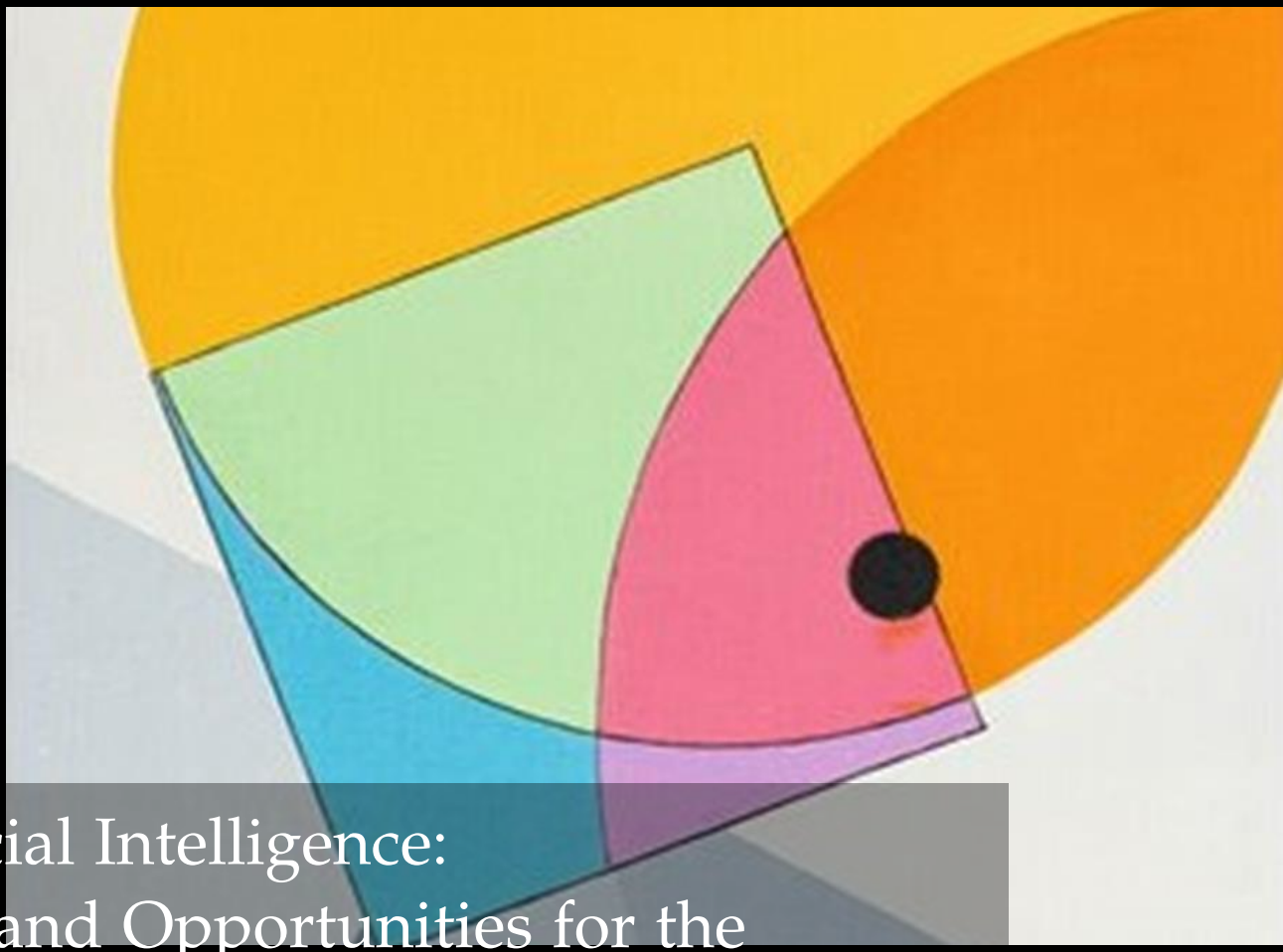
Research Paper Series



Artificial Intelligence: Risks and Opportunities for the Banking System

Marco Carminati

May 2024

iason
ESSENTIAL SERVICES FOR FINANCIAL INSTITUTIONS

iason
Research Paper Series
Year 2024 - Issue Number 66

Last published issues are available online:
http://www.iasonltd.com/research

Front Cover: **Luigi Veronesi**, *Costruzione KZ*, 1997.

Executive Summary

In recent years, there has been a significant **diffusion of Artificial Intelligence (AI) solutions** across various sectors. AI, denoting **computer systems' capability to mimic human intelligence and decision-making processes**, has seen increasing adoption due to the **availability of vast data sets** and **enhanced computing power**. Businesses stand to gain substantially from AI implementation, with **opportunities for automation, improved productivity, and profitability**. Similarly, society benefits from more **efficient services, advanced healthcare solutions**, and **heightened public safety measures**. The **banking and financial sector** is not exempt from AI's transformative potential, particularly in **customer interactions and risk management**, leveraging data-driven insights for informed decision-making. Despite the **clear advantages**, the widespread integration of AI brings **inherent risks**. **Complex decision-making models** pose challenges to **transparency and interpretability**, while the **reliance on data** increases the likelihood of **biases and privacy concerns**. Additionally, **cybersecurity threats** make AI solutions vulnerable, necessitating robust protective measures. **Regulatory authorities** worldwide are responding to these challenges with varying degrees of urgency and stringency. The **European Union**, for instance, has proposed **comprehensive regulations to govern AI usage**, emphasising risk assessment and compliance with EU fundamental values. In contrast, the regulatory frameworks in the **United Kingdom** and the **United States** are **less developed**, while **China boasts advanced AI regulations** tailored to specific applications. This paper seeks to highlight the **dual nature of AI adoption**, emphasising its **benefits** for businesses, and particularly for the **banking sector**, alongside the imperative of **addressing associated risks and challenges**. The paper furthermore underlines the **importance of ethical considerations and regulatory compliance** in harnessing AI's potential for **safe and responsible use**, ensuring the respect of **fundamental rights and ethical principles.**

About the Authors

Marco Carminati:
Senior Manager
After taking a Bachelor in Finance and a Master of Science in Economics, he specialised in Credit Risk Management and Modelling through almost 10 years of experience in both banking and consultancy industry with special focus on the development and validation of models for Credit Risk for regulatory and managerial purposes. Currently in charge of several projects in the Credit Risk Management area and Head of the iason Credit Risk Competence Center.

This document was prepared in collaboration Leonardo Bandini and Vincenzo Frasca who at the time were working for Iason Consulting.

Table of Content

Artificial Intelligence: Risks and Opportunities for the Banking System

Marco Carminati Leonardo Bandini
Vincenzo Frasca

In recent years, there has been a growing **diffusion of solutions and tools based on Artificial Intelligence**. The term **Artificial Intelligence** (hereinafter usually referred as AI) generally refers to the **ability of computer systems to develop knowledge and make decisions that would typically require human intelligence**, but **without human intervention**, relying on the observation and analysis of available data.
The **increasing adoption of AI techniques** observed in recent years is facilitated, on one hand, by the **continuous increase in data availability**, and on the other hand, by **improvements in the computing power of computers**, enabling the processing of large amounts of data more quickly.
The availability of **decision-making tools based on AI** represents an unprecedented **opportunity for businesses and society as a whole**. Companies can benefit from **intelligent tools** to support their **business processes**, allowing them to automate activities that would typically require human intervention, leading to **increased productivity and profitability**. From a broader perspective, individuals can benefit from the **availability of more efficient services**, **improved and more accurate healthcare solutions**, and a **higher level of public safety**.
The opportunities presented by AI for firms do not exclude the **banking and financial sector**. They are related, for instance, to the **streamlining of customer interactions** and of **risk mitigation activities**, relying more on **data-driven results and evidence**.
While the opportunities and potential benefits of the widespread adoption of AI are evident, there are also **risks that should not be underestimated** in the process of implementing such tools. The **increased complexity of AI decision-making models** poses **limits on the transparency and interpretability** of their results, which tend to be inherently more opaque. Additionally, there are challenges in verifying the **correctness** of and the **rationale** behind the decisions made by these systems. Furthermore, the **high dependence on underlying data** increases the **risk of amplifying potential distortions (bias)** present in the data, posing the risk of **distorted and potentially discriminatory results and decisions**. Moreover, the **analysis of vast amounts of data**, including information about individuals' activities and spending habits, poses **risks in terms of data privacy**. Finally, potential **cybersecurity risks** should not be neglected, as AI-based decision systems may be susceptible to **malicious attacks** aimed at **manipulating decisions** and producing possible discriminatory effects.
In light of the accelerated proliferation of AI-based solutions and in order to mitigate potential risks as listed above, **regulatory authorities** in major countries are taking **steps to regulate the adoption of AI** and promote its **safe and conscious use**. In the European Union (EU), the **European Commission** issued in 2021 a proposal for regulation (known as the "**Artificial Intelligence Act**") which, following a **risk-based approach**, includes a **ban on the use of certain forms of AI** deemed contrary to the values of the Union. It also introduces **specific requirements for the production and use of AI systems** considered to be of **higher risk**. The proposal has been recently subject to **agreement between the European Commission, Parliament and Council** in **December 2023**, and is currently awaiting for the publication of the official final text. In other countries, the state of

regulatory maturity varies significantly. In the **United Kingdom**, the current regulatory intervention is focused on **identifying principles and guidelines without legally binding regulatory provisions**. In the **United States**, **regulatory intervention is still in its early stages**, with an absence of specific and general initiatives regarding AI legislation. In **China**, on the other hand, **AI regulation is at an advanced stage**, with specific rules for each identified application of the technology currently in place.
Against this background, the **objective of this document** is, on one hand, to describe the **main benefits resulting from the adoption of Artificial Intelligence tools**, emphasising in particular for firms operating in the **banking sector**. On the other hand, it aims to **illustrate the potential risks** associated with such tools, which must be **assessed and managed consciously** to harness their potential **safely** and **ensure respect for ethical principles and fundamental rights of individuals**.
The document is organised as follows:

