Research Paper Series

AI Risk Management Frameworks

Nicola Mazzoni

Sara Martucci

Margherita Ranieri

November 2025

iason

ESSENTIAL SERVICES FOR
FINANCIAL INSTITUTIONS

Research Paper Series

Year 2025 - Issue Number 81

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

Front Cover: **Atanasio Soldati**, *Stanza*, 1951.

ESSENTIAL SERVICES FOR
FINANCIAL INSTITUTIONS

Executive Summary

This study aims to analyze AI Risk Management Frameworks (AI RMFs), exploring their role in promoting the safe, accountable, and transparent adoption of AI technologies within economic systems. The first part of the research provides a broad overview of the evolution of the AI market and its growing impact on strategic and operational processes, with a particular focus on the financial sector. The second part discusses the unique risks posed by AI systems, while the third part explores the regulatory responses to manage AI unique risks, with a particular focus on the EU AI Act. Finally, the fourth part analyzes several major AI RMFs developed by international and regional institutions, examining their guiding principles, technical requirements, and governance mechanisms. The study ultimately identifies common principles shared across regulations, guidelines, and AI RMFs, highlighting the strategic relevance of integrating AI governance into corporate strategy.

About the Author

Nicola Mazzoni:
Project Manager
He holds an MSc in Economics For Finance from the Catholic University of Milan and has furthered his expertise with postgraduate master's degrees in Business Analytics, Corporate Finance, and Innovation and Sustainability Management. He was actively engaged in Business and IT process implementation projects at one of Europe's most prominent banking institutions.

Sara Martucci:
Project Manager
She holds an MSc in Mathematical Engineering with a focus on Quantitative Finance from Politecnico di Milano. She has gained significant experience as a Project Manager, leading international projects and developing strong expertise in project governance and delivery. Currently, she is involved in engineering and automation projects within the CIB division of one of Europe's leading banking groups, where she contributes to the transformation of complex processes into scalable, reliable software solutions.

Margherita Ranieri:
Analyst
She holds an MSc in Management Engineering with a specialization in Finance from the Politecnico di Milano. She is currently working as a business analyst, actively engaged in Business and IT process implementation projects at one of Europe's leading banking institutions.

This document was prepared in collaboration with Nicola Mazzoni, who at the time was working for Iason Consulting.

Table of Content

AI Risk Management Frameworks

Nicola Mazzoni Sara Martucci Margherita Ranieri

Artificial Intelligence (AI), and particularly Generative AI (GenAI), is rapidly transforming industries worldwide, reshaping business models, and becoming a cornerstone of modern digital transformation. Driven by significant capital and human investments and a growing focus on productivity, efficiency, and innovation, AI technologies are being adopted across a wide range of sectors. However, this growing integration of AI also brings a complex set of challenges. As AI systems increasingly influence critical decisions, their unique characteristics, such as autonomy, opacity, and capacity for scale, raise important concerns around transparency, accountability, ethical use, and regulatory compliance. These challenges are amplified in highly regulated sectors like finance, where AI adoption must be carefully aligned with principles of financial stability, consumer protection, and institutional trust. Despite these concerns, the industry outlook for AI remains overwhelmingly positive. Barriers to adoption are progressively declining, supported by technological advances, cost reductions, and expanding availability of AI infrastructure. The continued rise in private investment, particularly from global leaders like the United States and China, further underlines the strategic relevance of AI for economic competitiveness and innovation leadership. The widespread deployment of AI technologies has intensified the need for coherent, harmonized regulatory frameworks capable of addressing the societal, ethical, legal, and economic risks associated with AI. Traditional regulatory approaches are increasingly proving insufficient, as AI's cross-sectoral and cross-border implications require a level of coordination that transcends national boundaries. In response, international and supranational bodies such as the OECD and UNESCO have developed guidelines to assist governments in navigating these challenges. At the same time, regional and national authorities have begun drafting or implementing specific regulatory mechanisms to manage AI risks, aiming to prevent legal fragmentation and ensure consistent oversight. A leading example of this evolution is the European Union's AI Act, which represents the first cross-jurisdictional regulatory framework explicitly dedicated to AI. The Act adopts a risk-based approach that classifies AI systems according to their potential impact on safety, fundamental rights, and societal stability. This paper aims to provide a concise overview of the evolution of the AI market and the corresponding regulatory responses to emerging challenges, with a particular focus on risk management frameworks. While offering context on the development of the AI market and the evolving regulatory landscape, it seeks to examine the methodological approaches developed to assess and manage AI-related risks. Specifically, the paper explores the key features and practical implementation of structured risk management frameworks to evaluate how institutions are addressing major AI-specific concerns such as bias, opacity, accountability gaps, and systemic instability.

1. AI: a Brief View

With its spread across various industries and daily activities, the term "AI" is becoming increasingly overused. While artificial intelligence is a branch of study that encompasses several types of models and algorithms, the term is often incorrectly used to indiscriminately refer to all potential AI applications that fall under the broader AI umbrella. Artificial Intelligence (AI) "is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable".[15] The words of J. McCharty[1]

[1] J.McCarthy is considered one of the founders of Computer Science. His 1956 speech at Dartmouth University introduced for the first time the term Artificial Intelligence.

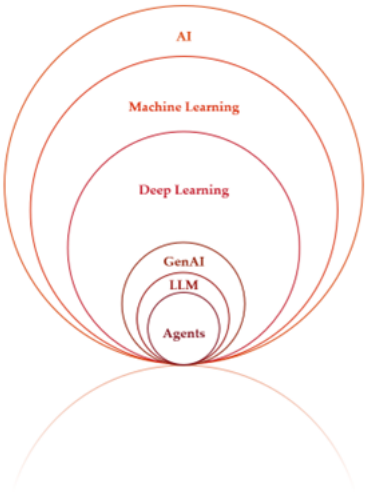

FIGURE 1: AI Methods

describe a branch of computer science focused on studying and developing systems capable of performing complex tasks typically requiring or associated with human intelligence. These include learning from data, understanding natural language, recognizing patterns, solving problems, and making autonomous decisions. The field of study of AI encompasses several different typologies of approaches and models that can serve different purposes:

- "**Machine Learning:** mathematical and statistical methods enabling machines to learn from data and improve with experience. It comprehends:
  - **Supervised Learning:** models that learn from input features and targets (training dataset) to generalize the model and make predictions on unseen data (e.g. identifying spam emails, image classification, etc.).
  - **Unsupervised Learning:** models that work with unlabeled data, aiming to discover patterns and relationships within datasets that lack predefined target labels (e.g. clustering, anomaly detection).
  - **Reinforcement Learning:** models that rely on agents that learn through their interaction with the environment, thanks to a reward system that assesses the quality of the agent's actions."[19]
- **Deep Learning:** a sub-branch of AI that refers to those machine learning models whose structure is based on the study of multiple successive layers, allowing algorithms to learn data representations at multiple levels of abstraction.
- **Generative AI:** GenAI is a subset of deep learning models focused on creating content. It is capable of generating diverse types of data, such as text, images, video, audio, and code, by learning patterns from large datasets and using that knowledge to produce new and original outputs. GenAI models are typically based on neural networks that are trained on massive amounts of data to understand the statistical relationships between elements such as words, pixels, or sounds. Once trained, these models can generate highly realistic and complex content that mimics human creativity, supports human labor by automating repetitive or time-consuming tasks, and serves as a powerful tool for enhancing productivity across multiple domains. They achieve this by predicting the next most likely element (e.g. word in a sentence) based on the patterns learned during training.
- **Large Language Models:** LLMs are a subset of GenAI models designed to understand and produce human-like language. Trained on vast text datasets, they aim to generate coherent and contextually relevant text by using statistical and probabilistic models to predict the next word or token in a sequence.

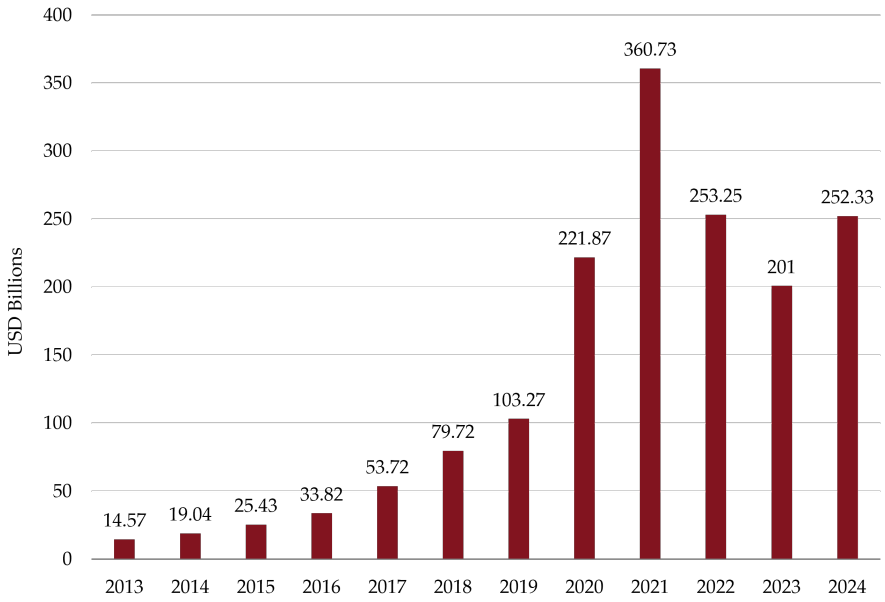

FIGURE 2: Global Corporate Investment in AI [18]

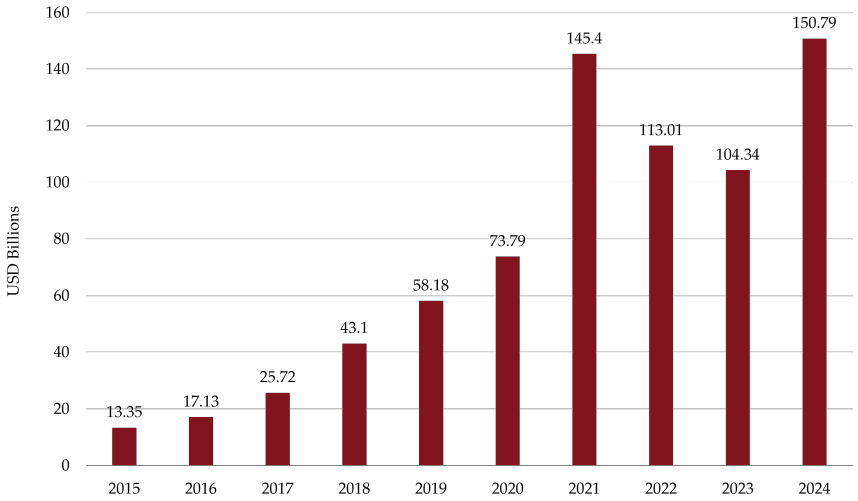

FIGURE 3: Global Private Investment in AI [18]

- **Agents:** agents are autonomous systems powered by Generative AI that can make decisions, complete tasks, and learn from experience. They can interact with their environment or users, adapt to changing conditions, and automate complex workflows across various domains such as customer service, robotics, and data analysis.

