cisco
isms forum
GIA GRUPO DE INTELIGENCIA ARTIFICIAL

AI Governance

Cisco, through its CDA ("Country Digital Acceleration") investment programme, known as DIGITALIZA in Spain, is committed to accelerating the digitalisation of our country across the various areas of our economy and society. In particular, the development of knowledge and awareness of cybersecurity practices within Spanish organisations is a key priority. In this context, the analysis focuses on establishing a comprehensive framework for the governance of Artificial Intelligence, addressing everything from ethical and regulatory principles to risk management, data protection and the definition of roles and responsibilities. The document provides practical guidelines to ensure trustworthy and secure AI that is aligned with current regulations, supporting strategic decision-making and organisational adaptation in the face of emerging technological and regulatory challenges.

Index

Project Members

Coordinators:
Angel Ortiz
Sergio Padilla

Participants:
Alberto López
Alberto Pinedo
Alberto Torralba
Carlos A. Sáiz
Elena Mora
Enrique Cervantes
Jose Ramón Monleón
Ignacio Cagiga
Jaime Requejo
Javier Pinillos
Jesús Alonso
Jesús Muñoz
Manuel Barrios
Manuel Ruiz del Corral
Mario Encinas
Marta Martínez
Rafael Fernández
Rafael Tenorio

Project Management:
Beatriz García

Design and Layout
Susana Marín

Review:
Angel Pérez
Carlos A. Sáiz
Francisco Lázaro

Introduction

01.

1.1. Objective of the AI Governance Document

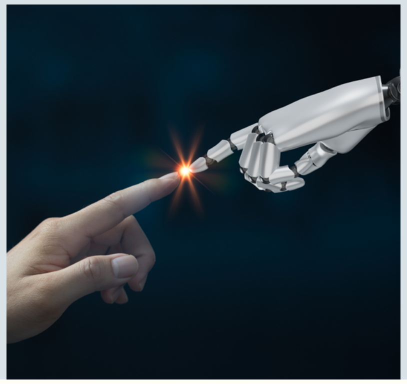
The objective of this document is to establish the fundamental framework and essential guidelines for **AI Governance (hereinafter, AI)**, serving as a structured and comprehensive roadmap for organisations to develop, deploy and use AI systems in a **safe, resilient, responsible, ethical and regulatory-compliant** manner.

It aims to be a **strategic and operational mandate** to ensure **business continuity** and **cyber resilience** in the age of AI. The rapid adoption of AI thanks, among other factors, to the rise of Large Language Models (LLMs) and Generative Artificial Intelligence (GenAI) is transforming key processes for business, operations, information security and customer service, with significant growth expected in adoption over the next three years. However, this exponential growth and its disruptive potential bring major legal, ethical and social challenges. Indeed, the vast majority of organisations perceive a **considerable potential risk** (90.8% rate it from moderate to extreme), with the most critical risks concentrated in **operations and information security.**

Thus, a key objective of this framework is to provide an **AI Governance Model** that not only enables the proper management of legal and ethical risks, but also ensures that the development and deployment of AI are carried out in a **secure, trustworthy and robust** manner, managing risks throughout the **entire service or system lifecycle.**

Promoting **Trustworthy, Ethical and Secure AI**

The document focuses on the need for AI systems to be trustworthy, which means they must be **lawful, ethical and robust at the same time.**



01.

Human-Centric and Ethical Approach

AI must focus on the welfare and safety of people. This requires the integration of ethical values from the early stages of design (ethics by design). The fundamental ethical principles to address include respect for **human autonomy**, **prevention of harm**, **fairness** (avoiding biases and discrimination) and **explainability and transparency**.

02.

Risk Mitigation

Continuous management of risks arising from and widely perceived in, the use of AI is necessary. Risks must be identified, assessed and addressed throughout the AI lifecycle, which is crucial for systems with a high degree of autonomy and decision-making capacity.

03.

Transparency and Explainability

AI systems must operate with a sufficient level of transparency so that those responsible for their deployment can correctly interpret and use their results. This obligation includes facilitating the understanding of the logic applied in decision-making, ensuring compliance with the GDPR transparency principle and avoiding "black box" systems.

Following on from the above, the inherent risk of AI requires an approach that integrates cybersecurity from the earliest stages of a project (security by design), proactively managing risk and minimising operational risk as much as possible. Thus, the security objectives for information security managers focus on:

Ensuring Robustness and Resilience

AI reliability (including validity, robustness and cybersecurity) is a an essential obligation, especially for high-risk systems (Art. 15 of the Artificial Intelligence Regulation, hereinafter referred to as RIA)[1]. The system must be resilient to errors, failures and inconsistencies and be able to withstand unauthorised attempts at alteration, such as **data poisoning** attacks or **adversarial examples** (model evasion).

Define Security Accountability (RACI)

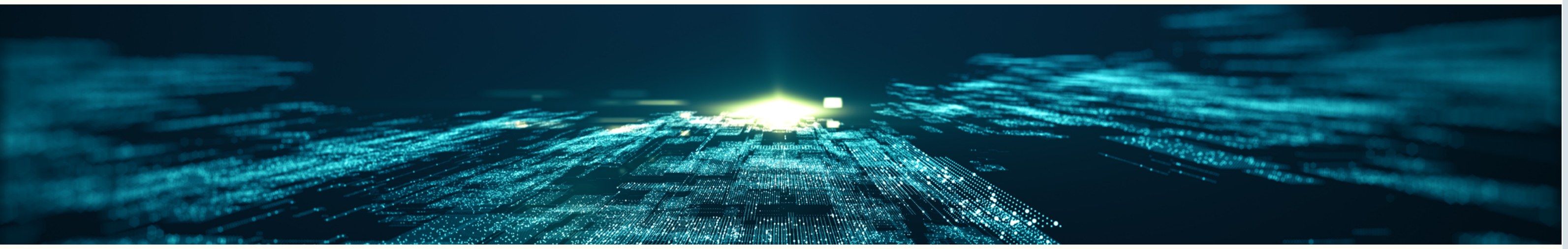
Establish a clear **Responsibility Matrix (RACI)** where an entity's information security teams are directly responsible (in English, **responsible and accountable - R, A -**), aligning monitoring, audit logging and risk management with best practices.

Mitigate Specific Threats

Going beyond traditional cyberattacks. It is imperative to develop a **comprehensive catalog of AI-specific attack scenarios** (such as those detailed by MITRE ATLAS and OWASP Top 10 for LLM) and to design realistic simulations to evaluate them.

Establish Continuous Control and Traceability

Need to implement **continuous monitoring mechanisms** to identify anomalous activities (such as Shadow AI or unauthorised use of public systems) and to guarantee the existence of automatic and detailed **log files** throughout the entire lifecycle, which is mandatory for high-risk systems (Art. 12 RIA).



[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council, of 13 June 2024, laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (OJ-L-2024-81079). https://www.boe.es/buscar/doc.php?id=DOUE-L-2024-81079

Alignment with the **Regulatory Framework and Legal Compliance**

As part of the document's objectives, it is important to guide the organisation toward compliance with the growing and demanding regulatory landscape, avoiding economic and reputational consequences from non-compliance and making a solid governance framework essential.

01.

Adaptation to the AI Regulation (RIA)

One of the main objectives of any company or organisation should be to ensure compliance with **Regulation (EU) 2024/1689 on Artificial Intelligence (RIA)**, considered the first comprehensive legislation worldwide in this field. The RIA establishes harmonised rules based on a risk approach (unacceptable, high, limited and minimal risk).

02.

High Risk Requirements

It seeks to ensure that systems classified as high-risk, which can **significantly impact safety, health or fundamental rights,** meet the mandatory requirements of the RIA, including risk management, data governance, technical documentation, traceability, transparency, human oversight and reliability. It is also highly relevant that we are very likely to use external AI models and services (procured as a consumer or as a SaaS offering), which makes it essential to clarify responsibility and require the provider to offer sufficient assurances.

03.

Regulatory Integration

It is not only necessary to ensure compliance with AI-specific regulations, but also with other essential legal frameworks such as the **General Data Protection Regulation (hereinafter, GDPR[2]),** Intellectual and **Industrial Property legislation and any mandatory rules or requirements relating to information security and cybersecurity.** The framework must demonstrate the previously mentioned obligation of accountability and the need to take proactive responsibility for the systems used or developed.

04.

Robust Contractual Framework

The document also seeks to establish a robust contractual framework to regulate the terms and responsibilities of parties involved in the contracting of AI systems and services, which is especially relevant for the management of AI system providers.

[2] European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L119, 1-88. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679

Provide a **Practical and Operational Framework**

The document is intended for a diverse audience (CISO, CAIO, CDO, CIO and other relevant and and involved CXOs within each organisation, DPO, engineers, developers, legal teams, etc.) and aims to serve as a practical and adaptable resource.

01. *Governance Structures and Internal Policies*

A key point of this guide is to help organisations define an **AI Governance Model** that includes adapting internal structures, defining clear roles and responsibilities and developing **internal policies and procedures.**

02. *Methodological Guide*

A methodology is provided to **embed governance throughout the entire AI lifecycle** (design, development, deployment, monitoring/control and final decommissioning). This includes guidelines on data quality, bias prevention and technical documentation.

03. *Culture and Training*

An aware, cross-sectional and empowered organisational culture must be promoted, with specific training programs on AI security, ethics and risks for all staff, since lack of internal capability and cultural resistance are common causes of failure in implementation.

In summary, the document aims to enable organisations to **harness the strengths of AI while mitigating its risks,** providing a clear and structured methodology for developing and using AI within a framework of trust, legality and responsibility.

Evolution Since the 2023 Guide 1.2.

The evolution of AI Governance since the publication of the "ISMS Forum 2023 Guide, AI Governance Model" to date has been characterised by a fundamental shift: moving from the **conceptual identification of the need** to govern AI to the **mandatory operational and regulatory implementation.**

Back in 2023, the debate focused on establishing the foundations of an **AI Management Strategy.** Today, the urgency lies in the **immediate adaptation** of governance, risk and compliance (GRC) structures to handle threats that have evolved in complexity and volume, aligning with a high-impact regulatory framework that has already been published and whose implementation timeline is underway.

The 2023 Guide recognised the potential of AI in business strategy and the need to establish guiding principles. The evolution is defined by the convergence of those principles being materialised into mandatory legal requirements (RIA), the technological acceleration driven by Large Language Models (LLM) and Generative AI and the emerging security risks inherent in these systems:

01. *The Consolidation of the Risk-Based Approach*

2023

Back in 2023, it was already stated that the governance model should be based on principles such as lawfulness, ethics and robustness and the need to manage AI-related risks on an ongoing basis was acknowledged.

2026

The main evolution is the ratification of the RIA, which transforms risk management into a structured and hierarchical legal obligation for high-risk systems (SAR) and systemic risk models by classifying general purpose AI models (GPAI or GPM) at standard and systemic risk levels.

[3] ISMS Forum. (2023). Artificial Intelligence Governance Model. Spanish Association for the Promotion of Information Security. https://www.ismsforum.es/ficheros/descargas/modelo-de-gobiernoconverted1700037593.pdf

02. *The Leap of Generative AI (GenAI)*

Although Generative AI had already emerged in 2023 (mainly with the ChatGPT phenomenon), its adoption and risks have quickly escalated:

Adoption Forecast

2023

The 2023 survey indicated that the vast majority of organisations **(88.89%)** expected an increase in GenAI usage in the next 3 years.

2026

The focus has shifted from a technological opportunity to also being a critical security challenge.

Emerging Personalised Risks

2023

Risks were grouped into general categories **(discrimination, lack of ethics, operational risk).**

2026

The evolution now requires a focus on risks specific to LLMs and GenAI, as defined in the **OWASP Top 10 for LLMs**, including ***Prompt Injection attacks, Model Extraction***, and ***Data Poisoning***, which are described in more detail in the relevant chapters below.

Governance by Design (AIGD)

2023

Recommendation to apply the principles of **"Privacy by Design"**.

2026

The complexity of autonomous systems (Agentic AI Systems) and GenAI has driven the approach of **AI Governance by Design (AIGD)**. This is a direct evolution of the principles of "Privacy by Design," now applied integrally. AIGD seeks to integrate ethical, legal and social considerations directly into system development from conception, avoiding costly adjustments typical of traditional reactive approaches.

Technological, **Regulatory** and **Organisational Advances**

1.3.

The current context of AI Governance is marked by a **confluence or convergence of exponential technological advances, intense global regulatory activity and the imperative need to restructure organisational capabilities** to manage the resulting risk. The situation has evolved from a theoretical debate on ethical principles to a **strategic imperative of legal compliance and cybersecurity** with defined deadlines.

Hence arises the need for a **profound reconfiguration of organisational maturity** to achieve cyber-resilience, as will be seen throughout the document.

Global Regulatory Landscpae: **From Ethics to Legal Imposition**

The regulatory context is radically transformed, from an initial approach of publishing ethical guidelines (such as UNESCO's or the OECD's) to the imposition of legally binding regulations, with the RIA being the catalyst for global change.

1. The centrality of Regulation (EU) 2024/1689 on Artificial Intelligence (RIA)

The RIA is considered the first comprehensive legislation in the world specifically designed to regulate AI systems. The risk-based approach dictates the compliance strategy of multinational organisations, as we will see.

2. The fragmented global regulatory framework

While the EU adopts a prescriptive approach, the rest of the world progresses at different speeds and with different philosophies:

United States

Historically, the approach has been more flexible and non-prescriptive, focused on sectoral self-regulation and encouraging innovation. The government has issued executive orders seeking global leadership in AI through international cooperation.

China and Brazil

Jurisdictions in response. China presented a draft law focused on research and development. Brazil also has legislation under debate (PL 2.338/2023), distinguishing between excessive risk and high-risk systems.

Multinational Challenge

This fragmentation forces organisations with global presence to develop a compliance approach across multiple jurisdictions, adapting contracts and technology to the highest standard (often the EU's RIA).

3. Normative convergence with existing laws

The current context is not limited to the RIA. AI governance is a transversal task that requires full integration with pre-existing regulations:

Privacy and Personal Data Protection

Historically, the approach has been more flexible and non-prescriptive, focused on sectoral self-regulation and encouraging innovation. The government has issued executive orders seeking global leadership in AI through international cooperation.

Data Governance and Access (Data Governance Framework):

Beyond personal data protection, it is crucial to ensure robust data governance, as data must be trustworthy and properly governed. This includes, for example, the need to consider new European regulations governing data exchange and access, such as the Data Governance Act (DGA) and the Data Act, which regulate the availability and sharing of information, being crucial for data quality and traceability.

Cybersecurity

Regulations such as **DORA** (Digital Operational Resilience Regulation for the Financial Sector), the **NIS 2 Directive** and the **Cyber Resilience Act** directly impact the robustness and security required of AI systems (Art. 15 RIA).

Intellectual Property, Trade Secrets and Civil Liability

An explicit contractual approach is required to protect **Intellectual Property** (software licensing, copyright over GenAI inputs and outputs) and **trade secrets**, especially when **confidential information** is introduced into third-party AI systems.

Sectoral Regulation

Undoubtedly, governance must be tailored to the industry. In highly regulated sectors, AI must comply with specific sectoral regulations.

For example, in the healthcare sector, any AI system involving the processing of health data is considered at least a high-risk system, subject to thorough validation under regulations such as the Medical Devices Regulation (Regulation (EU) 2017/745), in addition to strict requirements for access to medical records. The framework must demonstrate the obligation of accountability and proactively take responsibility for the systems used or developed.

Organisational Reconfiguration and **Proactive Governance:** Maturity, Roles and Supplier Management

It is a major challenge to move from intention to operationalisation of governance, which requires new structures, processes and an adapted culture.

The technological and regulatory context, characterised by technological volatility and RIA regulation, demands organisations rapidly mature their Governance, Risk and Compliance (GRC) processes. Strategy must pivot towards proactive governance and cyber resilience to avoid economic, reputational and operational consequences from disruption of the AI ecosystem.

Current Landscape of AI

02.

AI has established itself as a revolutionary technology that is redefining multiple sectors and aspects of daily life. A significant turning point has been the introduction of **generative AI and foundation models**, such as GPT (by OpenAI), Claude AI (by Anthropic) and Gemini (by Google), among others. These models, trained with vast amounts of data from different domains, have become essential pillars of the most innovative technologies, offering **extraordinary versatility and power** in tasks ranging from text and image generation to complex data analysis and natural user interaction. However, this increasing complexity also presents ***significant challenges*** in terms of interpretability of their decisions, data privacy, security, LLM opacity and the need for strong ethical and regulatory frameworks, such as the RIA, to ensure responsible and reliable development and use.

The current AI landscape is characterised by rapid and transformative evolution, driven by advances in data storage, processing and connectivity, which underscores the urgency of defining appropriate governance.

2.1. Transformative Impact of Foundation Models

The term foundation model, which today may seem familiar or everyday, was first coined by Stanford University in August 2021 and is used to describe AI models that have been trained with **large amounts of unlabeled data from different domains, usually through self-supervised learning.** This learning method involves the model learning from unlabeled data, generating its own supervision signals instead of relying on manual labels, which are costly and time-consuming, so the model is trained to predict part of the data from other parts of the same data.