- Chapter "Artificial Intelligence: Definitions and Main Diffusion Drivers" provides an **overview of the diffusion and definitions of AI** and describes some of the **main forms of Artificial Intelligence** available on the market;
- Chapter "Possible Benefits from the Use of Artificial Intelligence" describes the potential **benefits that the adoption of AI can bring to society and businesses**;
- Chapter "Potential Risks Associated with the Use of Artificial Intelligence"an overview of the **potential risks associated** with the use of AI;
- Chapter "Main Applications of Artificial Intelligence in Banking Sector" highlights the main **opportunities that the banking system can seize with the use of AI-based tools**, describing some examples of application;
- Chapter "Regulation of Artificial Intelligence: an Overview" illustrates the **regulatory evolution of AI in the European Union** and possible **regulatory trends in other globally relevant countries**;
- Chapter "Conclusions" provides some **concluding remarks**.

1. Artificial Intelligence: Definitions and Main Diffusion Drivers

1.1 Definitions and Diffusion of Artificial Intelligence

The term "**Artificial Intelligence**" (hereinafter also **AI** for brevity) generally refers to a **set of methodologies and techniques** that enable the design of **computer solutions capable of replicating human intelligence to varying extents**. More specifically, the term refers to the **ability of computer programs** to **acquire knowledge and make decisions without human intervention**, through the **observation and analysis of available data**. Artificial intelligence systems are characterised by their **ability to perform tasks that would typically require human intelligence**, such as understanding text, creating an image, or making a decision.
In recent years, thanks to the **continuous technological development** and the **increasing availability of data**, there has been a growing proliferation of techniques and solutions based on AI. In fact, AI as a field of computer research has **existed for decades** (for example, as early as 1950, the British mathematician Alan Turing formulated the so-called *Turing Test*, a criterion to determine if a machine could exhibit intelligent[1]). However, it is only in more recent years that solutions based on such techniques have accelerated their diffusion, mainly due to a series of **enabling factors**, such as:

- The **increasing availability of data**, of different nature and originating from various sources, is primarily linked to the **continuous growth in internet usage**, allowing for the increase in digital data production and availability. For instance, in a 2019 study, Deutsche Bank Research[12] highlighted how, in the last 10 years alone, the **amount of data generated globally has increased by a remarkable 17 times**, with estimates indicating a further fivefold

[1] See, for instance, "The Turing Test".

growth by 2025. This evolution provides a **vast amount of information** (commonly referred to as "***big data***"), which serves as the **primary informational source** to enable the use and maximisation of the potential of Artificial Intelligence solutions. In general, the term ***big data*** refers to data characterised by **specific features**, identified by the so-called **"3 Vs"**[2]: 1) **volume**, meaning a substantial amount of data, 2) **variety**, encompassing various types of data, including **structured**, **semi-structured**, and **unstructured data**[3], and 3) **velocity**, indicating that data is produced at high rates.

- The significant **increase in computing power** enables algorithms to **process information quickly**, contributing to the **accuracy of the decision-making process**. In this regard, the widespread adoption of **cloud computing solutions** serves as an accelerator for the use of advanced analytics techniques. The cloud infrastructure provides more space for the storage and processing of vast amounts of data in an efficient manner.
- Other evolutionary factors are related, for example, to the **reduction of costs associated with data storage solutions**, advancements in **data extraction and processing processes** (so-called "***data mining***"), and the increasing availability in the labour market of **IT experts specialised in the analysis of large quantities of data** (so-called "***data scientists***").

These factors have provided businesses with the **opportunity to employ advanced techniques**, which are currently used for various purposes across different sectors. For example, such solutions are already being applied in the **healthcare sector** (e.g., for the analysis of diagnostic images), in **marketing activities** (e.g., providing buyers with customised spending suggestions based on past purchasing experiences), and in the **financial sector** (see Chapter "Main Applications of Artificial Intelligence in Banking Sector" for detailed examples of possible AI applications in the banking sector).

1.2 Types of Artificial Intelligence

Despite being referred usually as Artificial Intelligence in a general way, AI encompasses a **broad set of techniques and methodologies**, each characterised by **specific features** that make them more suitable for application in certain areas of activity and for specific purposes. The **most commonly used forms of Artificial Intelligence** include, in a non exhaustive manner, the following:

- ***Machine Learning (ML)***;
- ***Natural Language Processing (NLP)***;
- ***Generative Artificial Intelligence (GAI)***.

These forms of AI are described in more detail in the following paragraphs.

1.2.1 Machine Learning

Machine Learning (ML, also known as "automatic learning") is a set of techniques that, based on a predefined set of rules (called ***hyperparameters***) and through a **learning process** (also called "learning" or "training"), allow for the identification of **relationships among data** and, based on these, formulate the best decisions and predictions. The learning process enables the generation of **predictive models** whose results can be used for various purposes (e.g., for classification problems, clustering, etc.). Thanks to these characteristics, Machine Learning techniques are often

[2]In this regard, see also EBA[14]. In some cases, these "3Vs" are accompanied by other 2 ("5Vs" paradigm), including also the following characteristics: veracity, which relates to the need to ensure that such data represent as accurately as possible the underlying reality, and value, a characteristic related to the need to be able to transform the data into useful business information. In this regard, see also Banca d'Italia[2].

[3]Structured data refers to data presented in an ordered and organised format within standard structures (e.g. tables within a database); semi-structured data, on the other hand, refers to data that contain some "tags" but do not respond to the structure associated with typical relational databases (e.g., data related to email or XML); finally, unstructured data are types of data that are not ordered or organised according to a predefined format, consisting of a wide variety of information that is inherently complex to navigate and process (think, for example, of information contained in audio or video files, or derived from surveillance cameras or social media).

Approach	Definition
Supervised Learning	In this context, the algorithm learns the **relationships between inputs and outputs through** a set of data that has been previously **labelled** by a human (referred to as a "**labelled dataset**"). For example, consider a dataset related to transactions for which, in addition to a series of descriptive features (such as transaction amount, business sector of the subject, etc.), there is also a label indicating whether a transaction was fraudulent or not. The algorithm identifies a **classification rule** for the outputs that can be used to predict the value of the label (in the previous example, identifying potentially fraudulent transactions) for new data.
Unsupervised Learning	In this case, the algorithm autonomously learns the model from the **available dataset** without the need for it to be preprocessed for label attribution. Here, the **algorithm identifies relationships between data by identifying groups ("clusters")** of similar observations, i.e., with **common characteristics.** These techniques can be used in various circumstances, such as for **clustering analysis** or the **detection of potentially anomalous observations** within a set of data.
Semi-supervised Learning	Semi-supervised learning represents an intermediate case between the two previous ones, where the **dataset available to the algorithm is only partially labeled.** These techniques are useful when only a **few labelled observations are available,** for example, in cases where **the label attribution process is particularly complex or costly.**
Reinforcement Learning	In this context, the algorithm doesn't **learn relationships** from observations in a dataset, but rather from its **interaction with the surrounding environment**. Specifically, it **takes an action and receives feedback on its correctness.** Based on this **feedback** (such as profit or loss in the case of a trading algorithm), the **algorithm improves its strategy to maximise positive feedback,** thereby **enhancing its accuracy**. An example of the application of these techniques could be trading or pricing. strategies.

TABLE 1: Types of Machine Learning Algorithms

used for the **development of predictive analysis solutions**. Compared to traditional statistical analysis techniques (e.g., linear regressions), these algorithms are capable of **identifying non-linear relationships among data**, thereby **increasing the accuracy of predictions**.
In general, Machine Learning algorithms can be divided into sub-categories, differentiated based on the learning mode used.
Regardless of the specific approach described above, there are **several families of ML algorithms** that can be used in each of the cases mentioned above, among which the following are notable due to their widespread use:

- ***Ensemble Learning***: these techniques are based on the **use of a set of models**, whose **predictions are then aggregated to determine the final prediction** (which can be, for example, determined as the average of the predictions of the individual models). A widely used type of technique of this kind is represented by **decision trees**, which can be used for **simultaneous learning of independent models** (called "bagging") or through **sequences of models that progressively refine the learning process**, reducing the error of the previous models (called "boosting").

- ***Deep Learning***: this is a family of ML techniques whose **learning process is structured into layers** (each composed of interconnected units, called "neurons") inspired by the **functioning of the human brain** (and therefore also called "neural network"). At each level of the network corresponds a phase of learning with increasingly complex concepts. Neural networks can have **very complex structures**, with hundreds or even thousands of layers, making it difficult to understand the rationale behind the decisions made; for this reason, they are often referred to as "black boxes".

