has moved from rule-based models to probabilistic models. Unlike traditional econometrics, the latter do not impose a specific functional form on the origin of the data. Instead, they directly infer what the underlying patterns are, thus improving the ability to predict or make decisions about a given problem.

These developments, along with a reduction in the cost of processing and storing information and the growing digitalisation of all types of data, have led to an increase in the number of sectors and use cases for which AI now provides solutions. The financial sector has been no exception to this phenomenon having, in recent decades, explored the possibility of implementing them in multiple scenarios (e.g. credit risk assessment and analysis) and internal management processes (e.g. automation) (Alonso-Robisco and Carbó, 2022).

However, users in general, and the financial system in particular, face many challenges. In practice, these solutions involve moving from a controlled scenario, where the gradual and selective implementation of rule-based models predominated, to one in which all business areas find potential applications that they would like to deploy as soon as possible, especially with the advent of proposals for GenAI models. The need thus arises for a framework that considers the risk implications (including possible regulatory requirements), analyses and optimises the management of the necessary resources and, ultimately, considers the strategic implications that the widespread use of this technology entails (Alonso-Robisco and Carbó, 2021).

In applying this technology, authorities and central banks face the same dilemmas as the private sector, with the particularity that understanding the risks and the mechanisms to contain them is paramount given their potential impact on financial stability. In particular, a dichotomy arises between the benefits of AI being adopted speedily and how this could in itself be a source of market imbalances, for example, for credit markets. Specifically, to highlight only some of the most prominent dangers, extensive AI use without proper safeguards could easily lead to cyber concentration risks and risks stemming from herd behaviours, with the consequent increase in market correlations. Therefore, when incorporating AI to comply with their respective obligations, a road map that takes into account and measures these implications must be defined, prioritising the areas where the impact is greater. For instance, there is a clearer need for AI to be adopted in the short term in fields such as the prevention of money laundering and financing of terrorism and the prevention or detection of cyber risks.

In addition, it is in the interest of central banks to be mindful of the impact that the widespread adoption of AI will ultimately have on issues such as the labour market, productivity and the degree of social inequality. To date, the preliminary analyses of this impact vary considerably. Some authors, such as Acemoglu (2024), consider the effect to be very modest, while others have arrived at the opposite conclusion (Cazzaniga, Pizzinelli, Rockal and Mendes, 2024). There is thus a wide range of possible outcomes, as shown in Aldasoro, Gambacorta, Korinek, Shreeti and Stein (2024), who propose analysing the effects using specific scenarios, and conclude that AI regulation must adapt as technological changes take form in the real economy.

This article does not address the latter aspects, but focuses specifically on the direct impact that AI has on the financial sector. After describing the changes that have taken place in the use of algorithms, it identifies the main activities where AI could have a significant impact and goes on to discuss the main risks posed and the responses that are being proposed by both regulators and sectoral authorities in general.

2 The world of AI in a nutshell[1]

Throughout its history, AI has seen periods of intense activity and progress combined with times of stagnation and even neglect (AI winters). As early as 1842, the mathematician and computer science pioneer Ada Lovelace could see that computers would go far beyond mere numerical processing (Carlucci Aiello, 2016). Later, Alan Turing would publish an article[2] in which he considered whether machines would be able to think or imitate the behaviour of the human mind. That was when the famous test of a machine's ability to convince a person that they are interacting with a human being was developed.[3] It was not until 2014 that a computer programme, the chatbot Eugene Goostman, managed to pass this test (Warwick and Shah, 2016).

[1] For a more detailed analysis, see Nayak and Walton (2024).

[2] "Computing Machinery and Intelligence" (1950).

[3] The Turing test.

Notwithstanding the above, the term “artificial intelligence” was first coined at the 1956 Dartmouth Conference to describe the science and engineering of making intelligent machines[4] (Moor, 2006). The concept was thus formally established, launching a new field of scientific study and trailblazing a first golden era that lasted until 1974.[5] The contrast between the inflated expectations about its possibilities and the practical limitations of the technology[6] led to the first winter, which was followed by a new boom (1980-1987).

The latter period was particularly fruitful with, among other things, the arrival of expert systems that, based on the explicit programming of “if-then” statements, were able to capture knowledge in certain subjects in order to be able to make informed decisions.[7] It also represented a new paradigm in the training of neural networks for probabilistic models, where they were able to minimise errors from learning training data, driving the resurgence of deep learning research.

The late 20th century and early 21st century saw significant new developments in AI, largely driven by an increase in computing power, more sophisticated machine learning algorithms and greater data availability to improve the training of models.[8] Lastly, the foundations of generative AI were laid in 2017, with the creation of a new neural network architecture (transformers) based on an “attention” mechanism (Vaswani et al., 2017), which afforded these algorithms renewed potential. This is based on models trained with a large set of heterogeneous data that, as a result, become capable of performing a variety of general tasks (foundation models), for example, the case of GPT (Generative Pre-trained Transformer), which can understand natural language. These models are subsequently used to train the algorithms that autonomously generate text, images or videos in response to questions made by the user through a prompt.[9]

The transition described above shows AI evolving from the use of analytical approaches towards generative models or, in other words, from applying logical rules and more specific programming to mirror human intelligence (symbolic AI) to trying to mimic the way the human brain works at a more structural level (connectionist AI) (see Figure 1). A clear example of this,

[4] The conference was organised by a group of computer scientists of the day, the original idea for which is attributed to John McCarthy.

[5] Notable research in this period includes the Perceptron, one of the first neural networks capable of recognising image and text patterns, developed by the psychologist Frank Rosenblatt in 1957, and ELIZA, the first-ever chatbot, invented by Joseph Weizenbaum between 1964 and 1966.

[6] For instance, during the Cold War, it was thought that AI would be able to produce machine translations, but this was not the case. Similarly, the results of the experiments conducted on neural networks fell short of expectations.

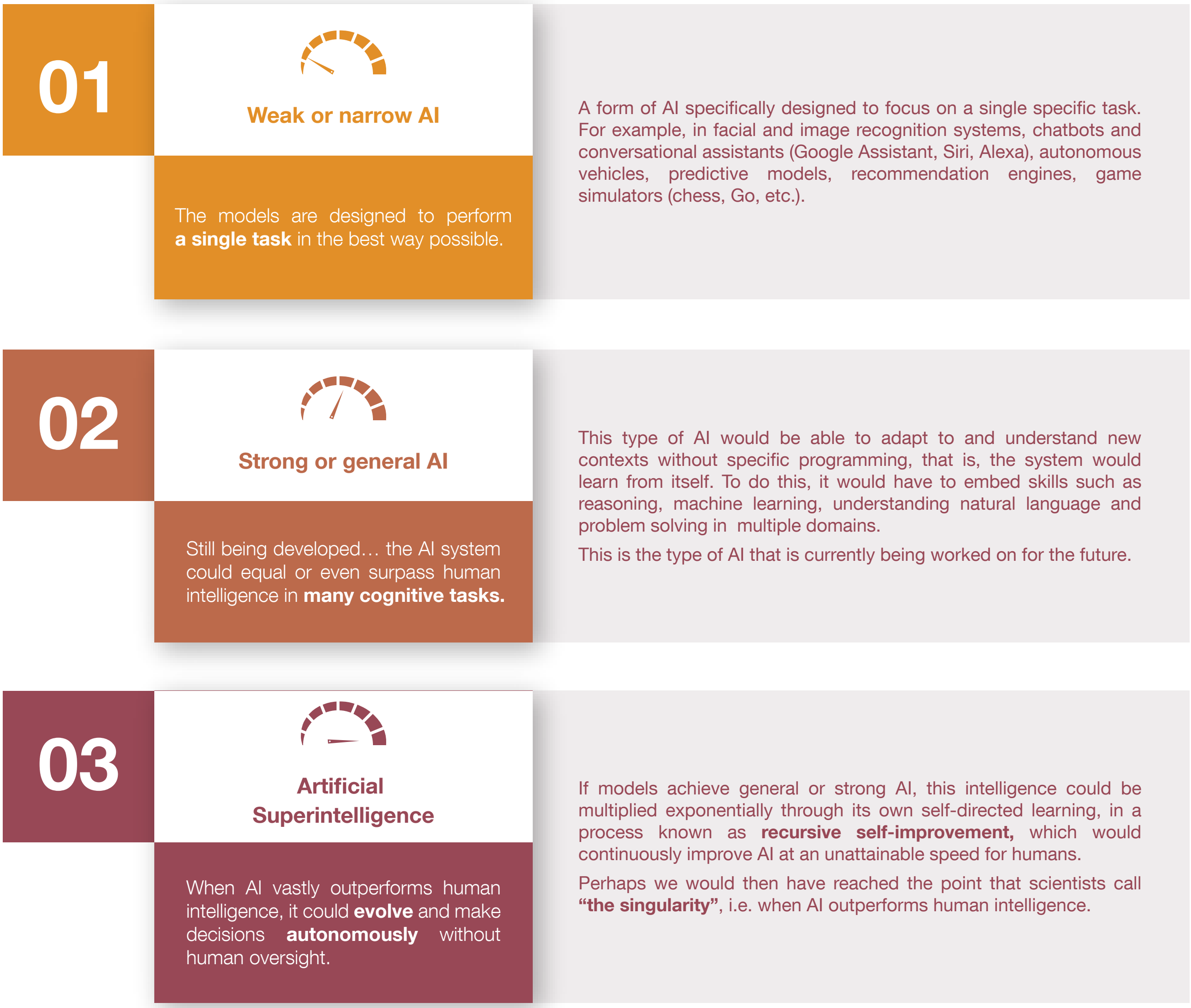
[7] They were very popular at the time and, as we will see later, they are at the core of what would come to be known as symbolic AI.