This type of learning gives them the ability to **adapt to a wide range of tasks** and learn general representations from data, which provides them with great **versatility and power.** Foundation models can then be fine-tuned with a minimal amount of labeled data for specific tasks (a process known as transfer learning), **thus drastically reducing the cost and development time of AI solutions.**

Setting aside the revolution they have brought about in the way new AI systems are developed today, the transformative impact of these models is substantial and can be observed across a wide range of areas. Some key issues to highlight include:

→ **Exponential evolution of AI usage:** the level of detail, completeness and richness in the responses of foundation models has qualitatively and quantitatively improved AI-based applications, driving exponential adoption by society, facilitating the digitisation process and becoming regarded as a disruptive phenomenon akin to the birth of the Internet.

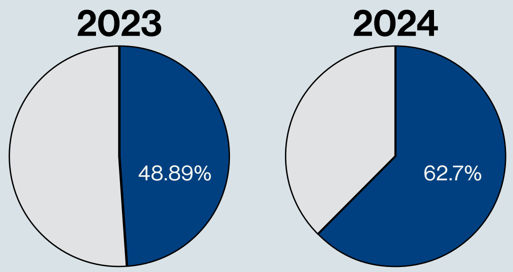
→ **Increased productivity:** their high processing and connectivity capacity in solving complex problems and their similarity to human reasoning, is becoming evident in boosting productivity and efficiency for professionals using them as daily support.

→ **Task versatility:** from their inception, foundation models are designed to be adaptable to a wide range of tasks, which has enabled their rapid adoption in everyday activities such as **text generation** (they excel at producing written content, answering questions, solving problems, and assisting with drafting or code generation), the **analysis and summarisation of complex texts, image generation and processing, the analysis of large volumes of data,** and **information search and research.**

→ **Competitive advantage:** AI is giving organisations a crucial competitive advantage, improving customer experience and operational efficiency.

→ **Cost reduction:** the adoption of foundation models and generative AI is directly linked to reduced operational and organisational costs, a key factor driving implementation in the business sector. In 2024[4], **62.7%** of companies were already using generative AI or planning to do so in the short term, compared to **48.89%** of companies that stated the same in 2023.

Use of Generative AI in companies



In summary, foundation models have redefined the AI landscape, making the technology more accessible versatile and capable of being integrated into a multitude of applications and processes, which leads to significant transformation in the way AI is developed, used and understood.

[4] ISMS Forum. (2024). Analysis II survey: Adoption and Governance of Artificial Intelligence. Spanish Association for the Promotion of Information Security. https://www.ismsforum.es/ficheros/descargas/encuesta-ia-final3-31731310534.pdf

2.2. Emerging Risks Associated with Advanced AI Use: Deepfakes, Attack Automation, LLM-Driven Exposure and the Evolution of Traditional Threats

By definition, an emerging risk is defined as those new or changing events or factors that, while having a low probability of materialising, can have a major impact because they present new forms of manifestation. With AI and its surge in many scenarios, **risks are emerging regarding cybersecurity and information protection in organisations.**

As AI engines improve, with new algorithms and optimised computing processes, it is easier to incorporate AI capabilities into all processes, whether legitimately and ethically, or less legitimately and very unethically. As a result, **traditional risks in the field of cybersecurity continue to evolve, both qualitatively and quantitatively.** Traditionally, a person was behind a threat or attack. Subsequently, the trend moved towards automation and scaling of tactics, which has clearly evolved thanks to the application of AI.

There are more and more associated risks and their presence is expected to increase, such as deepfakes, automation and sophistication of attacks using AI, the extended use of LLMs and the evolution of traditional threats under AI influence. We highlight some of them here, although they are developed in greater detail later in the guide.

01 | *DeepFakes*

Active since 2017-2018, the term developed mainly in connection with the creation of audiovisual content generated with a clear intent to deceive the recipient. It was already an existing practice, but it is nothing more than a traditional threat which is improved or enhanced through AI tools and not just with audio or messages, but now also videos and even real-time video calls, impersonating almost anyone's identity.

02 | *Automation and Sophistication of Attacks*

In many organisations, normal operations are being included or enhanced with the integration of AI tools, whether through process automation or as an improvement in results or achievement of previously unattainable goals.

With this premise, cybercrime groups incorporate the same AI tools to improve their processes and activities and just as a legal organisation has these technological accelerators, so does a cybercrime group. These improvements brought about by AI boost the success of attack campaigns and, worryingly, it should be noted that it is no longer just groups with advanced knowledge and very specific objectives, but the socialisation and generalisation of AI tools has allowed small groups to reach a large volume of targets with more elaborate and sophisticated mechanisms, achieving greater success in their frauds, especially among organisations with fewer cybersecurity resources or more vulnerable citizens in the digital realm.

03 | *Widespread Use of LLMs*

Specifically, the spread of LLM-type AI tools (Large Language Models), such as ChatGPT (OpenAI), Copilot (Microsoft), Gemini (Google), DeepSeek, LLaMA (META) and many others, is a basic element in the lives of new generations and is also an increasingly widespread addition in organisations and for people in their daily lives.

From an organisational standpoint, it can become a major risk if employees use these models widely and without filters, as they may compromise sensitive information and lose control over where the information or data resides. Although it is not about fencing in the open field either, it is important to establish strategies and analyse how these AIs are used to pursue the best integration strategies for these tools.

Given these risks associated with the use of AI, particularly LLMs, the best approach is to understand how they will be used within the organisation and to develop a strategy that, while meeting business needs and seeking to ensure security, leads to an informed decision on whether to block, adopt or integrate them. Organisations cannot remain detached from this context.

04 | *Traditional Threats*

As previously mentioned, cybercriminal groups enhance their capabilities by including AI tools in their operations, allowing them to reach objectives that without AI they could not, or that would be very costly in resources and/or time.

Traditional fraud attempts, such as phishing, targeted or otherwise, are already being enhanced with AI engines, automating these attacks, enabling more powerful strategies and more effective targeting of their victims. The success rate increases and allows cybercriminal groups to improve their capabilities and obtain greater profits.

For all these reasons, any organisation, regardless of size, already faces ever more sophisticated threats thanks to the use of AI, making them difficult to detect and with very high potential impact.

05 | *Misuse of AI Tools*

In some cases, the accelerated incorporation of AI functions into many existing technologies can lead to an increase in risks, with a growing likelihood that they will materialise. This can already be observed in autonomous systems, mobile phones, vehicles and countless commercial platforms and software products. These functionalities often enable basic decision-making, streamlining operations and workflows. Although this is clearly a benefit and a significant improvement to the products, it also means that errors in certain decisions made by these tools can result in serious consequences, especially when the impact concerns people's safety.

When addressing these solutions, it is necessary to analyse whether they are truly needed and if the benefits outweigh the possible damages or impacts, since often this is not the case.

Ecosystem of Roles in AI Governance

03.

Effective AI governance requires active and coordinated participation of different professional profiles. This chapter describes the key roles involved, their main responsibilities, the necessary collaboration dynamics between them and a proposed roles and responsibilities matrix.

Main **Roles Involved** 3.1.

AI governance is a cross-cutting discipline that requires smooth interaction between multiple strategic roles. While each organisation will have a governance model according to its nature, size or maturity, here, we pedagogically describe the main profiles that may make up this ecosystem.

Chief AI Officer (CAIO)

A relatively new role, already existing in modern organisations. Their primary responsibility is not only to oversee specific AI projects but to define the overall strategic vision for AI within the company. The CAIO answers questions such as: What are we going to use AI for? What value will it bring to the business? How will we do it ethically and sustainably?

The CAIO must ensure that AI is used safely, fairly and in line with the organisation's values. Their role includes connecting technical potential with business objectives, identifying opportunities, managing risks and ensuring that AI not only operates reliably but also responsibly.

A role that acts as a bridge between senior management and operational teams, collaborating closely with key profiles such as the CIO, CISO, CDO and DPO and other relevant roles, aligning AI strategy with corporate policies, legal regulations and the company's technological capabilities.

Their **functions** would include:

- **Being responsible for defining and implementing the organisation's overall AI strategy.**
- **Overseeing the ethical, safe and efficient use of AI systems.**
- **Collaborating, at a minimum, with the CIO, CISO, CDO and DPO to ensure alignment with corporate policies and regulatory requirements.**

Chief Information Officer (CIO)

The CIO is responsible for ensuring the company's technological infrastructure is ready to support AI.

Their responsibilities include ensuring the suitability, scalability, security and continuity of servers, networks, platforms and cloud and/or on-premises data centers for deploying AI models. They must also evaluate the interoperability of new and old systems, to facilitate seamless integration of AI-based solutions in business processes.

The CIO works with the CAIO to digitally transform the organisation, ensuring that AI is implemented effectively, without technical friction or bottlenecks.

Their main **functions** are:

- **Being responsible for the technological infrastructures supporting AI systems.**
- **Ensuring the viability, scalability and interoperability of the technology ecosystem.**
- **Coordinating with the CAIO to integrate AI into business processes.**

Chief Information Security Officer (CISO)

The CISO safeguards the organisation's information security, including all information systems and in the AI context their role becomes even more crucial. A poorly protected AI model can be manipulated, compromised or maliciously used, jeopardising data and/or the resulting decisions.

The CISO must implement security controls, ensuring the protection of models against adversarial attacks, information leaks or manipulation during training. They also actively participate in the assessment of risks inherent to AI use, establishing preventive measures and incident response plans.

In coordination with the CAIO and CDO, they ensure that the data used to feed the models is adequately protected and managed according to best cybersecurity practices.

Their main **functions** are:

- **Supervising the security of AI systems, including protection against threats such as adversarial attacks or model manipulation.**
- **Establishing security controls and participating in risk assessment.**
- **Coordinating with the CAIO and the CDO to protect the confidentiality, integrity and availability of data.**

Chief Data Officer (CDO)

The CDO is responsible for everything related to data within the organisation: from its acquisition and storage, to its quality, access, use and governance. In the AI context, data is the essential fuel for training models, making the CDO's mission critical.

They must define clear policies on how data is collected, classified, stored and used, especially sensitive or personal data. They also work to ensure that data is of high quality, relevant, unbiased and representative, all fundamental aspects for an AI model to function properly.

The CDO works closely with the CAIO to ensure models are built on appropriate data and with the CISO and DPO to ensure, beyond legal compliance, everything related to data security, privacy and personal data protection.

Their main **functions** are:

- **Being responsible for data governance and quality, a critical aspect in AI model training.**
- **Defining policies for management, access and ethical use of data.**
- **Collaborating closely with the CAIO, the CDO and the DPO to ensure regulatory compliance.**

Data Protection Officer (DPO)

The DPO is the guarantor of privacy within the organisation. Their main mission is to ensure that all processing of personal data—including what occurs in AI systems—complies with applicable regulations, such as the GDPR.

In AI projects, the DPO must assess whether personal data is being processed, the risks to individuals' rights and whether a Data Protection Impact Assessment (DPIA or RIA) is necessary.

Works closely with the CAIO, the CISO, the CDO and the legal team to prevent AI use from leading to privacy violations or regulatory penalties.

Their main **functions** are:

- **Supervising compliance with data protection regulations (e.g.: GDPR).**
- **Assessing privacy impacts of AI systems (e.g.: RIA - Risk Impact Assessments) and may perform Fundamental Rights Impact Assessments (FRIA) when required by the RIA.**
- **Collaborating with with the CAIO, the CDO and the legal team to ensure regulatory compliance.**

In addition, depending on the organisation and whether AI systems may process personal data, this will be the role that can assume functions related to RIA compliance, as well as ensuring responsible use of the AI system, leveraging existing structures and processes for GDPR compliance.

For detailed analysis of these issues, it is essential to consult the DPO White Papers[5] published by the ISMS Forum.

Compliance Officer and Legal Team

They are the regulatory compass within the AI ecosystem. They interpret and apply current legislation, assessing whether the intended uses of AI are within the applicable legal framework. They review contracts with technology providers, terms and conditions of generative AI platforms, intellectual property protection for models and data, as well as the legal implications of automating decisions that may affect people.

These professionals are key to ensuring the organisation does not incur legal or ethical risks, especially as new regulations come into force.

Their role is complementary to that of the DPO, but with a broader focus on legality, business ethics and regulatory compliance in general.

Their main **functions** are:

- **Ensuring that AI systems are developed and used according to current laws and regulations.**
- **Assessing contractual clauses, intellectual property rights and emerging and established regulatory frameworks.**
- **Advising on the legality of specific AI uses, including automated decisions.**

[5] ISMS Forum. (2019). The DPO White Paper: Functions, Challenges and Best Practices for the Data Protection Officer. https://www.ismsforum.es/ficheros/descargas/el-libro-blanco-del-dpo---isms-forum-y-data.pdf

ISMS Forum. (2024). The DPO White Paper: Functions, Challenges and Best Practices for the Data Protection Officer. https://www.ismsforum.es/ficheros/descargas/libroblancodpofinal1739382530.pdf

Collaboration Dynamics and **Interdepartmental Coordination** 3.2.

The cross-cutting nature of AI demands fluid dynamics between these roles.

To illustrate how these interactions materialise in practice, **Annex 1: Use Cases and Common Coordination Challenges** is included, where concrete examples are presented showing the need for effective collaboration between some of the mentioned areas.

Key collaboration mechanisms:

→ **Joint committees or groups:** regular meetings to review use cases, risks, compliance and progress.

→ **Escalation protocols:** for incident management or critical decisions.

→ **Collaborative workspaces:** use of shared tools like Teams for tracking AI projects.

→ **Culture of transparency:** accessible documentation on the design, training and use of AI models.

3.3. **RACI Matrix** of Roles and Responsibilities (Responsible, Accountable, Consulted, Informed)

Activity / Task	CAIO	CIO	CISO	CDO	DPO	Legal Team & Compliance	InfoSec Team	Technical Teams
AI strategy definition	R	C	C	C	C	I	I	I
Technical architecture design for AI solutions	R	A	C	I	I	I	I	R
AI security risk assessment	C	C	A	C	I	I	R	C
Regulatory compliance oversight (AI Act, GDPR)	C	I	I	C	A	R	I	I
Governance of data used for training	C	I	C	A	C	I	I	R
Development and implementation of AI models	A	C	C	C	I	I	I	R
Ethical and social impact assessment	A	I	I	C	C	R	I	I

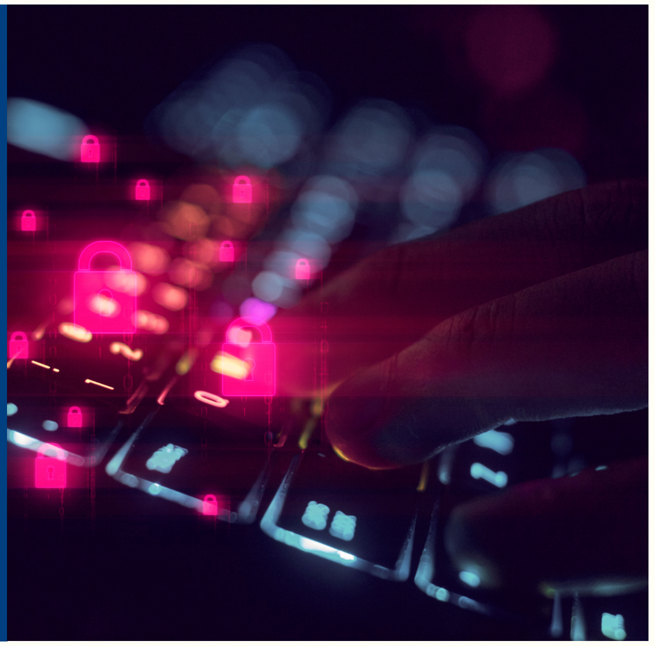
■ → **R (Responsible):** the one who executes the task

□ → **A (Accountable):** the one who makes the final decision and is responsible for it

▨ → **C (Consulted):** the one who must be consulted

▢ → **I (Informed):** the one who must be informed

This RACI matrix helps clarify expectations and responsibilities, reduce ambiguities and ensure efficient coordination among different actors involved in AI governance.



Governance **Indicators and Metrics**

AI governance requires measurement mechanisms that allow assessment of its maturity, compliance and effectiveness.

Indicators and metrics provide an objective basis to assess how ethical, legal, technical and security risks associated with AI are being managed and also serve to show traceability and continuous improvement of the governance framework, as well as to demonstrate compliance with the principles of transparency, traceability, human oversight and risk management throughout the entire AI system lifecycle.

Without measurable data, policies and principles remain declarative, lacking real capacity for monitoring or informed decision-making.

In this regard, having a metric framework enables these regulatory requirements to become mechanisms for continuous monitoring, providing an objective view of the level of control, effectiveness and improvement of the governance system. Furthermore, it facilitates alignment with international reference frameworks and supports accountability to auditors, authorities and stakeholders.

Practical **Recommendations**

Based on these metrics, it is possible to propose a set of recommendations to strengthen governance and align the use of AI with ethical and regulatory standards:

→ **Establish an AI policy aligned with RIA:** having a corporate policy that incorporates its requirements from the outset will allow organisations to anticipate legal obligations and avoid penalties.

→ **Create a multidisciplinary AI governance committee:** AI is not just a technical matter. Involving experts in ethics, law, security, business and technology ensures a comprehensive view, reduces biases and promotes more balanced decisions.

→ **Apply principles of transparency, fairness and accountability:** trust in AI is built with clear practices. This means not only documenting the models and their uses, but also ensuring auditable and responsible outcomes identified at every stage of the lifecycle.