1.2.2 Natural Language Processing

Another widely used form of Artificial Intelligence is ***Natural Language Processing (NLP)***, a subcategory of AI based on algorithms capable of **processing**, **understanding,** and **interpreting human language**, effectively **enabling communication between humans and machines**. The

application of such techniques allows, for example, the extraction of a subset of relevant information from a document, making content processing more efficient and reducing (or completely replacing) the need for human intervention. According to a **recent report** from the Alan Turing Institute[28], some of the most widely used NLP algorithms include:

- ***Sentiment Analysis***: algorithms that determine the emotional tone within a text by analysing textual data;
- ***Named Entity Recognition (NER)***: algorithms capable of identifying names (e.g., referring to people or places) within unstructured data;
- ***Machine Translation***: algorithms that translate text from one language to another while preserving the textual meaning of the processed sentences;
- ***Speech Recognition***: algorithms capable of listening to and understanding a text and transcribing it.

1.2.3 Generative AI

Another form of AI that has seen significant growth recently is ***Generative AI* (or *GAI*)**. It comprises a set of techniques and AI models that can **generate various types of digital content**, such as text, images, and music, after receiving specific instructions from a human. Generally, these models are based on ***Large Language Models (LLM)***, a form of **deep learning** that leverages techniques such as ***Recurrent Neural Networks (RNN)***, a type of neural network based on sequential data capable of **understanding and managing temporal sequences** in the context of **language translation** and **speech recognition.**[4]
One of the most prominent examples of Generative AI models is **ChatGPT**, a language model developed by the U.S.-based company OpenAI. It can **understand questions posed by humans** and **respond effectively** and almost **instantaneously**. Another example of GAI is represented by DALL-E 2, also developed by OpenAI, an AI model capable of **creating original and high-quality images** in response to a textual instruction[5].
While the potential **benefits** of such techniques are evident[6] (for example, enabling the rapid processing and analysis of a wide set of information and providing users with data-driven recommendations), the ability of these solutions to **create original and high-quality content** is **not without risks**. If used inappropriately, such technologies can pose **ethical risks**[7], allowing malicious actors to **create false and misleading content** to **manipulate the opinions and behaviors of individuals**. Additionally, from the perspective of **personal data protection**, there is a risk of u**nauthorised creation of "artificial" images or videos** depicting certain individuals without their consent. Furthermore, there is also the risk of **inappropriate diffusion of sensitive and non-disclosable data or information** uploaded or divulged by the user when using the system.

1.2.4 Other Forms of Artificial Intelligence

In addition to the forms of Artificial Intelligence described in the preceding paragraphs, other possible AI solutions of particular utility and prevalence in business activities include ***Expert Systems*** and ***Robotic Process Automation (RPA)*** tools. According to The Alan Turing Institute[28]:

- An **Expert System** is an Artificial Intelligence system that **replicates the decision-making capabilities of a human expert** in a specific field of application, using a **predefined set of rules and an inference engine** to solve problems that typically require human judgement. In the financial sector, such tools can be used for various applications, such as portfolio management or financial forecasting. These tools can be particularly useful in situations where a high level of knowledge is required and difficult to obtain or where there is limited availability of human experts in the field of analysis in question.

[4] See, for instance, IBM.¸
[5] For a detailed overview on Generative AI techniques, see Cao et al.[6].
[6] On the potential benefits of Generative AI, see [25].
[7] On the general risks posed by Generative AI, see [33], while for a more specific view related to the risks for the financial system, see [24].

- **Robotics Process Automation (RPA)** is a branch of Artificial Intelligence that enables the **automation of a set of repetitive actions and tasks** typically performed by human agents. Such solutions can be used in the financial sector to carry out a variety of routine activities, such as account opening and closure, complaint management, fraud detection, customer service activities, and reporting.

2. Possible Benefits From the Use of Artificial Intelligence

2.1 General Benefits

The adoption of solutions based on Artificial Intelligence can without doubt produce **benefits for society**. In particular, potential **advantages** related to the **diffusion of AI-based solutions** are linked, for example, to the **improvement of healthcare services** (think of potential improvements in diagnosis accuracy), the increase in the **accuracy of decision-making processes** (which can more effectively exploit the information contained in data), or the increase in precision in various stages of **agricultural cultivation** (which can allow for reduced land use and therefore mitigate the impacts of agriculture on climate change).
In its **White Paper** on Artificial Intelligence, the **European Commission**[17] identifies the **set of potential benefits** deriving from the adoption of Artificial Intelligence techniques, differentiating them according to the perspective of analysis:

- From the **citizens' perspective**, they can benefit, for example, from the availability of **better healthcare solutions, more efficient transportation, improved public services**, and greater efficiency of tools used in daily activities (e.g., smarter appliances);
- From an **economic development standpoint**, there is the potential for the diffusion of a **new generation of key products** and services in sectors where European Union countries excel, such as machinery, transportation, cybersecurity, as well as the agricultural sector and the **green and circular economy**, healthcare, and high-value-added sectors such as fashion and tourism;
- From the **perspective of public services**, AI-based solutions can **reduce costs related to the provision of key services** such as transportation, education, energy, and security, **improving the sustainability** of these products and services and equipping public safety authorities with appropriate and cutting-edge technologies to ensure **citizens' safety**.

In addition, Artificial Intelligence technologies can be a key tool for addressing some of the **greatest challenges of our era**, such as those related to **environmental protection and sustainability goals**. In this regard, the European Commission believes that the use of such systems can play an important role in achieving sustainability goals **linked to the United Nations' Sustainable Development Goals**, as well as in supporting democratic processes and respecting **human rights**.

2.2 Benefits for Businesses

The adoption of Artificial Intelligence techniques undoubtedly offers **benefits to businesses**. In particular, the application of these techniques can allow:

- **To reduce operating costs by automating activities** previously entirely reliant on human intervention;
- To **increase profitability**, for example, by retaining customers through personalised offerings, as close as possible to their purchasing preferences.

The table 2 provides a more detailed overview of the **main potential benefits** arising from the use of Artificial Intelligence techniques.

Benefit	Description
Reduction of operational costs	This is allowed by the automation of business processes, both routine ones (such as opening a bank account, massive pre-processing of documents for customers), and more specific ones (such as financial advisory services), also thanks to the so-called robo-advisors, digital assistants available at any time of the day. The use of these solutions allows companies to free up labor that can be dedicated to other activities with higher intellectual intensity.
Reduction of costs caused by fraud	The ability of Artificial Intelligence techniques to process, even instantly, large volumes of data from multiple sources, allows for more effective and faster detection of suspicious transactions, fraud schemes, signs of money laundering, and anomalous financial behaviors in general, reducing associated costs.
Improvement of sales strategies	The adoption of AI tools that track consumer purchasing preferences and provide personalized spending suggestions based on observed buying behavior can allow companies to better intercept the needs of potential buyers through targeted offerings.
Streamlining decision-making	This is enabled by the automated and almost instantaneous support that advanced technologies can offer in contexts where significant amounts of data (big data) are present, such as credit assessment and decisions regarding loans and investments.
Improvement in the accuracy of predictions generated by predictive models	This improvement is due to the ability of AI-based tools to identify connections between variables, even those that are less evident and highly complex, compared to what is possible through the use of traditional techniques. Additionally, AI techniques are significantly more efficient in handling larger volumes of data and factors that may seem less relevant at first glance to the specific context of financial risk assessment. Consequently, they enable the production of more accurate predictions.

TABLE 2: AI Benefits for Businesses

3. Potential Risks Associated with the Use of AI

In contrast to the benefits derived from the proliferation and use of AI-based solutions, **they are also accompanied by a series of disadvantages and potential risks that**, if not managed and mitigated properly, can **compromise the proper application of AI and**, in some cases, cause harm to its users and beyond. In particular, according to the White Paper on Artificial Intelligence published by the European Commission in 2020, **the risks associated with AI can be both material** (related, for example, to risks to the health and safety of individuals) **and immaterial** (related, for example, to limitations on data privacy or freedom of expression), depending on the actual use of such tools. Here is a more detailed overview of the **main potential risks** arising from the use of Artificial Intelligence techniques:

- **Transparency and Explainability**;
- **Distorsions (Bias) and Fairness**;
- **Accountability and Reliability**;
- **Data Privacy**;
- **Cybersecurity**.