[8] Some illustrative examples are the creation of IBM’s Deepblue chess programme, which beat chess champion Gary Kasparov in 1997; the presentation of IBM’s Watson system, which could respond to questions in natural language and outperformed contestants on the popular American quiz show Jeopardy!, or the launch of AlphaGo in 2016, which beat the world's top Go player Lee Seedol using deep neural networks and reinforcement learning techniques, in a historic five-game match.

[9] One of the milestones that has undoubtedly contributed to this new GenAI hype dates back to 2022, when the company OpenAI launched ChatGPT, an AI chatbot (far superior to the ELIZA chatbot of 1966) based on the GPT foundation model, which has become the fastest growing web app in history. In just two months it reached more than one hundred million active users, a figure that other types of technologies such as TikTok took nine months to attain (or up to two and a half years, in the case of Instagram).

Figure 1
Types of AI

Depending on the degree of development of the models, the generality of the algorithms developed and the ability of AI to outperform human reasoning in all facets of life, AI can be classified into three different types:



01

Weak or narrow AI

The models are designed to perform **a single task** in the best way possible.

A form of AI specifically designed to focus on a single specific task. For example, in facial and image recognition systems, chatbots and conversational assistants (Google Assistant, Siri, Alexa), autonomous vehicles, predictive models, recommendation engines, game simulators (chess, Go, etc.).

02

Strong or general AI

Still being developed… the AI system could equal or even surpass human intelligence in **many cognitive tasks.**

This type of AI would be able to adapt to and understand new contexts without specific programming, that is, the system would learn from itself. To do this, it would have to embed skills such as reasoning, machine learning, understanding natural language and problem solving in multiple domains.

This is the type of AI that is currently being worked on for the future.

03

Artificial Superintelligence

When AI vastly outperforms human intelligence, it could **evolve** and make decisions **autonomously** without human oversight.

If models achieve general or strong AI, this intelligence could be multiplied exponentially through its own self-directed learning, in a process known as **recursive self-improvement,** which would continuously improve AI at an unattainable speed for humans.

Perhaps we would then have reached the point that scientists call **“the singularity”**, i.e. when AI outperforms human intelligence.

SOURCE: European Commission (2024).

within the financial system, is the use of symbolic AI models by marketing departments to determine aspects such as bank customers’ propensity to purchase certain financial assets based on their banking history, or anticipating the loss of bank customers in order to take preventive measures (churning). The financial system is also using connectionist AI systems to employ models that better detect fraud in the areas of compliance and fraud, to develop chatbots as internal support tools for the management of customer portfolios by sales department staff, and to have generative models that boost efficiency in software engineering when writing and documenting code or writing test cases, among other uses.

3 The opportunities of artificial intelligence for the financial system: a snapshot

In recent years (especially following the outbreak of the pandemic and the resulting transition to a new digital scenario), AI has gradually taken hold as an important pillar of the financial sector's value chain.[10] The use of AI varies geographically and across institutions but has always been linked to the specific nature of the banking business, very much centred on data exploitation and analysis.

In the case of the financial system, in addition to the technological factors described in the previous section, certain idiosyncratic elements have played an important role, such as the steady narrowing of profit margins and the consequent need to find new sources of efficiencies and income or the competition from fintechs (Boukherouaa et al., 2021).

Similarly, the emergence of GenAI foundation models has given renewed impetus to the adoption of tools that provide affordable access to pre-trained models with a broader purpose,[11] and also enable user-machine interaction formulas based on natural language rather than on mastering the notions of programming.

As has happened in other sectors, there are many and diverse reasons for the interest that AI has elicited among financial institutions, mainly associated with the promise of productivity gains, lower operating costs and enhancing the quality and safety of products, services and processes. In the same vein, the appetite shown by these players has much to do with the potential for optimising returns on investments, raising customer satisfaction rates and boosting the levels of financial inclusion (Fernández, 2019). Here, AI cannot only help complement credit assessment where other, traditional methods have limitations, but can also be harnessed to increase banking penetration levels, insofar as it can provide assistance to carry out formal procedures or select the most appropriate service providers for each case, aspects that, along with the costs involved, are often the main obstacles to access.

In this setting, banks' use of machine learning tools is particularly notable on the following fronts, to name a few:

— To facilitate compliance with banking regulations (for example, statistical reports, prevention of money laundering or terrorist financing), helping to minimise errors thanks to workflow automation, meet deadlines for submission, and provide better quality, clearer, more granular and precise information.

[10] In fact, the European Banking Authority estimated that almost 80% of the region's banks reported using AI tools proactively for different purposes in 2022. This figure was over 98% for institutions that were considering the use of AI or were in the early stages of AI development (European Banking Authority, 2023). However, other sources based on quantitative indicators point to still-low adoption rates.

[11] This leads to significant savings in terms of the investment required to develop and adapt the models internally, the costs associated with training them and the estimated time-to-market to successfully complete their effective implementation.

- To optimise, more generally, the internal business processes with a view to cutting costs, either by reducing the number of manual tasks and freeing up resources for other activities, or by improving staff performance through the provision of informed assistance for the processes they manage (e.g. advisory services to customers).

- To contribute to more effective and diligent control of banking risks, both from an operational (e.g. early detection of and response to potential fraud or mistakes) and financial standpoint (e.g. assessment and monitoring of solvency and probability of default or prediction of cash flows)[12] and in terms of cyber security (e.g. reducing false positive rates and identifying non-trivial anomalies or correlations).

- To boost analytical and predictive power in relation to market events in order to maximise investment returns amid uncertain volatility. These tools enable more accurate and robust estimates based on unstructured data from non-traditional sources and incorporate real-time information to generate knowledge that enables dynamic decision-making.[13]

- To enhance user experience, including the marketing and customisation of products and services to attain higher levels of customer satisfaction and retention, and speeding up of pre- and post-sale processes and expanding general access to banking services through the automation of certain interaction channels (chatbots).

Despite the apparent breadth of uses, the initiatives currently in production are mostly centred on back-office and middle-office functions. This is a logical decision is informed by the intention to contain potential risks associated with shortcomings in or disputes over the use of this technology with end-customers (Aldasoro, Gambacorta, Korinek, Shreeti and Stein, 2024), in an environment in which the regulatory framework for using these techniques has not yet been consolidated and fully implemented.

Central banks and sectoral authorities have also shown a clear desire to adopt AI tools with a view to better performing their functions.[14] In addition, they are interested in exploring, through first-hand experience, the inherent opportunities and risks to be better placed to understand and assess the real impact of these tools on the financial sector.

[12] Potential gains include reducing potential losses (Khandani, Kim and Lo, 2010), the optimisation of capital (Fraisse and Laporte, 2022), automating decision-making (Owolabi, Uche, Adeniken, Ihejirika, Bin Islam and Chhetri, 2024) and increasing turnover by expanding the base of approved credit applications (Sadok, Sakka and Maknouzi, 2022).

[13] By way of example, GenAI is capable of strengthening price discovery mechanisms or lowering the barriers to entry to less liquid markets, such as corporate debt or emerging markets.

[14] Both in their traditional roles (microprudential and macroprudential supervision, conduct supervision, oversight of payment systems or economic analysis) and in newly emerging roles, such as financial innovation, financial inclusion and environmental protection (Carstens, 2019).

Against this backdrop, the ongoing access to large amounts of complex and granular information, combined with the increase in the technological resources available to them (Cipollone, 2024), and the attainment of other strategic goals,[15] have all been key drivers in this transformation. Hence, areas such as statistics, macroeconomic analysis, oversight of payment systems and supervision are precisely those that have seen the most significant progress (Araujo, Doerr, Gambacorta and Tissot, 2024).