→ **Adopt frameworks such as NIST AI RMF or ISO/IEC 42001:** relying on international standards facilitates practice uniformity, enhances interoperability and provides legitimacy in the eyes of auditors and regulators.

→ **Train teams on AI risks and best practices:** governance cannot depend solely on regulations or formal processes. Understanding risk across the entire workforce and their active role in responsible AI management is key.

These recommendations constitute a practical roadmap for organisations to move towards robust governance, not only complying with regulations, but also generating trust and sustainable value in the use of AI.

AI Governance: Principles, Models and Operational Frameworks

04.

AI governance encompasses principles, rules, processes and controls that guide the design, development, deployment and oversight of this technology within organisations. Its purpose is twofold: on one hand, **to drive the economic and social benefits** that this technology provides; and on the other hand, **to contain the ethical, technical and legal risks** that may arise when algorithms and models influence decisions affecting people, services and public or private resources. This chapter provides a structured vision enabling an organisation to move from general ideas to daily practice. References to international standards are included, such as ISO/IEC 42001 for AI management systems, ISO/IEC 23894:2023 which provides specific guidance for managing AI risks in organisations of any sector, ISO/IEC TR 24368 on ethics and social concerns in AI, the IEEE 7000 series, the AI Risk Management Framework from the U.S. National Institute of Standards and Technology (hereinafter, "NIST AI RMF"), the Principles of the Organisation for Economic Cooperation and Development (hereinafter, "OECD") and the RIA.

Fundamental Principles of AI Governance 4.1.

The basis of AI governance rests on a set of fundamental principles designed to guide the design, development and use of intelligent systems. Various international organisations, such as the OECD, UNESCO, or the European Union, have proposed a series of converging values and ethical guidelines. These principles aim to ensure that AI is trustworthy, that is, it respects fundamental human rights, is technically robust, transparent, fair and subject to accountability:

01.

Transparency and Explainability

Clarity in algorithmic decision-making, understandable to human beings. Transparency also facilitates traceability in automated decisions and enables the identification of errors or biases. Explainable AI means providing meaningful information, tailored to the context, that identifies the data sources it relies on and shows the criteria and logic used by the model.

02.

Fairness and Non-Discrimination

Promote social justice and inclusion, avoiding the reinforcement of biases or inequalities. Systems shall not treat unfairly based on race, gender, socioeconomic background, disability, or other categories protected by law. Additionally, organisations must be able to detect and mitigate algorithmic biases, ensuring that the benefits of AI are accessible to all sectors of society. This principle is linked to the idea of human-centered AI.

03.

Privacy and Data Protection

The use of large volumes of data may include personal data, so it is necessary to safeguard the right to privacy by complying with existing data protection regulations (such as the aforementioned GDPR in previous chapters). Ethical handling of AI includes data minimisation, that is, using only the necessary data, anonymization when appropriate and conducting privacy impact assessments before deploying high-risk systems.

04.

Security, Robustness and Resilience

Systems must be robust and secure, behave as intended (reliability) and be resilient to failures, cyberattacks, or even malicious manipulations. Security risks must be anticipated and mitigated from the design phase. A robust AI system must also properly handle any erroneous or adverse data, maintaining its integrity and preventing undesirable or dangerous consequences. In critical environments (such as healthcare or essential and important entities), AI safety is essential to prevent physical or material harm.

Human Oversight and Control

Mechanisms for human intervention and control must be enabled, especially when facing decisions that may be high-risk, ensuring that people retain ultimate authority over such AI systems and can override or modify decisions when necessary. This principle guarantees human autonomy and ethical judgment over machine autonomy.

05.

Responsibility and Accountability

Every AI system must have clearly defined responsibilities throughout its entire lifecycle. Organisations must take responsibility for the outcomes of their systems, establishing mechanisms for review, audit and intervention in case of undesirable behavior. This requires conducting algorithmic impact assessments, external audits and complying with certification schemes. It is essential to document design decisions, maintain logs of model operations and undertake due diligence before adopting any system. If harm or bias occurs, there must be transparency about the causes and corrective actions must be taken to mitigate it. Additionally, accessible channels should exist for claims and for correcting errors detected by users or authorities.

06.

Sustainability and Social Wellbeing

It is essential to measure and manage the environmental and social impact of AI, including the carbon footprint of the models and their effects on employment and social cohesion. Sustainability involves aligning AI development with sustainable development goals, minimising negative impacts on the environment. Furthermore, AI must make a positive contribution to social welfare, improving services such as health, education and accessibility and not exacerbating issues such as misinformation or polarization.

07.

Governance Models for AI

4.2.

When translating principles into concrete mechanisms, various AI governance models have emerged, with countries and entities addressing the challenge of not hindering innovation. Among these models, three stand out that prioritise different aspects, depending on whether binding legal norms or voluntary guidelines are favored.

Risk-Based Governance Model

This is the predominant model in Europe and is aligned with the AIA. It should be noted, however, that the AIA does not expressly regulate a governance model or specific professional roles. Nonetheless, there are national and local laws that may result in such regulations, so this should be considered when defining the most appropriate model for each organisation.

Risk-based models focus on proactively identifying, assessing and mitigating risks, applying safeguards proportional to the level of risk each AI system poses to people or society. AI systems are classified by **categories or taxonomies of risk** (for example, “unacceptable risk”, “high risk”, “limited risk”, or “minimal”), with different requirements for each category. The AIA adopts precisely this pyramid scheme by prohibiting inadmissible risk systems (for example, widespread social scoring systems) and proportionally regulating the rest (with high-risk systems, such as AI for hiring, credit evaluation, or medical diagnostics, being subject to very strict requirements on transparency, traceability, data management, etc.); finally, medium- or low-risk systems have simpler obligations, such as adhering to codes of conduct or providing transparency to users. The objective is to foster trust in AI by ensuring that controls are appropriate to the potential for harm, without imposing unnecessary burdens on low-risk systems.

Rights-Based Governance Model

Rights-based approach, guaranteeing the development and use of AI does not violate fundamental rights (privacy, non-discrimination, freedom of expression, due process, etc.), while also promoting participation of people and communities potentially affected by AI in regulatory discussions.

In essence, it shifts the question from **"what risks does this technology have?"** to **"how is AI impacting the rights and dignity of people?"**

This approach holds that certain values are non-negotiable and cannot be graduated by risk level; for instance, if an AI application affects equality before the law or privacy, it must be addressed through positive obligations to protect those rights, rather than merely as a "risk to be mitigated". Organisations such as the Council of Europe and the UN have advocated for this vision, highlighting the need to avoid so-called "ethics washing" and move towards binding legal guarantees. An example is UNESCO's Recommendation on the Ethics of AI (adopted by 193 States in 2021), which defines itself as a human rights-centered approach.

In practice, this approach and the risk-based approach are not mutually exclusive. EU legislation combines a risk assessment logic with explicit requirements to protect rights, so we can say the AIA strikes a good balance.

Principle-Based **Approaches and Self-Regulatory Approaches ("Soft-Law")**

Approaches or models that are voluntary or non-binding, including various ethical guidelines developed by expert committees, professional associations, or by the technology companies themselves.

In 2019, the European Commission published the **Ethics Guidelines for Trustworthy AI**, defined by a group of experts, which listed seven voluntary requirements to guide companies in the self-certification of their AI systems, promoting practices such as ethical impact assessments and transparent communication with users.

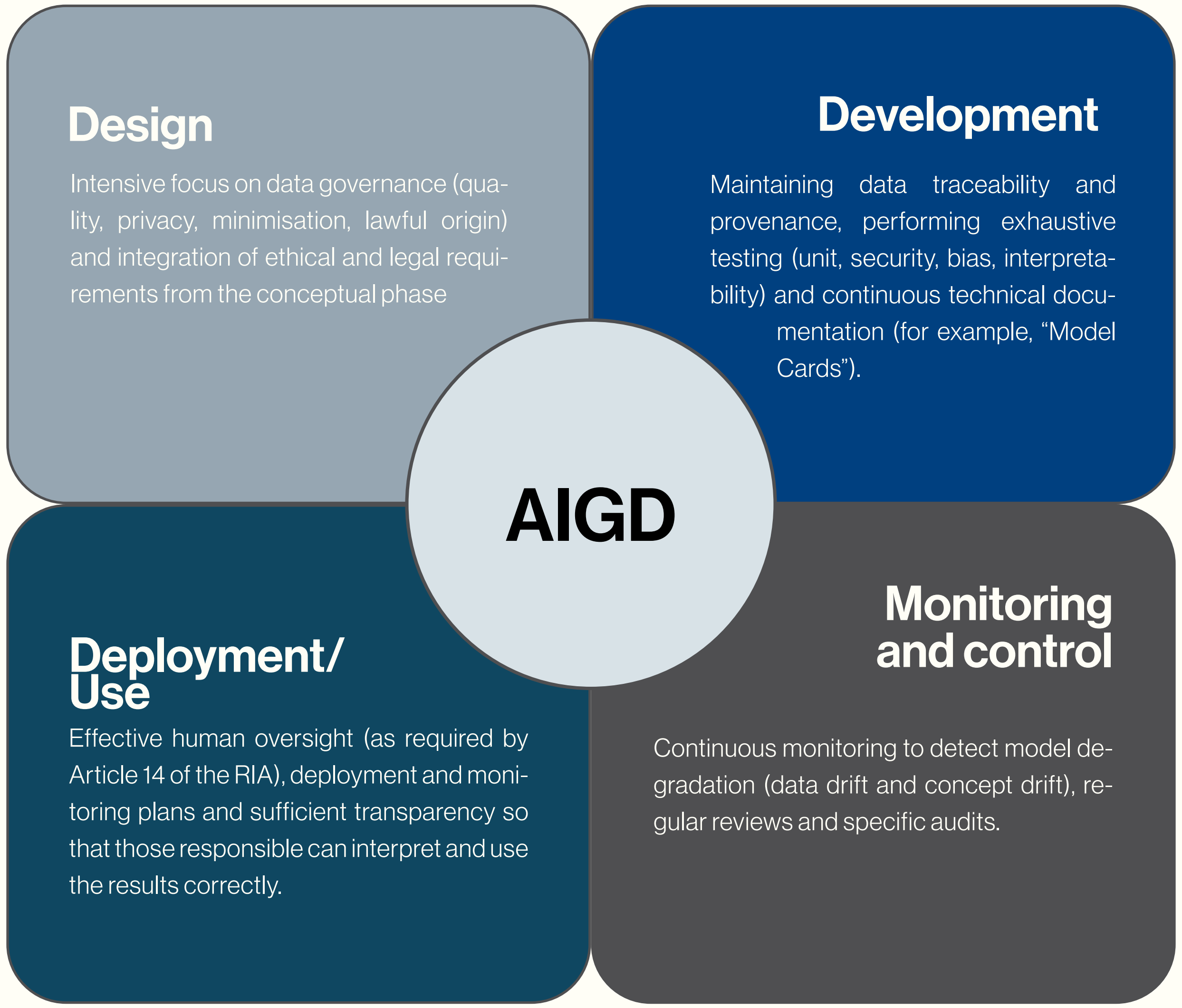
Many leading technology companies in their respective fields (Microsoft, IBM, Telefónica, Google, etc.) have published their own governance models, which include AI principles and practices and have also established internal responsible AI committees to oversee the systems they market.

Operational Frameworks of Governance 4.3.

AI Governance by Design (AIGD)

The most advanced trend in the field is the integration of governance from the very design of the systems, known as "AI Governance by Design" (AIGD). This approach means that ethical, legal and social considerations are not introduced afterwards, but are part of the AI lifecycle from inception, ensuring proactive and effective governance.

AIGD means that governance is not a later addition, but an essential component from the conception of the system. This translates into:



Design

Intensive focus on data governance (quality, privacy, minimisation, lawful origin) and integration of ethical and legal requirements from the conceptual phase

Development

Maintaining data traceability and provenance, performing exhaustive testing (unit, security, bias, interpretability) and continuous technical documentation (for example, "Model Cards").

Deployment/ Use

Effective human oversight (as required by Article 14 of the RIA), deployment and monitoring plans and sufficient transparency so that those responsible can interpret and use the results correctly.

Monitoring and control

Continuous monitoring to detect model degradation (data drift and concept drift), regular reviews and specific audits.

Governance Frameworks

Below is a review of the main standards, rules and frameworks, both international and national, that help to achieve a good AI governance framework, as well as recommended organisational practices for their adoption.

ISO/IEC 42001 (2023)

for AI Management Systems: sets the requirements for implementing an Artificial Intelligence Management System (AIMS) in organisations. Similar to quality management standards (ISO 9001) or information security (ISO 27001), ISO 42001 provides a structured framework based on the Deming cycle (PLAN, DO, CHECK, ACT) to incorporate governance in all phases of the AI lifecycle.

ISO 42001 is a certifiable standard that allows organisations to demonstrate, through an independent audit, that they manage their AI responsibly and according to a set of verifiable criteria. It lists a series of key requirements, the most relevant being to define the context and scope of AI use; implement ongoing risk management processes; carry out impact assessments before deployment; ensure data quality and traceability; control suppliers and third parties (such as cloud services or pre-trained models); continuously monitor and improve systems; and establish ethical design criteria and metrics for monitoring them. Essentially, ISO 42001 requires integrating AI into corporate governance with committed leadership, clear policies and responsibilities, training plans, thorough documentation and regular evaluation of both AI systems and the personnel managing them.

The EU itself foresees that compliance with ISO 42001 could facilitate conformity with the AIA, making its adoption within an organisation's governance framework highly recommended to anticipate legal requirements and facilitate compliance.

ISO/IEC TR 24368 (2022)

Focused exclusively on the ethical and social aspects of AI and the key principles to address them. This guide is advisory in nature, it does not prescribe specific values, but rather presents common considerations to help evaluators, technologists and regulators understand and mitigate the ethical challenges of AI.

The importance of ISO TR 24368 lies in that it synthesises the existing global consensus on AI ethical risks and provides a structured and inclusive approach, proactively addressing issues such as algorithmic bias, the opacity of deep learning models, or possible privacy violations by systems analyzing large sets of personal data. It also promotes the participation of different disciplines and stakeholders (business, engineers, ethicists, jurists, sociologists, etc.) in the evaluation of AI systems, so that results are fair, transparent and responsible.

In practice, organisations can use ISO TR 24368 as a reference guide to draft their ethical AI codes or when conducting ethical impact assessments (EIA) of AI projects. In fact, UNESCO has developed an Ethical Impact Assessment tool[6] inspired by similar principles, to help project teams identify the social and ethical impacts of an AI system before making it available to end users.

NIST AI RMF from NIST (National Institute of Standards and Technology of the USA)[7]

Structured guidance to identify, measure, mitigate and monitor risks throughout the lifecycle of AI systems. This framework is organised into four main functions: Map (context, risk mapping), Measure (risk analysis and evaluation), Manage (response and mitigation) and Govern (overarching function of continuous oversight).

It also defines the characteristics of trustworthy AI to be pursued, including system validity (working according to its purpose), safety, resilience, responsibility, explainability, privacy and absence of biases. It is a framework aligned with other internationally accepted frameworks and standards (ISO/IEC 5338, ISO/IEC 38507, ISO/IEC 22989, ISO/IEC 24028, ISO/IEC DIS 42001 and ISO/IEC NP 42005) and is very similar to the EU's approach to addressing risks. It starts from the premise that managing the incremental risk of each AI application is the way to avoid harms, instead of prohibiting the technology.

[6] UNESCO. (2023). Ethical impact assessment. https://unesdoc.unesco.org/ark:/48223/pf0000386276

[7] National Institute of Standards and Technology (NIST). (2023). AI Risk Management Framework (AI RMF 1.0). https://airc.nist.gov/airmf-resources/airmf

ISO/IEC 23894:2023 (Artificial Intelligence - Guidance on risk management)[8]

It provides specific guidance to manage AI risks in organisations from any sector. ISO 23894 aligns with the general principles of ISO 31000 (corporate risk management) but adapts them to the AI context, describing processes for identifying, assessing, treating and monitoring algorithmic risks throughout the entire lifecycle of an AI system.

For example, this standard places strong emphasis on considering emerging risks, such as adversarial model manipulation (adversarial attacks), lack of context in models (commonly known as hallucinations) in generative AI, or rights impacts as an integral part of the required risk analysis. This guide is also very clear and aims to be easy to understand for non-technical managers, facilitating organisational communication about AI risks.

IEEE (Institute of Electrical and Electronics Engineers)

In parallel to the ISO standards and initiatives promoted by public bodies such as NIST, IEEE (Institute of Electrical and Electronics Engineers) has promoted a fairly extensive repertoire of standards to help guide responsible development of autonomous and intelligent systems, as well as their governance. The IEEE 7000 series stands out, covering very specific guides on ethical aspects.

[8] ISO/IEC 23894:2023 – ISO (2023) "Artificial Intelligence - Guidance on risk management" https://www.iso.org/es/contents/data/standard/07/73/77304.html

Practical Implementation of AI Governance 5.5.

AI governance is not limited to regulatory frameworks or theoretical principles; its true value lies in the ability to turn them into operational practices within organisations. These frameworks offer general guidelines and structures, but the challenge lies in integrating them into internal processes and daily routines, which is sometimes referred to as corporate culture. In this way, we ensure that each AI project is developed, deployed and supervised in accordance with applicable technical, ethical and legal requirements.