1.1 AI Market Evolution

Since 2022, with the first public release of ChatGPT, AI has undergone a profound shift, transitioning from a specialized domain to a central focus of public discourse and strategic planning across several industries. The rapid adoption of these technologies (in particular GenAI) has led to widespread interest among both businesses and policymakers, driven by AI's demonstrably transformative impact on productivity and economic development. Enterprises across diverse sectors, recognizing the transformative potential of AI, have seized the momentum to increasingly integrate it into both their operational frameworks and strategic planning. This burgeoning interest is mirrored by a significant increase in global AI investment over the past decade, now amounting to hundreds of billions of dollars worldwide. Investments in the industry (ref. Figure 2) have seen a substantial

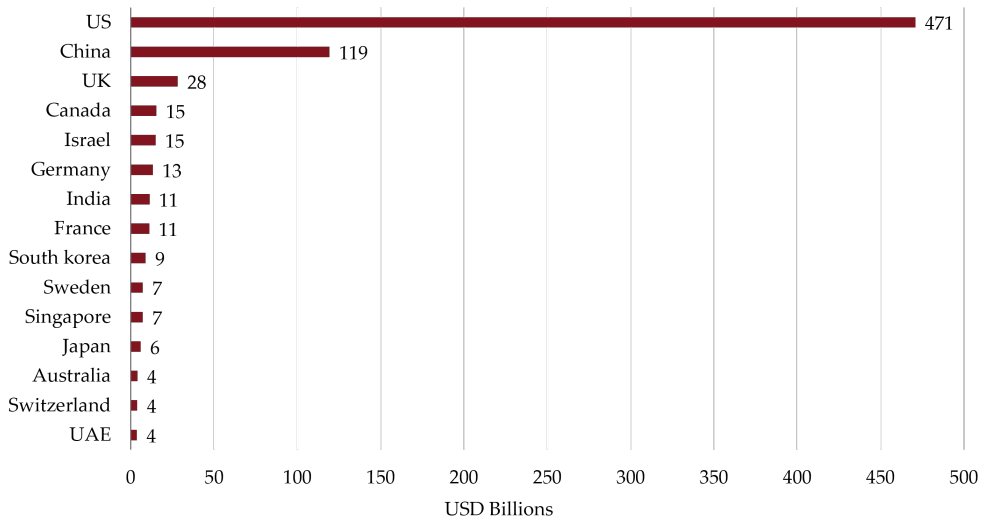

FIGURE 4: 2013-2024 AI Private Investment by Country [18]

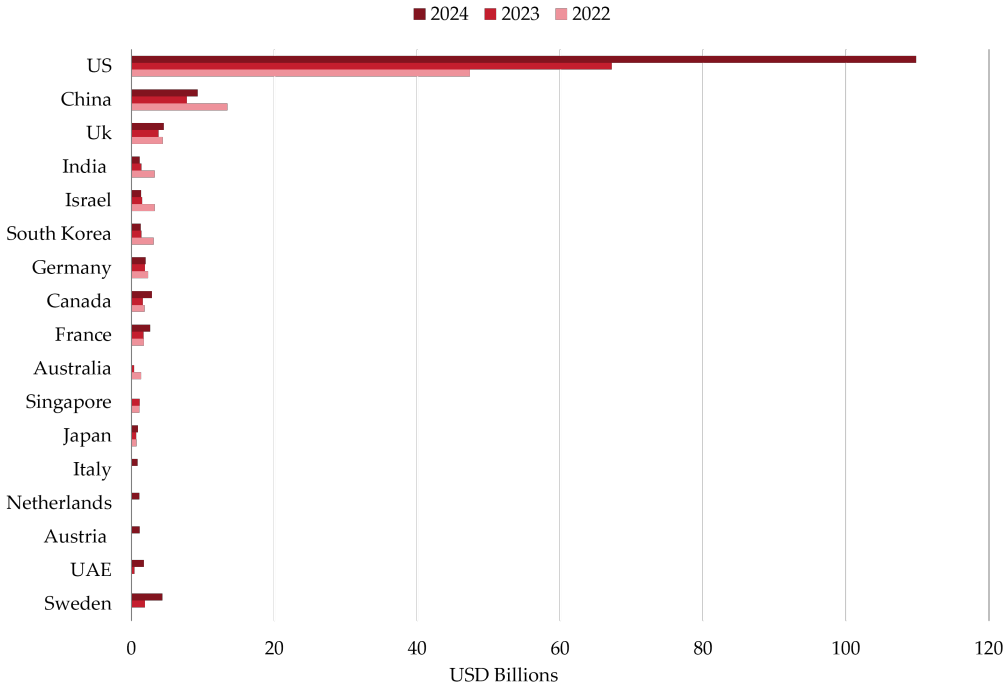

FIGURE 5: 2022, 2023, 2024 AI Private Investment by Country Comparison [18] [17] [16]

increase: from 2023 to 2024, investments in AI grew by approximately 26%, rising from 201$ billion to 252$ billion. Notably, over the past eleven years, total investment in AI surged from around 14$ billion in 2013 to 252$ billion in 2024, with a historical peak of 360$ billion reached in 2021, prior to the public release of ChatGPT. [18]
In the same period, the value of private investment (defined as the share of total investment excluding M&A, public offerings, and minority stakes) has also seen a substantial increase over the past year (ref. Figure 3), rising from 104$ billion to slightly over 150$ billion, marking a growth of approximately +45%. Looking at the period since 2013, private investment, despite a temporary slowdown in 2022 and 2023, has followed an overall upward trend, with its value more than tenfold over the past eleven years, growing from 13,34$ billion in 2013 to 150$ billion in 2024. [18]

Taking a deeper look at who is driving most of the private investment in AI (ref. Figure 4), it's not

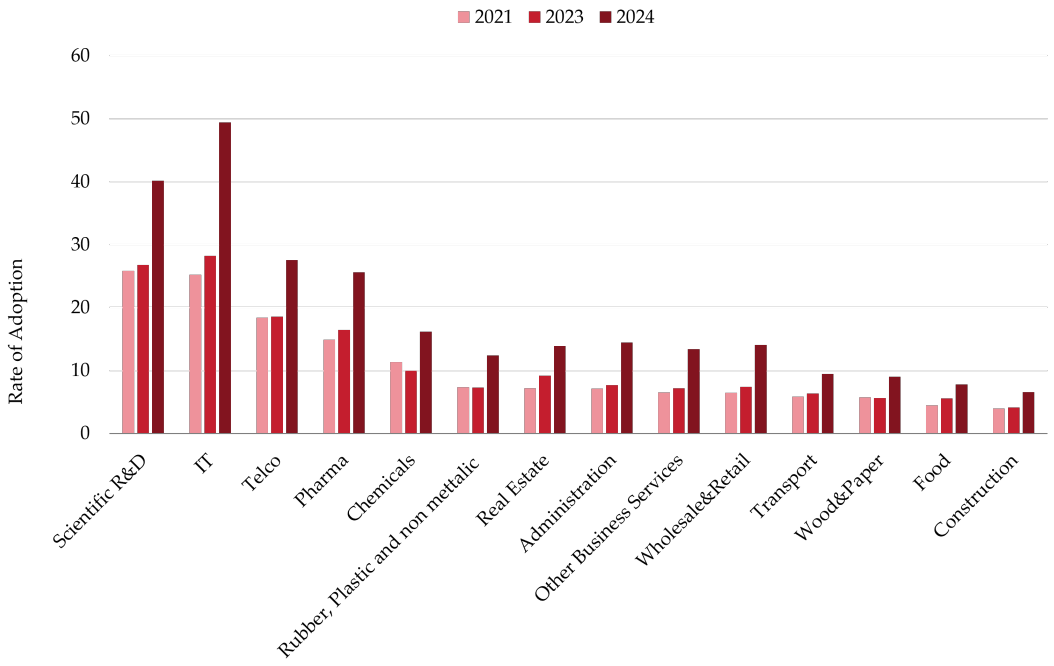

FIGURE 6: *EU Cross-Industry AI Adoption Between 2021-2024 [28]*

surprising that over the last eleven years, the United States and China have led the field, accounting for a combined 81% (590$ billion) of the total aggregate investment among the top 15 countries investing in AI. [18]

In the last 11 years, only three EU countries, Germany, France, and Sweden, have consistently carried out investments that placed them among the top providers of private AI investment. However, a closer look at the past three years reveals interesting insights. While the United States and China continue to lead AI investments, other European countries such as Italy (0,86$ billion), Austria (1,15$ billion), and the Netherlands (1,09$ billion) have significantly increased their share of private AI investment entering among the global top 15 countries for private AI investment in 2024 (ref. Figure 5). In the same period, Singapore and Australia, which were among the top 15 investors in 2022 and 2023, fell out of the top 15 in 2024, although they still rank among the top investors in terms of aggregate investment over the past 11 years. [18] [17] [16]
The data [18] confirms that AI is increasingly becoming a key growth market, as evidenced by the ever-growing flow of investment capital. Market outlooks project a compound annual growth rate (CAGR) of over 34% for the next five years, with the total market value expected to exceed 3.000$ billion[21].
The growing importance of AI across industries is further demonstrated by rising adoption rates. In the EU (ref. Figure 6) [28], between 2021 and 2024, AI adoption has shown a consistent upward trend across various sectors, with an average increase of approximately 42% between 2021 and 2024. Notably, industries traditionally characterized by technology-intensive operations, such as IT (+96%) and pharmaceuticals (+72%), have experienced a steep increase in AI integration. However, it is particularly significant that sectors like administrative services (+103%), business services (+106%), and wholesale and retail commerce (+117%) have shown an even more substantial rise in AI adoption. [28]
It is clear that AI has expanded through both private and professional adoption. However, the emergence of ChatGPT has placed a strong spotlight on Generative AI. GenAI has gathered substantial and growing attention in recent years, primarily because of its accelerated adoption across an increasingly diverse array of industries. In a relatively short timeframe, GenAI technologies have exhibited a potential capacity to redefine operational workflows, optimize resource allocation, and support the automation of complex tasks. This rapid diffusion underscores GenAI's transformative potential, not only in terms of productivity gains and cost efficiency but also in enabling entirely

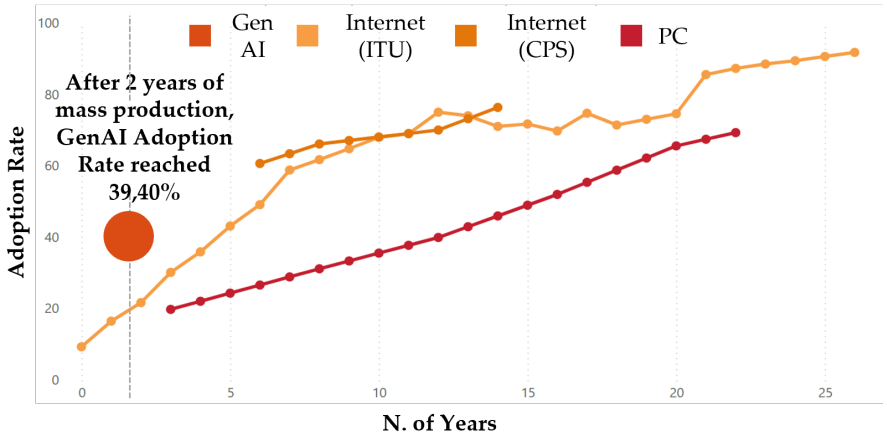

FIGURE 7: Technologies Adoption Rate, US [21], Data [2]