3.1 Transparency and Explainability

As described in Chapter "Possible Benefits from the Use of Artificial Intelligence" regarding the potential benefits of AI solutions, **they are capable of accurately modeling phenomena for which they are used**, often leveraging complex relationships among data that are **not identified by traditional models** and techniques. While this characteristic improves the accuracy of predictive models and the decisions based on them, **it also complicates the interpretation of the relationships** on which these decisions are based, making them opaque and difficult to explain. This phenomenon is particularly relevant with higher levels of complexity in AI techniques, such as **deep learning based on complex neural networks**, which essentially act as **black boxes**. This usually results in a

trade-off between the performance (accuracy) of AI systems on one hand and **transparency** on the other, necessitating developers to strike an appropriate balance between the two dimensions. When analysing problems related to the complexity or even the impossibility of understanding how a particular AI system operates and the **rationales behind its decisions**, the concepts of **explainability** or **interpretability** typically arise. According to EBA[14], **an AI model is considered "explainable" when its internal dynamics can be directly understood by humans** (interpretability) or when explanations can be provided regarding the main factors that have led to its results.
The **lack of transparency** in AI systems not only **negatively impacts the explanation of the results they produce** but also **affects the ability to identify potential malfunctions** or areas for improvement. In general, issues related to the transparency and explainability of AI systems are more relevant when their results and decisions **impact humans**. For example, in the case of adopting an AI model to determine loan approval for a credit applicant, **it is crucial for the bank to identify the rationales that led the model to deny the approval**. In this regard, the **European Regulation on the protection and processing of personal data (GDPR)**[20] stipulates, in **Article 13**, the right of data subjects (and thus customers) to receive information about "the **existence of automated decision-making** [...] and the **logic used**, as well as the significance and the envisaged consequences of such processing for the data subject." According to Banca d'Italia[2], the request for "meaningful information on the logic used "***implies an obligation for intermediaries to provide so-called "local" explanations*** *to applicants, inclusive of the details of the main variables that contributed to a specific outcome regarding the loan approval or denial*". The risks associated with the **lack of explainability** of AI systems are further **accentuated** in the case of **solutions purchased by companies from third parties**. In such cases, the EBA recommends that the purchasing institution have **adequate tools** available to **explain** and **validate** the results produced by the purchased system without being heavily dependent on the third-party provider.
In light of the above, it is evident that for those using such solutions for **decision-making purposes** that **impact individuals** (e.g., credit scoring, fraud detection, etc.), such as financial institutions, there is a need to **balance the trade-off** between model performance and explainability, **accepting lower accuracy for greater transparency**. A contribution in this regard is linked to the increasing **availability** of explainability tools (so-called "**Explainable Artificial Intelligence**," or **XAI**) that can be adopted to **reduce the opacity level** of AI models. Academic research on XAI is continuously evolving and already provides developers with **explainability libraries** for the most common programming languages used in AI[8].

3.2 Distorsions (Bias) and Fairness

A **widely discussed topic** in the realm of AI-based decision-making applications with impact on individual persons is the **potential presence of biases in the data** used for model training. If **not adequately addressed** during the development and/or application of the AI model, such biases can lead to **discrimination against individuals or groups of individuals** (e.g., discrimination based on age, disability, gender, sexual orientation, political, religious affiliations, etc.).
In general, as reported by Castelnovo et al.[9], the concept of bias can be **distinguished** into:

- **Statistical or representation bias**: this form of **data distortion** occurs when the data is **not representative** of the **true population** it refers to. It can be generated by forms of **selection bias**, where the individuals represented in the available data **do not belong to a random selection of the true population**. For example, consider the case of data related to **loan repayment capacity**, which is **observed** and **available** only for individuals who have actually been granted **credit in the past**, and therefore **not representative** of individuals who have applied for credit but have been denied it.

- **Historical or societal bias**: even if the data is not affected by statistical bias, there may be a **form of distortion** in the data that reflects distorted **behaviors or decisions in the past**. This can typically be attributed to a **bias in label assignment**, where there is a phenomenon of **systematically favorable/unfavorable treatment towards certain groups of individuals** when the target variable on which the AI system is trained is created (for example, consider the case

[8]For an overview regarding the most popular XAI techniques currently available, see [26].

where a credit scoring model is developed based on credit decisions made by managers in the past).

Regardless of their nature, **biases in data** can lead a model developed on such data to produce results, and **ultimately decisions**, that may be **discriminatory** towards certain individuals or homogeneous groups of individuals. This phenomenon, which generally represents a risk even in the case of traditional statistical models, **is exacerbated in the case of AI models**, as they **tend to replicate or even amplify any biases** present in the data, for example through **non-linear connections** between different data sources[9].
The **principle of non-discrimination** is generally referred to as the principle of **fairness**. The topic of fairness represents a deeply explored and **debated element** in the literature concerning the **ethical implications** arising from the **application of AI** techniques. One of the **main issues** to address when analysing the **fairness of a model** relates to **defining the characteristics** for which the model under consideration is **deemed fair**. Indeed, there are **various definitions** and **interpretations** of fairness, primarily based on the distinction between **individual fairness** and **group fairness**. Before proceeding with the definitions, however, it is important to define the concept of **protected** or **sensitive attributes**, typically used in articulating possible definitions of fairness. Protected or sensitive attributes are defined as those **personal characteristics of individuals** (or groups of them) **on which the absence of discrimination is to be verified**, such as gender, ethnicity, age, sexual orientation, political or religious affiliation, etc. Consequently**, it is possible to define the concepts of individual and group fairness** (for a broader and more detailed examination of possible definitions and articulations of fairness metrics, refer to Castelnovo et al. (2021)[8], Castelnovo et al. (2022)[9] , and Rubicondo and Rosato[27]):

- *Individual fairness*: this principle focuses on **comparing individual persons** and posits that similar individuals (i.e., those differing only with respect to their characterisation of sensitive attributes) should receive **equal** or at least similar treatment.
- *Group fairness*: this principle of fairness focuses on **comparing groups** of individuals and requires that these groups receive similar treatment **regardless of their sensitive characteristics**. The principle of group fairness is typically expressed by demanding **equality** of certain **statistical metrics** across different groups. In the literature, three possible notions of group fairness are proposed, as follows:
  - *Independence*: according to this criterion, the model's outcomes should be **independent** of the characterisation of **sensitive attributes** across different groups analysed. Taking the decision regarding the approval of a loan for individuals of different genders (male and female) as an example, this principle implies that the **probability of a positive decision** (credit approval) should be **equal** between males and females.
  - *Separation*: this criterion entails that the model's outcomes are, as above, **independent of the characterisation** of sensitive attributes, but **conditioned** on the value of the **target variable** considered. In the previous example regarding credit approval, this means that any **differences** in the probability of a positive decision between males and females are entirely justified by **differences in observed riskiness** between the two groups (for example, the historically observed default frequency).
  - *Sufficiency*: this criterion **reverses** what is expected by the **concept of separation** and requires that, among subjects belonging to **different sensitive groups** and receiving the **same decision** from the model, the **probability of having the same realisation** of the true target variable considered is **equal**. Continuing with the previous example regarding credit approval, this implies that, for two subjects of different genders who have received the same decision from the model (for example, loan denial), **the actual riskiness** (risk of default) **is equal or at least similar**.

Beyond the precise **definition of fairness** among those listed above, a model can generally be deemed **fair** if it **does not produce discrimination in outcomes** based on the values of protected or **sensitive attributes** as defined above. **Biases in data** and potential risks of discrimination in

[9] European Banking Authority, *Report on Big Data and Advanced Analytics*, January 2020, see [14].

outcomes can be **mitigated** by addressing specific aspects in the **model development** process, such as **removing information** on sensitive attributes from the **training dataset** (so-called "Fairness Through Unawareness"), or intervening in the dataset composition using **oversampling techniques** to mitigate issues of poor representativeness towards certain groups.
The **ethical implications** related to fairness represent one of the **main concerns** highlighted by the **European Commission** in its Ethics Guidelines for Trustworthy AI[18] and its White Paper on Artificial Intelligence, which are reflected in its proposed regulation on Artificial Intelligence (described in more detail in Chapter "Regulation of Artificial Intelligence: an Overview", to which reference is made).

3.3 Accountability and Reliability

An additional **element of complexity** associated with AI systems lies in their **accountability**, namely the ability to **trace back** to the origin of a particular decision made by the system and ultimately define **responsibility** in case of erroneous or **harmful decisions**. The concept of **accountability** is therefore **closely linked** to that of transparency and explainability discussed in the preceding paragraphs, and the related issues are of **increasing importance** as the volume of decisions delegated to such systems grows, which must therefore be as **transparent** and **justifiable** as possible. The need to ensure a **sufficient level** of accountability for the results of AI systems is among the principles of the **Ethics Guidelines for Trustworthy AI** issued by the European Commission, which require **suppliers and/or users** of such solutions to establish **adequate mechanisms** to ensure the accountability of AI systems and their results, both during **development** and **in their use**. To achieve this, AI systems must be **auditable**, for example by providing traceability and **logging mechanisms** to ensure that their operation can be **independently audited**. Another **potential risk** associated with AI techniques and connected to accountability is that of the **reliability of the results** produced. According to Rubicondo and Rosato [27], reliability refers to the **poor robustness** of **AI-based models** with respect to possible **variations in input data**, which, even if limited, can produce **significant variations** in the results of such models. Indeed, as these solutions rely on data as their **strength**, they are more **vulnerable** to possible **distributive shifts** in the underlying data and to potential variations in the **actual relationships** among them. Issues related to reliability are further exacerbated by the difficulty of **monitoring** and **investigating** the results produced by AI models, confirming the **need to ensure transparency and auditability** in their operation.

3.4 Data Privacy

AI-based models **benefit** enormously from the **abundance of data** at their disposal, which allows them to model phenomena under analysis **more accurately** and precisely. However, the **ability** of these tools to analyse data from **diverse sources**, such as internet browsing data, social media data, online purchasing experiences, or through payment cards, poses potential **risks** regarding the **protection of customers' personal data**. Therefore, it is important for those who develop and use such tools to **ensure compliance** with **principles** related to the protection of personal data, ensuring that categories of **protected data are not actually used** by the analysis models. This principle generally applies to **analytics solutions** based on **traditional techniques** as well; however, it becomes more relevant in the context of using **AI techniques** because they, by exploiting **correlation relationships** between data, may be able to **reconstruct** protected information (such as sensitive attributes exemplified in the section dedicated to fairness) even if they are **previously removed** from the training dataset.
In this regard, EBA[14] recommends that **financial institutions** ensure **compliance with the data protection principles** established by the **GDPR** in **all stages** of the life cycle of big data and advanced analytics-based models (both in development and production phases). Among these **data protection principles**, the EBA emphasises the need for **customer consent** to the processing of personal data and the **obligation to inform** the customer about any form of data processing carried out on such data. Regarding data used for **credit assessment** purposes, it is also worth noting the **European Commission's proposal** for a new **directive on Consumer Credit**, which highlights the **prohibition** of using certain categories of personal data for **creditworthiness assessment** (CWA), including

data derived from **social media** and data related to the **health conditions** of customers[10]. Lastly, concerning **data privacy issues**, it's worth noting that when using Generative AI solutions (like **ChatGPT**), there's a **risk of disclosing sensitive data** (personal data or internal company data) during system queries. There's indeed a risk that such data might be **acquired** by the system, potentially **violating** the principle of **confidentiality** and **protection**.