Starting with the latter, the benefits of AI will materialise insofar as, in a context marked by changing risks and more heterogeneous banking service providers, these tools truly serve to boost the effectiveness of actions aimed at monitoring the soundness, solvency and conduct of financial institutions to serve the objective of safeguarding the stability of the entire system (Beerman, Prenio and Zamil, 2021). Thus, the early detection of latent risks by capturing the hidden anomalies/signals in the reported data or greater accuracy when simulating and identifying the consequences of adverse scenarios (stress testing) are inherent advantages that all these authorities agree should be investigated.

Similarly, this greater analytical capacity benefits modelling activities aimed at determining the nature, scale and potential impact of structural imbalances on the economy, thanks, in particular, to its effectiveness in identifying non-linear relationships (Hellwig, 2021). Specifically, AI helps to more accurately establish the patterns that may underlie the relationships between variables without, however, inferring their causality,[16] and, hence, to construct more sophisticated and representative models of agents' behaviour (Atashbar and Shi, 2023). The exploitation of granular, unstructured[17] and continuously accessible data is essential for the production of useful economic indicators that enable public policies to be deployed when they will be most effective (Doerr, Gambacorta and Serena-Garralda, 2021).

In the same vein, AI creates the conditions for strengthening the monitoring of payment circuits, supporting the search for unexpected patterns in transactions in order to be able to warn of the presence of fraudulent or unlawful activity or to anticipate liquidity or operational problems whether at the level of individual banks or in the system as a whole (Rubio, Barucca, Gage, Arroyo and Morales-Resendiz, 2020). Although their implementation is as yet limited, these models make it easier to identify the channels through which systemic shocks can be transmitted, thus helping to counter their potential effects by accelerating the response measures and targeting the heart of the problem.

[15] For instance, the Banco de España's Strategic Plan 2020-2024 specifically included as one of its priorities the promotion of technological innovation with a view to modernising the institution, focusing especially on advancing the digital transformation, integrated information management and managing cyber security risk. Developing and rolling out suptech tools was just one of the actions launched for this purpose.

[16] Particularly in non-linear cases, as shown in Bahrammirzaee (2010).

[17] For instance, the use of web scraping techniques to collect information about prices, the exploitation of satellite data to proxy economic activity or the analysis of non-economic microdata shared daily by firms and individuals on social media (Shabsigh and Boukherouaa, 2023).

Enhancing the quality and functionality of data is another area in which AI's differential value is clear. This has become increasingly important as the volume, detail, complexity and frequency of information has grown, favouring the implementation of automatic validation processes to detect errors, outliers or significant omissions (Araujo, Bruno, Marcucci, Schmidt and Tissot, 2022). In addition to identifying information gaps, these tools can be used to try to remedy the negative effects linked to the information gathering process itself, such as the shortcomings associated with small data samples.[18] By combining different techniques, machine learning enables the cleansing, imputation and modelling of missing data, including their interpolation. This improves the models' robustness and potentially prevents the "overfitting" that may arise in the training phase (Rebuffi, Gowal, Calian, Stimberg, Wiles and Mann, 2021).

Lastly, AI has a great deal of scope for enhancing the effectiveness of central banks' interactions with the general public. This includes, for example, adapting their communication strategies, paying attention to the language and content used (Bholat, Broughton, Parker, Ter Meer and Walczak, 2018), broadening their scope thanks to the machine translation opportunities provided by recent developments (Cipollone, 2024) and simplifying the legislation they are responsible for drafting (Moreno, Gorjón and Hernáez, 2021).

As regards the deployment of solutions based on GenAI models, a different scenario emerges, given that the scope, potential and implementation of GenAI, as we have mentioned earlier, require a strategic approach and a greater degree of coordination, while managing the challenges and risks described in the following section. Current developments are mostly limited to tests or pilots, and their application is confined to internal use cases, aimed at improving business processes. In general, these tools are used in isolated environments for reasons of security and protection of sensitive data, and their role is limited to that of virtual assistants, seeking to enhance human decision-making capacity (Organisation for Economic Co-operation and Development, 2019).

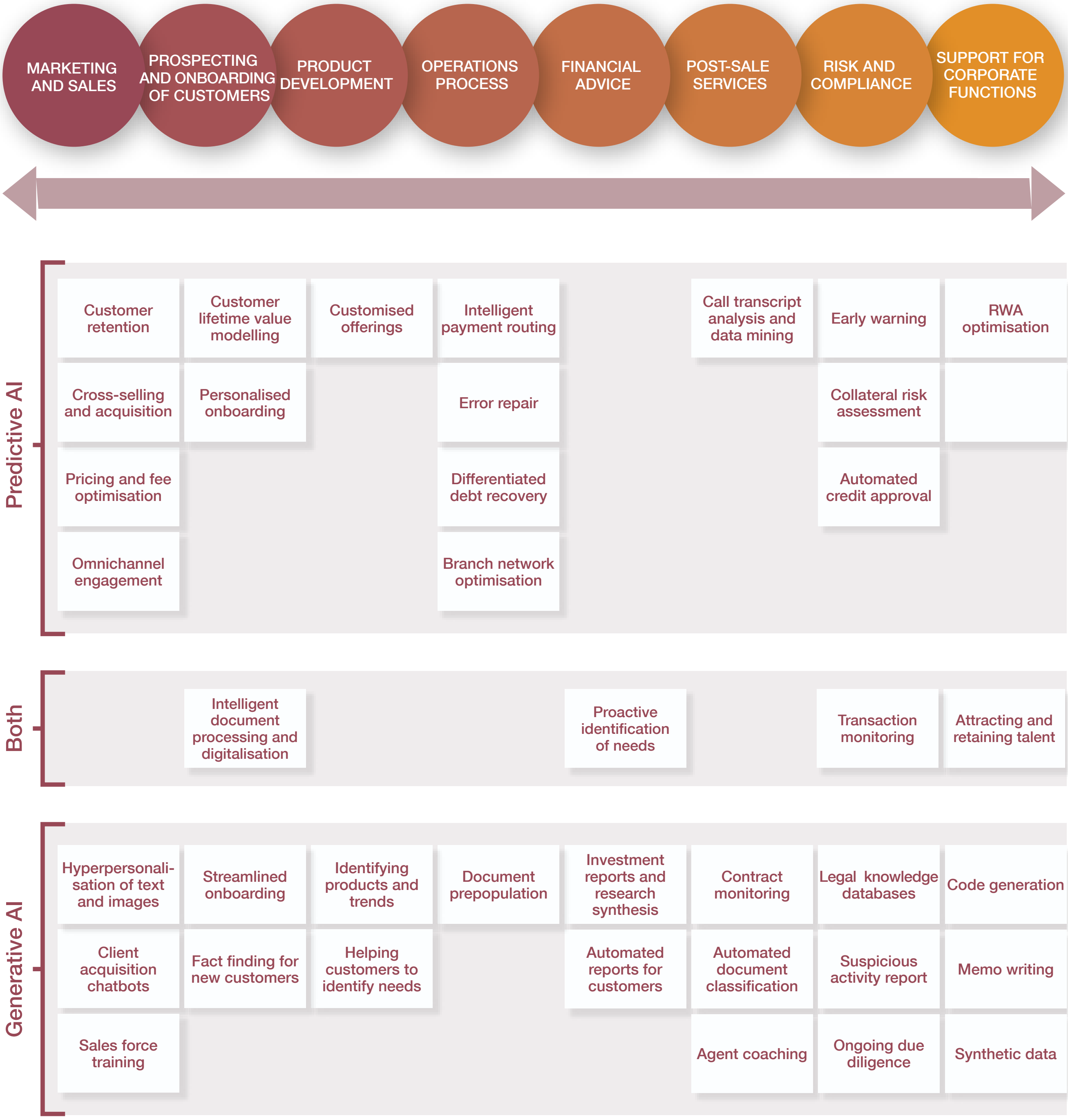
Other examples include the automated generation of programming code, as well as the testing and debugging of existing ones. Similarly, there are, at present, a large number of initiatives in the areas of text summarisation and machine translation.

Finally, more and more institutions are capitalising on these tools to transcribe and analyse conversations with customers in real time and assist managers in resolving incidents or marketing more customised products and services. Nonetheless, GenAI's potential goes much further and, once key governance and data management-related issues have been addressed, its influence will in all likelihood increase and spread to other areas rapidly (see Figure 2).

In this connection, initiatives aimed at optimising marketing and sales efforts have already been documented, both at the level of communication and the design of the commercial

[18] Bias and noise, among others.