To achieve practical implementation, organisations must undertake a series of tasks. First, it is essential to design **corporate AI policies**, covering from the **ideation phase** (including ***legal checks and ethical requirements***) to **production deployment** (with ***change controls*** and continuous oversight) and these are usually integrated with existing data management and information technology (IT) policies.

Second, the creation of **ethics or AI governance committees**. These committees, formed by management, legal experts, risk specialists, AI experts and sometimes external consultants, evaluate whether the projects **comply with organisational values and current regulations**, being able to recommend adjustments or veto initiatives.

Finally, conducting **algorithmic impact assessments (AIA)** is essential, especially in AI systems that can significantly affect people. Inspired by the privacy impact assessments (DPIA) required under the GDPR and required by ISO 42001 for critical systems, the AIA allows analysis of the **possible consequences for rights and values**, such as **discrimination**, **opacity**, **biases** or the **privacy** of people and their data and allows identification of mitigation measures. Tools like Canada's Algorithmic Impact Assessment[9] or the AI Impact Assessment proposed by the EU and required under Article 27 of the RIA exemplify this approach, complementing technical assessments with a deep analysis of the social, ethical and legal impacts of AI.

[9] Government of Canada. (2024). Algorithmic Impact Assessment. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Data Governance and Management in AI

05.

In his work American Ideals, Theodore Roosevelt said ***"Nothing in this world is worth having or worth doing unless it means effort, pain, difficulty."***

The use of AI for our strategic objectives is no exception.

It is true that the siren songs we often hear (and, worse, many managers often hear) do not mention too many complications, focusing instead on the last mile: the development of an analytical model that will solve our problems. That is the part where the athlete reaches the finish line and takes the glory. What is usually omitted is all the effort involved.

Fortunately, those responsible for data in organisations know well that getting value from data, consistently, requires a series of tasks that are not so easy to explain in a PowerPoint.

And that is what **Governance** means: aligning people and technology around the process of generating value from data, with the goal of optimizing that value. And for this, a strategy is necessary: a **Data Strategy.**

Moreover, the arrival of generative AI means that what was once the domain of a minority of highly trained developers is now easily accessible to any employee with some basic skills, including less experienced managers.

However, this is causing certain governance problems—previously limited to a controllable group—to spread throughout the organisation. There are organisations that seek to help address these problems effectively. A clear example is the CDO Club[10], which has designed an analytics maturity diagnostic tool called dataMat. It comes with a second tool (dataToolkit) that, based on the dataMat diagnosis, provides a system for prioritizing initiatives; that is, it designs your organisation's data strategy. Both tools consist of three layers: overall maturity diagnosis, data ethics maturity and generative AI maturity.

Indeed, in the course of these maturity analyses and especially in Data Ethics, a number of good practices for responsible AI governance are identified, some of which are summarised in the following points, grouped into five blocks: strategy, risk management, governance, model operation and culture.

Strategy and Ethical Framework

→ Make ethical aspects part of the company's strategy, including those related to responsible AI.

→ Design AI governance to be tailored to the criticality of the use case.

→ Have policies approved at the highest level, with defined roles, clear responsibilities and a suitable committee structure.

→ Have the ethical dilemmas associated with each business initiative identified.

Identification and Management of Ethical Risks

→ Have criteria for ethical risk tiering.

→ Include ethical risk in the company's risk catalog.

→ Have a measure of ethical risk appetite.

→ Carry out periodic ethical risk analyses, linked to the dilemmas in the catalog.

→ Have a register of ethical dilemmas, associated with each AI model or system.

→ Have KPIs for measuring ethical impact.

→ Have measurements of inherent risk and residual risk (after mitigation actions are applied).

[10] Club Chief Data Officer. www.clubcdo.com

Organisational Governance and Operations

→ Existence of specialised Model Governance teams, acting across the entire organisation.

→ Carry out independent external validations on the design and fulfillment of the data ethics strategy.

→ Extend ethical governance to both data suppliers and customers, insights and models.

→ Creation of internal independent validation teams (second line of defense).

→ Include aspects of responsible AI within whistleblower channels.

Lifecycle of Models and Data

→ Have a **model inventory**, with designated responsible parties and an associated risk level.

→ Know the **trace** in the informational processes that feed analytical models.

→ Ensure the **explainability of models**, both locally (individual level) and globally (entire model).

→ Have an analysis based on **ethical criteria** in model development, including bias analysis.

→ Having a **continuous monitoring** of model degradation.

→ Use corporate tools for **model, data and process control**.

→ Have **Near Real Time** processes for controlling bias in mass data loads, especially with sensitive groups.

→ Have a **conformity assessment** on all models.

→ Design model development processes as **Ethics by Design**.

→ Include **human oversight** in automated decision-making processes.

Culture, Training and Awareness

→ Have a **training and communication plan** to raise awareness of responsible AI at all participating levels and promote discussion forums.

→ Promote **internal discussion forums** about the ethical risks of business objectives.

In summary, integrating ethical and responsible processes into the development and management of artificial intelligence models is essential to ensure their transparency, reliability and alignment with organisational values. Continuous training, human oversight and measurement of ethical impact help build a strong culture of responsible AI, where risks are proactively managed and ethical dilemmas are handled with judgement and rigor. Only in this way can organisations move toward a sustainable and socially committed business strategy. And, of course, without a clear strategy sponsored at the highest level, none of this will be achieved.

AI as a Strategic and Operational Risk

06.

In recent decades, the field of Information Technologies has undergone various technological revolutions: the emergence of the Internet, mobile telephony, or cloud services, among others. Each of these has forced us to adopt new security measures and rethink our operational processes.

The current AI revolution follows the same path, but introduces additional dimensions that directly impact competitiveness, resilience, regulatory compliance and user and customer trust.

For example, AI is capable of issuing responses autonomously. This implies the risk that such responses may be incorrect, generating unwanted consequences at operational, reputational, or economic levels. As in a company offering prices through an automated assistant: if the AI provides an incorrect value, it could incur an economic loss or harm the corporate image.

It is important to emphasise that AI is a system that learns and evolves. Even with optimal training, there is no guarantee of absolute accuracy, as it can always produce errors.

In the documents **"Introduction to Artificial Intelligence for Information Security Professionals"**[11] and **"AI Governance Model"** the risks of AI to rights and freedoms are already addressed, including aspects such as ethics and legal responsibility.

This constitutes a differentiating factor compared to previous technological revolutions, in which the ethical dimension was rarely foregrounded.

For a deeper analysis of the risks associated with the use of AI and strategies to address them, it is recommended to read **"Artificial Intelligence and Cybersecurity"[12]** published by ISMS Forum.

Finally, it is important to highlight that AI is not merely a technological matter, with approaches viewing it as another business component, no doubt with a strong dependency on data, among other issues.

[11] ISMS Forum. (2023). Introduction to Artificial Intelligence for Information Security Professionals. Spanish Association for the Promotion of Information Security. https://www.ismsforum.es/ficheros/descargas/isms-gt-ia---01-introl-a-la-ia1701173559.pdf

[12] ISMS Forum. (2024). Ethics and Compliance in the use of Artificial Intelligence. Spanish Association for the Promotion of Information Security. https://www.ismsforum.es/ficheros/descargas/isms-gt-ia-021707141605.pdf

Definition and Scope of AI Risk 6.1.

AI risk is the possibility of losses, damages, or breaches at any stage of the system lifecycle: design, development, deployment, operation, or retirement. This risk is cumulative and evolves along with technology, data and the environment.

When we talk about AI risk, it can take several major forms:

⟶ **Strategic risk:** impacts the value proposition, business model, reputation and compliance with corporate strategy. ***Example: inadequate use of generative AI can erode customer trust.***

⟶ **Operational risk:** affects processes, people, systems and suppliers, including information security, data protection, business continuity and cost control. ***Example: a failure in a customer service model may overwhelm channels and increase costs.***

Transversality: AI impacts, directly or indirectly, all areas of the organisation, from strategic to operational functions.

Evolving nature: this is a rapidly transforming technology, which means identified risks today may change, amplify, or even disappear in a matter of months.

Breadth of exposure: risk exposure is not limited only to AI models developed internally. It also encompasses the use of third-party solutions and even open or freely accessible models that are integrated into corporate processes.

⟶ **Technological risks:** those most associated with classical ICT risks, information security and/or cybersecurity.

AI risk management requires a comprehensive approach, considering its transversality, dynamism and variety of models used. Failure to manage it properly can impact both technology and business strategy.

6.2. **Strategic** Risk Map

These risks in the adoption and use of AI are not merely technical, but can affect corporate objectives, competitive market position and relationships with clients, regulators and other stakeholders. Some of the most significant risks:

01. Strategic **Misalignment**

Investing in AI initiatives without a clear business case or explicit alignment with strategic objectives measured through OKR (Objectives and Key Results[13]) or KPI (Key Performance Indicators[14]). This can lead to low-return projects, initiatives that never integrate into critical processes, or the proliferation of pilots that are never scaled—a situation commonly referred to as "pilot purgatory." This situation consumes resources, creates internal frustration and erodes AI credibility as a strategic lever.

02. Supplier Dependence and **Lock-In**

The use of closed models or APIs can severely condition the organisation's ability to change technology or provider in the future. This entails risks such as restriction of portability, unforeseen cost increases, or changes in terms of use, as well as excessive concentration on a single provider or region (single supplier, single region), which increases the likelihood of disruptions or unilateral decisions by the provider.

03. Regulatory Compliance and **Licensing**

Compliance assessment processes, technical documentation development and regular audits are basic requirements for high-risk systems. There is also the risk of using data or results with intellectual property restrictions, or trained with datasets under licenses that limit exploitation or are illegitimate at source, which could lead to sanctions or legal disputes.

04. Reputation **and Trust**

Brand image can be severely affected if AI generates false, inaccurate, or misleading content (for example, deepfakes or hallucinated responses) that is publicly attributed to the organisation. Lack of transparency about model functioning, absence of explainability, or lack of effective complaint mechanisms can weaken the confidence of customers, employees and strategic partners.

05. Competitive Advantage **and Time to Market**

The pace of AI adoption is a critical factor, potentially resulting in market share loss to more agile competitors in AI industrialization. However, adopting immature technologies too quickly also carries risks (cost overruns, underperformance, or redoing developments). The key is to synchronise implementation with technological maturity and the organisational capacity to absorb change.

06. Ethics and **Social Responsibility**

Algorithmic biases or discriminatory decision-making can directly impact clients, employees, or communities. Furthermore, using AI in activities that do not align with corporate values or with ethical and responsible AI can harm the organisation's public image and social commitment.

07. Continuity and **Resilience**

Dependency on AI systems involves taking on the risk of severe interruptions for reasons such as changes to external models in use, cloud service outages, or withdrawal of functionalities by unilateral provider decision. The resilience of critical processes will depend on the ability to anticipate and manage this type of situation.

[13] Panchadsaram, R. (n.d.). What is an OKR? OKR Meaning, Definition & Examples. What Matters. https://www.whatmatters.com/faqs/okr-meaning-definition-exampl

[14] KPI.org. (n.d.). KPI Basics. https://www.kpi.org/kpi-basics/

6.3. **Operational** Risk Map

The operational risks of AI are related to the daily operation of systems, the quality of the data feeding them, the security of their lifecycle, their interaction with people and processes, as well as the management of third parties involved. Below are detailed the main types of risk, each with a potential impact on service continuity, quality and security.

01. Data

AI depends heavily on the quality, integrity and provenance of the data. Problems such as incomplete, outdated, or unverifiable origin information can degrade model performance or generate wrong results.

In the field of privacy, risks include the use of personal data without a legitimate basis, identification of individuals from seemingly anonymous data, or leakage of sensitive information through prompts.

Data sovereignty adds another dimension: cross-border transfer and restrictions on the physical location of information (data residency) can create legal and compliance conflicts, especially in multinational settings.

02. Security **(AI Security)**

AI introduces specific attack vectors. Among these are model attacks such as prompt injection (injection of malicious instructions), jailbreaks (bypassing restrictions), data poisoning (introducing corrupted examples during training), or model extraction (theft of parameters or weights).

There are also risks in the technology supply chain, where external dependencies, model weights, containers, or libraries can be compromised.

API exposure is another critical front, with risks of abuse, bypass of consumption limits (rate limits), or inclusion of secrets in prompts that could be exfiltrated.

03. Model Risk

Models may lose validity due to phenomena such as data or concept drift, overfitting, or simply from lack of robustness under unseen conditions.

Poor governance, lacking clear documentation (for example, datasheets or model cards), traceability and version control, complicates management and auditing.

In sensitive cases, explainability is critical: opaque or non-interpretable models can prevent justifying decisions to regulators or customers.

04. Operations **(MLOps)**

Continuous integration and deployment (CI/CD/ML) pipelines without risk control points, or with insufficient testing, can result in defective models in production. There are also risks in deployments without "canary release" or fast rollback strategies.

In operation, monitoring must go beyond technical availability: metrics such as latency, cost/token, quality of responses, detection of toxicity, bias and plagiarism must be monitored to react and correct in time.

05. Third Parties and **Procurement**

The use of third-party AI solutions without a rigorous assessment of their security, compliance, service level agreements (SLAs) and audit procedures exposes the organisation to risks that may not be apparent until an incident occurs.

06. People and **Processes**

A common risk is excessive dependency on AI-generated responses (automation bias), which can lead to a gradual loss of human expertise and a decline in the critical ability to validate results.

There are also risks associated with prompt "ergonomics": inefficient or poorly designed interactions can reduce productivity.

At the organisational level, the introduction of AI requires adapting roles, training plans, incentives and accountability frameworks. If not managed clearly, this can lead to diffuse responsibility, where it is unclear who is responsible for decisions made with AI support.

07. Legal/ **Contractual**

The legal framework for AI is still evolving, but clear challenges already exist: limitations on contractual indemnities for generated results, assignment of liability for AI-created content and risks from terms of use that do not adequately cover incidents such as illicit material production or unauthorised data distribution, as well as causing harm to collectives using AI without respecting fundamental rights.

6.4. ICT and **Cybersecurity Risks**

The risk associated with AI can be defined as the possibility of losses or damages that generate an impact. Among the most widely recognised risks are:

DeepFakes

Creation of fake audiovisual content enabling identity impersonation.

Prompt Injection

Prompt manipulation to bypass restrictions or access sensitive information.

Sensitive Information Disclosure

Accidental or forced disclosure of confidential information.

Model Theft

Model extraction or improper replication of its operation.

Bias and Algorithmic Discrimination

Reproduction or amplification of biases present in data.

Training Data Poisoning

Altering training data that compromises performance or introduces biases.

Insecure Output Handling

Unintended inclusion of malicious code in model outputs.

Excessive Agency / Overreliance

Excessive delegation of functions or critical operational reliance on the model.

Intellectual Property Risk

Generation or training with materials subject to copyright.

Disinformation

Production of inaccurate or misleading information that can erode trust.

In summary, once the different risks associated with the use of AI are identified and understood, it becomes easier to define a structured process for their analysis, implementation and monitoring, as well as to determine suitable metrics for each use case. Specific indicators should also be established, along with specific KPIs and KRIs of different types linked to the uses of AI in the organisation.

Management and Assessment of AI Risks

07.

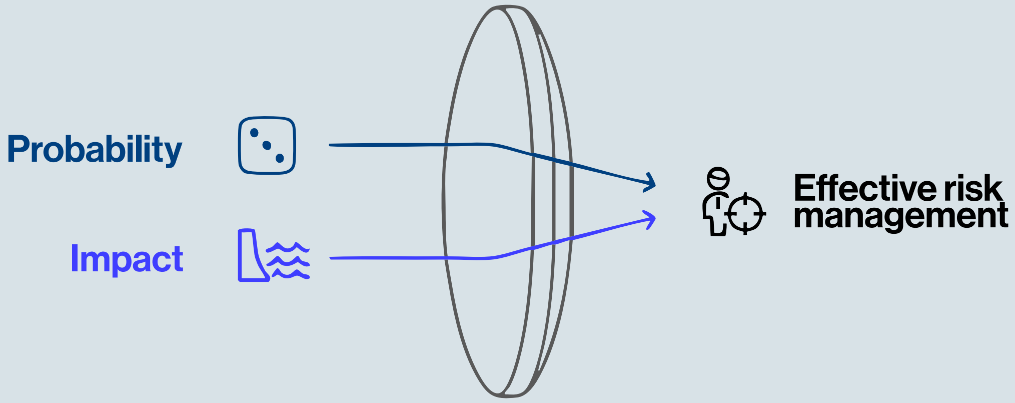
In order to understand how AI works and to be able to apply it effectively in our organisations, as well as support its implementation, it is necessary to understand which risks are present and how to manage them adequately.

While other chapters of this guide cover strategic, operational and more specific and emerging risks, such as algorithms, their implications for use, training and development, among others, this chapter aims to explain how they should be managed.

AI risk management is the process of: identifying, analysing, evaluating and treating the risks associated with AI technologies.

Before deciding how to address a risk (mitigate, accept, etc.), we need to quantify or qualify its severity.