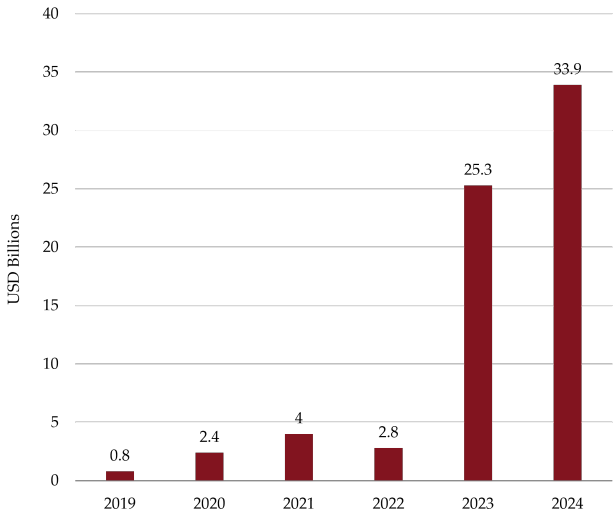

FIGURE 8: Total Private Investment in GenAI [18]

new modes of value creation. Looking ahead, the continued integration of GenAI into organizational structures is anticipated to drive fundamental shifts in business and institutional models, fostering enhanced innovation, scalability, and adaptability across both the private and public sectors. To better understand the extent of GenAI rapid breakthrough, it is useful to compare its adoption rate with that of other major disruptive technologies (ref. Figure 7)[2]. Within just two years of becoming widely available, GenAI has reached a workplace adoption rate close to 40% (U.S. data). To better understand the magnitude of this data, it is useful to consider that the internet required approximately five years to reach a comparable level of adoption, while personal computers (PCs) took nearly twelve years to achieve the same diffusion.[2]
Following the broader trend in Artificial Intelligence, Generative AI has attracted substantial private investment in recent years (ref. Figure 8)[18]. Notably, the growth in GenAI-related investments has significantly outpaced the average trend observed across the AI sector. Since 2019, private investment in GenAI has increased by 41 times within just five years. From accounting for a modest 0,8% of total private AI investment in 2019, equivalent to 0,8$ billion, GenAI has rapidly expanded its share to nearly 14% in 2024, reaching a total value of almost 34$ billion.[18]

1.2 AI Adoption in Financial Services

The financial services industry has historically served as a frontrunner in the adoption of technological innovations, leveraging them to enhance operational efficiency and maximize returns. In recent years, the integration of various forms of AI has expanded rapidly[12], with financial institutions

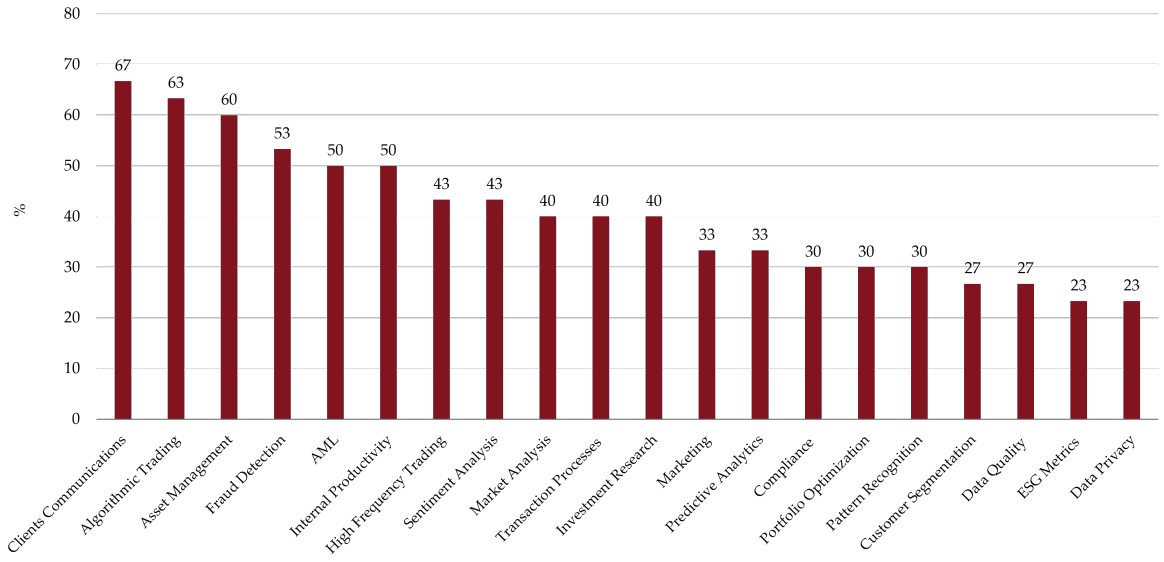

***FIGURE 9**: Total FI AI Application Across Business Functions[12]*

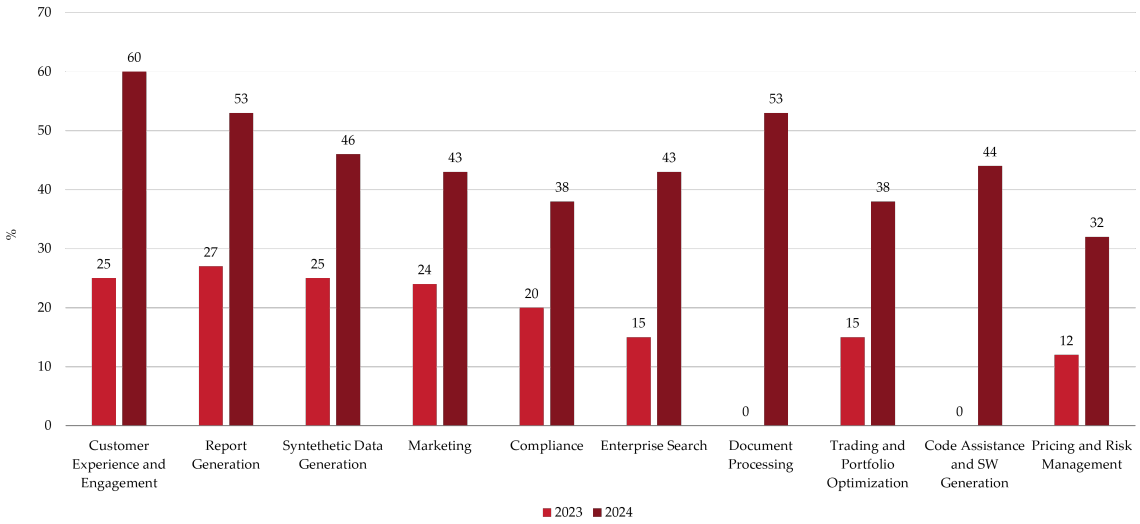

***FIGURE 10**: Top Generative AI Use Cases in Financial Services Industry [25]*

deploying AI across a wide range of operational and business processes, from generic administrative tasks to highly specialized, industry-specific functions (ref. Figure 9) [12].
This trend holds true also in the adoption of GenAI, which has experienced a rapid and widespread acceleration within the financial services sector. Over the past two years (ref. Figure 10), the adoption rate of GenAI has increased by more than 300%[20], reflecting its integration across a wide range of business functions. This substantial growth is not confined to general-purpose applications; rather, it encompasses both support activities and critical operational areas. Data[25] confirms this broad diffusion and highlights a growing reliance on GenAI for high-value tasks such as software engineering, code generation, and the automation of complex processes. This development points to an increasing awareness among financial institutions of GenAI's potential to foster innovation, enhance operational performance, and support strategic differentiation in a highly competitive market environment. Crucially, two of the three fastest-growing use cases, pricing and risk management, which increased by +167% between 2023 and 2024, and trading and portfolio optimization, which increased by +153% between 2023 and 2024, are core, industry-specific functions. This emphasizes the strategic relevance of GenAI in the financial sector and signals a transition from exploratory adoption to its institutionalization as a tool for sustained competitive advantage.[25]
Looking ahead, the industry outlook confirms a sustained commitment to increasing investment

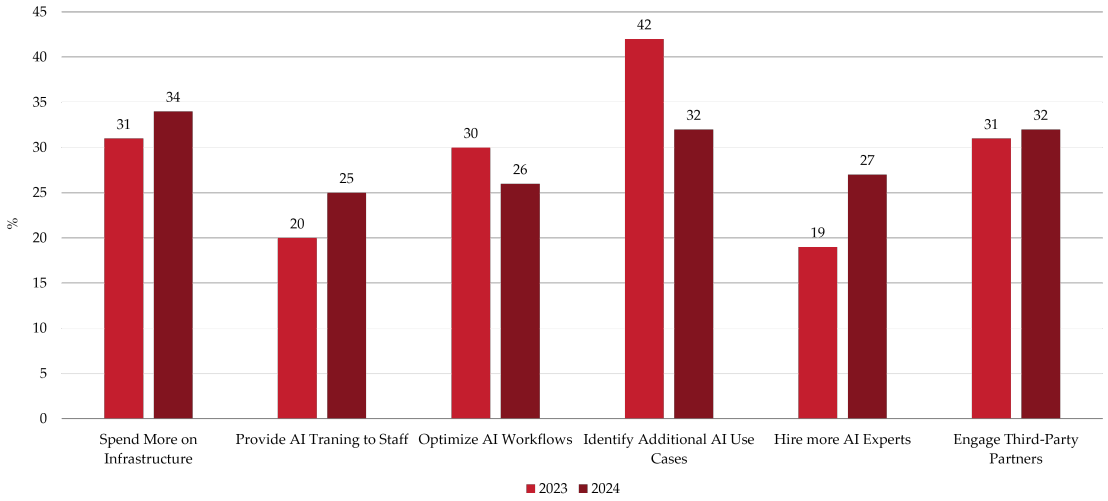

FIGURE 11: Financial Services Industry Investment Plans for 1 Year [25]

in AI and GenAI technologies. These technologies are increasingly regarded as strategic levers for business development, supported by the expansion of an AI-specialized workforce and the growing reliance on third-party partners to improve process efficiency and accelerate solution development. While overall investment levels are expected to grow, emerging evidence suggests a shift in budget allocation priorities (ref. Figure 11)[25]. Specifically, there appears to be a reduction in funding directed toward exploratory research into novel AI applications. Instead, strategic emphasis is being placed on talent acquisition, the reinforcement of collaborations with external service providers, and the enhancement of technological infrastructure required to develop, deploy, and maintain advanced AI systems at scale.[25]

2. Riks and the Increasing Need for Regulatory and Risk Management Frameworks

As shown in previous chapters, AI, and particularly GenAI, are ushering a profound transformation in how organizations across several industries operate, compete, and deliver value. From automating decision-making processes to enabling scalable customer engagement, AI is becoming a foundational component of enterprise strategy across nearly every industry. As adoption accelerates, organizations are not only integrating AI into discrete functions but are increasingly embedding it into the very structure of their business models. However, the growing reliance on AI technologies also introduces a range of complex and potentially high-impact risks. In this context, the development of clear regulatory frameworks and the implementation of robust AI Risk Management Frameworks (AI RMFs) has become essential. These efforts are not only necessary to mitigate potential harm but also to support the long-term sustainability, resilience, and trustworthiness of economic systems across all sectors. AI systems, while powerful, are not inherently neutral. They are built, trained, and operated by humans, and as such, they are susceptible to the full spectrum of human error, bias, and oversight. The risks arising from the deployment of AI technologies are multifaceted and increasingly material, comprising:

- **Inaccuracy and Hallucinations:** generative models may produce outputs that appear plausible but are factually incorrect or misleading, posing significant risks in contexts requiring accuracy and reliability;
- **Algorithmic Bias and Discrimination:** AI systems trained on biased data can replicate or even exacerbate social inequalities, affecting decisions related to employment, lending, healthcare, or law enforcement;
- **Data Privacy and Security:** AI often relies on sensitive personal or proprietary data, which, if mismanaged, can result in regulatory violations or large-scale data breaches;

- **Intellectual property concerns:** the use of AI tools to generate content based on vast datasets raises legal and ethical issues related to ownership, copyright, and content originality;
- **Lack of explainability:** the decision-making processes of many AI systems are opaque, limiting transparency and accountability;
- **Cybersecurity Vulnerabilities:** AI systems can be exploited through adversarial attacks or model manipulation, increasing the surface area for cyber threats;
- **Reputational Damage and Stakeholder Mistrust:** public failures or misuse of AI can erode trust in an organization, affecting customer loyalty, investor confidence, and employee morale.