Figure 2
Areas to which IA can be applied in the financial sector



	MARKETING AND SALES	PROSPECTING AND ONBOARDING OF CUSTOMERS	PRODUCT DEVELOPMENT	OPERATIONS PROCESS	FINANCIAL ADVICE	POST-SALE SERVICES	RISK AND COMPLIANCE	SUPPORT FOR CORPORATE FUNCTIONS
Predictive AI	Customer retention	Customer lifetime value modelling	Customised offerings	Intelligent payment routing		Call transcript analysis and data mining	Early warning	RWA optimisation
	Cross-selling and acquisition	Personalised onboarding		Error repair			Collateral risk assessment	
	Pricing and fee optimisation			Differentiated debt recovery			Automated credit approval	
	Omnichannel engagement			Branch network optimisation				
Both		Intelligent document processing and digitalisation			Proactive identification of needs		Transaction monitoring	Attracting and retaining talent
Generative AI	Hyperpersonalisation of text and images	Streamlined onboarding	Identifying products and trends	Document prepopulation	Investment reports and research synthesis	Contract monitoring	Legal knowledge databases	Code generation
	Client acquisition chatbots	Fact finding for new customers	Helping customers to identify needs		Automated reports for customers	Automated document classification	Suspicious activity report	Memo writing
	Sales force training					Agent coaching	Ongoing due diligence	Synthetic data

SOURCE: Devised by authors drawing on Riemer et al. (2023).

offering. The idea is to tailor both aspects to individual customer experiences. Likewise, improving the capacity to identify possible fraud or unlawful activity is increasingly accompanied by the automatic completion of the associated reports. That said, in the future, front office is where it could make most difference, provided its interactions become indistinguishable from human ones.

4 Most significant risks and barriers

As already mentioned in Section 2 above, 2022 marked a turning point in the resurgence of GenAI. Since then, a number of LLMs have been launched, including ChatGPT3, ChatGPT4, Claude, Gemini, Llama or Mistral, with advanced capabilities in natural language processing (NLP), text and image generation and, ultimately, creative content.

The impact that this modern technology could have soon became apparent despite it still being in its infancy. The race for companies to offer a competitive edge with this type of solution should be accompanied by management of the inherent risks. In the United States, the National Institute of Standards and Technology published its *AI Risk Management Framework* (AI RMF 1.0) as early as 2023, to help companies efficiently manage such risks.

An article by Gartner, published in 2023, on the requirements for using AI safely and effectively, highlights that: "By 2026, organizations that operationalize AI transparency, trust and security will see their AI models achieve a 50% improvement in terms of adoption, business goals and user acceptance."

Although more traditional AI models may present challenges in this direction (Alonso-Robisco and Carbó, 2022), and given the novelty and potential scope of GenAI, we will focus in this section on the specific risks its use entails. Specifically, the different types of potential problems posed by GenAI notably include:

— *Accuracy of model output:* according to an article published by McKinsey in May 2024, entitled "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value", the main risk that organisations encounter when using GenAI models is their lack of accuracy. What happens when a GenAI system generates text, images and other output that are inaccurate, incorrect, misleading or inappropriate? This may be due to the poor quality of the data used to train the model, or even to the configuration of the system itself or the inherent lack of explainability of some GenAI models, all of which can lead to the wrong decisions, with the implications this could have for an organisation's business process.[19] To mitigate these situations, it is especially important to have sandboxes or frameworks to check that the two most significant components of any GenAI system function correctly: that which provides context information (retriever) and that which processes and generates the response (generator).

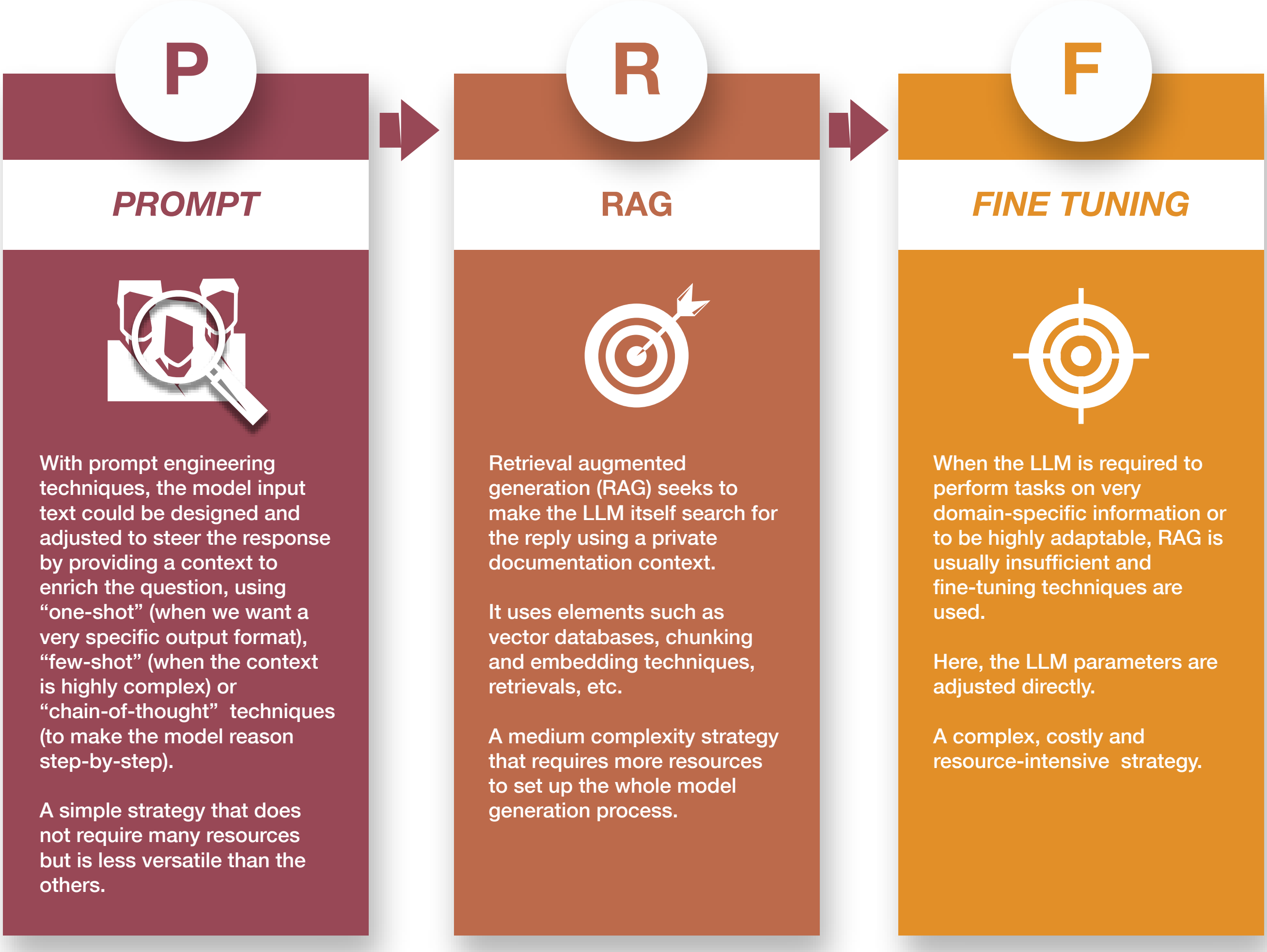
— *Model hallucinations:* this phenomenon occurs when a model generates information that appears to be plausible, but which is in fact made-up and not based on the information drawn from the model's training data. Owing to their nature, LLMs will

[19] By way of example, worth noting is the case of Hong Kong business tycoon Li Kin-kan, who in 2017 lost up to 20 million dollars a day through a trading bot called K1, due to the decisions that the GenAI model generated based on various sources of information and which it subsequently transferred to market operations. The tycoon sued the London-based investment firm Tyndaris Investment, which marketed this smart trading bot.

Figure 3

Strategies to reduce LLM hallucinations

Three strategies can be applied to reduce the risk of hallucinations in LLM outputs. The most straightforward strategy is prompt engineering (“P”), followed by retrieval augmented generation (RAG) (“R”), and the most complex, fine-tuning (“F”) of the model itself.



P

PROMPT

With prompt engineering techniques, the model input text could be designed and adjusted to steer the response by providing a context to enrich the question, using “one-shot” (when we want a very specific output format), “few-shot” (when the context is highly complex) or “chain-of-thought” techniques (to make the model reason step-by-step).

A simple strategy that does not require many resources but is less versatile than the others.

R

RAG

Retrieval augmented generation (RAG) seeks to make the LLM itself search for the reply using a private documentation context.