Risk = Probability x Impact

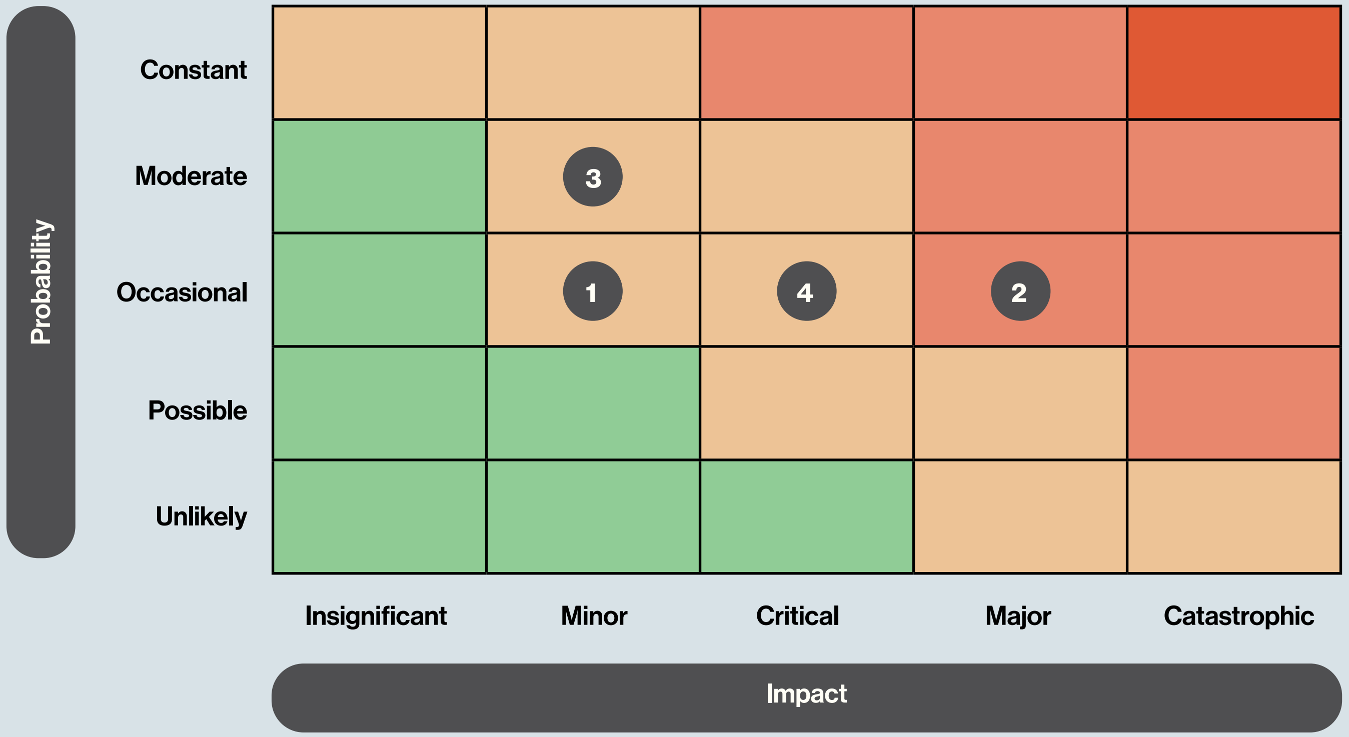


The methods for evaluating these risks are qualitative and quantitative analyses.

→ **Qualitative Analysis:** This is a subjective method, based on expert judgment, experience and predefined scales to assess:
- Probability
- Impact
- Risk level

→ **Quantitative Analysis:** This is an objective method, assigning real numeric values
- Probability (percentage)
- Impact (economic costs, time, etc.)
- Annualised risk exposure (Annualised Loss Expectancy, ALE)

Once risks are evaluated, we will have a system that allows us to understand the overall risk situation, identifying the most critical ones.



Probability	Insignificant	Minor	Critical	Major	Catastrophic
Constant	🟨	🟨	🟧	🟧	🟥
Moderate	🟩	🟨 3	🟨	🟧	🟧
Occasional	🟩	🟨 1	🟨 4	🟧 2	🟧
Possible	🟩	🟩	🟨	🟨	🟧
Unlikely	🟩	🟩	🟩	🟨	🟨

Once identified, we will begin mitigation tasks.

7.1. ISO **27005**

ISO 27005 provides a framework for information security risk management, detailing a systematic process for identifying, analyzing, evaluating and treating risks. This process, which uses a risk-based approach and complements ISO 27001, helps organisations efficiently protect their information assets and make informed security decisions. The process is cyclical and is based on the **PDCA (Plan, Do, Check, Act)** model to ensure continuous improvement.

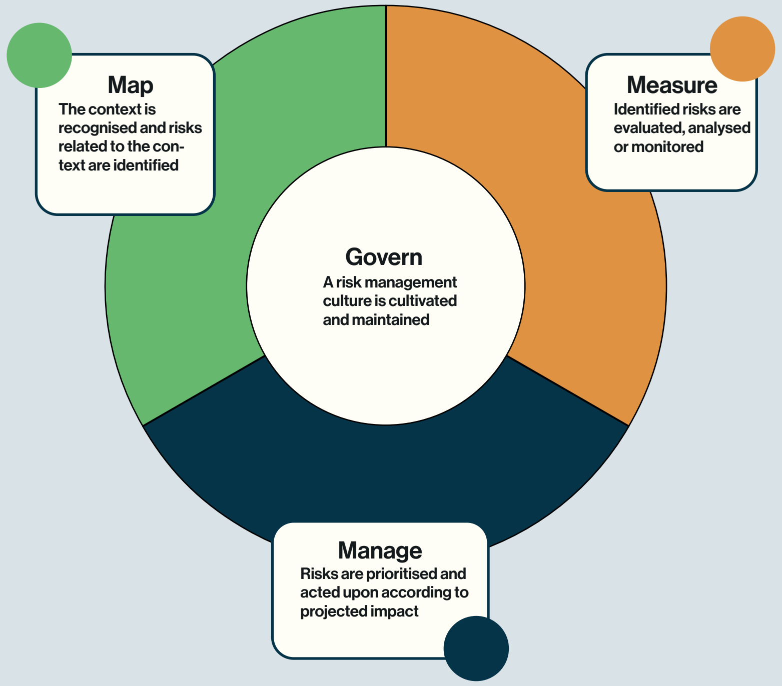
- **Establish the context:** define the objectives, criteria and scope of risk management, as well as the reference framework.
- **Identify the risks:** locate and describe potential risks. The standard proposes two approaches:
    1. *Event-based:* focuses on general scenarios and threats.
    2. Asset-based: focuses on specific information assets and associated vulnerabilities.
- **Analyse the risks:** determine the probability that the risk occurs and the impact it would have on the organisation if realised.
- **Evaluate the risks:** compare the analysed risks to the acceptance criteria defined in the context phase to prioritise them.
- **Treat the risks:** select and implement one or more options to mitigate, avoid, accept, or transfer risks that exceed the acceptable level.
- **Communicate and consult:** keep all stakeholders informed about the process and its outcomes.
- **Monitor and review:** continuously monitor the risks, threats and effectiveness of implemented controls, as well as periodically review the process to adapt it to changes.

7.2. **NIST AI 100-1** (AI RMF 1.0)

The NIST AI Risk Management Framework (AI RMF) is a voluntary guide published in January 2023 to help organisations manage AI risks. Its purpose is to promote responsible development and deployment of AI systems, focusing on characteristics such as reliability, fairness and transparency. The core of the framework is based on four functions: govern (create a risk management culture), map (identify and frame risks in business contexts), measure (analyse and evaluate risks) and manage (address identified risks).

Key Components of the NIST AI RMF

- **Objective:** provide a structured approach to identifying, assessing, managing and mitigating AI system risks to ensure they are trustworthy and safe.
- **Main functions:** the four main functions of the framework are:
  → Govern: establish a culture of AI risk management within the organisation.
  → Map: contextualise AI risks according to business operations and needs.
  → Measure: systematically analyse and evaluate risks.
  → Manage: take action to address risks that have been mapped and measured.
- **Applicability:** it is voluntary and designed for use by a wide range of organisations in various industries and geographies.
- **Benefits:** helps organisations leverage the benefits of AI while mitigating potential harms, ensures consumer trust in their personal data and promotes the continuous development of the technology.



Map
The context is recognised and risks related to the context are identified

Measure
Identified risks are evaluated, analysed or monitored

Govern
A risk management culture is cultivated and maintained

Manage
Risks are prioritised and acted upon according to projected impact

7.3. **Artificial Intelligence** Regulation

The RIA establishes a set of enforceable obligations for those who develop, implement, or use AI systems in the EU. Among these, the obligation to conduct a risk analysis, especially if the system falls into the high-risk category.
The RIA is based on the USE we make of AI, with the regulation defining its own set of criteria to categorise AI. The evidence that it is a risk-based regulation is seen in the fact that the word "risks" is mentioned 181 times.
The RIA proposal establishes a risk-based approach to regulating the use of AI systems. This classification determines which obligations and restrictions apply to each type of system.
Risk analysis is performed for High-Risk AIs.

Risk Level	Description and examples	Legal Treatment
Unacceptable Risk 🔴	Systems that contravene fundamental values	**Prohibited**
High Risk 🟠	They affect fundamental rights, health or safety.	**Strictly regulated**
Limited Risk 🟡	They may influence user decisions.	**Require transparency**
Minimal or No Risk 🟢	They pose no significant threat.	**Free use**

Risk analysis is carried out for High-risk AI systems.

Prohibited AI Practices
Article 5 of **RIA**

- AI systems that use subliminal techniques.
- AI systems that exploit any of the vulnerabilities of a natural person or a particular group of people.
- AI systems to evaluate or classify natural persons or groups of people.
- AI systems to perform risk assessments of natural persons committing a crime (to assess or predict).
- AI systems that create or expand facial recognition databases.
- AI systems to infer the emotions of a natural person in the workplace and in educational centers.
- Biometric categorization systems that individually classify natural persons.

High-Risk AI Practices
Annex III, Articles 6 and 9 of **RIA**

These systems are not prohibited but are subject to demanding obligations regarding design, oversight, documentation and above all, risk management.

Providers of high-risk AI systems must:

- Ensure that their high-risk AI systems comply with the requirements set out in **section 2.**
- Indicate in the high-risk AI system or, where not possible, on its packaging or accompanying documentation, as appropriate, its name, registered trade name or trademark and the address where they can be contacted.
- Have a quality management system that is in line with the provisions of **article 17.**
- Maintain the documentation referred to in **article 18.**
- When under their control, retain the automatically generated records from their high-risk AI systems referred to in **article 19.**
- Ensure that the high-risk AI system undergoes the appropriate conformity assessment procedure referred to in **article 43** before it is marketed or put into service.
- Draft an EU Declaration of Conformity in accordance with **article 47.**
- Affix the CE marking on the high-risk AI system, or, where this is not possible, on its packaging or accompanying documentation, to indicate its compliance with this Regulation, in accordance with **article 48.**
- Comply with the registration obligations referred to in **paragraph 1 of article 49.**
- Take the necessary corrective measures and provide the information required in **article 20.**
- Upon a reasoned request from a competent national authority, demonstrate the compliance of the high-risk AI system with the requirements set out in **section 2.**
- Ensure that the high-risk AI system complies with accessibility requirements in accordance with **Directives (EU) 2016/2102** and **(EU) 2019/882**.
- An accountability framework establishing the responsibilities of management and other staff with respect to all the aspects listed in this section.

Limited Risk AI Practices
Articles 50 and 52 of **RIA**

The obligation is to inform the user that they are interacting with an AI.

7.4. **The Risk Management** System of RIA

The AIA transforms risk management into a structured and hierarchical legal obligation:

- **High-risk systems (HRS)**

A set of mandatory requirements for HRS is determined. These include **risk management** throughout the system's lifecycle, **data governance** (quality and representativeness), technical documentation, **traceability** and **human oversight**.

- **Systemic risk model**

Legislation has gone beyond high-risk systems by classifying **general-purpose AI models (GPM or GPAI)** in two levels: ordinary and **systemic risk**. This distinction, based on factors such as the number of parameters, the volume of training data or computing capacity, adds a complex regulatory layer that was not fully anticipated in early 2023 discussions, which were more focused on traditional AI systems.

Therefore, the risk management system applies to high-risk systems and is defined in Article 9 of the AIA

1. **A risk management system will be established, implemented, documented and maintained in relation to high-risk AI systems.**
2. **The risk management system will be understood as a continuous iterative process planned and executed throughout the lifecycle of a high-risk AI system, requiring systematic regular reviews and updates. It will consist of the following stages:**

    a. The identification and analysis of known and foreseeable risks that the high-risk AI system may pose to health, safety, or fundamental rights when the system is used as intended;

    b. The estimation and assessment of the risks that could arise when the high-risk AI system is used as intended and when it is reasonably foreseeable to be misused;

    c. The assessment of other risks that could arise, based on the analysis of the data collected with the post-market monitoring system referred to in Article 72;

    d. The adoption of appropriate and specific risk management measures designed to address the risks identified in point a).

3. **The risks referred to in this Article are only those that can be reasonably mitigated or eliminated through the development or design of the high-risk AI system or the provision of appropriate technical information.** When determining the most appropriate risk management measures, the following shall be sought:

    a. Eliminate or reduce the identified and assessed risks in accordance with paragraph 2 as far as technically feasible by proper design and development of the high-risk AI system;

    a. Implement, where appropriate, suitable mitigation and control measures to address risks that cannot be eliminated;

    a. Provide the information required under Article 13 and, where appropriate, provide training to those responsible for deployment.

    With a view to eliminating or reducing the risks associated with the use of the high-risk AI system, due consideration will be given to the technical knowledge, experience, education and training expected of the responsible deployer, as well as the context in which the system is expected to be used.

4. **High-risk AI systems shall undergo testing to determine which risk management measures are most appropriate and specific.**

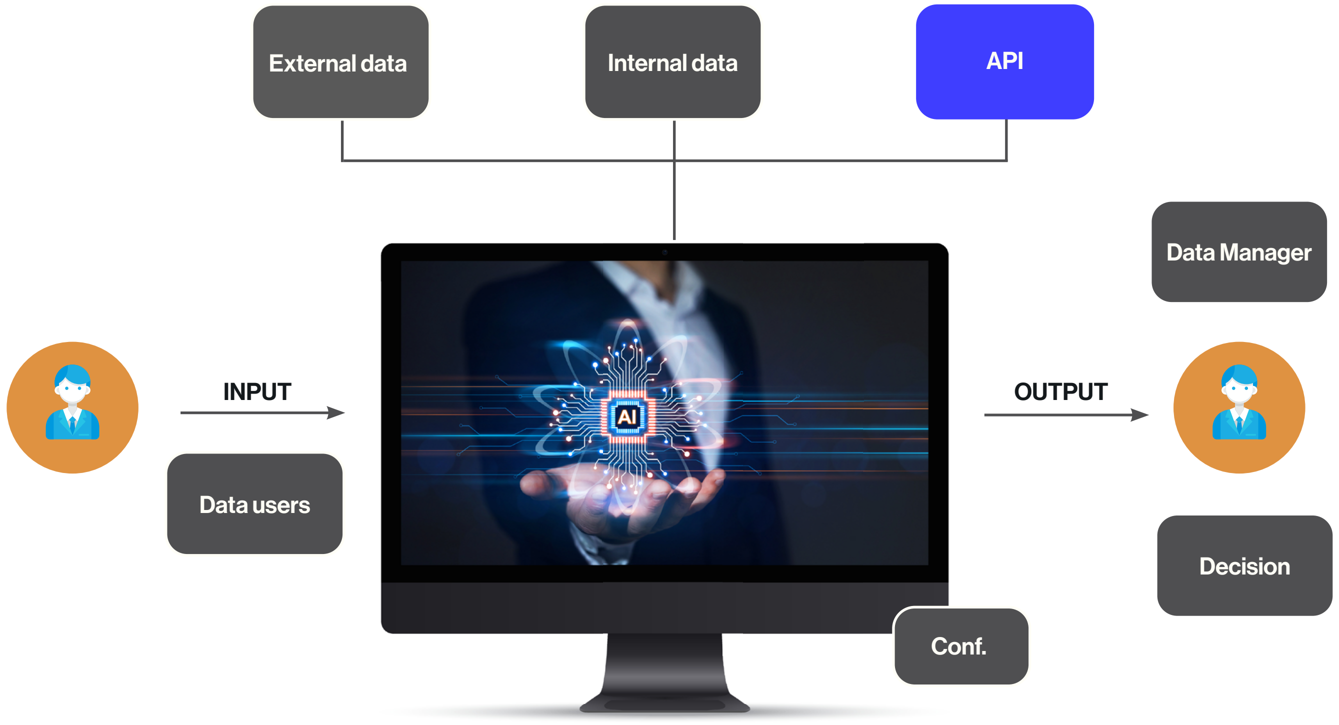
    Translation into risk language:

    - Identify the AI system.
    - Identification of potential risks.
    - Assessment of each risk: probability and impact.
    - Define and implement mitigation measures.
    - Monitor and review periodically.
    - Document the entire management process.
    - Conduct continuous testing.

7.5. Designing **the AI Model**

This section is about structuring the AI model in order to properly analyse the risks, fitting the solution within the template, so that it will be much easier to identify the risks.

Being a template, it does not fit all AI models, so it can be updated according to the model.



Architecture/Infrastructure

The elements that make up this simple model are:

- **The input data/actions.**
- **The output data/actions.**
- **The training data.**
- **The operating data.**
- **The configuration data.**

In this simple way we can characterise the model.

Once we have the details of our AI solution within the model, we apply the main risks according to a previously defined taxonomy. Currently there are several proposals.