Without proactive governance, these risks can lead to severe consequences, including regulatory sanctions, operational disruption, financial losses, and societal harm. To mitigate these risks and fully harness the transformative potential of artificial intelligence, organizations must transition from reactive risk responses to a proactive and structured governance strategy. This requires the implementation of cross-functional, enterprise-wide systems that embed responsible AI principles across every stage of the AI lifecycle, from initial design and data sourcing to model development, deployment, post-deployment monitoring, and eventual system decommissioning or retirement. This is where AI Risk Management Frameworks (AI RMFs) become indispensable. These frameworks provide a systematic and repeatable structure for integrating risk identification, mitigation, and oversight into an organization's broader technology and governance ecosystems. When properly implemented, AI RMFs enable organizations to:

- Identify and assess AI-specific risks early in the development cycle, ensuring that potential harms related to bias, inaccuracy, misuse, or non-compliance are detected before models are deployed at scale;
- Implement technical, ethical, and procedural safeguards that are tailored to the organization's risk tolerance and legal obligations, such as differential privacy measures, adversarial robustness, fairness constraints, or explainability features;
- Ensure alignment with applicable laws, standards, and ethical guidelines, including emerging national and international AI regulations, industry codes of conduct, and human rights frameworks;
- Establish formal accountability structures by clearly defining ownership and responsibilities for AI oversight across business units, data science teams, legal departments, and executive leadership, including escalation procedures for adverse events;
- Continuously monitor AI systems post-deployment through performance metrics, fairness audits, real-time alerts, and incident reporting mechanisms to ensure that the system continues to operate in a safe, effective, and equitable manner throughout its lifecycle.

Guidelines issued by supranational bodies could play a pivotal role in supporting the establishment of robust and coherent regulatory frameworks for AI. These regulatory frameworks, grounded in shared principles of safety, transparency, ethics, and accountability, could serve as essential references for national legislators and supervisory authorities, thereby potentially ensuring a harmonized and comprehensive approach to AI risk management at the global level. In turn, these regulatory frameworks could provide the foundation for the development and implementation of AI RMFs within organizations. On the other hand, well-designed AI RMF could enable organizations to translate regulatory guidelines into concrete operational practices, facilitating the identification, assessment, and mitigation of risks associated with AI deployment. When designed strategically, they become powerful enablers of innovation. By embedding trust, transparency, and ethical integrity into AI development, these frameworks provide organizations with the confidence and legitimacy needed to scale AI adoption responsibly. They help streamline decision-making, eliminate uncertainty, and foster internal alignment, allowing cross-disciplinary teams to collaborate more effectively and accelerate delivery cycles without compromising risk standards. Furthermore, by demonstrating a proactive commitment to ethical and responsible AI, organizations can strengthen stakeholder trust, enhance brand reputation, and distinguish themselves in increasingly competitive

and scrutinized markets. This trust becomes particularly valuable in customer-facing industries or regulated sectors, where public perception and compliance readiness are essential to long-term success.

2.1 AI Risks: A View on the Financial Industry

As shown in the previous chapter, financial institutions have already started integrating AI and GenAI into both operational processes and business-specific tasks, and they plan to continue investing in their development and strategic integration in the coming years. However, AI, and GenAI in particular, raise several concerns regarding potential risks for industry players. In 2025, both IOSCO[12] and the Japan FSA[13] conducted comprehensive analyses exploring the current use and outlook of AI in the financial sector through surveys of various financial entities. The results highlighted that, despite the significant potential benefits for business innovation and efficiency, it is essential to thoroughly assess and manage the risks associated with these emerging technologies. These risks and challenges can be summarized as follows:

- **Lack of Explainability:** many AI systems operate as "black boxes," making it difficult for institutions to interpret or justify decisions, posing challenges for transparency and accountability;
- **Data Privacy and Quality:** AI's performance heavily depends on the quality, relevance, and accuracy of data. Poor data management can result in flawed outputs, while over-reliance on personal data raises privacy and ethical concerns;
- **Cybersecurity and Operational Risk:** AI systems may introduce new vulnerabilities, including susceptibility to adversarial attacks, model drift, and technical failures that disrupt financial operations or cause regulatory breaches;
- **Third-Party Risk:** many institutions rely on external vendors for AI tools, raising concerns about oversight, vendor lock-in, and exposure to unregulated service providers;
- **Difficulties in Model Governance:** AI introduces unique model risks due to its broad applicability, non-deterministic behavior, and reliance on externally managed foundation models. These characteristics pose several challenges to traditional risk management frameworks, potentially resulting in an insufficient ability to fully assess, control, and explain the risks associated with generative AI systems.

Other than those mentioned above, the Japan FSA report[13] underlined some new risks and concerns arising from the adoption of GenAI models:

- **Hallucination:** hallucination risks raise several concerns due to the potential for misleading outputs and information when generative AI models are used in business applications or decision-making processes. It is therefore essential to implement systems that maintain a "human-in-the-loop" to ensure robust oversight and avoid potential critical damages.
- **Financial Crime:** criminal methods are becoming increasingly sophisticated due to the adoption of AI and GenAI, particularly amplifying the potential risks to financial institutions and their customers. For example, the advent of generative AI further increases risks by automating the production of highly believable written text, audio, and images.
- **Systemic Risk:** the growing integration of GenAI into business functions and decision-making processes may increase the risk of highly correlated behaviors of market players, as similar AI-generated signals lead to uniform decisions. This convergence could foster herd behavior, amplifying market volatility and significantly increasing systemic risk.

Most jurisdictions have not adopted AI-specific regulations for the financial sector. Instead, they apply existing technology-neutral regulatory frameworks, which already cover key areas such as risk management, governance, cybersecurity, data protection, and consumer protection. As a result, many of the cross-sectoral themes relevant to AI are broadly addressed under current financial regulations, making the need for dedicated AI-specific financial rules debatable. As reported by

BIS[4] his regulatory stance likely explains why financial authorities are not planning new AI-specific rules in the near term, while actively evaluating whether additional measures are needed to address AI-specific risks in the currently developed frameworks[27].

3. AI Regulatory Framework

As illustrated in previous sections, AI adoption has accelerated significantly, increasing the urgency for robust oversight frameworks capable of addressing its complex challenges and potential risks. In response, and in a context marked by fragmented national policies and limited institutional expertise, supranational bodies have begun implementing targeted frameworks and proposing guidelines aimed at fostering openness in AI practices, strengthening governance responsibilities, and aligning supervisory approaches across jurisdictions. Within these efforts [21] led by supranational bodies, the first globally recognized initiative to establish a normative framework for AI governance came from the OECD, which adopted the OECD Principles on Artificial Intelligence in May 2019 (updated in 2024) [29]. Endorsed by over 40 countries, including all OECD members and several non-member states, these principles are intended to promote the responsible stewardship of trustworthy AI. The OECD Recommendation on AI seeks to advance the development and use of AI that is trustworthy and human-centric, to support responsible innovation, to safeguard human rights and democratic values, to foster international cooperation, and to guide public policy through a shared global framework. The Recommendation is structured around five key principles for responsible AI and five corresponding policy recommendations for national and international action. The five value-based principles for the responsible development and use of AI are [29]:

1. **Inclusive Growth, Sustainable Development, and Well-Being:** AI should benefit people and the planet by driving inclusive economic growth and sustainable development;
2. **Human-Centered Values and Fairness:** AI systems should be designed in a way that respects human rights and democratic values, including privacy, liberty, and equality;
3. **Transparency and Explainability:** the functioning of AI systems should be transparent to users and regulators, and decisions should be explainable where possible;
4. **Robustness, Security, and Safety:** AI systems must be technically robust and secure, and should function appropriately throughout their lifecycle;
5. **Accountability:** organizations and individuals developing, deploying, or operating AI systems should be accountable for their proper functioning.

In addition to these principles, the OECD outlines five strategic recommendations for policymakers [29]:

1. Promote investment in AI research and development;
2. Foster a digital ecosystem for AI;
3. Ensure a policy environment that promotes trustworthy AI;
4. Equip people with the necessary skills to interact with and benefit from AI;
5. Encourage international cooperation to ensure the global alignment of AI governance.

These principles and recommendations serve as a foundational reference for the development of numerous national and international AI policy frameworks that followed.

3.1 EU AI Act

Inspired by the OECD principles [29], the European Union has been pioneering in the definition of an AI regulatory framework. The AI Act[2], which represents the first cross-jurisdictional regulatory

[2]For an extensive analysis, we suggest[5].

framework focused on artificial intelligence, establishing a harmonized set of rules for development, market introduction, deployment, and use of AI across the EU, entered into force in August 2024. It introduces a regulatory roadmap that outlines the gradual application of several rules, which will come fully into effect by August 2026. Key innovations of the AI Act include:

- Binding rules for high-risk applications (e.g. biometric identification, credit scoring, recruitment tools);
- Strict requirements for data quality, documentation, human oversight, and cybersecurity;
- Creation of national supervisory authorities and a European AI Office;
- Enforcement mechanisms with fines up to 7% of global turnover[19].

3.1.1 An Harmonized Regulatory Framework

The AI Act represents a significant regulatory milestone within the EU's digital policy agenda. It is part of the broader initiative Europe Fit for the Digital Age, which, among other objectives, seeks to position the European Union as a global leader in the development of trustworthy and human-centric AI. Unlike voluntary guidelines or fragmented national regulations, the AI Act establishes uniform, directly applicable rules across all Member States. Its legal basis derives primarily from Article 114 of the Treaty on the Functioning of the European Union (TFEU)[9], which enables the EU to adopt measures for the approximation of laws to ensure the functioning of the internal market. However, the AI Act goes beyond market concerns by explicitly incorporating the protection of fundamental rights, as enshrined in the Charter of Fundamental Rights of the European Union[5]. In doing so, the Act reflects the EU's dual objective: fostering innovation in AI while ensuring that such innovation does not come at the expense of safety, human dignity, and democratic values. Before describing regulatory measures, it is important to highlight a fundamental challenge: the lack of a universally accepted definition of Artificial Intelligence. As emphasized in the OECD 2024 report "Regulatory Approaches to Artificial Intelligence in Finance"[27], the absence of a shared understanding of what constitutes AI complicates regulation and supervision. The European Union has adopted a formal definition of AI, while many countries rely on non-binding, non-prescriptive definitions or lack a standardized description altogether. Against this backdrop, the AI Act provides a formal and legally binding definition of Artificial Intelligence, aiming to bring clarity and legal certainty to stakeholders operating across the EU. The definition adopted by the Act is closely aligned with the approach developed by the OECD in its 2023 revision, reflecting international efforts towards regulatory convergence. According to Article 3[9] of the AI Act, an AI system is defined as: "A machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments". [9] This definition does not aim to establish a fixed list of AI systems but rather provides a flexible, technology-neutral guideline to respond to the rapid technological evolution of the market. As defined in Recital 12 of the AI Act, this definition should not be applied mechanically to every AI system; instead, each system must be assessed based on its specific characteristics. AI Act Recital 12[9], together with the "Guidelines on the definition of an artificial intelligence system"[8], specifies that a system qualifies as AI only if it meets certain criteria. AI systems must be able to infer outputs from the input data they receive, enabling them to generate predictions, recommendations, and decisions by using models and algorithms (e.g., machine learning approaches, logic- and knowledge-based approaches). Conventional, rule-based software or deterministic algorithms that do not perform such inference particularly if they lack autonomy or adaptiveness (e.g. systems for improving mathematical optimization, basic data processing systems, systems based on classical heuristics, simple prediction systems), fall outside the scope of this definition. By adopting this definition, the AI Act ensures that its provisions apply not only to fully autonomous AI systems but also to systems that operate with partial autonomy or human oversight, if their outputs have the potential to influence decision-making processes or produce effects in physical or digital environments. This definitional clarity represents a crucial step towards ensuring legal consistency, preventing regulatory loopholes, and enabling effective enforcement of the rules across the internal market.