It uses elements such as vector databases, chunking and embedding techniques, retrievals, etc.

A medium complexity strategy that requires more resources to set up the whole model generation process.

F

FINE TUNING

When the LLM is required to perform tasks on very domain-specific information or to be highly adaptable, RAG is usually insufficient and fine-tuning techniques are used.

Here, the LLM parameters are adjusted directly.

A complex, costly and resource-intensive strategy.

SOURCE: Devised by authors.

always return a text sequence unless explicitly told otherwise. There are three complementary strategies to avoid these hallucinations: prompt construction engineering techniques to communicate with the model, using techniques such as Retrieval Augmented Generation (RAG) and/or carrying out a process of calibration or fine-tuning of the model itself. All three have their pros and cons (see Figure 3).

— *Data privacy:* Here, the following situations should be addressed:

- GenAI is essentially driven by an ever-expanding amount of available data in an increasingly digitalised world. The problem is that some of the information used to train the models could contain personally identifiable information (PII). As a result, simply using such data to train models or including them in the output could disclose confidential information about individuals and lead to a breach of privacy and possible misuse. Today, different solutions and techniques to mitigate this risk are available on the market.[20]

- In addition, a technique called "membership inference attack", used to attack LLMs, can reveal the original data used to train the AI models (including personal information) without having information about the architecture of the model or the parameters that have been used within the model itself, which could lead to a breach of data privacy (Shokri et al., 2017).

- Moreover, when a user of these public GenAI models writes or pastes confidential information in a prompt mistakenly, negligently or intentionally, this information automatically becomes available to the creator/manager of these models and, potentially, to other users.[21]

— *Copyright, intellectual property:* as mentioned above, GenAI models are trained using huge volumes of data. These data could potentially include copyright protected material, which would entail complying with the copyright legislation in force in order to use such data. If the owner of the GenAI model fails to do so, they would be infringing intellectual property rights. The creation of new content that very closely resembles existing works could open the door to legal disputes over its originality and ownership in the event of possible plagiarism.[22]

— *Biased results:* most GenAI models are mainly trained on data of more than questionable quality from the internet. It is important to stress that a GenAI model can produce biased results, generally because of bias found in the training data. As a first step towards mitigating this issue, companies should prioritise the use of high-quality reliable data sources when training their GenAI models, avoiding to the extent possible datasets that might include racist[23] or sexist approaches. In addition, the opacity of GenAI models does not exactly help to mitigate this risk. However, fields such as explainable AI (XAI) that address a model's interpretability and explainability have emerged (Arrieta et al., 2019).

[20] Google, for example, has a data loss prevention solution to ensure that all PII is automatically identified and anonymised before training the models. In addition, techniques such as differential privacy, homomorphic encryption, secure multi-party computation or federated learning can be used to protect privacy when using AI models. Another concept that is currently being used to protect privacy is the data privacy vault, e.g. the Protecto Vault service provided by the firm AI Protecto, which uses tokenisation.

[21] In 2023 a Samsung engineer uploaded sensitive internal code to ChatGPT, automatically leading to a disclosure of confidential information to third parties, with the consequent reputational impact on the company.

[22] A clear example is the legal action brought by *The New York Daily News, The Denver Post* and *The Chicago Tribune* against OpenAI and Microsoft for copyright infringement.

[23] For example, in 2015 Amazon's software developer recruitment model, which was trained on biased data, assigned a greater weight to men's CVs, skewing the model's final results.

— *Model security:* GenAI models can be used for a wide range of security applications. While the first thought that springs to mind is these models being used to buttress organisations' current defences, it is also true that hackers can use them to make their attacks on third-party infrastructure more sophisticated. The most relevant security risks are summarised below:

- Phishing attacks on users have become exponentially more sophisticated as a result of GenAI being used to improve the quality of the drafting (tone, style, format) and content of the texts of malicious emails sent to third parties. GenAI models more than make up for poor grammar or drafting, meaning that it is increasingly difficult to distinguish between a legitimate and a bogus email. In addition, attackers can use GenAI to write, develop, fine-tune and overhaul malware capable of bypassing traditional security measures.

- Given that GenAI systems generally comprise a variety of different data sources, application programming interfaces and other systems, and since integration tends to be complex, the attack surface can be larger and create vulnerabilities that attackers can capitalise on to access a company's confidential information. The OWASP Top 10 for Large Language Model Applications, a list of recommendations that organisations can take into account to minimise this attack risk, are especially important (see Figure 4).

- Deepfakes are another interesting case. Current GenAI models are clearly capable of generating manipulated images and videos, voice cloning and all manner of hyperrealistic content that can be very hard to distinguish from real content, which could cause incorrect information to be shared intentionally and even to influence public opinion.[24] These issues raise ethical concerns that should be considered in all GenAI model risk governance processes.

- It should be borne in mind that, irrespective of the possible external attacks that these types of models might suffer, care should also be taken with internal security systems to avoid employees using data exfiltration techniques to "extract" information from the organisation for their own gain (e.g. by using prompt injection).

- Another point to be considered is assessing how and with what data LLMs have been trained, thereby avoiding "backdoors" that could be used to provide an attacker with unrestricted illegal access to systems where LLMs are stored.

— *Third party conformity:* when designing a GenAI model that manages above all confidential information and is deployed in the infrastructure of an external AI service provider, it is key that such provider have compliance certificates to at least provide

[24] For example, in 2024 US President Joe Biden's voice was used in fake robocalls to discourage voting in the New Hampshire primaries.

Figure 4
Main vulnerabilities of LLM use

LLM01:
Prompt injection
When an attacker manipulates an LLM through crafted inputs, causing the LLM to unknowingly execute the attacker's intentions.

LLM02:
Insecure output handling
Refers specifically to insufficient validation of the outputs generated by LLMs before they are passed downstream to other components and systems.

LLM03:
Training data poisoning
Manipulation of training data to embed attack vectors (backdoors, etc.) or false information.

LLM04:
Model denial of service
A user manages to cause a decline in the quality of the service or crash it by increasing the resources consumed.

LLM05:
Supply chain vulnerabilities
Vulnerabilities within all model construction, training and deployment processes.

LLM06:
Sensitive information disclosure
Vulnerability resulting in unauthorised access to sensitive data, intellectual property and privacy violations.

LLM07:
Insecure plugin design
Vulnerabilities found when LLMs use insecure third-party plugins.

LLM08:
Excessive agency
Vulnerabilities that exist because the model or the agents the model uses have excessive functionality, excessive permissions or excessive autonomy.

LLM09:
Overreliance
Vulnerabilities that occur when an LLM hallucinates if the LLMs are not cross-checked against reliable sources.

LLM010:
Model theft
Intellectual property vulnerability that occurs when an LLM has been "copied" unlawfully, extracting data from the training parameters to simulate the workings of the original LLM.

SOURCE: OWASP (2023).

assurance that, a priori, they are not in breach of any compliance or data protection rules and that the possible uses of the GenAI models do not serve to retrain them.

— *Cost of experience and computing:* when developing, using and maintaining GenAI models it should be borne in mind that these systems require specialised hardware, typically cloud services, which can be costly. In addition, given that it is a "relatively new" technology, profiles such as data scientists, machine learning engineers and prompt engineers demand higher salaries. If we also factor in that such profiles are quite thin on the ground, there are significant barriers to entry for many organisations.

There is no doubt that there are inherent risks in developing and using any emerging and disruptive technology, as is the case of GenAI. This means that companies must identify, assess and mitigate in advance, in order to be aligned with each of the above-mentioned aspects.

Further, the main technical challenges should also be considered in order to implement AI effectively in organisations.

Prior to the resurgence of GenAI, when organisations were undertaking machine learning projects, they followed the MLOps paradigm to improve processes in which machine learning-based software was developed, implemented and monitored. In this work environment, data scientists, machine learning model engineers, data engineers and experts in development, continuous integration, security and privacy work together, from defining data quality processes to model training, the roll-out in central repositories and the launch of systems.

However, with the arrival of GenAI the MLOps concept itself has evolved, and a new term – LLMOps, focused mainly on the management, implementation and maintenance of LLMs – has been coined. Due to their complexity and resource requirements, LLMs pose unique Ops challenges, such as selecting foundation models, prompting, benchmarking the models' results under a new system of indicators different from those used in machine learning, fine-tuning processes, model governance and observability, among others.

Undoubtedly, all these new paradigms that organisations have to implement do nothing but add to the list of complex tasks to integrate LLMs into their proprietary systems. This is compounded by the new computing needs, not just to train the models, but also to implement them either on-site or through cloud providers, which will certainly be key to the organisation successfully deploying them internally and externally.