3.5 Cybersecurity

Another potential **risk** associated with the application of **AI-based tools** is related to **cybersecurity** issues. Given that these are **decision-making** systems, AI models are susceptible to potential **malicious data manipulation** aimed at **distorting** their results. For example, **Deutsche Bank Research**[12] highlights the risk of potential **hacker attacks** aimed at **altering the database** used by an AI system (for example, by **spreading false news**) in order to manipulate its results in a particular direction. Other examples of possible manipulation using AI include the alteration or even **creation of fake media content** (such as images or videos) designed to influence **individuals' opinions** on various aspects, such as political views or consumer preferences, through **false representation** of reality, with significant ethical and social repercussions[11].
According to the **European Banking Authority** (EBA) in its ***Report on Big Data and Advanced Analytics***, some of the **main types of attacks** to which **AI** solutions are susceptible include, for example:

- ***Model stealing***: a type of attack aimed at "**stealing**" the models by **replicating their functioning** (for example, requesting a model to **provide predictions** on a wide variety of different inputs and using these predictions to develop a new model, which will effectively tend to **replicate** the first one).
- ***Poisoning attacks***: attacks in which attempts are made to **influence and manipulate the training data** in order to **distort the results** and decisions made by the model.
- ***Adversarial attacks***: in these cases, the attack consists of providing the model with a **slightly perturbed** input data sample in order to **alter its predictive power**. These attacks are typically aimed at inducing the model to **avoid detecting** a particular element (so-called "evasion attack").

In light of the above, it is of **paramount importance** for businesses employing AI tools to **equip** themselves with adequate **cybersecurity** measures, maintaining a level of protection and **technical surveillance** that is sufficiently robust, and **continuously monitoring** potential cybersecurity attacks and the corresponding defensive techniques available.

4. Main Applications of Artificial Intelligence in Banking Sector

4.1 Opportunities for the Banking Sector

As anticipated in Chapter "Possible Benefits from the Use of Artificial Intelligence", the adoption of **AI solutions** can bring **undeniable benefits** to businesses, assisting them in their **digital transformation** processes and increasing **profitability**. This also applies to companies in the **banking** and **financial** sector, which have been actively employing AI techniques in various areas of their operations for several years. In fact, **AI-based tools** are already used for different purposes and **multiple applications** by **banking intermediaries**. A survey conducted by The Economist Intelligence Unit in 2022, targeting the **major global banks** on the **use of AI** in the banking context, highlighted that **almost all financial institutions** resort to Artificial Intelligence to some extent. Specifically, **more**

[10] European Commission, *Proposal for a Directive of the European Parliament and of the Council on Consumer Credit*, 2021 [15]. It reports at paragraph 37: "[...] what categories of data may be used for the processing of personal data for creditworthiness purposes, which include evidence of income or other sources of repayment, information on financial assets and liabilities, or information on other financial commitments. Personal data, such as personal data found on social media platforms or health data, including cancer data, should not be used when conducting a creditworthiness assessment."
[11] Rubicondo,D., Rosato,L., *AI Fairness: Addressing Ethical and Reliability Concerns in AI Adoption, Iason Research Paper Series*, March 2022, see [27].

than half of the participants in the study reported using **AI technologies**, for example, for fraud detection, optimisation of IT operations, and digital marketing (in terms of purchase suggestions and recommendations based on purchase history). Other areas where these technologies are **widely applied** include risk assessment and credit scoring, customisation of customer experience (including marketing and sales aspects), product design optimisation, and personalised investment strategies. In the **United Kingdom**, a survey conducted by the **Bank of England**[5] in 2022 on a sample of national banks illustrated that about **two-thirds of the responding institutions** already use **Machine Learning** techniques for various purposes, with **customer engagement** and areas of **risk management** and compliance predominating.
In particular, based on various studies and reports (including the aforementioned survey by the **Bank of England**, as well as reports published by the **Financial Stability Board**[21] and the European Banking Authority[14][12]) the following is an **overview of the main areas of application of AI in the banking sector**.
While the above provides an overview of the potential applications of AI-based tools by banks, it's important to note that, from a **supervisory perspective**, regulatory authorities have also embraced the adoption of such techniques. In this regard, the emergence of so-called "***SupTech***," the adoption process by financial sector supervisors of technological and digital tools, including AI[13], is noteworthy. According to The Alan Turing Institute's aforementioned study, AI tools are already being used by some **regulatory authorities**, for instance, to identify the **risk of illicit conduct by financial advisors** or to independently **verify the appropriateness of risk models** of supervised intermediaries. The Financial Stability Board[21], for example, highlights the launch by the Bank of Italy of a **textual sentiment analysis model** to monitor depositors' confidence levels by analysing Twitter posts, for challenging the funding models of supervised banks. In a recent blog post[14], the **European Central Bank (ECB)** provided some details regarding certain areas of activity where it is already **applying AI tools** (both in its role as a monetary policy authority and as a banking sector supervisory authority). To cite a few examples, albeit not exhaustively, the ECB states, for instance, that it uses **Machine Learning techniques** for classifying data used for **statistical and decision-making purposes** across various reference sectors. It also employs **Natural Language Processing techniques** to analyse vast amounts of documents and extract relevant information (for example, applying entity recognition tools to identify information associated with a specific bank among news articles or newspapers).

4.2 Potential Challenges for the Banking Sector

In addition to the opportunities that Artificial Intelligence offers to the banking system, there are also notable **potential challenges and difficulties** that a conscious adoption of AI entails. Among the possible challenges for the banking system, the following points of concern are highlighted.

5. Regulation of Artificial Intelligence: an Overview

As described in previous chapters, the phenomenon of AI has seen increasing diffusion and application only in recent years. Its use and adherence to existing regulatory frameworks are therefore not yet fully defined, nor are they fully covered (except partially) by current regulations. In fact, the risks described in previous chapters make some form of **regulation necessary**, partly new, to **ensure the security, reliability, and fair dissemination of AI techniques**. This regulation should ensure both a correct and functional use by companies and guarantee that customers and society as a whole can benefit from **fair and transparent technology**.

5.1 The Regulation in the European Context

Regarding the European Union, the European Commission has initiated a regulatory process on the application of Artificial Intelligence a few years ago, ensuring its **safe and reliable development**.

[12] Further insights into the main applications of AI tools in banking and finance can be found in the aforementioned reports by The Alan Turing Institute[28] and Deutsche Bank Research[12].
[13] Regarding the applications of AI for central bank, see [4].
[14] See The ECB Blog.

Type of application	Examples
Front-office and customer interaction	AI tools can be used for automating customer service and assistance activities, which can be managed through the use of chatbots (virtual assistants capable of generating personalised responses based on available data) and the application of Robotic Process Automation (RPA) techniques for providing automated and customised recommendations to clients regarding investment choices (robo-advice) or for delivering basic services (e.g., guiding a new client in opening a bank account).
Know Your Customer	Another area of application for AI solutions concerns streamlining Know Your Customer (KYC) processes, reducing the time associated with document verification by leveraging a broader spectrum of data to verify the reliability of the information provided and the background of customers. AI techniques can be used, for example, to verify the authenticity of images contained in customer documents (performing cross-checks with other documents pertaining to the same subject) and identifying and flagging any potentially anomalous cases requiring human review.
Credit risk scoring	One of the most significant applications of AI in the banking sector concerning risk management activities relates to credit risk scoring of customers, which can be executed using predictive models based on AI to support lending decisions. Banks can benefit from an increasing amount of additional data, including data on current account movements or web interactions, for assessing creditworthiness[13]. The ability to leverage additional data allows for a quicker and more informed assessment of the potential borrower's creditworthiness, while also ensuring better access to financing for creditworthy counterparts whose historical data may not be extensive enough to be included in traditional statistical models. For an investigation into the adoption of AI techniques in credit assessment models in the Italian banking sector, reference is made to a recent study conducted by the Bank of Italy in 2022[2].
Anti-money laundering and fraud detection	Another area of significant application of AI solutions for risk mitigation purposes is related to fraud detection and anti-money laundering activities. AI techniques can be used to analyse (even in real-time) vast volumes of customer transaction data, identifying any anomalies and flagging potentially suspicious transactions for further scrutiny. Furthermore, AI techniques can be employed to analyse corporate financial data and intercept any financial irregularities.
Mitigation of the risk of illicit conduct	AI tools can also be used to identify potential insider trading situations. In particular, through clustering algorithms, it is possible to analyse the trading activities of a specific actor, identifying any discontinuities in their behavior, and detecting the presence of groups of investors acting similarly even in the absence of price-sensitive public information, thus highlighting potential illicit conduct.
Model validation	AI techniques can be used by the Internal Validation Functions of financial institutions to develop challenger models to test the robustness of models developed within the bank.
Algorithmic Trading	AI systems can be used to develop algorithmic trading tools to identify anomalies and/or predict market dynamics. Some examples include AI tools used to analyse investor sentiment or investigate the impact of social media on the performance of financial instruments.
Calculation of Capital Requirements	The same techniques used for credit scoring purposes can be employed to estimate credit risk parameters for calculating capital requirements through Internal-Ratings Based (IRB) models. In this regard, the Follow-up Report on the use of Machine Learning for IRB Models published by the EBA in August 2023[14] illustrates the state-of-the-art practices among European banks regarding the usage of Artificial Intelligence for IRB models. Specifically, the report highlights that, despite some challenges faced by banks, such as issues related to the interpretability of such models or the lack of adequate skills for their development and validation, models based on AI techniques are already widespread. They are predominantly used in specific phases of the IRB approach, including the core development of modeling, validation activities using "challenger" models, and collateral assessment activities.
Regulatory Compliance	Among these, the possibility of streamlining regulatory reporting activities is highlighted. AI tools can be used to support the management of the increasing volumes of data to be reported within the regulatory reporting requirements of financial institutions. For example, such tools can be used to expedite data processing and verify its quality, thus identifying any errors and other data quality issues and anomalies to be reported to data analysts.