In summary, the RIA is based on risk management and its analysis in systems of high criticality. For this, it is necessary to have adequate knowledge of risk management in order to conduct analyses correctly and to be able to apply mitigation measures.

Currently there are multiple risk analysis methodologies, derived from security and cybersecurity, so these are the most commonly used, although some that are specific to AI are already appearing.

Data Protection and Privacy in AI

08.

AI has posed great challenges in the process of innovation and digital transformation from different spheres at the social, cultural and economic levels. Without a doubt, one of these challenges is creating an ecosystem that allows its integration while guaranteeing the protection of a fundamental right such as data protection, given that the development and use of AI systems may involve the processing of personal data.

The processing of personal data may occur at different points in the lifecycle of the AI system and with different functionalities, therefore, from the perspective of data protection, the development of AI presents major challenges to be analysed:

01. Compliance with the Data Minimisation Principle

Among the principles of the GDPR, already mentioned in previous chapters from an AI perspective, we must address the data minimisation principle, which is set out in Article 5(1)(c).

According to the GDPR itself, this principle is based on assessing and establishing an analysis that restricts the collection and processing of data to those strictly necessary for the intended purpose, requiring a proportionality assessment of the processing. As mentioned earlier, AI systems will often require the processing of personal data[15].

[15] Cotino Hueso, L., & Simón Castellano, P. (Eds.). (2024). Treatise on the European Union Artificial Intelligence Regulation. Aranzadi. https://roderic.uv.es/items/cec02ec1-44a4-41b6-9df7-de3ed7d6a223

Therefore, prior to processing it is necessary to perform an analysis in which the following issues should be considered:

- → **The processing** is suitable and appropriate for the intended purpose.
- → **Relevance**: the data are directly related to the purpose.
- → **Limitation of processing:** there is no collection of irrelevant and unnecessary personal data for the purpose.

Thus, in the case of AI systems that process personal data, a prior study must be carried out to assess the proportionality and necessity of processing, based on the following considerations:

- → Seek to limit the categories of data used, as well as their degree of accuracy to what is actually needed for processing.
- → Limit the volume of data and the subjects whose data are processed.
- → Establish mechanisms to limit access to the different categories of data.

To this end, the AEPD reminds us that there are various techniques to enable data minimisation, including anonymization and pseudonymisation [16], as well as the generation of synthetic data (which must not contain identifiable information) and also others such as suppression of irrelevant conclusions associated with personal information during the training process, for example in the case of unsupervised training.

[16] Court of Justice of the European Union. (2025, September 4). Judgment in case C-413/23 P: EDPS v. SRB (Concept of personal data and pseudonymisation). https://curia.europa.eu/jcms/upload/docs/application/pdf/2025-09/cp250107en.pdf
Spanish Data Protection Agency (AEPD). (2020). GDPR compliance of processing incorporating Artificial Intelligence. https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf

02. Transparency and **Explainability**

The principle of transparency is a point of convergence between the GDPR and the RIA, establishing harmonised rules for AI (hereinafter, RIA), as they address this principle from different but complementary perspectives.

Thus, in the case of the GDPR framework, the purpose of the transparency principle is to provide individuals with information that allows them to understand the processing of their data and give them greater autonomy in decision-making. It is established that processing must be transparent (art. 5. 1st) so that the right of data subjects to be properly informed about the collection of their data, its use and its retention is guaranteed in a clear, accessible and understandable manner, but, in addition, art. 22 regulates the right not to be subject to decisions based solely on automated processing that may produce significant legal effects.

However, the duty of transparency in the RIA is viewed from a technical and organisational obligation perspective, to ensure that users understand when they are interacting with AI systems and how results are generated and to achieve trustworthiness, security and accountability.

However, the technical opacity of certain models, as well as the protection of trade secrets covering algorithms, makes it difficult to fulfill the information and transparency rights. Therefore and in order to facilitate its management, measures can be established to combine the information obligations of the GDPR with the reinforcements established by the AIA, within the obligations set out in that body of law, which establish obligations arising from the duty of information and the principle of transparency:

→ **AI identification**. People must be able to tell when they are interacting with an AI system.

→ **Labeling synthetic content.**

→ **Explainability and traceability.** High-risk systems must be designed so that their operation allows users to interpret and use them appropriately. This information may include technical documentation regarding instructions, limitations and risks.

In this regard and in relation to compliance with transparency, information and explainability obligations, the interpretation issued by Judgment of the Third Section of the Contentious-Administrative Chamber of the Supreme Court of September 11, 2025[17] is of interest.

[17] Based on articles 14 and 16 of Law 19/2013 on transparency, access to information and good governance, access to the source code of the BOSCO computer application, launched by the Ministry for Ecological Transition to evaluate applicants for the social electricity bonus to be considered vulnerable consumers, is supported (however, it should be noted that the BOSCO system does not introduce its own criteria but mechanises operations to apply criteria without making autonomous decisions). This access is weighed considering the duty of transparency, given that it is an administrative action by a foundation defending vulnerable groups' rights.

03. Avoidance of Bias and **Discrimination**

One of the challenges faced in the design and training of AI systems with impacts on data protection is the avoidance of bias, but to do this we must understand what types of bias may occur:

→ **Dataset bias.** This risk occurs when the data used for training AI systems do not reflect accurate real-world data, such as sampling or measurement bias.

→ **Personal biases**, as the biases of the developers may influence the model's results by inadvertently introducing their own biases and perceptions.

→ **Algorithmic bias.** When, through design, certain characteristics are prioritised over others, resulting in unfair conclusions.

Regarding the treatment and mitigation of bias, it should be noted that the European Commission's expert group developed Guidelines for Trustworthy AI[18], which emphasised the need for AI systems to be based on the commitment to use them in the service of humanity and the common good, avoiding unfair bias, as it could have multiple negative consequences, from the marginalization of vulnerable groups to the exacerbation of prejudice and discrimination.

To prevent the appearance of biases, we must address their prevention and establish mechanisms from a holistic approach at different stages of the process:

[18] https://digital-strategy.ec.europa.eu/es/library/ethics-guidelines-trustworthy-ai

Design and Planning:

At this stage, clear objectives of the system and the groups that may be affected should be established, identifying sensitive variables such as gender or ethnicity and analyzing possible discriminatory impacts.

System files should be prepared to include information about the intended use, risks and limitations. As well as enabling and designing mechanisms that allow us to manage the rights of data subjects in relation to data protection beyond the right of access and transparency, by setting up procedures to address rights such as deletion, objection, or not being subject to automated decisions.

01.

Data

At the processing stage, anonymization and data minimisation techniques will be applied, based on the aforementioned criteria and documenting the data sets: origin, purpose, limitations, etc.

02.

Modeling and Training

The selection of algorithms will be made according to criteria that allow auditability and explainability, in order to comply with the principle of transparency.

Notwithstanding, design decisions and assessments used in this selection should be documented to allow traceability of adopted decisions and facilitate the identification of errors.

03.

Evaluation and Testing

Prior to implementation, conduct internal bias audits and establish testing environments to evaluate the system. For this, it is necessary to establish metrics that allow results to be reviewed. In this regard, it is useful to review the ISMS guide which includes the most commonly used metrics as evaluation and validation techniques[19]. Subsequently, the results of audits and identified limitations will be documented.

04.

Implementation and Deployment

At the implementation stage, mechanisms for human oversight will be established, including training and education of system users, as well as mechanisms and processes to help correct or modify results.

05.

Maintenance and Monitoring

User feedback and criteria will be included to enable corrections, as well as periodic bias and performance audits to evaluate the system.

06.

[19] ISMS Forum. (n.d.). Introduction to artificial intelligence. isms-gt-ia---01-introl-a-la-ia1701173559.pdf

04. **Integrated Impact** Assessments

As part of the obligations arising from proactive responsibility and risk management resulting from the use of AI systems, we must consider impact assessments relating to data protection **(DPIA)** as required by **Article 35 of the GDPR** and the **fundamental rights impact assessment under the RIA**. Both assessments are distinct pieces that converge and complement each other within the framework of AI governance in organisations, as recognised by the AIA itself.

Thus, **the DPIA focuses on risks related to the protection of personal data**. Fundamental rights impact assessments (known as FRIA) focus on analyzing non-discrimination, freedom of expression, security and health among others, when high-risk systems are deployed.

DPIAs under Article 35 of the GDPR must be carried out when data processing poses a high risk to the rights and freedoms of individuals. Therefore, DPIAs must be conducted when several of these factors are met:

→ **Innovative processing or use of new technologies**

→ **Automated decisions with significant effects**

→ **Combining data sets in ways unexpected by the user**

→ **Evaluation of people to assess behavior, productivity, or credit risk**

→ **Processing of data of vulnerable individuals.**

For the preparation of DPIAs, it is especially relevant to review the Guidelines of the Article 29 Working Party of the European Commission and the guide prepared by the AEPD, which provides a list of types of processing requiring DPIA[20].

Minimum content. The DPIA must include:

→ Description of processing operations and purposes.

→ Assessment of necessity and proportionality.

→ Risk assessment.

→ The measures and safeguards foreseen to prevent the occurrence of the risks or to establish their mitigation. For this, it is possible to seek the advice of the DPO (if any) and to consult with the authority in advance if, after the measures, a high residual risk persists.

[20] Spanish Data Protection Agency (AEPD). (2019). Evaluate-GDPR Risk. https://www.aepd.es/documento/wp248rev01-es.pdf

Without prejudice to the above, **Article 27 of the AIA** includes the obligation to carry out an impact assessment concerning fundamental rights before deploying certain high-risk AI systems, defining the scope of application for deployers that are public bodies or private entities providing public services (for high-risk systems under **Article 6.2, except Annex III, point 2** for critical infrastructures) and for deployers of systems in **Annex III, point 5, letters b) and c)** (related to creditworthiness/score and life and health insurance).

Likewise, the provision stipulates that when a DPIA already exists, the FRIA should complement it so that synergies between both assessments can be used.

Therefore, the coexistence of the Data Protection Impact Assessment (DPIA) detailed in Article 35 of the GDPR and the fundamental rights impact assessments under the AIA means organisations will have to develop joint methodologies to optimise and efficiently identify, manage and monitor risk, establishing a more coherent supervision and control process.

This obligation to prepare FRIA, which will take effect on August 2, 2026, requires that the minimum content to be included in its preparation will contain:

→ Description of the process in which the system will be used and its alignment with the intended purpose.

→ Categories of persons and groups affected, with particular consideration to vulnerable groups.

→ Specific risks of harm for the groups affected.

→ Time window and frequency of use.

→ Mitigation measures in the event that internal governance risks materialise, together with the definition of complaint and redress mechanisms, which should incorporate procedures to document traceability and monitoring, and to record any changes or updates as they occur. To this end, it is advisable to establish methodologies that enable impacts to be assessed in a parameterised manner..

→ Notification to the supervisory authority. Upon completion of the FRIA, the result must be notified to the supervisory authority. It is relevant to mention the methodology promoted by the Council of Europe for conducting the FRIA, called "HUDERIA[21]".

→ Human oversight measures to be applied, insofar as it is necessary to define how such oversight will be carried out at the different stages of the system's life cycle (before, during and after use), in a way that integrates with technical measures for robustness, data quality and explainability.

Therefore, in conclusion, it should be highlighted that data protection should be understood as a strategic value that allows its integration into organisational culture, identifying the protection of personal data as an extension of human identity and the appropriate protection thereof.

Protecting data is not only about complying with legal requirements, but also about the need to integrate privacy by design and by default to respect human dignity.

Conceptually, this vision has been expressed in international ethical frameworks such as the **UNESCO Recommendation on AI Ethics**[22] or the **OECD Guidelines** which highlight privacy, fairness and accountability as pillars of trustworthy AI.

For this and for the integration of this principle in a conceptual and cross-cutting manner within organisations, ***we must develop training and awareness programs, internal policies and committees*** to facilitate this transformation process and to protect the individual and their data.

[21] Council of Europe. (n.d.). HUDERIA – Risk and impact assessment of AI systems. https://www.coe.int/en/web/artificial-intelligence/huderia-risk-and-impact-assessment-of-ai-systems

[22] Recommendation on the Ethics of Artificial Intelligence | UNESCO https://www.unesco.org/es/articles/recomendacion-sobre-la-etica-de-la-inteligencia-artificial

Regulatory Compliance in AI 09.

The regulatory context has undergone a radical transformation, moving from an initial focus on publishing ethical guidelines (such as those from UNESCO or the OECD) to the **imposition of regulations with the force of law.**

AI systems have become fundamental drivers of economic, social and political transformation. From recommendation algorithms on social networks and facial recognition applications to automated decision-making systems in sectors such as healthcare, employment, or finance, AI is redefining the boundaries of what is possible. However, this accelerated expansion has brought with it growing **public, institutional and governmental concern about the ethical, legal and social risks associated with its use.** In this context, regulatory compliance in AI has become an essential pillar to ensure trust, security and legitimacy in the contemporary technological ecosystem. Its objective is to ensure that the development, implementation and use of AI systems are carried out in accordance with current legal provisions, internationally recognised ethical principles and technical standards of quality, safety and transparency.

The first reason for the growing importance of regulatory compliance in AI is the **expansion of regulatory pressure at the global level.** As governments recognise the transformative—and also disruptive—potential of AI, national and international institutions are adopting increasingly strict legal frameworks for its responsible development and use.

The most paradigmatic example of this new context is the aforementioned RIA. This pioneering regulatory instrument worldwide establishes a classification of AI systems according to their risk level (minimal, limited, high, or prohibited) and imposes specific obligations for transparency, security, traceability and human oversight. In this way, the EU aims to ensure that high-risk systems—such as those used in critical infrastructure, personnel selection, or medical diagnosis—are developed and used in compliance with specific requirements.

But this regulation has not been the only one; there are many regulatory initiatives and guidelines related to AI.

Thus, countries such as the United States, Canada, Japan, China, Brazil, Peru, Italy and the United Kingdom are developing their own regulatory frameworks or best practice guidelines to control the impacts of AI. The OECD, the Council of Europe and the United Nations are also promoting common principles of algorithmic governance, focusing on transparency, fairness, accountability and respect for human rights. This phenomenon has created increasing regulatory pressure, whereby organisations – both public and private – must adapt simultaneously to multiple legal frameworks, technical standards and social expectations.

The current situation is analogous to what technology companies experienced with the entry into force of the GDPR in 2018. At that time, the privacy of personal data became a strategic requirement, not just a legal one. Today, with AI, we are witnessing a similar process: **entities that do not integrate specific AI compliance programs are exposed to financial penalties, reputational loss and operational restrictions.**

The use of AI in the automation of decisions that directly affect the rights and opportunities of people – such as hiring, granting credit, or police surveillance – has demonstrated the need for control and accountability mechanisms.

Among the main **concerns** are:

→ The **opacity of algorithms:** due to their operation as a "black box", with internal processes that are difficult even for their own developers to explain. This lack of transparency undermines public trust and makes it difficult to assign responsibilities in case of errors or discrimination.

→ **Bias and algorithmic discrimination:** algorithms learn from the data they are trained on and if these contain social, historical, or demographic prejudices, AI can replicate or amplify them. There are numerous documented cases of systems displaying gender, race, or age bias in selection or evaluation processes.

→ **Information manipulation and disinformation:** with the rise of generative models that produce synthetic texts, images, or videos, new risks have emerged in the political, media and cultural spheres. The improper use of these technologies can affect the integrity of democratic processes or harm the reputation of individuals and institutions.

→ **Loss of human control and machine autonomy:** the increasing complexity of AI systems raises questions about oversight, correction and human control to reverse automated decisions.

These concerns have generated a social demand for accountability, requiring AI developers, providers and users to guarantee respect for fundamental values: human dignity, non-discrimination, justice, privacy and transparency. Consequently, **compliance programs are becoming essential tools to prevent risks, demonstrate due diligence and build trust among citizens, regulators and investors.**

An AI compliance program does not merely aim to formally comply with the law; it means designing an organisational and technical architecture that integrates ethics, governance and accountability throughout the entire AI system lifecycle.

In practice, **these programs must include,** among others:

01. **Mapping and classification of AI systems according to their level of risk,** in accordance with current regulations.

02. **Algorithmic impact assessments** that identify potential adverse effects on fundamental rights.

03. **Transparency and explainability protocols** that make it possible to understand how and why a system makes certain decisions.

04. **Effective human oversight mechanisms** to ensure that the final decision remains controllable.

05. **Internal reporting and audit channels** to detect and correct non-compliance before they cause harm or penalties.

06. **Ongoing staff training** and the creation of ethics or algorithmic governance committees.

AI compliance must also be understood as a competitive advantage, enabling the reduction of exposure to legal risks and strengthening an organisation's reputation and credibility with clients, investors and public authorities. In a market increasingly aware of the social impact of technology, trust becomes a strategic asset.

That is why the real challenge does not lie solely in creating compliance documents or isolated protocols, but in fostering an organisational culture based on responsibility. Ethics and compliance must be part of the DNA of entities developing or using AI, integrating them from the design phase (known as ethics by design or compliance by design).

In this regard, AI compliance must be conceived as a dynamic and cross-cutting process, evolving at the same pace as technology and regulation. It is not enough to comply with current regulations; it is necessary to anticipate future regulatory trends, participate in the drafting of sectoral standards and promote proactive transparency.