The AI Act applies to a broad range of actors involved in the lifecycle of AI systems, regardless of their geographical location, provided their activities have an impact within the EU. The regulation covers:

- **Providers:** any natural or legal person, public authority, agency, or other body that develops an AI system or has it developed and places it on the market or puts it into service under their name or trademark;
- **Users:** individuals or organizations deploying AI systems within the EU in the context of their professional activities;
- **Importers and Distributors:** entities responsible for ensuring compliance when AI systems from outside the EU are introduced into the European market;
- **Third-Country Providers:** AI system providers established outside the EU whose products or services are offered to users within the Union.

This extraterritorial scope mirrors similar approaches adopted in other EU regulations, such as the General Data Protection Regulation (GDPR), reinforcing the EU's ambition to influence global AI governance standards.

3.1.2 Risk-Based Classification of AI Systems

The AI Act introduces a tiered regulatory approach based on the potential risks AI systems pose to health, safety, and fundamental rights. This risk-based framework allows for proportional obligations, ensuring that regulatory intervention corresponds to the level of risk. The four levels of risk identified are:

1. **Unacceptable Risk;**
2. **High-Risk;**
3. **Limited-Risk;**
4. **Lower-Risk.**

3.1.3 AI Systems Prohibited Due to Unacceptable Risk

Certain AI applications are considered fundamentally incompatible with EU values, such as human dignity and democracy, and are therefore prohibited outright. These include:

- AI systems that use subliminal techniques to manipulate individuals' behavior in a manner that may cause harm are strictly prohibited. This restriction is designed to safeguard individual autonomy and ensure freedom of thought and decision-making.
- Exploitative AI that takes advantage of vulnerable groups, such as children, individuals with disabilities or groups defined by socio-economic status.
- The deployment of real-time biometric identification systems by law enforcement in publicly accessible areas is generally not allowed. However, narrowly defined exceptions apply, limited to situations where such technology is strictly required to locate victims of crimes such as abduction, human trafficking, or sexual exploitation; to prevent an imminent and serious threat to individuals' lives or physical safety, including terrorist threats; or to identify suspects involved in criminal offences, for the purposes of investigation, prosecution, or enforcing judicial decisions.
- AI-based social scoring by public authorities, particularly when it leads to discriminatory or unjustified treatment of individuals.

- In addition to the restrictions already mentioned, the AI Act also prohibits the use of AI systems for emotion recognition in sensitive contexts such as the workplace and educational institutions, unless strictly necessary for health or safety purposes. This measure is intended to safeguard individuals from intrusive evaluations that could result in discrimination or negative consequences based on inferred emotional states.

Since February 2025[19], the use of AI systems classified as posing an unacceptable risk has been officially prohibited under Article 5 of the AI Act [9]. This prohibition reflects the EU's emphasis on preserving human autonomy, dignity, and protection from undue surveillance or manipulation. To assist stakeholders in identifying such prohibited AI practices, the European Commission published the Guidelines on Prohibited AI Practices[7] in February 2025, providing further clarity on the types of applications deemed fundamentally incompatible with EU values. These aim to clarify the scope and concrete application of the prohibitions set out in the mentioned Article 5, ensuring consistent enforcement across the European Union. Specifically, the Guidelines provide:

- A set of practical indicators for determining whether an AI system uses subliminal techniques in a way that may significantly impair an individual's ability to make autonomous decisions;
- Clarifications on what constitutes exploitation of vulnerabilities, including illustrative scenarios involving minors, persons with disabilities, or socio-economically disadvantaged groups;
- Concrete examples of AI-based social scoring practices by public authorities that are likely to produce unjustified or disproportionate negative effects on individuals, thereby falling under the prohibition.
- Specific parameters for assessing the use of real-time remote biometric identification by law enforcement, including how to evaluate the existence of exceptional circumstances that justify its deployment, as well as procedural safeguards required under such circumstances;
- Criteria for identifying prohibited emotion recognition applications in sensitive environments, including how to distinguish between acceptable uses for health or safety reasons and practices that may result in invasive monitoring or discriminatory consequences.

3.1.4 High-Risk AI Systems and Compliance Requirements

Systems are classified as high-risk when their use may significantly affect the health, safety, or fundamental rights of individuals, as set out in AI Act Article 6 [9]. This category includes AI systems intended to serve as safety components of products subject to third-party conformity assessments, as well as AI systems performing profiling of natural persons. Moreover, AI systems are considered high-risk when their deployment, due to their intended purpose or the specific context of use, carries a tangible risk of harm or adverse impact on fundamental rights. Conversely, systems that merely perform supporting or preparatory tasks, without substantially influencing decision-making or replacing human judgment, are generally excluded from the high-risk classification. The list of high-risk AI use cases, defined primarily in Annex III of the AI Act, may be updated by the European Commission in light of technological developments or emerging risks. Regarding so, at the beginning of June 2025, the European Commission's AI Office has launched a 6 weeks targeted stakeholder consultation in order to "to collect input from stakeholders on practical examples of AI systems and issues to be clarified in the Commission's guidelines on the classification of high-risk AI systems and future guidelines on high-risk requirements and obligations, as well as responsibilities along the AI value chain". [10] To mitigate potential risks, high-risk AI systems must comply with a comprehensive set of obligations throughout their lifecycle. These include:

- **Risk Management and Mitigation:** providers are required to establish and maintain a comprehensive risk management system that applies throughout the entire lifecycle of the AI system. This system must ensure the continuous identification, assessment, and mitigation of risks, taking into account both the intended use and reasonably foreseeable misuse scenarios. Moreover, the risk management process must be subject to regular, systematic review and, where necessary, updated to reflect new information, evolving use cases, or emerging risks.

- **Data Governance and Quality:** AI providers are required to implement robust data governance and quality management processes covering all phases of the system's lifecycle. These processes must guarantee that datasets used for training, validation, and testing are sufficiently relevant, representative, and appropriately curated to minimize biases and inaccuracies. The objective is to ensure that AI systems operate in a fair, reliable, and non-discriminatory manner, fully aligned with the requirements of the regulatory framework.
- **Technical Documentation and Record-Keeping:** providers are required to prepare and maintain comprehensive, up-to-date technical documentation demonstrating the AI system's compliance with the obligations set out in the AI Act. This documentation must cover all relevant aspects of the system, including its design, development processes, intended purpose, risk management measures, and performance evaluation. In addition, AI systems must be designed to enable systematic record-keeping of key operational events, including those that may have an impact on safety, fundamental rights, or that indicate substantial modifications to the system. These requirements ensure that competent authorities can effectively assess compliance and investigate potential incidents or risks.
- **Transparency and User Instructions:** to ensure transparency and promote responsible use, providers must supply clear, accurate, and accessible information to users regarding the AI system. This includes comprehensive instructions for use, a description of the system's capabilities and intended purpose, as well as its known limitations and potential risks. Users must also be informed about any conditions or constraints under which the system may or may not perform reliably. These requirements are essential to enable users to make informed decisions, use the system appropriately, and avoid unintended consequences.
- **Human Oversight:** systems must incorporate safeguards that allow effective human monitoring and intervention, including the possibility to override or disable automated operations.
- **Robustness, Accuracy, and Cybersecurity:** AI systems must meet high technical standards for reliability, accuracy, resilience against manipulation, and protection against cyberattacks.

Before being placed on the market, high-risk systems must undergo a conformity assessment, typically conducted internally, but requiring third-party certification by notified bodies for specific applications. After deployment, providers must establish post-market monitoring systems and report serious incidents or malfunctions to competent authorities.

3.1.5 Regulatory Approach to Limited and Minimal Risk AI Systems

In addition to the stringent requirements imposed on high-risk and prohibited AI applications, the AI Act introduces a differentiated regulatory regime for AI systems classified as posing limited or minimal risk. This approach reflects the EU's intention to strike a balance between safeguarding fundamental rights and promoting innovation, applying obligations proportionate to the potential risks associated with the use of AI.

Limited Risk AI Systems

AI systems falling under the category of limited risk are not subject to strict compliance requirements but must adhere to specific transparency obligations, as explicitly outlined in Article 52 of the AI Act. These obligations are designed to ensure that users are aware when they are interacting with an AI system, particularly in cases where the AI might influence their perceptions, decisions, or behavior without their full awareness. Examples of limited risk AI systems include:

- **AI Chatbots and Virtual Assistants:** users must be clearly informed that they are engaging with an AI-driven system rather than a human. This is intended to avoid confusion or deception in digital interactions.
- **AI-Generated Contents (e.g. deepfakes):** when an AI system produces synthetic audio, images, video, or text intended to resemble authentic content, it must be explicitly labeled as artificially generated or manipulated. This measure aims to prevent misinformation and protect individuals from deception.

- **Emotion Recognition and Biometric Categorization Systems (outside of high-risk contexts):** in these cases, transparency requirements apply to inform individuals about the use of such technologies. It is important to note, however, that when these systems are deployed in sensitive environments, such as workplaces, educational settings, or law enforcement, their risk classification may be elevated to high-risk, triggering more stringent obligations.

The transparency requirements for limited risk AI systems do not extend to imposing technical or organizational controls beyond the obligation to inform users. Nonetheless, these provisions play a crucial role in promoting trust and user awareness in AI interactions.

Minimal Risk AI Systems
Minimal, or lower risk, AI systems are those whose use is considered to entail negligible or no risk to fundamental rights, safety, or public interests. These applications are largely excluded from binding legal obligations under the AI Act. Examples include:

- Spam filters, which use AI to automatically detect and filter unsolicited communications;
- Recommendation algorithms in entertainment platforms, such as AI systems suggesting movies, music, or video games based on user preferences;
- AI-based functionalities in video games, including non-player character (NPC) behaviors or adaptive difficulty systems;
- Autocorrect or grammar suggestion tools integrated into word processors or messaging applications.

While these systems are not subject to specific mandatory requirements under the AI Act, the European Commission and other regulatory bodies encourage providers of minimal risk AI to voluntarily adhere to codes of conduct, industry best practices, and principles for trustworthy AI.