TABLE 3: *Examples of AI applications in the banking sector*

Below is a brief chronology based on **key regulatory milestones**:

1. The publication of the ***GDPR (General Data Protection Regulation)* in 2018**, although not directly related to AI regulation, is considered a foundational starting point for the development and application of these technologies. As seen, AI relies heavily on data usage, and it's impossible to separate the fundamental issues of its correct and fair use from those related to the treatment and protection of sensitive data.

2. The release of the ***Ethics Guidelines for Trustworthy AI***[18] in 2019 and the ***Assessment List for Trustworthy Artificial Intelligence*** (ALTAI)[19] in 2020 was carried out by the Independent High-Level Expert Group (HLEG) on Artificial Intelligence appointed by the European Commission. Both publications focus on additional **fundamental aspects** necessary for creating a conducive environment for the development and application of AI techniques that are safe and reliable. Specifically, they suggest conditions useful for **implementing a reliable and ethical framework** that takes into account all possible aspects of respecting individuals' **fundamental rights, preventing potential harms, ensuring fairness, and enhancing transparency and interpretability** (as already mentioned among the potential risks in Chapter "Potential Risks Associated with the Use of Artificial Intelligence"), with particular attention to impacts on minorities and vulnerable groups. Additionally, within the ALTAI, a **checklist** is proposed for the first time, which is useful for evaluating, through **self-assessment** by each entity that - regardless of the development stage - uses AI, the **level of compliance** with the mentioned requirements. The outlined requirements generally mirror those listed within the proposed AI Act, namely:

    - ***Human Agency and Oversight***: requirements aimed at ensuring that the **decision-making**

Type of application	Examples
Availability and quality of input data	Banks face the following challenges related to input data: • Availability of a sufficiently large volume of data: training AI models requires vast amounts of data, enabling them to identify relationships that may not be evident in smaller datasets and contribute to increasing the accuracy of the estimates provided. The use of large datasets is also essential for reducing the risks of overfitting, where a model fits too closely to the known data during training and fails to generalize its predictive ability to new data. The growth in the quantity of available data, driven by the increasing digitisation of the financial sector, should over time help mitigate limitations due to this obstacle; • Availability of high-quality data: in addition to the volume of data required to produce reliable AI solutions, the data itself must be of good quality. If the data used is incomplete, inaccurate, distorted, or inconsistent, the model's predictions can be imprecise or distorted, exacerbating the potential risks of bias and discrimination described in Chapter 3; • Compliance with data privacy requirements, which can nullify the potential effect of data availability and quality (as well as expose the bank to legal liabilities). In this regard, Art. 5 of the aforementioned European regulation on the protection of personal data (GDPR) requires banks to demonstrate compliance with the principles of personal data protection. This requirement becomes even more relevant in the use of AI techniques, as they are capable of reconstructing "protected" information from their relationship with other variables in the dataset, even in the absence of sensitive information among the data available. It is therefore evident that banks need to adopt appropriate measures for responsible data management in compliance with regulatory requirements, ensuring that models do not leverage sensitive personal data where its use is not permitted[16]. The relevance of these requirements is greater when the decisions based on such data have consequences for individuals' access to fundamental rights, such as access to credit or the possibility of obtaining a certain job[17]. This need is also reflected in one of the four pillars outlined by EBA [13] for the use of big data and advanced analytics, which involves banks identifying and categorising the data protection constraints used and anticipating the definition and implementation of technical and organizational measures aimed at ensuring data security.
Interpretability and Explainability	As mentioned in Chapter 3, the complexity of relationships modeled by AI solutions can make it challenging to interpret the results. Banks must therefore balance the trade-off between model performance and interpretability. Additionally, they need to equip themselves with appropriate interpretability techniques to increase the transparency of these models and explain the relationships between inputs and outputs that have produced a particular result.
IT Infrastructure	Banks need to adapt their IT infrastructure, which, if outdated, can hinder the introduction of modern artificial intelligence techniques, with negative effects on their diffusion. Particularly, the hardware and software systems currently in use by banks may lack the computational power and storage capacity needed to effectively develop and manage next-generation models, which heavily rely on these resources. One of the initial responses adopted by banks to address the challenges of potential technical obsolescence, which are difficult to resolve quickly and in line with market competitiveness, is the adoption of cloud systems. This involves purchasing remote servers to outsource computationally intensive processes and hardware requirements. The adequacy of IT infrastructure is also one of the pillars outlined by EBA for the use of big data and advanced analytics, envisioning a technological structure based on three components: infrastructure, data platform, and processing component. The infrastructure includes network resources for data transmission, computing resources, and storage resources. The data platform manages all the data used by the analytics system, allowing access to them, while the processing component supports the software required for analysis processing, in line with their volume and speed[18].
Availability of adequately skilled personnel	The use of Artificial Intelligence-based tools poses a challenge in terms of skills, as many of these systems require advanced mathematical-statistical knowledge as well as specific programming and data analysis skills. Without these competencies, company resources may not fully leverage the potential of such systems, encountering difficulties in understanding and interpreting the results provided by the models, or failing to identify and manage aspects of model correctness or bias and distortions in the results appropriately. Additionally, ensuring an adequate level of understanding of the results produced by these systems at all levels of the organisation, starting from top management, is necessary. Therefore, it is essential for banks to anticipate and plan for these changes and provide their structures with the necessary support and training to adapt to technological advancements effectively.
Uncertain and evolving regulatory landscape	Given the novelty factor and considering the paradigm shift brought about by Artificial Intelligence techniques, the level of regulation on these issues in the near future remains uncertain. Currently, what emerges is an evolving and partially heterogeneous regulatory forecast across different jurisdictions. The evolving regulatory landscape thus poses a challenge for financial institutions looking to embark on their AI adoption journey, as they must strive to anticipate regulatory developments in the sector to avoid being caught unprepared for potential future regulatory scenarios. A dedicated overview of AI regulation is provided in the next chapter, which offers an insight into the current state of regulation and potential future developments in this area, with a predominant focus on the European context before briefly analysing regulatory trends at the international level.
Regulatory Body Recommendations	With regard to all the challenges described above, Banca d'Italia [3] recommends ensuring the centrality of governance to adequately manage the risks associated with AI, strengthening data governance measures, and ensuring continuity and integrity of information systems. Furthermore, Banca d'Italia recommends keeping human responsibility as the ultimate decision-maker, with AI serving as a supportive tool rather than a guiding force in decision-making processes. Additionally, concerning the possibility for banks to use AI techniques for calculating capital requirements in credit risk (IRB), the recommendations provided by the EBA [14] are highlighted, reaffirming the challenges and critical issues mentioned in earlier chapters of this document.
Algorithmic Trading	AI systems can be used to develop algorithmic trading tools to identify anomalies and/or predict market dynamics. Some examples include AI tools used to analyse investor sentiment or investigate the impact of social media on the performance of financial instruments.

***TABLE 4:** Challenges in AI Adoption by Banks*

process of AI systems is supervised by humans to ensure that the resulting decisions respect individuals' fundamental rights;

- ***Technical Robustness and Safety***: this requirement is aimed at ensuring the **technical safety and reliability of AI systems** and minimising any unexpected and unintentional damages;
- ***Privacy and Data Governance***: data governance requirements aimed at **ensuring the quality and integrity of the data used** and compliance with the principles of protecting sensitive personal data;
- ***Transparency***: this requirement aims to **ensure the transparency and explainability** of the results produced by AI systems and full communication of their potential limitations;
- ***Diversity, Non-discrimination and Fairness***: these requirements aim to ensure the **absence of biases** in the data supporting AI systems and the **absence of discrimination** in the decisions resulting from them, also ensuring the **fair and non-discriminatory accessibility of AI solutions**;
- ***Societal and Environmental Well-being***: these requirements aim to **ensure that AI systems are compatible with sustainability principles**, both environmentally and socially;
- ***Accountability***: this final requirement aims to **trace back the causes of any potential negative effects** resulting from the application of AI systems.

3. The ***White Paper on Artificial Intelligence***[17] published in 2020 by the European Commission, continues the aim of outlining a **competitive and reliable framework for the development and application of Artificial Intelligence** within the European Union. The document defines the objective of creating both an "ecosystem of excellence, "which encourages the adoption of AI-based solutions, and an "ecosystem of trust," which ensures compliance with regulations by institutions and guarantees citizens the necessary level of trust for the widespread adoption of such technologies".