5 Public and regulatory policy responses: some considerations

Considering the above-mentioned risks, and AI's potential impact on society, it should come as no surprise that the authorities have started to develop a public policy and regulatory framework geared to mitigating them. Both at national and at international level, many of the

Table 1

Notable aspects of the main regional regulatory initiatives

Europe	United States	Asia
Cross-cutting European regulation applicable both to providers of AI systems susceptible of being used in the EU (irrespective of where they come from) and to their respective users It establishes requirements (or, where appropriate, prohibitions) for AI systems that are proportionate to the risk posed by their intended use (specific technologies are not regulated) Obligations to which providers are subject include quality management systems, the preparation of technical documentation, record-keeping and conformity assessments Obligations to which users are subject include adequate human oversight, reporting serious incidents and malfunctioning and compliance with other legal requirements such as those under the General Data Protection Regulation	The National Artificial Intelligence Initiative Office was established in 2020 There is still no federal law regulating, prohibiting or restricting the development and use of AI Instead different types of legal acts have been enacted in different areas that address specific matters related to the application of AI The White House has launched some measures, with particular focus on matters such as access and fair use of AI systems or the development of foundation models and matters related to security The debate continues over legislative initiatives that address, for example, automated decision-making systems (transparency, right to not participate, non-discrimination), consider the possibility of requiring certain service providers to be licensed or protect intellectual property	Singapore has an AI governance framework that fosters explainability, transparency, fairness and the safeguarding of civil rights It has also issued guidelines on the use of personal data by AI In addition, it is finalising a set of recommendations on GenAI to create a trustworthy ecosystem The Monetary Authority of Singapore is developing a framework to manage GenAI risks in the financial system China has adopted rules aimed at specific technologies that address different types of AI risks There are provisions regulating the use of recommendation algorithms, banning them for minors and giving users the option to opt out Other provisions regulate the providers and users of deep synthesis-capable technologies (labelling content created using such technologies and restricting certain applications) Rules on GenAI have recently been enacted that protect intellectual property rights and require that measures be applied to ensure data quality, accuracy and reliability

SOURCE: Devised by authors.

measures are general in nature, and by extension therefore affect the financial system. However, for the time being, there are not many specific regulations for this sector. While there are differences in both the detail and the level of requirements, all these initiatives reflect common goals and ultimately aim to ensure that this technology is rolled out in as orderly a manner as possible (see Table 1).

One of the first focal points is the ethics of AI or, in other words, providing a framework that enables responsible ecosystems to be built to ensure that AI's outcomes are fair, inclusive, sustainable and non-discriminatory (United Nations Educational, Scientific and Cultural Organization, 2021). To this end, one of the major international benchmarks is the OECD's 2019 Recommendation,[25] which in turn gave rise to the G20 AI Principles. Its aim has been to provide a specific global standard, supplementing other more general standards equally applicable to this field,[26] that could inform the corresponding actions of the national authorities.

[25] Since then the Recommendation has undergone several revisions to facilitate its implementation and incorporate the technical and policy changes that have taken place, such as those stemming from the eruption of GenAI, to thereby ensure it remains valid.

[26] Mainly privacy and data protection, digital information security and conduct.

The OECD Recommendation is implemented through a set of principles that aim to facilitate the implementation of national policies and foster cross-border cooperation.[27] The G20 AI Principles encourage investment in the research and development of an open science that fosters knowledge and information sharing. They also advocate for the creation of a governance framework and public policies that spur innovation and help the transition to the deployment stage. Governments are also expected to pay special attention to workforce retraining and foster interdisciplinary dialogue, both domestically and across borders, to reach major agreements.

The universality of these premises has meant they have been echoed in numerous initiatives, such as the EU's Ethics guidelines for trustworthy AI or, subsequently, in the EU's Artificial Intelligence Act. Something similar can be said for other countries at the forefront of this field, such as the United States,[28] China,[29] Japan[30] and the United Kingdom,[31] to name a few. Despite their differing approaches, they all acknowledge the importance of holding an open dialogue that fosters cooperation and promotes adjusting the practical requirements to the actual level of risk of each specific application of AI.

Against this backdrop, Singapore has blazed a trail by reinterpreting the Recommendation to facilitate its application to the financial sector, which now has some specific principles.[32] This appears to be the route preferred by financial sector operators demanding specific criteria that provide them with the necessary assurance about the validity of the implementations to be made.

Another of the key components of the emerging public policy framework for AI is the above-mentioned Artificial Intelligence Act. With this initiative, Europe is stealing a march on other countries in terms of regulating potential AI uses in order to strengthen its open strategic autonomy in the development of the digital market and, above all, safeguard social well-being. To do so, it is adopting a human rights-based approach consistent with the region's values (Calderaro and Blumfelde, 2022).

The AI Act is cross-cutting and affects both the providers of AI systems that could be used in the EU (irrespective of where they come from) and their respective users. Its definition of AI is

[27] These include dimensions such as those related to AI contributing to improved social well-being, the transparency of AI systems for users and ensuring AI system robustness.

[28] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023). Further, starting with the 2022 Algorithmic Accountability Act, the US Congress appears to be outlining a definitive roadmap that could lay the foundations for a joint effort that helps shape a regulatory package on specific AI matters.

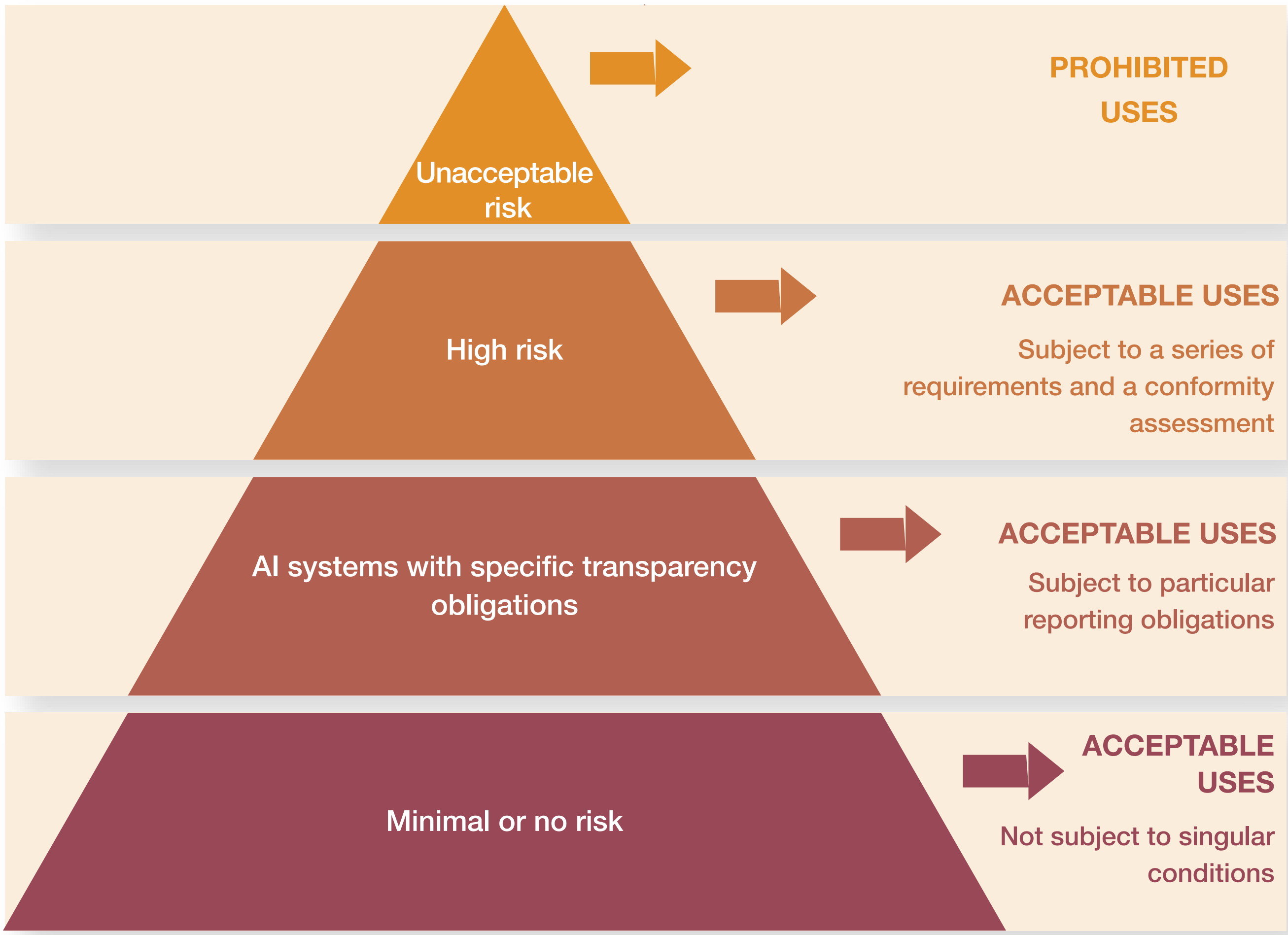
[29] Characterised by a gradual roll-out of laws focused on specific aspects of AI use, which, in each area, address ethics-related facets. Of note are the 2021 regulation on recommendation algorithms, the 2023 regulation on deep synthesis – to prevent deepfakes – and the 2023 law on generative AI. This set of regulatory initiatives was based, among others, on the 2019 general governance framework that laid the foundations for the principles applicable to a new generation of responsible AI (Sheehan, 2023).