Likewise, compliance programs can serve as a bridge between innovation and regulation, proving that it is possible to develop advanced AI responsibly. Faced with the false dichotomy between innovation and control, experience shows that ethical innovation can also be sustainable.

9.1. Regulatory Compliance **with RIA**

In the current regulatory and normative landscape, it is essential to focus attention on the RIA. This regulatory framework stands out not only for being the first of a global character to establish a harmonised approach to the development and use of these technologies, but also for its direct impact on the Member States of the European Union and, in particular, Spain.

Its relevance lies in the fact that it is designed to **ensure a balance between technological innovation and the protection of fundamental rights, health and safety,** which are essential aspects in an increasingly digitalised and automated environment, while also aiming to support innovation and prevent market fragmentation. Furthermore, the RIA is becoming an international reference point for AI regulation, setting a standard that will influence the development of technology policies beyond European borders.

The Artificial Intelligence Regulation (RIA) or AI Act (Regulation (EU) 2024/1689, of June 13, 2024) was published in the Official Journal of the European Union on July 12, 2024 and entered into force on August 1 of the same year and, as we have indicated, marked a milestone as the **first binding regulation** of general scope on AI in the world. It establishes a horizontal legal framework governing the development, commercialization and use of AI systems within the territory of the Union, with extraterritorial effects on non-European actors whose systems are used in the EU.

It applies to any company that purchases, develops, customises, or uses AI systems that may **affect an EU citizen**, including providers established outside the Union if their results are intended to be used within the EU.

Excluded from its application are AI systems or models used exclusively for **military, defense, or national security purposes**, as well as **scientific research, testing, or development** prior to their introduction into the market.

The structure of the RIA is based on a risk-based approach, so that compliance obligations increase as the potential harm or impact of the AI system on fundamental rights, security, or democratic values increases. The RIA distinguishes four levels of risk:

01.

Unacceptable Risk

Covers prohibited uses of AI, such as cognitive manipulation of vulnerable people, social scoring systems by governments, biometric surveillance in real time in public spaces (except for very limited exceptions), or emotion inference in workplaces and educational centers.

02.

High Risk

Includes systems used in critical areas (health, employment, justice, education, credit, essential infrastructure, security, among others) and those that can have a significant impact on health, safety, or fundamental rights.

Examples of high risk include systems used in: employment and worker management (hiring, performance evaluation), law enforcement (crime risk assessment, evidence reliability), migration, asylum and border control (polygraphs, risk assessment) and access to essential services (public and private).

03.

Limited Risk

Where transparency obligations are imposed (for example, chatbots or generative systems that must disclose their artificial nature).

04.

Minimal or No Risk

Where compliance is based on best practices or voluntary codes.

This comprehensive regulatory framework, which will be implemented in stages, translates into practical implications and specific responsibilities for all operators in the AI value chain, with significant consequences for non-compliance with its provisions:

- **Phased implementation process:** the different obligations come into force at different times, not all at once, with full application extended until 2027. This phased approach gives companies and organisations time to adapt gradually, prioritizing the highest risks first.
- **Professionalisation of AI:** organisations are professionalizing the AI strategy, creating roles or departments dedicated to defining and managing its adoption and deployment in a coherent and centralised manner.
- **Transparency and documentation:** a significant level of use-case inventory, technical documentation, activity logs and transparency regarding system operations is required, as well as clear identification of AI-generated content.
- **Cybersecurity:** it is essential to strengthen cybersecurity measures to protect AI systems from new vectors, techniques, tactics and attack procedures that are already emerging, as well as privacy breaches.
- **Training and knowledge:** given the complexity of the RIA, it is crucial to train staff on its implications for security and data protection, as a large percentage are still unfamiliar with it.
- **Evaluation and adaptation:** companies must identify if they use or develop AI systems and if so, determine the role they play and their level of risk.
- **Rigorous data management:** it is essential to establish robust data governance policies to ensure quality, representativeness and bias minimisation in training data sets, especially when handling personal data.
- **Human and ethical oversight:** organisations must ensure effective human oversight and conduct assessments of ethical impact and fundamental rights, with staff trained in AI and safety.
- **Copyright and Intellectual Property:** GPAI providers must respect copyright and provide detailed summaries of the training data.
- **Risks and Responsibility:** the law introduces clear responsibilities for all actors in the AI value chain, requiring companies to reassess their processes, the value of human resources and the mitigation of operational, regulatory and bias risks.

Ultimately, the RIA does not just seek to regulate technology, but also to encourage responsible development and use that builds trust and benefits society, marking a milestone in global AI governance.

Compliance Requirements for High-Risk Systems

The core of regulatory compliance in the European Union falls on AI systems classified as high risk, regulated in **articles 8 to 51 of the RIA and in Annexes II and III** thereof.

Providers and, in some cases, deployers of these systems must implement a quality management system covering the entire product lifecycle, from design to post-market surveillance. Said system must document, among other elements, the following aspects:

- **Risk management:** continuous process of identifying, analyzing and mitigating risks arising from the use of the system, which must be updated throughout the model's useful life. It includes both foreseen and emerging risks, as well as preventive and corrective measures.
- **Data governance:** obligation to ensure the quality, relevance and representativeness of data used to train, validate and test models. Data must be free from undue bias, properly labeled and subject to statistical control mechanisms.

→ **Technical documentation:** maintenance of comprehensive technical documentation describing the system design, architecture, algorithms used, data sources, performance metrics, model limits and security measures implemented. If the system is provided by a supplier, they must provide the necessary technical documentation.

→ **Logging and traceability:** systems must allow for the automatic recording of events (log files) throughout their lifecycle. This obligation entails keeping automatic records of system operations enabling reconstruction of decisions or results, facilitating audits and subsequent evaluations.

→ **Transparency and user information:** providers must provide clear instructions about the intended use of the system, its accuracy level, limitations and the necessary measures for human oversight.

→ **Effective human supervision:** intervention and human oversight are required through measures that enable operators to manually override decisions. Thus, the system design must ensure that responsible persons can understand, control and, if necessary, interrupt the operation of the AI system to prevent harm or erroneous decisions.

→ **Robustness, accuracy and cybersecurity:** the system must demonstrate resistance to failures, adversarial attacks, manipulation, data poisoning, or unintended outcomes.

→ **Post-market surveillance:** providers must establish a continuous monitoring plan, collect information on the actual system behavior and notify authorities of serious incidents or anomalies.

The legal framework is already in force (since August 2024) and the application is staggered. **Prohibitions** entered into force in February 2025 and the rules for GPMs in August 2025, while most obligations for SAR are expected in August 2026. The threat of penalties is severe: **up to 35 million euros or 7% of annual global turnover** for non-compliance. They may also include prohibitions on use or withdrawal of certifications. This makes the RIA one of the strictest regulations on technology, comparable to the GDPR in its sanctioning regime.

Sanctions:

35 million euros

7% of the annual global turnover

Conformity Assessment **and CE Marking**

Before placing a high-risk system on the market or into service, the provider must subject it to a conformity assessment procedure, which can be internal or through an independent notified body, depending on the system category and the applicable sectoral framework (for example, medical devices, transport, or products subject to harmonised legislation).

Once compliance with RIA requirements has been verified, the provider must issue an EU declaration of conformity and affix the CE marking to the product or digital interface, ensuring compliance with Union legislation. Any substantial modification of the system requires repeating the evaluation.

In addition, providers must register high-risk systems in a European AI database managed by the European Commission, which reinforces the transparency and traceability of products in the internal market. Note also that the EC has the ability to expand the list of high-risk use cases under the RIA, requiring all entities to review their inventories when these potential new inclusions are published in the future.

Additional Requirements for **General-Purpose AI Models (GPAI)**

The RIA also introduces a specific category for general-purpose AI models (GPAI), such as large language or multimodal models that underpin multiple applications.

From August 2, 2025, GPAI providers must:

→ **Document** the architecture, data sources and training methods used.

→ **Ensure transparency** regarding energy use, model size and computing resources.

→ **Adopt measures** to prevent misuse and facilitate downstream compliance by those integrating these models into specific applications.

→ In cases of GPAI with systemic risk, **enhanced obligations** are imposed, such as independent audits, robustness testing, regular technical reports to the AI Office and global risk mitigation plans.

Roles around AI

The RIA defines a number of roles and responsibilities across the AI value chain, including the role of Manufacturer, Provider, Deployer, Importer, Distributor and Authorised Representative.

It is crucial to understand these roles and which role each player assumes in each use case in order to know the responsibilities corresponding to each, especially for the aforementioned high-risk systems.

Supervision Mechanisms and Competent Authorities

The RIA creates a European institutional supervisory architecture composed of several bodies:

- → The **European Artificial Intelligence Office (European AI Office)**, responsible for coordinating the application of the RIA, overseeing GPAI and promoting regulatory coherence among Member States.
- → The national competent authorities designated by each Member State, responsible for market surveillance, enforcement of penalties and conformity assessment. In Spain, the **AESIA, the Spanish Agency for the Supervision of Artificial Intelligence**, has been created for this purpose, although a national and even regional regulation is foreseen, which could result in additional Control Authorities.
- → **Notified bodies**, which act as independent certification and technical audit entities.
- → The European Artificial Intelligence Committee, with a consultative nature, which brings together representatives of Member States to ensure harmonised interpretations and the exchange of good practices.

This institutional framework seeks a balance between legal certainty, the protection of fundamental rights and responsible innovation.

Organisational Compliance: Practical Implications

From a corporate governance perspective, RIA compliance requires organisations to develop an internal AI Compliance Framework. This framework must include:

- → An inventory and classification of the AI systems used or developed by the entity.
- → Pre-deployment impact assessments analyzing ethical, social and fundamental rights risks.
- → Data governance policies and information quality control for training models.
- → Transparency protocols and technical documentation, enabling demonstration of compliance during regulatory audits.
- → Human oversight mechanisms and incident response plans.
- → Regular internal and external audits to verify the effectiveness of the management system and its alignment with regulatory requirements.
- → Ongoing staff training in AI, digital ethics and regulatory compliance.

Integration with International Frameworks and Other Regulations

Compliance with the RIA can be complemented by adopting international AI management standards, such as the NIST AI Risk Management Framework (2023), ISO/IEC 42001:2023, ISO/IEC TR 24368:2022, ISO/IEC 23894:2023, Artificial Intelligence - Guidance on Risk Management, among others already detailed in chapter 4 of this document.

On the other hand, the current regulatory and risk management framework (GRC) is not limited to the RIA and GDPR framework described; AI governance is a cross-cutting task that requires full integration with convergent regulations from different areas:

01.

Data Protection and Privacy

Regarding AI, the GDPR continues to apply critically at every stage of its lifecycle (training, validation, inference, decommissioning), as indicated in the specific privacy and data protection chapter of this guide.

For this reason, concepts such as privacy by design, transparency and explainability, or guaranteeing user rights, as well as carrying out data protection impact assessments (DPIA), which are integrated with FRIA, are essential for proactive risk management and compliance with the privacy by design principle that must be intrinsic to AI system implementation strategies.

Details can be found in the corresponding Privacy and Data Protection chapter of this guide, since, given the importance of these aspects when talking about an AI system and since their importance goes beyond compliance with the GDPR or any privacy and data protection regulation, these aspects have their own chapter in the present guide.

02.

Information Security and Cybersecurity

AI systems must also comply with necessary cybersecurity requirements.

Thus, products or services may be subject to Regulation (EU) 2024/2847 of the European Parliament and of the Council, of October 23, 2024, also known as the Cyber Resilience Act (CRA) and entities must comply with regulatory requirements on information security such as Directive (EU) 2022/2555 of the European Parliament and Council of December 14, 2022 on measures to ensure a high common level of cybersecurity across the Union (NIS2), Regulation (EU) 2025/38 of the European Parliament and Council, of December 19, 2024 (Cyber Solidarity Regulation), which aims to strengthen EU resilience against cybersecurity incidents, Regulation (EU) 2022/2554 of December 2022 on digital operational resilience for the financial sector (DORA) and the Spanish National Security Framework at the national level.

All of this regulatory framework converges and aims to create an ecosystem that allows AI governance in entities to also be carried out under the principles of security by design and resilience.

03.

Data

A basic cornerstone for success in the implementation of AI systems. Data and quality. To this end, the European Union approved Regulation 2022/868 (Data Governance Act), also known as the Data Act, which aims to promote access to and reuse of data and to foster data interoperability between sectors and service providers among States in accordance with the provisions established in the applicable regulation as of September 2025.

In this regard, it is also worth mentioning the recent Regulation (EU) 2025/327 of the European Parliament and of the Council, of February 11, 2025, on the European Health Data Space, as a sector-specific regulation that establishes data interoperability measures for primary use and efficiency of healthcare resources and a system that allows secondary use of such data.

04. Intellectual Property, Trade Secrets and Civil Liability:

An explicit contractual approach is required to protect Intellectual Property (software licensing, copyright over GenAI inputs and outputs) and trade secrets, especially when confidential information is introduced into third-party AI systems.

LLMs present enormous challenges regarding authorship and protection of AI-generated content (outputs) and the use of copyright-protected data for training (inputs) and we must also consider the regulatory framework from the perspective of harm and the possible non-contractual liability of damages caused by AI.

To this effect, it is worth remembering that the European Court of Justice, the Judgment of the Court of Justice of June 20, 2024, AT and BT v PS GbR and others, C-590/22[23] establishes the non-applicability of the **non bis in idem** principle in administrative sanctioning schemes and damage law.

Regarding damage law, in 2025 we saw how the Commission withdrew, for lack of consensus among the different Member States and regulatory pressure in recent years, the Directive on non-contractual liability for damages caused by AI systems. Therefore, from the perspective of damage law at the European level, we have Directive (EU) 2024/2853 of the European Parliament and of the Council of October 23, 2024, on liability for damages caused by defective products, the Consumer and User Law, as well as the Civil Code for determining possible compensation for damages.

In this Judgment, the right of the affected party to receive compensation is recognised, with the imposition of sanctions in response to different purposes. Insofar as the CJEU understands that compensation responds to the goal of providing "total and effective" compensation for suffered damages and sanctions are a deterrent or punitive measure for non-compliance or regulatory violations.

These are just some of the regulations that must also be taken into consideration to ensure that the AI system complies with current regulations, since any regulation can directly or indirectly affect AI systems, as they deal with organisational data and can be used for multiple purposes. Therefore, to guarantee that the system is AI compliant, it will be necessary to analyse both the system and the use case and identify the various regulations that may apply in order to ensure full compliance.

[23] European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of June 13, 2024 establishing harmonised rules on artificial intelligence and amending the indicated Regulations and Directives (Artificial Intelligence Regulation). Official Journal of the European Union, L 2024/1689. https://op.europa.eu/en/publication-detail/-/publication/72688445-2ee7-11ef-a61b-01aa75ed71a1/language-en

Complying with these regulations is not merely a matter of compliance or avoiding penalties; it is a way to ensure that AI systems operate within a framework of trust, transparency and effective human oversight. The early implementation of internal AI compliance structures, aligned with the RIA and international management standards, is today an essential condition for the competitiveness, sustainability and legitimacy of any organisation that uses AI.

05. Other Applicable Regulations

In addition to the already mentioned regulations and the comprehensive framework established by the RIA, organisations must also consider three additional aspects to fully understand the regulatory framework that may affect them:

→ At the beginning of 2025, the Government of Spain presented a **draft bill on Artificial Intelligence (AI) governance**, which aims to ensure ethical, inclusive and beneficial use of Artificial Intelligence for people. One of the main aspects of its text is the regulation of the penalty regime, as well as the powers that different Control Authorities will have in this regard.

→ Spain's territorial reality also makes it necessary to know when **regional laws are published that regulate AI in some way** and who the obligated parties are, such as Law 2/2025, of April 2, for the development and promotion of artificial intelligence in Galicia.

→ Other types of state regulations that partially or specifically regulate uses of AI. It is worth mentioning:

- The creation of the **regulatory Sandbox**, by which a controlled environment for AI systems testing is created to observe their behavior under the RIA requirements **(regulated in Royal Decree 817/2023, of November 8).**
- **Law 12/2021 of September 28**, which regulates the use of algorithms in the workplace, establishes the right of workers to be informed by the company of the parameters, rules and instructions on which the algorithms or artificial intelligence systems affecting decisions on working conditions, access and job retention are based, including profiling.

Ethics in Artificial Intelligence 10.

10.1. Current Context of **AI Ethics**

Ethics as a Requirement for Trust in AI

A system can be functionally brilliant and still be unacceptable if it erodes rights or discriminates. Talking about ethics in AI means building trust. Because an organisation that deploys AI without a clear ethical framework asks users, clients and regulators to blindly trust complex systems that may also handle sensitive data, hallucinate, contain biases, or have security gaps. Trust is built with very concrete measures.

Quoting our previous ISMS Guide "Ethics and Compliance in the Use of Artificial Intelligence" (2023), these measures should enable individuals to understand the behavior of the system and the degree of human intervention in it. Actual intervention, with information and control mechanisms and the ability to review and correct. Ethics also helps manage priorities when there are divergent objectives and reduces risk. Documenting data and decisions, auditing biases, enabling complaint channels and planning rapid corrections decreases the likelihood of reputational, regulatory, or legal incidents. Moreover, ethics enables adoption: users and teams trust and make better use of technology when they perceive clear rules, proper limits and recourse channels. With generative AI this is evident: with clear controls and responsibilities, people integrate it productively into their daily lives. Ethics does not slow innovation; it makes it sustainable and reliable.