4. The ***Artificial Intelligence (AI) Act***[16], whose proposal dates back to April 2021 and was agreed upon in December 2023 following discussions between the Commission, Parliament, and the European Council.[15] , is aimed at the entire spectrum of businesses and activities that use or intend to use AI methodologies for their purposes. This regulation applies to all AI tools[16] used within the **European Union**, even if produced by providers from third countries. The AI Act follows a **risk-based approach**, distinguishing various levels of inherent risk associated with potential AI tools based on their nature and use. It identifies different requirements and obligations accordingly:

   - AI systems characterised by an **unacceptable risk**, and therefore **prohibited** (so-called "**prohibited practices**"). This category includes all systems whose use is deemed **unacceptable as it contradicts the values of the Union**, namely those that **violate fundamental rights**. These are practices that could potentially **manipulate individuals** through subliminal techniques without their awareness or methods used to exploit the vulnerabilities of specific individuals to materially distort their behavior in ways that could cause personal harm to themselves or others. Prohibited practices include those related to the **attribution of a social score** generated by AI for general purposes by public authorities and the use of real-time remote **biometric identification systems** in public spaces for surveillance purposes.
   - **High-risk AI systems**, meaning those AI systems that pose a **high-risk to health and safety or to the fundamental rights of natural persons**. In line with the risk-based approach, such systems are **allowed on the European market only if they meet certain**

[15] See Artificial Intelligence (AI) Act.

[16] According to Annex 1 of the proposal of *AI Act*, the following AI techniques and approaches fall within the scope of regulation: "Machine learning approaches, including supervised learning, unsupervised learning, and reinforcement learning, with use of a wide range of methods, including deep learning; b) logic-based and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inferential and deductive engines, (symbolic) reasoning, and expert systems; c) statistical approaches, Bayesian estimation, search methods, and optimisation".

mandatory requirements and pass a conformity assessment. The classification of an AI system as high risk is based on its intended purpose, following current product safety legislation. Consequently, the classification as high risk depends not only on the function performed by the AI system but also on its **specific purpose and how it is used**. Annex III of the regulation identifies the list of AI systems classified as high risk, including, for example, systems used to determine **access to essential services** (including access to credit), systems used for **recruitment and human resources** management purposes (e.g., determining termination of employment), and systems used for **critical infrastructure management** purposes (including traffic management and the provision of water, gas, and electricity).

According to the regulation, the European Commission will **maintain a list of high-risk AI systems**, containing a limited number of these systems whose risks have already materialised or may materialise in the near future. Following the **risk-based approach**, the Commission may **expand the list of high-risk systems** by applying a series of criteria and a risk assessment methodology. Regarding the specific requirements for systems classified as high risk, the regulation broadly follows what has been outlined in the ALTAI document (previously described), with requirements namely for:

- ***Data and data governance***: requirements aimed at ensuring that the data used for model development and training meet **quality standards**, particularly regarding **model design choices, data collection and preparation processes,** assumptions about the data, a prior assessment of data availability, suitability, and deficiencies, as well as an assessment of potential biases;
- ***Technical documentation***: this requirement aims to ensure that each high-risk AI system intended for commercial use is accompanied by **accurate and up-to-date informational documentation**, enabling authorities to verify the system's compliance with regulatory requirements;
- ***Record keeping***: this requirement ensures that all AI systems in use maintain a **continuous and systematic record of events and operations performed**;
- ***Transparency and Provision of information to users***: according to these requirements, AI systems must be structured to ensure that users can **interpret and use their outputs correctly**, if necessary by providing concise and clear instructions to users;
- ***Human oversight***: this requirement aims to ensure that high-risk AI systems are designed to **allow human supervision** through measures such as incorporating controls into the system. Such supervision aims to prevent risks to health, safety, and fundamental rights during the use of AI, assigning these oversight activities to individuals fully capable of understanding the system's operation and intervening as necessary;
- ***Robustness, Accuracy and Security***: these requirements aim to ensure that high-risk AI systems are **accurate, robust, and computationally secure** throughout their entire lifecycle. At the same time, models on the market must be resilient to biases, data inconsistencies, and interferences with other systems.

- **Low-risk AI systems**, for which **compliance with minimum transparency standards** is required, for example, concerning the need to **inform users that they are interacting with an AI system** and, in the case of content generated by AI tools that may be misconstrued as authentic, to **disclose that the content has been manipulated or generated using AI tools**.

5.2 Worldwide AI Regulatory Trends

In its proposal for the AI Act, the European Commission states that the requirements established for high-risk AI systems are broadly **consistent with other international recommendations and principles**, ensuring that the European regulatory framework for AI is compatible with those adopted by the international partners of the European Union. Below is a brief overview of the **trends and possible regulatory developments on Artificial Intelligence** in the other main countries of the world, namely the **United Kingdom**, the **United States of America**, and **China**.

In the United Kingdom, the **policy paper** published by the Department for Science, Innovation, and Technology promotes an approach to AI regulation aimed at fostering an **environment of innovation and achieving specific goals** to make the country a scientific and technological "superpower"[17]. In terms of **AI governance approach**, the main difference from the rule-based approach of the European Union lies in the proposal of a "sectoral regulatory framework", anchored to its existing and widespread network of regulators and laws, which is therefore **less centralised**. The UK's approach is based on **two main elements**: the **AI principles** that existing regulators will be tasked with implementing, and a series of **new "central functions"** to support this implementation activity. In particular, the **principles** underlying the framework to guide responsible AI use are as follows:

- ***Safety, security and robustness***: AI systems must function robustly and securely throughout their lifecycle, and their risks must be identified, monitored, and managed continuously;
- ***Appropriate transparency and explainability***: AI systems must provide adequate levels of transparency, meaning sufficient information to the parties involved in using an AI system, and explainability, allowing the involved parties to access, interpret, and understand the decision-making process of the system;
- ***Fairness***: AI systems must not limit the rights of individuals or organisations, and must not create unjustified discrimination against certain individuals;
- ***Accountability and governance***: adequate governance measures must be provided to ensure an appropriate level of supervision over the provision and use of AI systems, with clear lines of responsibility established throughout the lifecycle of such systems;
- ***Contestability and redress***: where appropriate, users or affected third parties must be able to contest a decision or result produced by an AI system where it causes harm or there is a material risk of harm.

While these requirements may appear in line with those outlined in European regulation, a substantial **difference** lies in the fact that, for the UK, their implementation will not be initially based on legislation but rather **progressively implemented** by existing authorities. This approach aims to prevent new rigid legislative requirements from stifling innovation and reducing the ability to respond quickly and proportionately to future technological advancements.
Regarding regulatory trends in the United States, a **lighter approach** to AI regulation is observed, which is less extensive and impactful compared to what has been seen in the EU and the UK. Despite the expressed intent to introduce federal legislation to regulate AI specifics, the country's regulatory framework largely relies on **voluntary guidelines** at present. For instance, the ***AI Risk Management framework*** [31] published in early 2023 by the *US National Institute of Standards and Technology* (NIST) or **industry self-regulation**, mainly targeted at major tech companies pioneering this field. In this context, noteworthy is the publication of the ***Blueprint for an AI Bill of Rights****[30]*, by the **Office of Science and Technology Policy (OSTP) of the White House**. This white paper aims to outline a **minimum set of guiding principles** for the development and conscious adoption of AI techniques. While these principles are non-binding, they are intended as **guidelines**, and they include:

- ***Safe and Effective Systems***: AI systems must be developed and used safely and effectively, with independent evaluation confirming compliance with these characteristics.
- ***Algorithmic Discrimination Protections***: AI systems must be developed with adequate safeguards to prevent discriminatory effects and violations of fairness principles.
- ***Data Privacy***: AI systems must ensure data usage that is consistent with and compliant with the right to personal data protection, with data collection and usage only permitted with the consent of the individuals concerned.

[17] See in this regard Department for Science, Innovation and Technology, *A pro-innovation approach to AI regulation*, March 2023, see [11].

- ***Notice and Explanation***: the use of an AI system for certain decision-making purposes must be disclosed to the individual beforehand, who has the right to information about the reasons behind a specific decision made by the system.
- ***Human Alternatives, Consideration, and Fallback***: where appropriate, there must be the option to replace judgment based on the AI system with a human assessment, as well as the need for human intervention in cases where automatic systems produce clear errors.

At the federal level, it is also worth considering the publication, in October 2022, of a series of guidelines on **consumer protection** in the use of AI. However, these guidelines serve as general indications and do not have legally binding value. In general, therefore, there is currently **no legislative process** aimed at introducing a set of **legally binding rules** on the development and application of AI, which must therefore rely on the principles and guidelines mentioned above.
In the opposite direction is the regulation of Artificial Intelligence in China, which has adopted **new proposals and laws**[18], in recent years, including a set of **rules on recommendation algorithms** that came into effect in 2022 (introducing requirements for regular assessment of these algorithms regarding their effectiveness, fairness, and security[19]), In 2022, **regulations on synthetically generated content** (deep synthesis) were also introduced, obliging clear disclosure if content was created using AI. Finally, in July 2023, a draft regulatory package on **Generative Artificial Intelligence** was formulated[20]. This package includes licensing requirements for providers of generative AI solutions and aligns with aspects already present in the European regulatory proposal, such as **safety, transparency, and non-discrimination**. It also includes specific requirements, such as adherence to socialist values and a prohibition on producing content that incites against the state[21]. Due to its tendency to foresee updates and further expansions of the regulatory framework in response to new technological developments, China's AI regulations are described as **iterative**. Compared to European AI regulations, China's approach can be considered more "vertical," focusing on individual applications of the technology and differing from the European regulation analysed earlier, which takes a more cross-cutting approach covering all possible applications of AI technologies.