4. AI Risk Management Frameworks

The AI Act has served as an inspiration for regulatory frameworks on AI, laying the foundation for a technology and sector-neutral approach. It addresses one of the key issues raised by both the OECD[27] and BIS[4], the need for a shared cross-jurisdictional definition of AI systems and establishes a clear risk-based classification of AI systems into different risk categories. This growing regulatory focus on AI requires organizations to implement robust frameworks and practices, not only to ensure compliance but also to demonstrate AI governance maturity and maintain public trust. In this context, several supranational bodies have proposed and developed guidelines and frameworks to support organizations in establishing procedures and practices that span the entire organization and effectively manage AI-related risks. The successful deployment of AI systems requires not only technical integration but also substantial organizational adaptation. AI is not a standalone tool or isolated technology; it is a systemic capability that touches every facet of an organization's operations, decision-making, and stakeholder engagement. As such, its governance cannot be relegated to individual departments or ad hoc teams. It must be institutionalized through deliberate structural realignment. Leading organizations are beginning to recognize that to realize the full benefits of AI, they must embed it deeply into their strategic architecture and align it with their governance, culture, and talent management systems. This transformation is unfolding through several critical changes:

- **Redesigning Workflows:** companies are revisiting existing processes to determine where AI can enhance speed, accuracy, or efficiency. This includes automating repetitive tasks, augmenting human judgment in complex decisions, and re-engineering customer service, supply chain, or analytics functions to be AI-enabled by default.
- **Appointing Executive Leadership:** AI governance is increasingly being elevated to the C-suite. Organizations are naming Chief AI Officers or designating senior executives with formal authority to oversee AI strategy, risk management, and ethical compliance. This leadership is

critical for securing resources, aligning cross-functional teams, and ensuring that AI initiatives support the organization's mission and risk appetite.

- **Centralizing Governance Functions:** functions such as data governance, algorithmic accountability, and risk oversight are being consolidated into Centers of Excellence or transformation offices. These structures serve as custodians of best practices and ensure that AI initiatives across business units adhere to consistent standards.

- **Creating Cross-Functional AI Governance Bodies:** effective AI governance requires a blend of perspectives. Legal experts, data scientists, cybersecurity professionals, ethicists, and business leaders are increasingly collaborating through formal committees or steering groups to assess model performance, regulatory exposure, and ethical implications.

- **Developing Adoption Roadmaps:** rather than deploying AI in an uncoordinated fashion, organizations are crafting strategic roll-out plans that define where and how AI will be introduced. These roadmaps include phased adoption schedules, integration milestones, and mechanisms for evaluation and iteration.

- **Institutionalizing Role-Based Training Programs:** as AI reshapes the nature of work, employees at all levels must be equipped with the knowledge to understand and interact with AI systems responsibly. Training is being tailored by function, for example, developers on fairness auditing, marketing teams on content validation, and compliance officers on risk classification, ensuring that each stakeholder understands both the capabilities and limitations of the AI tools they use.

These structural changes do more than reduce risk; they signal a broader organizational evolution. They reflect a growing recognition that responsible innovation must be an organizational value, not just a technical feature. Companies that succeed in operationalizing these changes are building a foundation where AI is not only scalable but also trustworthy, explainable, and socially acceptable. While awareness of AI's opportunities and challenges is growing, the implementation of risk management best practices remains immature in many organizations. Despite the expansion of AI use cases, significant gaps persist in how companies measure, monitor, and govern AI systems. Many companies still lack of:

1. Defined KPIs tailored to AI;

2. Formalized adoption and risk mitigation strategies;

3. Centralized governance structures for AI oversight;

4. Processes for reviewing AI-generated outputs for accuracy and ethical compliance.

To build maturity in AI governance, organizations must move from experimentation to institutionalization. This involves embedding a set of structured and repeatable best practices across the AI lifecycle. Key practices include:

- **Executive-Level Risk Ownership:** assigning responsibility for AI governance at the highest levels, such as the CEO, CRO, CIO or Board of Directors, ensures accountability and signals strategic importance. This helps align AI initiatives with the organization's broader risk framework.

- **Tracking Performance Through Measurable KPIs:** establishing indicators related to accuracy, fairness, interpretability, and real-world impact is essential for both transparency and optimization. These metrics should be updated regularly and tied to business objectives and ethical commitments.

- **Phased Deployment with Pilot Programs:** introducing AI incrementally allows organizations to test, refine, and scale technologies with greater control. Pilot programs help surface potential risks before full implementation and allow for stakeholder feedback to inform adjustments.

- **Comprehensive Risk Assessments:** evaluations must cover technical, legal, ethical, and societal dimensions. This includes impact assessments on privacy, discrimination, security vulnerabilities, explainability, and regulatory exposure, both before deployment and during continuous operation.
- **Real-Time Monitoring and Alert Systems:** AI models are dynamic; without ongoing surveillance, their behavior may drift or degrade. Monitoring mechanisms, such as automated alerts for anomalies or bias shifts, are essential to ensure consistent performance and safety.
- **Transparent Internal Communication:** employees must understand the purpose, limitations, and oversight protocols of AI systems. Internal transparency fosters alignment and ensures that AI is not perceived as opaque or arbitrary.
- **Ongoing Employee Education:** training programs should evolve alongside AI tools and regulations. This includes both technical training (e.g. model validation) and ethical training (e.g. human-in-the-loop decision-making).
- **Customer-Facing Trust Strategies:** public trust is essential. Organizations should implement mechanisms such as user disclosures, consent protocols, explainable interfaces, and opt-out functionalities to ensure end-users understand when and how AI is being applied.

By institutionalizing these practices, organizations ensure that AI systems remain aligned with organizational values and stakeholder expectations even as technologies and markets evolve.

4.1 OECD Framework for the Classification and Risk Management of AI Systems

The OECD Framework for the Classification of AI Systems[26], developed within the broader context of the OECD AI Principles [29], has become a foundational reference for organizations aiming to implement effective procedures and processes to address AI-related challenges. Notably, regulations such as the EU AI Act and AI RMFs like the U.S. National Institute of Standards and Technology (NIST) AI Risk Management Framework [23] have incorporated many of the practices and principles that align with the OECD's vision and objectives, offering a foundational instrument for both public and private sectors to evaluate and manage AI technologies with greater clarity and foresight. This framework serves a unique function: rather than prescribing technical or legal obligations, it provides a structured and policy-relevant lens for categorizing AI systems according to their core characteristics, contexts of application, and potential impacts. Rooted in the OECD AI Principles [29], which emphasize human-centered values, transparency, accountability, and robustness, this classification tool is designed to support the development of proportionate, evidence-based governance strategies. The OECD Framework provides a consistent method to understand and compare the risks, benefits, and operational realities of AI systems by offering a non-normative but comprehensive classification structure. It empowers policymakers and organizations to align AI deployments with public interest, mitigate potential harms, and guide innovation in a direction that supports democratic values, sustainable development, and economic resilience.

4.1.1 Core Content

The OECD Framework for the Classification of AI Systems is structured around five foundational dimensions (ref. Figure 12) [26]. These dimensions work synergistically to facilitate consistent classification and comparative analysis, serving as the backbone for public policy design, institutional accountability, and ethical oversight: [26]

- **People and Planet:** explores the interface between AI systems, human well-being, and environmental sustainability. It captures the roles of actors involved in the development, deployment, and use of AI technologies, such as providers, end-users, impacted communities, and vulnerable groups, and examines how the system aligns with democratic values, fundamental rights, and sustainable development goals. This dimension assesses the potential for system-wide harms, such as those arising from biased outcomes, power asymmetries, labor displacement, or environmental degradation. It considers whether the system is deployed in contexts of

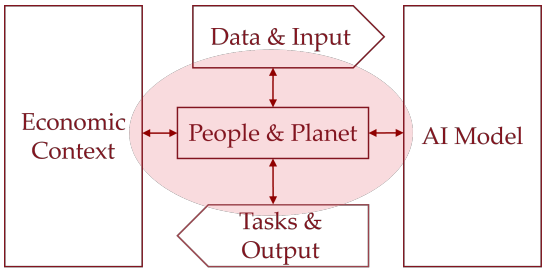

FIGURE 12: The Five Dimensions of OECD Framework [26]

power imbalance, such as law enforcement or employment screening, where contestability and recourse mechanisms are essential. Additionally, it addresses the degree of user dependency and control, evaluating whether humans can opt out or meaningfully override automated decisions. It also touches on transparency of communication, including disclosures to users about the presence and functioning of AI systems, and whether redress mechanisms are available when harm occurs.

- **Economic Context:** analyzes the sectoral and institutional environment in which the AI system operates. It distinguishes between sectors with high criticality and public service functions, such as healthcare, education, finance, transportation, and justice, and sectors with relatively lower systemic risk. The dimension also considers the economic role of the AI system: whether it supports productivity, cost reduction, strategic decision-making, or personalized services. Importantly, it accounts for the business model dependencies, such as monetization through user data or third-party AI licensing, and the potential for market concentration or vendor lock-in. The framework invites analysis of whether the system is central to mission-critical operations or infrastructure, and if its failure would result in disproportionate economic or societal disruption. Moreover, it incorporates the scalability and diffusion potential of AI systems, recognizing that models with high replicability across markets may pose amplified systemic risks or lead to widespread behavioral impacts.

- **Data and Input:** examines the lifecycle of the data used in AI systems, focusing not only on type and provenance, but also on the mechanisms of collection, annotation, transformation, and use. It distinguishes among first-party, third-party, synthetic, derived, and public data, emphasizing the importance of data integrity and representativeness in minimizing algorithmic bias. The framework places particular emphasis on whether sensitive data, such as biometric, financial, or health-related information, is involved, and whether appropriate safeguards are in place to ensure lawful and ethical processing. It also assesses the degree of automation in data collection, the use of sensor-driven input streams (such as from wearables or IoT devices), and the system's ability to generate or infer additional data. This dimension is key to evaluating compliance with privacy regulations, exposure to adversarial data attacks, and the traceability of input-output linkages. Furthermore, it supports evaluation of documentation practices, including dataset documentation standards (e.g. datasheets for datasets) and versioning policies.

- **AI Model:** delves into the system's internal logic and technical properties. It characterizes the model type, statistical, symbolic, hybrid, and the learning paradigm used, such as supervised, unsupervised, reinforcement, or transfer learning. It also considers whether the system is trained once and then fixed, or whether it is dynamic and self-learning, which significantly affects its risk profile and auditability. The framework emphasizes the degree of explainability and interpretability, considering whether stakeholders can understand, challenge, or replicate the rationale behind model outputs. Models deployed in high-stakes decisions are expected to offer some level of intelligibility, either inherently or through post-hoc methods. Furthermore, the dimension evaluates model transparency, including the disclosure of architecture, parameters, training data, and design decisions, especially when the system is offered commercially or as open source. Attention is also paid to the model's robustness to perturbations, security vulnerabilities, and capacity to generalize beyond the training environment. Where relevant,

the framework supports the assessment of model cards or system documentation, especially for foundation models that may be reused across multiple downstream applications.

- **Task and Output:** describes the purpose, behavior, and real-world implications of the AI system. It categorizes the types of tasks the system performs, such as generation, prediction, optimization, detection, personalization, or decision support, and distinguishes between systems designed for supportive use and those intended for fully autonomous execution. This dimension incorporates an assessment of the level of human oversight, including whether human intervention is active, passive, or entirely absent at the point of decision-making. It evaluates the consequences of erroneous outputs, particularly in critical domains such as medical diagnostics or autonomous driving, where output reliability has life-and-death implications. The framework also encourages consideration of feedback mechanisms, such as whether the system's outputs are monitored, logged, and corrected post-deployment. Another relevant factor is the contextual use of outputs, including whether the AI results feed into final decisions or are mediated by human judgment. Finally, this dimension accounts for whether the task supports core public interest functions, which may raise regulatory obligations or justify heightened scrutiny.