[30] Social Principles of Human-Centric AI (2019).

[31] Ethics, Transparency and Accountability Framework for Automated Decision-Making (2023).

[32] Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector.

Figure 5
Types of AI risks in the AI Act


Unacceptable risk
PROHIBITED USES
High risk
ACCEPTABLE USES
Subject to a series of requirements and a conformity assessment
AI systems with specific transparency obligations
ACCEPTABLE USES
Subject to particular reporting obligations
Minimal or no risk
ACCEPTABLE USES
Not subject to singular conditions

SOURCE: European Commission (2024).

quite broad to thus accommodate developments such as foundation models.[33] The AI Act also establishes levels of risk based on the intended use of these tools, enabling the identification of a series of categories on which proportionate obligations are placed (see Figure 5).[34]

Turning to the financial system, high-risk AI systems and, in particular, those related to biometric identification[35] and credit scoring,[36] are particularly important. These, like any other in this category, will undergo a conformity assessment and be subject to certain particular requirements.

[33] During negotiations over the regulation this was subject to some debate because of the possibility of such a broad definition also including more traditional inference models. However, finally, there seems to be a consensus that this is not the case.

[34] Potentially abusive uses of AI, e.g. those that can distort human behaviour or lead to discriminatory outcomes, are prohibited.

[35] Without prejudice to the foregoing, biometric identification will remain subject to the General Data Protection Regulation (Regulation (EU) 2016/679) and the Law Enforcement Directive (Directive (EU) 2016/680).

[36] Risk assessment and pricing in the insurance sector are other affected areas.

Among other key aspects, the above-mentioned systems must guarantee: i) the deployment of specific risk management policies; ii) the implementation of a training and testing data management and governance model that is robust; iii) the preparation of technical documentation that provides sufficient evidence of compliance with requirements; iv) the transparency and traceability of decisions; and v) their appropriate supervision by ensuring human intervention. Further, the specific transparency risk systems must inform the user that they are interacting with a bot and not a person.

Although the principles underpinning the AI Act establish the general playing field, the secondary legislation will be tasked with clarifying its more practical aspects. This, combined with how effective the established mechanisms for coordinating the different supervisory authorities are, will without doubt be key to ensuring that the AI Act is implemented with the utmost consistency, that it avoids unnecessary frictions with the existing sectoral legislation and that it provides enough flexibility to enable the development of a technology still largely considered to be in its infancy.

The Ministry of Digital Transformation and the Civil Service's regulatory sandbox pilot project, which seeks to provide guidelines and best practices to facilitate the optimal application of the AI Act, is expected to contribute to achieving this same objective. Despite the AI Act's above-mentioned cross-cutting nature, various use cases exist that affect the financial system in particular. In this respect, it should be noted that the European Commission recently launched a public consultation on AI in the financial sector, to provide guidance to the financial sector for the implementation of the AI Act, given its connections with other legislation, such as that related to outsourcing and operational risk control.

Lastly, the varying approaches followed by each international jurisdiction are viewed as a potential source of fragmentation in an industry that, by its very nature, is not bound by borders. Therefore, initiatives seeking to advance coordinated efforts, like the recent agreement between the OECD and the Global Partnership on Artificial Intelligence, are on the rise.

6 Conclusions

Tech firms aside, the financial industry is perhaps the sector that, most broadly and profoundly, is harnessing the huge opportunities offered by AI. As part of the roadmap designed to transform information into knowledge and knowledge into intelligence, institutions are rapidly shifting from more traditional techniques, based on analytical approaches, to others that replicate human behaviour at a more structural level. Consequently, at present, these tools are expected to be able to address increasingly complex problems.

In this regard, AI is currently one of the drivers behind institutions being able to meet their most typical goals (e.g. productivity gains, better efficiency, lower costs and higher quality or safer products and services). In particular, with regard to the latter, deploying these tools to combat fraud or cyber threats is a matter of urgency, as the number of criminals using these same techniques to make significant ill-gotten gains is on the rise. In addition, a new generation of GenAI-based tools is gradually making its way. These tools are having a significant impact and could modify not only internal processes but also how organisations interact with customers and employees.

This scenario opens the door to the development and exploitation of complementary sources of income and offers new ways of maximising operating profits in a changing environment. However, seeing as institutions are deploying these new tools cautiously, with internal processes and applications largely the focus, it is still hard to gauge how important this contribution will be. As legislative changes take hold and are implemented, the impact is expected to be much more disruptive.

Similarly, financial authorities can also harness the potential of these technologies to perform their responsibilities and thereby help improve social well-being. Hence why many of them already have ambitious strategies and programmes to explore these technologies and eventually pave the way for their implementation.

However, the widespread use of these techniques also poses sizeable risks that need to be managed. In this respect, the main practical challenge facing authorities and users consists of deploying an appropriate and robust governance model that ensures that the technology is transparent and safe. This is the only way to guarantee that the public trusts it enough to smooth its adoption and acceptance en masse.

This is the reason for the proliferation of initiatives aimed at establishing a regulatory and supervisory framework that is conducive to the most orderly deployment possible. In short, the aim is to ensure that wherever it is applied, AI always provides fair, inclusive, sustainable and non-discriminatory results; this is something that affects the financial system in particular. In this respect, perhaps one of the aspects in which more headway needs to be made in the future is for legislative efforts at international level to converge somewhat, so as to prevent a phenomenon that by its very nature is global from being hampered by a multiplicity of fragmentary national rules.

REFERENCES

Acemoglu, Daron. (2024). "The simple macroeconomics of AI". *Economic Policy.* https://doi.org/10.1093/epolic/eiae042

Aldasoro, Iñaki, Leonardo Gambacorta, Anton Korinek, Vatsala Shreeti and Merlin Stein. (2024). "Intelligent financial system: how AI is transforming finance". BIS Working Papers, 1194, Bank for International Settlements. https://www.bis.org/publ/work1194.pdf

Alonso-Robisco, Andrés, and José Manuel Carbó. (2021). "Understanding the Performance of Machine Learning Models to Predict Credit Default: A Novel Approach for Supervisory Evaluation". Documentos Ocasionales, 2105, Banco de España. https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/21/Files/dt2105e.pdf

Alonso-Robisco, Andrés, and José Manuel Carbó. (2022). "Inteligencia artificial y finanzas: una alianza estratégica", Documentos Ocasionales, 2222, Banco de España. https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/22/Fich/do2222.pdf

Araujo, Douglas, Giuseppe Bruno, Juri Marcucci, Rafael Schmidt and Bruno Tissot. (2022). "Machine learning applications in central banking". Irvin Fisher Committee on Central Bank Statistics Bulletin, 58, Bank for International Settlements. https://www.bis.org/ifc/publ/ifcb57_01_rh.pdf

Araujo, Douglas, Sebastian Doerr, Leonardo Gambacorta and Bruno Tissot. (2024). "Artificial intelligence in central banking: Executive Summary". BIS Bulletin, 84, Bank for International Settlements. https://www.bis.org/publ/bisbull84.pdf

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila and Francisco Herrera. (2019). "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". *Information Fusion,* 58. https://doi.org/10.1016/j.inffus.2019.12.012

Atashbar, Tohid, and Rui Aruhan Shi. (2023). "AI and Macroeconomic Modeling: Deep Reinforcement Learning in an RBC Model". IMF Working Paper, 40, International Monetary Fund. https://doi.org/10.5089/9798400235252.001

Bahrammirzaee, Arash. (2010). "A comparative survey of artificial intelligence applications in finance. Artificial neural networks, expert system and hybrid intelligent systems". *Neural Computing & Applications,* 19, pp. 1165-1195. https://doi.org/10.1007/s00521-010-0362-z