6. Conclusions

In the previous pages, an overview has been provided regarding the **definition and diffusion of Artificial Intelligence**, along with the **potential benefits and related risks** that can arise from its increasing adoption.
As seen, AI solutions offer the potential for businesses to **enhance their efficiency and productivity** and to **guide them in the transformation processes** dictated by their digital strategies. At the same time, these benefits are accompanied by **risks and challenges**, which must be managed carefully and consciously in order to **maximise the long-term benefits** associated with the use of such new technologies. To do this, companies, including those in the banking sector, will be required to **invest in acquiring adequate skills and IT infrastructure** that allow them to make the most of AI in a secure manner while ensuring respect for **fundamental rights and principles of ethics and fairness**. Furthermore, the continuously **evolving regulatory landscape**, as well as being partially misaligned with the jurisdictions of the major global economic players, will require companies to remain vigilant about what the **future regulatory scenario** may entail, anticipating possible developments. Regarding this, considering the risks associated with AI as described within the document and beyond the regulatory requirements applicable to AI following the conclusion of the ongoing regulatory process, a set of **guiding principles** is outlined below to be evaluated preventively in the development and use of Artificial Intelligence solutions:

- **Accuracy and replicability**: ensure that the results produced by the AI system are reliable and that the system provides indications of the probability of errors. Also, ensure that the results of the AI system can be reproduced by third parties.

[18] For an overview of these regulatory interventions, see [7].
[19] Friedrich Ebert Stiftung, *China's Regulations on Algorithms. Context, impact and comparisons with the EU*, see [23].
[20] Cyberspace Administration of China, *Interim Measures for the Management of Generative Artificial Intelligence Services*, see [10].
[21] Forbes, *How Does China's Approach To AI Regulation Differ From The US And EU?*, see [22].

- **Communication**: clearly and proactively communicate the benefits and limitations of the AI system to all stakeholders. In the case of systems that interact autonomously with users, transparently communicate to users that they are interacting with an AI system.
- **Data privacy**: during the development of an AI system, ensure that the data used or collected by it is processed and stored consistently and in compliance with applicable data protection regulations. Additionally, ensure that any sensitive data collected by the system is not used to produce discriminatory decisions against users. During the use of an AI system, especially within the scope of using Generative AI solutions, ensure that sensitive and/or confidential data is not improperly disclosed to the system.
- **Data quality**: during the development of an AI system, ensure that the data used for training the system is of adequate quality and does not contain errors, inaccuracies, or distortions that could alter the decision-making process of the system.
- **Equity and non-discrimination**: ensure respect for human dignity, individual freedoms, and principles of fairness and equality in the development and use of an AI system. Minimise the risk of generating unjustified discrimination based on individuals' "sensitive" characteristics by adopting measures to ensure that the data used by the systems are accurate, unbiased, and representative of the context in which the system results are used. Pay particular attention to cases where the use of the system impacts vulnerable individuals or minority groups at risk of discrimination.
- **Technical skills**: promote training and awareness among stakeholders involved in the development or impacted by the use of the AI system, at all levels of the organisational structure, to ensure adequate understanding of the results produced by the system and management of associated risks.
- **Technical robustness and security**: adopt appropriate measures to ensure the technical robustness of the AI system through testing activities, continuous monitoring of stability and proper functioning over time, and verification of its compliance with expected behavior. Additionally, ensure the system's resilience to malicious attacks (cyberattacks) that could alter its operation through appropriate cybersecurity measures (e.g., data protection measures against unauthorised or illicit processing and monitoring processes during both training and post-release phases).
- **Human oversight**: provide a mechanism for human oversight during the development and use of an AI system through the presence of adequate controls and the possibility of manual intervention during the decision-making process and system operation. Clearly document the roles and responsibilities of individuals involved throughout the AI system's lifecycle.
- **Transparency, explainability, and traceability**: adopt appropriate measures to ensure the explainability of the results produced by the AI system to transparently communicate them to involved stakeholders (e.g., apply adequate Explainable AI techniques to elucidate the relationships between the system's inputs and outputs). Additionally, adopt measures to ensure the traceability of the system and any errors produced by it, clearly and comprehensively documenting the data used by the system and its development process.

References

[1] **AIFIRM, Position Paper No. 33.** *Artificial Intelligence and Credit Risk. Possible uses of alternative methodologies and data in internal rating systems*, January 2022.

[2] **Banca d'Italia.** *Intelligenza artificiale nel credit scoring. Analisi di alcune esperienze nel sistema finanziario italiano, Questioni di Economia e Finanza,* October 2022.

[3] **Banca d'Italia, Intervento di Giuseppe Siani, Capo del Dipartimento Vigilanza bancaria e finanziaria della Banca d'Italia, Boston Consulting Group.** *AI-driven bank: Opportunità e sfide strategiche per il sistema finanziario e la vigilanza,* October 2023.

[4] **Bank for International Settlements.** *Artificial Intelligence in Central Banking - Bulletin No. 84*, January 2024.

[5] **Bank of England.** *Machine Learning in UK financial services*, October 2022.

[6] **Cao,Y., Li,S., Liu,Y., Yan,Z., Dai,Y., Yu,P.S., Sun,L.** *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT,* August 2018.

[7] **Carnegie Endowment for International Peace.** *China's AI Regulations and How They Get Made*, July 2023.

[8] **Castelnovo, A., Crupi, R., Del Gamba, G., Greco, G. Naseer, A., Regoli, D.and San Miguel Gonzalez, B.** *Addressing Fairness in the Banking Sector*, February 2021.

[9] **Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I.G. and Cosentini, A.C.** *A clarification of the nuances in the fairness metrics landscape, Scientific Reports Volume 12, No. 4209*, March 2022.

[10] **Cyberspace Administration of China.** *Interim Measures for the Management of Generative Artificial Intelligence Services*, July 2023.

[11] **Department for Science, Innovation and Technology.** *A pro-innovation approach to AI regulation,* March 2023.

[12] **Deutsche Bank Research.** *Artificial intelligence in banking. A lever for profitability with limited implementation to date,* June 2019.

[13] **European Banking Authority.** *Follow-up Report from the Consultation on the Discussion Paper on Machine Learning for IRB Models*, August 2023.

[14] **European Banking Authority.** *Report on Big Data and Advanced Analytics*, January 2020.

[15] **European Commission.** *Proposal for a Directive of the European Parliament and of the Council on Consumer Credit*, June 2021.

[16] **European Commission.** *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts*, April 2021.

[17] **European Commission.** *White Paper on Artificial Intelligence - A European approach to excellence and trust*, February 2020.

[18] **European Commission - High Level Expert Group (HLEG) on Artificial Intelligence.** *Ethics Guidelines for Trustworthy AI*, April 2019.

[19] **European Commission - High Level Expert Group (HLEG) on Artificial Intelligence.** *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*, July 2020.

[20] **EU Regulation 2016/679.** *EU Regulation 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation)*, April 2016.

[21] **Financial Stability Board.** *Artificial intelligence and machine learning in financial services. Market developments and financial stability implications*, November 2017.

[22] **Forbes.** *How Does China's Approach To AI Regulation Differ From The US And EU?*, July 2023.

[23] **Friedrich Ebert Stiftung.** *China's Regulations on Algorithms. Context, impact and comparisons with the EU*, January 2023.

[24] **International Monetary Fund.** *Generative Artificial Intelligence in Finance: Risk Considerations*, August 2023.

[25] **Korzynski, P., Kozminski, A. and Baczynska, A.** *Navigating leadership challenges with technology: Uncovering the potential of ChatGPT, virtual reality, human capital management systems, robotic process automation, and social media*, May 2023.

[26] **Molnar, C.** *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*, August 2023.

[27] **Rubicondo, D. and Rosato, L.** *AI Fairness: Addressing Ethical and Reliability Concerns in AI Adoption, Iason Research Paper Series*, March 2022.

[28] **The Alan Turing Institute.** *The AI Revolution: Opportunities and Challenges for the Financial Sector,* August 2023.

[29] **The Economist Intelligence Unit.** *Banking on a game-changer: AI Risk Management Framework*, 2022.

[30] **The White House - Office of Science and Technology Policy.** *Blueprint for an AI Bill of Rights*, October 2022.

[31] **US National Institute of Standards and Technology.** *AI Risk Management Framework*, January 2023.

[32] **US National Institute of Standards and Technology.** *Big Data Interoperability Framework*, June 2018.

[33] **Wach, K., Duong, C. D., Ejdys, J., Kazlauskaité, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J.and Ziemba, E.** *The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT*, May 2023.

Iason is an international firm that consults Financial Institutions on Risk Management. Iason is a leader in quantitative analysis and advanced risk methodology, offering a unique mix of know-how and expertise on the pricing of complex financial products and the management of financial, credit and liquidity risks. In addition Iason provides a suite of essential solutions to meet the fundamental needs of Financial Institutions.

ESSENTIAL SERVICES FOR
FINANCIAL INSTITUTIONS