Each of these dimensions is not static; they are intentionally designed to be interoperable and adaptable across sectors, jurisdictions, and levels of AI maturity. Their application enables a granular and multidimensional characterization of AI systems, fostering clarity in governance, consistency in comparative analysis, and foresight in risk mitigation.

4.2 NIST Artificial Intelligence Risk Management Framework

In response to the challenges posed by AI usage proliferation, the U.S. National Institute of Standards and Technology (NIST) developed the AI RMF [23] to assist organizations in designing, developing, deploying, and using AI systems in ways that are trustworthy and responsible. Released in 2023, the NIST AI RMF is a voluntary, sector-agnostic, and use-case-neutral framework, structured to promote flexibility and broad applicability across diverse organizational contexts. This framework serves to help organizations "to better manage risks across the AI lifecycle, aiming to:

- The development of innovative approaches to address characteristics of trustworthiness, including accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security, and mitigation of unintended and/or harmful bias, as well as of harmful uses;
- Consider and encompass principles such as transparency, fairness, and accountability during design, deployment, use, and evaluation of AI technologies and systems;
- Consider risks from unintentional, unanticipated, or harmful outcomes that arise from intended uses, secondary uses, and misuses of the AI."[6]

The NIST Framework is complemented by the NIST AI RMF Playbook [22] , a practical guide that provides operational recommendations for implementing the framework within organizations. This document helps stakeholders understand how to translate the core principles of the AI RMF into actionable steps within different organizational contexts. The Playbook includes suggested actions, references, and related guidance to support the achievement of desired outcomes across the AI lifecycle. In 2024, NIST also released the Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile[24], a guideline specifically designed to address the unique risks associated with generative AI. This profile adapts the NIST AI RMF to the specific challenges posed by GenAI systems, offering tailored risk assessment and mitigation strategies for developers, deployers, and users of large-scale generative models.

4.2.1 Core Content

The framework is built around four interdependent core functions (ref. Figure 13):

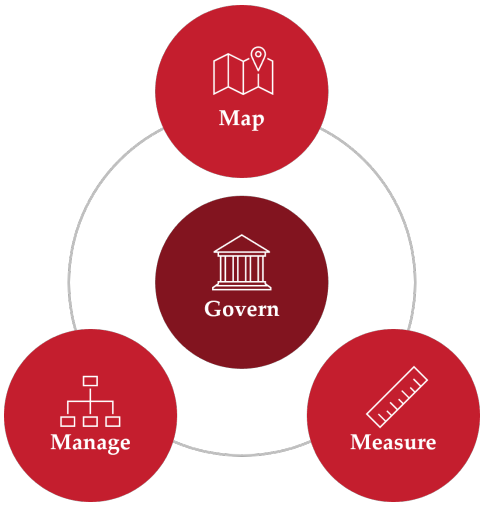

FIGURE 13: Four Interdependent Core Functions [23]

- **Govern:** the Govern function cultivates and implements a culture of risk management within organizations involved in designing, developing, deploying, evaluating, or acquiring AI systems. It establishes clear processes, documentation, organizational frameworks, and procedures to achieve desired risk management outcomes while incorporating assessments of potential impacts. This function provides a structure that ensures all activities align with the organization's principles, policies, and strategic priorities. By connecting the technical aspects of AI with organizational values and principles, it supports individuals responsible for acquiring, training, deploying, and monitoring AI systems. Additionally, it addresses the full product lifecycle and associated processes, including the management of any issues that may arise. The playbook[22] distinguishes six main domains related to the Govern function, which everyone comprehends as a subset of practices that should be put in place:
  - **Govern 1**: policies, processes, procedures, and practices across the organization related to the mapping, measuring, and managing of AI risks are in place, transparent, and implemented effectively;
  - **Govern 2**: accountability structures are in place so that the appropriate teams and individuals are empowered, responsible, and trained for mapping, measuring, and managing AI risks;
  - **Govern 3**: workforce diversity, equity, inclusion, and accessibility processes are prioritized in the mapping, measuring, and managing of AI risks throughout the lifecycle;
  - **Govern 4**: organizational teams should be committed to a culture that considers and communicates AI risk;
  - **Govern 5**: processes are in place for robust engagement with relevant AI actors;
  - **Govern 6**: policies and procedures are in place to address AI risks and benefits arising from third-party software and data, and other supply chain issues.
- **Map**: The Map function defines the context required to identify and assess risks associated with an AI system across its lifecycle. AI development involves multiple interdependent activities, often managed by different actors who may lack full visibility or control over the entire process. This fragmentation can result in unforeseen impacts, as early design decisions may influence how the system behaves and how it interacts with its deployment environment. Such complexity introduces uncertainty into risk management, which the Map function seeks to reduce by gathering contextual knowledge and identifying potential sources of negative risk. This information supports informed decision-making and lays the foundation for the Measure and Manage functions. The Map function also encourages inclusion of diverse internal perspectives and, where relevant, engagement with external stakeholders such as users, affected communities, or collaborators. The engagement of different levels of

stakeholders could help organizations to better understand the context of use and recognize both potential benefits and foreseeable negative impacts. In the end, the Map function will provide sufficient insight to determine whether the development or deployment of an AI system is appropriate. If the decision is to proceed, organizations should continue to apply the Map function throughout the system's lifecycle as risks, capabilities, and contexts evolve. The playbook[22] distinguishes five main domains related to the Map function:

  - **Map 1**: context is established and understood;
  - **Map 2**: categorization of the AI system is performed;
  - **Map 3**: AI capabilities, targeted usage, goals, and expected benefits and costs compared with appropriate benchmarks are understood;
  - **Map 4**: risks and benefits are mapped for all components of the AI system, including third-party software and data;
  - **Map 5**: impacts to individuals, groups, communities, organizations, and society are characterized.

- **Measure**: The Measure function applies quantitative, qualitative, or mixed-method approaches to analyze, assess, benchmark, and monitor AI risks and their impacts. It builds on the contextual understanding gained through the Map function and provides critical input to the Manage function. AI systems should be continuously tested and evaluated not only for performance but also for their social impact, human-AI interactions, and alignment with trustworthy characteristics. Where trade-offs among these characteristics occur, measurement serves as a traceable foundation to support informed management decisions. The playbook[22] distinguishes four main domains related to the Measure function:

  - **Measure 1**: appropriate methods and metrics are identified and applied;
  - **Measure 2**: AI systems are evaluated for trustworthy characteristics;
  - **Measure 3**: mechanisms for tracking identified AI risks over time are in place;
  - **Measure 4**: feedback about the efficacy of measurement is gathered and assessed.

- **Manage**: The Manage function enables organizations to prioritize, mitigate, accept, or avoid AI risks. It supports feedback loops, incident response planning, and risk communication strategies, allowing continuous improvement of AI systems. The playbook[22] distinguishes four main domains related to the Manage function:

  - **Manage 1**: AI risks based on assessments and other analytical output from the Map and Measure functions are prioritized, responded to, and managed;
  - **Manage 2**: strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, documented, and informed by input from relevant AI actors;
  - **Manage 3**: AI risks and benefits from third-party entities are managed;
  - **Manage 4**: risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly.

The AI RMF is designed to be both scalable and iterative, offering practical guidance for organizations ranging from early-stage AI developers to large, mature enterprises. One of its primary utilities lies in enhancing AI trustworthiness by embedding principles such as reliability, fairness, transparency, and privacy into the design and deployment processes. These safeguards not only improve the integrity of AI systems but also help reduce reputational and operational risks. Another key strength of the framework is its support for regulatory readiness; it aligns closely with global standards and emerging legal frameworks, including the OECD classification and the EU AI Act[19], making it a valuable foundation for organizations seeking to meet compliance requirements across jurisdictions. Furthermore, the AI RMF promotes meaningful cross-functional collaboration by encouraging active participation from legal, technical, governance, and policy stakeholders. Organizations that adopt the AI RMF typically experience several significant improvements across their governance and operational practices. One of the key benefits is the enhanced clarity in defining

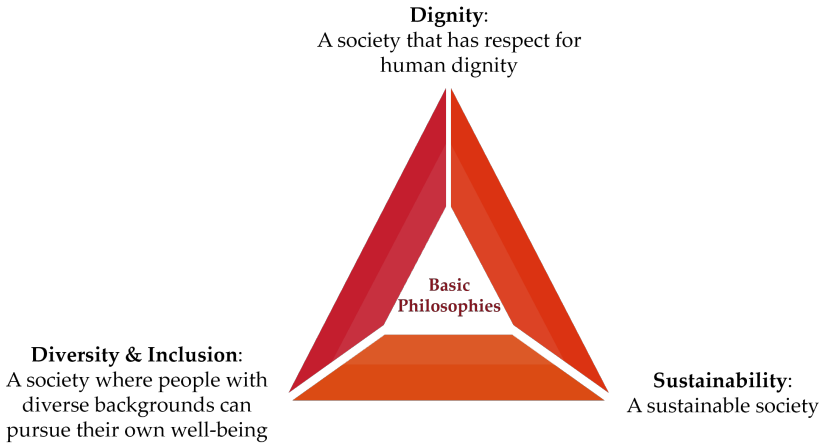

FIGURE 14: Basic Philosophies [14]

internal roles and responsibilities, ensuring that each team involved in the AI lifecycle, whether in development, oversight, or deployment, understands its specific accountability in managing AI-related risks. The framework also promotes more robust documentation and traceability of risks, enabling organizations to track, audit, and respond to potential issues systematically over time. In addition, it strengthens processes for evaluating AI safety and preparing for incident response, helping institutions anticipate, detect, and mitigate adverse events more effectively. Perhaps most importantly, the adoption of the NIST AI RMF encourages a cultural shift toward ethical deployment, fostering greater alignment between the outcomes of AI systems and broader human values, societal expectations, and fundamental rights.

4.3 The Japanese AI Guidelines for Business

The AI Guidelines for Business[14], developed within the Japanese policy ecosystem and aligned with the national vision of Society 5.0[3], represent a forward-looking effort to foster responsible AI adoption across industrial sectors. The guidelines serve as a comprehensive, non-binding framework that encourages companies to voluntarily adopt risk-based governance principles, spanning the full life cycle of AI systems. Japan has adopted a goal-based, soft-law approach, whereby ethical, technical, and organizational recommendations support practical governance without imposing rigid constraints. The guidelines are informed by both international discussions (e.g. OECD and G7 Hiroshima Process) and domestic principles developed in earlier documents, such as the Social Principles for Human-Centric AI and the Governance Guidelines for Implementation of AI Principles ver. 1.1 [1].