Beerman, Kenton, Jermy Prenio and Raihan Zamil. (2021). "Suptech tools for prudential supervision and their use during the pandemic". Financial Stability Institute Insights on Policy Implementation, 37, Bank for International Settlements. https://www.bis.org/fsi/publ/insights37.pdf

Bholat, David, Nida Broughton, Alice Parker, Janna Ter Meer and Eryk Walczak. (2018). "Enhancing central bank communications with behavioural insights". Bank of England Staff Working Paper, 750, Bank of England. https://doi.org/10.2139/ssrn.3233695

Boukherouaa, El Bachir, Ghiath Shabsigh, Khaled AlAjmi, Jose Deodoro, Aquiles Farias, Ebru S. Iskender, Alin T. Mirestean and Rangachary Ravikumar. (2021). "Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance". IMF Departmental Paper, 2021/024, International Monetary Fund. https://doi.org/10.5089/9781589063952.087

Calderaro, Andrea, and Stella Blumfelde. (2022). "Artificial intelligence and EU security: the false promise of digital sovereignty". *European Security,* 31(3), pp. 415-434. https://doi.org/10.1080/09662839.2022.2101885

Carlucci Aiello, Luigia. (2016). "The multifaceted impact of Ada Lovelace in the digital age". *Artificial Intelligence,* 235, pp. 58-62. https://doi.org/10.1016/j.artint.2016.02.003

Carstens, Agustín. (2019). "The new role of central banks". Financial Stability Institute's 20th Anniversary Conference, Basel, 12 March. https://www.bis.org/speeches/sp190314.pdf

Cazzaniga, Mauro, Carlo Pizzinelli, Emma Rockal and Marina Mendes Tavares. (2024). "Exposure to Artificial Intelligence and Occupational Mobility: A Cross-Country Analysis". IMF Working Paper, 116, International Monetary Fund. https://www.elibrary.imf.org/view/journals/001/2024/116/article-A001-en.xml

Cipollone, Piero. (2024). *Artificial intelligence - a central bank's view.* National Conference of Statistics, Rome, 4 July. https://www.bis.org/review/r240709c.pdf

Congress of the United States of America. (2022). *Algorithmic Accountability Act of 2022.* https://www.congress.gov/bill/117th-congress/senate-bill/3572/text

Doerr, Sebastian, Leonardo Gambacorta and Jose Maria Serena-Garralda. (2021). “Big data and machine learning in central banking”. BIS Working Paper, 930, Bank for International Settlements. https://www.bis.org/publ/work930.pdf

European Banking Authority. (2023). *Machine learning for IRB models: Follow-up report from the consultation on the discussion paper on machine learning for IRB models.* https://www.eba.europa.eu/sites/default/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf

European Parliament and the Council. (2024). 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,* 1689. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

Fernández Bedoya, Ana. (2019). “Artificial intelligence in financial services”. *Economic Bulletin - Banco de España,* 2/2019, Analytical Articles. https://repositorio.bde.es/handle/123456789/9047

Fraisse, Henri, and Matthias Laporte. (2022). “Return on investment on artificial intelligence: The case of bank capital requirement”. *Journal of Banking & Finance,* 138(106401). https://doi.org/10.1016/j.jbankfin.2022.106401

High-Level Expert Group on Artificial Intelligence. (2018). *Ethics guidelines for trustworthy AI.* European Commission. https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

Hellwig, Klaus-Peter. (2021). “Predicting Fiscal Crises: A Machine Learning Approach”. IMF Working Paper, 150, International Monetary Fund. https://doi.org/10.2139/ssrn.4026328

Japan Cabinet. (2019). *Social Principles of Human-Centric AI.* https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf

Khandani, Amir, Adlar J. Kim and Andrew Lo. (2010). “Consumer Credit Risk Models Via Machine-Learning Algorithms”. American Finance Association 2011 Denver Meetings Papers. https://doi.org/10.2139/ssrn.1568864

Lorenz, Philippe, Karine Perset and Jamie Berryhill. (2023). *Initial Policy considerations for Generative Artificial Intelligence.* Organisation for Economic Co-Operation and Development Artificial Intelligence Papers, 1. https://doi.org/10.1787/fae2d1e6-en.

Monetary Authority of Singapore. (2023). *Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector.* https://www.mas.gov.sg/-/media/mas/news-and-publications/monographs-and-information-papers/feat-principles-updated-7-feb-19.pdf

Moor, James. (2006). “The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years”. *AI Magazine,* 27(4), 87. https://doi.org/10.1609/aimag.v27i4.1911

Moreno, Ángel-Iván, Sergio Gorjón and Joaquín Hernáez. (2021). *Computerized text analysis for assessing legal complexity: the practical example of the Circulars of the Banco de España.* 5th International Conference on Public Policy, Barcelona, 5-9 July. https://www.ippapublicpolicy.org/file/paper/60c0beea33381.pdf

Nayak, B., and Nigel Walton. (2024). *Political Economy of Artificial Intelligence: Critical Reflections on Big Data, Economic Development and Data Society.* Palgrave Macmillan.

Organisation for Economic Co-operation and Development. (2019). *Recommendation of the Council on Artificial Intelligence.* https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Owolabi, Omoshola S., Prince C. Uche, Nathaniel T. Adeniken, Christopher Ihejirika, Riyad Bin Islam and Bishal Jung Thapa Chhetri. (2024). “Ethical Implication of Artificial Intelligence (AI) Adoption in Financial Decision Making”. *Computer and Information Science,* 17(1), pp. 49-56. https://doi.org/10.5539/cis.v17n1p49

Rebuffi, Sylvestre-Alvise, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles and Timothy Mann. (2021). *Fixing Data Augmentation to Improve Adversarial Robustness.* Thirty-fifth Annual Conference on Neural Information Processing Systems, 6-14 December. https://doi.org/10.48550/arXiv.2103.01946

Riemer, Stiene, Michael Strauß, Ella Rabener, Jeanne Kwong Bickford, Pim Hilbers, Nipun Kalra, Aparna Kapoor, Julian King, Silvio Palumbo, Neil Pardasani, Marc Pauly, Kirsten Rulf and Michael Widowitz. (2023). *A Generative AI Roadmap for Financial Institutions.* Boston Consulting Group. https://www.bcg.com/publications/2023/a-genai-roadmap-for-fis

Rubio, Jennifer, Paolo Barucca, Gerardo Gage, John Arroyo and Raúl Morales-Resendiz. (2020). “Classifying payment patterns with artificial neural networks: An autoencoder approach”. *Latin American Journal of Central Banking,* 1(1). https://doi.org/10.1016/j.latcb.2020.100013

Sadok, Hicham, Fadi Sakka and Mohammed El Hadi El Maknouzi. (2022). “Artificial intelligence and bank credit analysis: A review”. *Cogent Economics & Finance,* 10(1), 2023262. https://doi.org/10.1080/23322039.2021.2023262

Shabsigh, Ghiath, and El Bachir Boukherouaa. (2023). *Generative Artificial Intelligence in Finance: Risk Considerations.* IMF Fintech Notes, 2023/006, International Monetary Fund. https://doi.org/10.5089/9798400251092.063

Sheehan, Matt. (2023). “China’s AI Regulations and How They Get Made”. *Horizons,* Summer 2023(24), pp. 108-125. https://www.cirsd.org/files/000/000/010/82/21e461a985f43655b1731b3c1b50cdccb631afaf.pdf.

Shokri, Reza, Marco Stronati, Congzheng Song and Vitaly Shmatikov. (2017). “Membership Inference Attacks Against Machine Learning Models”. 2017 IEEE Symposium on Security and Privacy (SP), pp. 3-18. https://doi.org/10.1109/SP.2017.41

Turing, Alan. (1950). “Computing Machinery and Intelligence”. *Mind,* 59(236), pp. 433-460. https://doi.org/10.1093/mind/LIX.236.433

United Nations Educational, Scientific and Cultural Organization. (2021). *Recommendation on the Ethics of Artificial Intelligence.* https://unesdoc.unesco.org/ark:/48223/pf0000380455

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin (2017). “Attention Is All You Need”. Cornell University. https://arxiv.org/abs/1706.03762

Warwick, K., and Huma Shah. (2015). “Can machines think? A report on Turing test experiments at the Royal Society”, *Journal of Experimental & Theoretical Artificial Intelligence,* 28(6), pp. 989–1007. https://doi.org/10.1080/0952813X.2015.1055826

White House, The. (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.* https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf

How to cite this document

Balsategui, Iván, Sergio Gorjón and José Manuel Marqués. (2024). “Artificial intelligence in the financial system: implications and progress from a central bank perspective”. *Financial Stability Review - Banco de España,* 47, Autumn. https://doi.org/10.53479/38942