4.3.1 Core Content

The AI Guidelines for Business are organized into five principal sections [14], forming a governance framework for entities involved in the development, provision, and use of artificial intelligence systems. The normative foundation of the document outlines a shared societal vision for AI aligned with Japan's strategic concept of Society 5.0, anchored in three foundational philosophies: Human Dignity, Diversity and Inclusion, and Sustainability (ref. Figure 14). These principles emphasize the role of AI to support social advancement, individual autonomy, and inclusive development, while ensuring that its integration into society contributes to long-term well-being and equitable access to its benefits.
These principles encompass human-centricity, privacy protection, safety, fairness, transparency, accountability, and education. To facilitate their implementation, the guidelines introduce a range of concrete measures intended to mitigate societal risks. These include the prevention of manipulative system behaviors, attention to the informational impacts of filter bubbles and disinformation, the promotion of explainability through mechanisms that trace decision-making processes, and the adoption of documentation practices that support both auditability and external validation. Beyond

these general principles, the framework addresses the governance of advanced AI systems, including generative and autonomous technologies. In alignment with the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems[11], the guidelines recommend a lifecycle-based approach to AI oversight. This encompasses pre-deployment testing methods such as red-teaming, incident documentation protocols, post-deployment monitoring, and public disclosure of system capabilities, limitations, and intended uses. The guidelines also advocate for secure development practices, multistakeholder collaboration, and the use of content authentication technologies, such as watermarking, to detect and prevent malicious use of AI-generated outputs. The core content of the framework is structured into 5 parts [14]:

- **Part 1 - Definitions:** introduces key definitions and conceptual distinctions, establishing a shared vocabulary that clarifies the scope and application of the framework. This groundwork enables a structured understanding of the differentiated responsibilities that follow in Parts 3 to 5, which delineate the roles of developers, providers, and users within the AI value chain.
- **Part 2 - Society to aim for with AI and matters each AI business actor works on:** it sets the normative foundation of the guidelines by articulating the societal vision for AI based on three core philosophies and by establishing a set of common guiding principles accompanied by operational recommendations to promote responsible AI development and use.
- **Part 3 - Matters related to AI Developers:** outlines the requirements for AI developers, defined as entities responsible for creating AI systems, including the design of algorithms, models, and training pipelines. Given their upstream position in the AI lifecycle, developers carry significant responsibility in shaping the behavior, reliability, and risks of AI systems. They are expected to ensure the quality, appropriateness, and legal compliance of training data; to implement bias detection and mitigation techniques during model development; and to apply privacy- and security-by-design principles from the earliest phases of system construction. Developers must document their design decisions, training procedures, data handling protocols, and evaluation methodologies to enable transparency and future auditability. Furthermore, they are encouraged to assess potential downstream impacts of the technologies they produce, especially in high-risk applications, and to engage in continuous research and collaboration to align their practices with emerging technical standards and evolving societal needs.
- **Part 4 - Matters related to AI Providers:** outlines the responsibilities of AI providers, who act as intermediaries between developers and end users by embedding AI models into applications, systems, or services and distributing them for practical use. Providers are tasked with ensuring that AI systems are properly configured, integrated, and validated for the intended use cases. They must test and verify the performance, accuracy, robustness, and resilience of AI systems under real-world conditions, and define clear operational parameters, including intended purpose, constraints, and potential risks. Providers are also required to develop comprehensive user documentation, including guidelines for appropriate use, explanation of system functionality, and information about known limitations or possible failure modes. In cases where retraining or updates are necessary, providers should establish maintenance protocols and communicate with developers and users to coordinate improvements. Their role extends to implementing incident-handling systems, facilitating post-deployment monitoring, and ensuring that end users receive adequate support and training. In addition, providers are expected to uphold transparency obligations by communicating essential information in a manner accessible and appropriate to the technical capacities of the users.
- **Part 5 - Matters related to AI Business Users:** addresses AI business users, defined as organizations or entities that apply AI systems within their internal operations or customer-facing processes. As the closest actors to the end effects of AI deployment, business users are responsible for ensuring that AI tools are used in accordance with the provider's specifications and within the bounds of ethical, legal, and sectoral norms. Users must continuously monitor the behavior and outputs of deployed AI systems, identify irregularities or deviations, and report them through established feedback channels. They are also required to evaluate the potential impact of AI use on individuals, institutions, or society, particularly in contexts involving decisions about employment, credit, healthcare, law enforcement, or public administration.

Business users must implement appropriate safeguards to prevent unintended harm, including mechanisms for human oversight, redress, and the protection of fundamental rights. Internally, they must ensure that relevant staff are properly trained in the use of AI systems and that operational procedures align with the governance principles established in the earlier parts of the guidelines. Where AI deployment intersects with public-facing services, users are also encouraged to engage with affected stakeholders, maintain transparency, and uphold accountability regarding how AI is used and governed within the organization.

Taken together, Parts 3 to 5 outline a role-specific governance architecture that distributes accountability and responsibility across the full AI lifecycle. A distinguishing feature of the guidelines is the promotion of an agile governance model, which encourages organizations to move beyond static, rule-based compliance toward dynamic, iterative oversight. This approach is grounded in continuous risk assessment, regular updates of governance protocols, and responsiveness to changes in technology, regulation, and stakeholder expectations. The guidelines emphasize that effective governance cannot rely on uniform rules alone but must be tailored to the roles and influence of each actor across the AI value chain. They advocate for embedding governance within broader business strategies and institutional cultures, promoting coordination, risk sensitivity, and proactive engagement with evolving AI-related challenges. This integrated approach aims to ensure long-term resilience, accountability, and alignment with the public interest in a rapidly evolving technological landscape.

5. Conclusions

The analysis has shown how the increasing and widespread adoption of AI across industries has accelerated the need to establish both regulatory frameworks and robust governance mechanisms capable of addressing emerging risks and promoting the responsible use of the technology. In recent years, several frameworks have been developed to manage and mitigate AI-specific risks. Despite differences in organizational structures and implementation approaches, these models converge around a common set of core principles, the same principles, inspired by the OECD and embedded in the EU AI Act, and can be summarized as follows:

- **Centrality of Ethics and Human Rights:** guidelines and regulatory frameworks are driven by the need to protect human rights, requiring high standards of transparency, accountability, non-discrimination, privacy, robustness, and security.
- **AI Lifecycle as the Foundation of Governance:** effective AI risk management demands a holistic approach that spans the entire lifecycle of AI systems, from development to decommissioning. Frameworks emphasize the need for continuous risk monitoring and mitigation across all the life-cycle phases.
- **Human Oversight as a Safeguard:** the necessity of maintaining human-in-the-loop oversight over AI systems, especially in high-risk or decision-critical applications, is a crucial pillar across all frameworks. This requires the development of mechanisms that enable humans to oversight on AI systems, ensuring accountability, safety, and control throughout the system's lifecycle.
- **Risk Tiering and Proportionality:** a key principle is the assessment and classification of AI systems based on their risk level, ensuring that the intensity of controls is proportionate to the potential impact of the system.
- **Organizational Approaches:** the importance of embedding AI governance across organizational structures is underlined by all the frameworks. This requires the adoption of AI risk management not only as procedures and processes but as a corporate mindset involving different functions and teams to ensure effective oversight and alignment with enterprise goals.

To conclude, too often, risk management is perceived as a compliance exercise, a necessary but limiting set of controls. In the context of AI, this mindset is not only outdated but also strategically

shortsighted. When designed and executed effectively, AI governance Frameworks become a powerful driver of competitive advantage. Robust governance empowers organizations to:

- **Accelerate Safe Innovation:** by establishing clear boundaries, escalation paths, and validation protocols, governance reduces uncertainty and enables faster experimentation and deployment.
- **Build Stakeholder Trust:** consumers, investors, regulators, and the public are increasingly concerned about the ethical use of AI. Demonstrating a credible governance framework enhances reputation, supports brand integrity, and attracts ethically conscious partners and clients.
- **Enable Internal Coherence:** a standardized approach to AI governance facilitates cross-departmental collaboration, minimizes duplication of efforts, and ensures consistency in how decisions are made and risks are managed.
- **Enhance Regulatory Readiness:** as AI regulation evolves across jurisdictions, proactive governance allows organizations to anticipate requirements, reduce compliance burdens, and respond swiftly to legal changes.
- **Foster Long-term Adaptability:** with technology evolving rapidly, static or informal practices are insufficient. A governance model that is scalable, flexible, and principle-based equips organizations to manage future use cases, risks, and opportunities more effectively.

References

[1] **AI Strategy Council, Japan.** *Governance Guidelines for the Implementation of AI Principles ver. 1.1.* January 2022.

[2] **Bick, A., Blandin, A. and Deming, D.** *The Rapid Adoption of Generative AI.* NBER Working Paper Series N. 32966, February 2025.

[3] **Cabinet Office, Government of Japan.** *Social Principles of Human-Centric AI.* Official Publication (Japan), March 2019.

[4] **Cristanto, J. C., Leuterio, C. B., Prenio, J. and Yong, J.** *Regulating AI in the Financial Sector: Recent Developments and Main Challenges.* BIS FIS Insight n. 63, December 2024.

[5] **Esposito, D., Carrozino, P., Ghilardi, B. and Cecchin, M.** *Artificial Intelligence Act (AI Act).* Argo N. 26, August 2024.

[6] **Esposito, D., Ghilardi, B. and Cecchin, M.** *AI Risk Management Framework: National Institute of Standards and Technology (NIST).* Just in Time, February 2023.

[7] **European Commission.** *Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act).* February 2025.

[8] **European Commission.** *Guidelines on the definition of an artificial intelligence system established by Regulation (EU) 2024/1689 (AI Act).* February 2025.

[9] **European Commission.** *Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act).* Official Journal of the European Union, July 2024.

[10] **European Commission.** *Targeted stakeholder consultation on classification of AI systems as high-risk.* June 2025.

[11] **G7 Leaders.** *Hiroshima Process - International Guiding Principles for Organizations Developing Advanced AI Systems*. G7, October 2023.

[12] **IOSCO.** *Artificial Intelligence in Capital Markets: Use Cases, Risks, and Challenges.* Report of the Board of IOSCO, March 2025.

[13] **Japan FSA.** *Preliminary Discussion Points for Promoting the Sound Utilization of AI in the Financial Sector.* AI Discussion paper, March 2025.

[14] **Japan Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry.** *AI Guidelines for Business Ver1.0.* Official Publication (Japan), April 2024.

[15] **McCarthy, J.** *What is Artificial Intelligence?* Stanford Press, November 2007.

[16] **Maslej et al.** *The AI Index 2023 Annual Report.* AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2023.

[17] **Maslej, N. et al.** *The AI Index 2024 Annual Report.* AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2024.

[18] **Maslej, N. et al.** *The AI Index 2025 Annual Report.* AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2025.

[19] **Mazzoni, N.** *AI Act: AI System Definition and Prohibited AI Practices.* Just in Time, April 2025.

[20] **Mazzoni, N., Figuriello, J. and Campaniolo, G.** *Artificial Intelligence Financial Industry Market Overview.* Just in Time, July 2025.

[21] **Mazzoni, N., Ranieri, M., Gentalavigna, F. and Bandini, L.** *AI: Regulation, Rise and Challenge.* Just in Time, April 2025.

[22] **National Institute of Standards and Technology.** *AI RMF Playbook.* January 2023.

[23] **National Institute of Standards and Technology.** *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* January 2023.

[24] **National Institute of Standards and Technology.** *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.* July 2024.

[25] **NVIDIA.** *State of AI in Financial Services: 2025 Trends.* NVIDIA Survey Report, February 2025.

[26] **OECD.** *OECD Framework for the Classification of AI Systems.* OECD Digital Economy Papers No. 323, February 2022.

[27] **OECD.** *Regulatory Approaches to Artificial Intelligence in Finance.* OECD Artificial Intelligence Papers, September 2024.

6. Sitography

[28] **Eurostat.** *Website.*

[29] **OECD AI Principles.** *Website.*