Beyond **Technical Compliance:** Emerging Dilemmas in Generative and Agentic AI

Regulation and traditional compliance have enabled reasonably predictable technologies to be organised—requirements, certifications and audits—but are increasingly shifting towards a risk-based approach. And generative AI and autonomous agents are scaling to another level. Generators of narratives, images, videos or code and with the capacity to perform actions without human intervention, not limited to data processing or prediction of outcomes.

One of the most pressing issues is the impact on the **cognitive dimension of people.** Generative models produce quick, compliant and convincing answers, even when they are false, which can erode critical thinking and encourage passive acceptance of any response, risking the loss of analytical and reflective skills because it is easier to delegate them.

Technical transparency is not enough; the ethical challenge is how we preserve people's ability to question, contrast and decide for themselves in an environment saturated with artificially generated texts, images and videos. To this end, it is necessary to reorient training and education of people to this new environment, as well as to introduce cognitive guardrails into AI tools.

Likewise, although a question-answering model may seem harmless, the same system integrated with external tools—capable of buying, booking, sending messages or scheduling tasks—turns the reflection from theoretical to practical:

Which tasks are legitimate to hand over to an agent? What does "human supervision" mean when the system operates at machine speed?

Another exponentially expanding conflict is the alteration of trust and public authenticity: hyper-realistic deepfakes or large-scale persuasive texts threaten the democratic foundations, proof of truth and the reputation of people and institutions.

Technical compliance standards mention watermarks or user notices, but ethics goes further: in which contexts is it legitimate to use synthetic content? What duties of diligence do we have before sharing a generated image or audio? The issue is no longer just technical, but cultural: preserving trust in what we see and hear.

Compliance provides rules, but ethics must give meaning and protect people. And this reflection is not trivial: as an extreme and illustrative example, in mid-2025[24] the case became known of a Californian teenager who took his own life after months of intense conversations with ChatGPT and according to the lawsuit filed by his parents against the manufacturer, it not only validated his hopeless thoughts, but even went so far as to suggest suicide methods, review self-harm photos and reinforce the idea of keeping his intentions secret. An episode that shows we are not talking about theoretical abstractions, but about real impacts on human lives

[24] Reuters. (2025, August 26). OpenAI, Altman sued over ChatGPT's role in California teen's suicide. https://www.reuters.com/sustainability/boards-policy-regulation/openai-altman-sued-over-chatgpts-role-california-teens-suicide-2025-08-26

10.2. **Applicable Ethical Context** and Guidelines

Regulatory Framework and Reference Standards

Over recent years, regulatory and governance frameworks have been developed that translate principles, guidelines and ethical recommendations into a more operational plan. Among the most relevant are:

- **AI Regulation, RIA (European Union, 2024)**: risk-based framework. It turns principles like transparency, human oversight, non-discrimination and robustness into legal obligations.
- **GDPR (European Union, 2016):** regulates privacy, data minimisation and the right to protection against automated decisions across all domains.
- **NIST AI Risk Management Framework (USA, 2023)**[25]**:** not binding. De facto standard for AI risk management. It includes principles such as explainability, reliability, traceability and alignment with social values.
- **UNESCO Recommendation on the Ethics of AI (2021)**[26]**:** first global instrument adopted by 193 Member States setting universal ethical principles: human rights, inclusion, cultural diversity, environmental sustainability.
- **OECD Principles on AI (2019)**[27]**:** reference framework for OECD and G20 countries, incorporating values such as inclusive growth, transparency, robustness and accountability.

This context includes binding regulations (RIA, GDPR) and governance guides adopted by international consensus (OECD, UNESCO, NIST). Taken together, they encompass the applicable ethical axes that every organisation should consider when designing, implementing and supervising AI systems.

[25] National Institute of Standards and Technology (NIST). (2023). AI Risk Management Framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework

[26] UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

[27] Organisation for Economic Co-operation and Development (OECD). (^{2019}). OECD AI Principles. https://oecd.ai/en/ai-principles

Understanding the Traditional **Risk-Based Approach** for Regulated Systems

The RIA embodies Europe's risk-based approach by classifying systems into categories according to their potential risk. But it goes further, establishing in its Article 5 a list of unacceptable practices, such as subliminal manipulation or the exploitation of minors' vulnerabilities (rooted in the Charter of Fundamental Rights of the European Union and the UNESCO Recommendation on the Ethics of AI, which stresses that technological innovation cannot erode human dignity).

Not everything that is technically possible is legally acceptable or ethically legitimate. Respect for the law and for human rights marks a boundary: excluding from the outset uses that violate people's freedom and autonomy.

On the other hand, the area of **privacy by design and data governance** is primarily addressed through the GDPR, but it is also integrated into the RIA (article 10). The principles of minimisation, purpose limitation and control by the data subject are as relevant in AI as for any other data processing and are more important than ever.

Another fundamental axis is the demand for **transparency and explainability**. Europe requires this as a duty, imposing documentation requirements on high-risk systems and for GPAI the obligation to report on training and label artificially generated content. In parallel, the **NIST AI Risk Management Framework** in the United States highlights that the features of trustworthy AI include systems being “responsible and transparent, explainable and interpretable”, implying explanations suitable to each type of audience and not just technical. Similarly, **OECD** incorporates as a guiding principle “transparency and responsible disclosure around AI systems to ensure that people understand when they are interacting with them and can challenge the results”.

Justice, fairness and non-discrimination are another key axis. The RIA requires data governance and measures to mitigate bias in high-risk systems. UNESCO and the OECD, meanwhile, insist that AI should serve the common good and social inclusion.

It is not just about reviewing datasets for errors, but about continuously evaluating whether results are generating unjust inequalities.

Organisations must design systems with diverse teams, test with data from representative groups and ensure algorithms do not reproduce discrimination. The RIA also devotes Article 14 to **human oversight in high-risk systems,** which NIST calls human-in-the-loop or human-on-the-loop: oversight is not just symbolic and must guarantee real power to review or cancel. In practice, this should mean that in sensitive fields like justice or health, a human decision can never be entirely replaced.

More recently, issues such as **sustainability and proportionality** have entered the debate and deserve attention. The development of ever larger models requires huge amounts of energy and resources: for example, according to the International Energy Agency (IEA), the electricity consumption of AI-linked data centers could double by 2030, reaching 945 TWh per year, equivalent to Japan's total consumption. Proportionality requires calibrating the social value of each application against its ecological footprint.

GPAI and Agents: Limits of the Current Regulatory Framework and the Need for New Approaches

Unlike traditional systems, where it was possible to assign a purpose and proportional obligations, GPAI go beyond this classification logic. Their open, multifunctional nature makes it unfeasible to pigeonhole them into a specific risk: they can both drive medical applications and generate disinformation campaigns, draft a legal report, or program malware.

Therefore, in GPAI there is not a single purpose that allows risks to be anticipated and regulated proportionally. What once was a framework of defined purposes becomes a blurred scenario of legitimate uses, abuses and emerging threats.

What purpose can be assigned to a model that writes an essay, codes, generates images, translates legal text, or simulates a human voice? The RIA imposes additional obligations on GPAI—technical documentation, summaries on training data, labeling synthetic content. These are important measures, but risks remain: models released to the market and integrated into third-party products connect with other apps and are reused in contexts far removed from the developers' objectives, creating whole ecosystems on the same technological core, where the boundaries between innovation, abuse and unforeseen risk are blurred.

That is the nature of an agentic AI ecosystem: systems that not only generate content but also act, plan, interact with external tools, execute tasks and make decisions autonomously.

There is a need to conceptualise new approaches and transfer them into the legal order. First, **protection and cognitive autonomy**. Because beyond technical transparency, organisations must protect people's ability to think and decide for themselves. This implies designing interfaces that foster critical verification and offering clear notices about the generated nature of content.

Also d**istributed responsibility and shared traceability.** Ethics here demands a clear distribution of roles and obligations, avoiding both responsibility vacuums and dilution that leaves the end user unprotected. But this shared responsibility is only effective if accompanied by traceability mechanisms that allow auditing the model's trajectory: what data were used, what changes were made and in what contexts it was applied. In other words, verifiable accountability.

[25] National Institute of Standards and Technology (NIST). (2023). AI Risk Management Framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework

[26] UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

[27] Organisation for Economic Co-operation and Development (OECD). ([2019]). OECD AI Principles. https://oecd.ai/en/ai-principles

Intellectual Property and the Vulnerability of Truth

The tension between intellectual property and the generative AI's ability to replicate, recombine and recreate protected content is one of the current ethical focal points, since trust in the authenticity of what we see, hear, or read is threatened when AI results can mix protected works with real content in an indistinguishable way. The training of large models with huge volumes of data raises questions about the use of copyrighted materials without authorisation and, although there are arguments regarding fair use or data mining exceptions, many authors feel their works become free raw material for systems that later generate content competing with them, adding the dilemma of authorship:

Who is the holder of a song composed by an algorithm trained with thousands of pieces of human music? The programmer, the company, the user who requested the work, or no one at all? The lack of clear answers threatens to discourage creation, as misappropriation and inadvertent plagiarism become constant risks.

This conflict does not stop at the legal realm. The ability of generative AI to produce hyper-realistic images, convincing audio or large-scale persuasive texts puts **public trust in truth** in check, making us stop trusting what we see or hear: before, a photograph served as proof, a recording as testimony, a document as reliable record. Today, that trust is shattered by deepfakes, mass-generated texts fueling disinformation, or manipulated videos.

What is at stake is not just the protection of copyright, but the very infrastructure on which coexistence and trust in the existence of an accessible and verifiable truth are based.

From an ethical perspective, both dilemmas are deeply intertwined, eroding the very basis of culture and even democracy. Technical solutions—such as digital watermarks or generated-content labels—are necessary but not sufficient. The real challenge is normative and cultural, redefining the rules of intellectual property and strengthening citizens' digital literacy so they can distinguish, contrast and avoid manipulation.

Beyond these reflections, legal battles are already being fought, showing how relevant and urgent this issue is. from a lawsuit by a digital publisher against OpenAI in Delaware for use of their content without permission (Reuters[28]), to an action led by The New York Times against OpenAI and Microsoft, joining over a dozen other author lawsuits for the unauthorised use of their work (The Guardian[29]).

[28] Reuters. (2025, April 24). Publisher Ziff Davis sues OpenAI for copyright infringement. https://www.reuters.com/business/publisher-ziff-davis-sues-openai-copyright-infringement-2025-04-24/

[29] The Guardian. (2025, April 4). US authors' copyright lawsuits against OpenAI and Microsoft combined in New York with newspaper actions. https://www.theguardian.com/books/2025/apr/04/us-authors-copyright-lawsuits-against-openai-and-microsoft-combined-in-new-york-with-newspaper-actions?utm_source=chatgpt.com

AI Ethics and Neuro-Rights: The Future?

Soon we will have technologies capable of directly interacting with brain activity, which raises an unprecedented ethical issue: **neuro-rights**. Indeed, the human mind and its most intimate processes deserve specific protection against the advance of neurotechnologies and AI applications that, when combined with sensors and predictive models, can access, infer or even alter mental states.

Here, ethics is no longer limited to ensuring responsible use of personal data or algorithms in decision-making: we are talking about the defense of **cognitive liberty**, the right to keep our thoughts private and not to be manipulated in our emotions or behaviors through stimuli directed by intelligent systems.

Some countries have begun taking concrete steps. Chile, for example, amended its Constitution in 2021 to recognise neuro-rights as part of basic human rights, setting an international precedent (Norm 1166983[30]). Organisations such as UNESCO and the Council of Europe have also warned about the need to extend the protection of human dignity to this new area, which raises unprecedented dilemmas regarding informed consent, subliminal manipulation and social control (UNESCO NEUROTECH[31])

What happens if a company can infer a person's emotional state in real time and tailor advertising to it? What are the implications of a system that, connected to a virtual reality headset, can alter or induce sensory experiences that the user perceives as real?

[30] Chile. (2023). Law No. 21.595 on computer crime. Library of the National Congress of Chile. https://www.bcn.cl/leychile/navegar?idNorma=1166983

[31] UNESCO. (2023). Recommendation on the ethics of neurotechnology. https://www.unesco.org/en/ethics-neurotech/recommendation

Practical Guidelines and Governance Mechanisms for Companies that Develop, Use or Adopt AI

10.3.

Ethical AI governance in companies requires integrating practices and responsibilities, not just complying with regulation.

To this end, this guide distinguishes three complementary axes:

01. Companies that develop AI, responsible for introducing safeguards from the design, training and deployment of models.

02. AI users—employees, clients or GPAI consumers—who need clear guidelines for critical use, verification and continuous training.

03. Leaders with decision-making capacity and those responsible for adopting AI in their organisations. They must ensure coherence between ethics and strategy, provide resources for governance and assume accountability.

Reference Guidelines **for Companies that Develop AI**

→ Anonymise and minimize data from the beginning, training systems only with strictly necessary information, deleting or transforming personal or sensitive identifiers from the design phase. For example, a company that develops a model to analyse work absenteeism patterns will replace employee names, addresses, etc. with anonymous identifiers, keeping only relevant information such as age, job category and contract type.

→ Document and trace data and models, i.e., identify, inventory and document the origin of data, their type of license/permissions/rights, as well as the modifications made. Traceability, diligence regarding intellectual property and data protection and ease of regulatory audits. Thus, a startup training an image generation model can document that 60% of its dataset comes from Creative Commons license repositories, 30% from photographer contracts and 10% from proprietary material, keeping evidence in case of legal dispute.

→ Make conscious decisions about the purpose of the system and about which data to exclude, which objectives to prioritise and which limits to set before beginning training. Companies are responsible for explicitly defining the scenarios in which their systems should not be used. An example: when designing a job recommendation algorithm, the company decides from the start not to include variables such as gender or marital status, which could bias opportunity allocation. Another example is a company marketing a voice generation model that may expressly prohibit its use to imitate minors or political figures and document this exclusion as part of its ethics-by-design policy.

→ Ensure human supervision with real power to act. Based on three main elements: access to understandable explanations of system decisions, specific training to interpret those explanations and clear action buttons like "confirm", "modify", "accountability", etc. within the tools themselves. For instance, when using a credit risk system, the analyst sees on screen not only the score proposed by the algorithm but also the main factors that motivate it (e.g., income level or payment history). In addition, the interface should include direct options to approve, reject, or modify the automatic decision, so that the person retains control.

→ Establish agile and responsible mechanisms for user interaction: listen to users and integrators to identify unintended biases, errors, or harms. To do so, companies must offer simple and accessible complaint channels, as well as commitments for response. A company offering an AI API for image classification includes a button in the control panel allowing users to report "misclassification" or "inappropriate content" and commits to review and respond within a maximum period.

→ Attack and stress-test the system, making it resistant to malicious use or unexpected inputs. It is key to simulate prompt injection attacks and deliberate manipulation scenarios before deployment. For example, a company launches a medical chatbot where a user tries to force illegal prescriptions or requests self-harm instructions, simulating how the system detects and blocks these behaviors.

→ Document decisions and training processes: what decisions were made, why and with what criteria, clearly documenting the data used, model versions, optimization goals and any ethical dilemmas encountered. An example is a company developing a recruitment system that archives the criteria for which it chose to weigh work experience over academic qualifications, to justify this choice if a bias is detected in the future.

→ Conduct regular assessments of ethical and social impact and plan cycles of continuous monitoring. When relevant, it is important to anticipate the social, cognitive and environmental impact of AI systems. These assessments help identify less obvious risks, such as loss of user autonomy or excessive energy consumption. For example, in education, a generative model may reduce students' critical capacity if the tool is overused. Continuous monitoring enables detection of deviations, biases or vulnerabilities and adaptation of the system to normative or social changes, such as adjusting a recommendation engine that favors sensationalist content.

Reference Guidelines **for AI Users (employees, clients, consumers)**

01 Train in critical use of AI: learning how models work, their limitations and reliability, reducing the risk of accepting erroneous outputs, hallucinations or manipulations.

02 Avoid introducing sensitive data into external systems, as prompts containing sensitive information may be stored or reused without control.

03 Set delegation limits for AI, deciding in advance which tasks can be assisted by AI and which cannot, using professional judgment and preventing critical decisions from falling entirely to an automatic system.

04 Develop habits of cognitive contrast, encouraging questioning of what the tool returns, comparing alternatives and not accepting the fastest answers without question. Preserving critical thinking and avoiding "intellectual comfort" from convincing outputs.

05 Use AI as support, not as a substitute, combining human judgment with model capability rather than delegating everything. In this way, the risk of excessive dependency is reduced and the quality of the result is reinforced.

06 Be transparent in AI use, communicating to clients or interlocutors when a text, image or report has been generated with the help of a model. This practice builds trust and avoids misunderstandings.

07 Keep copies and backups of results, so as not to depend exclusively on the tool and avoid losing traceability of outputs. This allows you to show what was generated, when and how it was used.

Reference Guidelines for Leaders (CEOs, Senior Executives, Decision-Makers) Adopting AI within Their Organisations

01.

Define an AI strategy aligned with organisational values, so adoption is not only based on efficiency or cost reduction.

02.

Assess the impact of AI on the workforce, anticipating which tasks will be automated and which roles will be transformed, with the aim of planning the transition and avoiding unmanaged mass layoffs.

03.

Invest in reskilling and upskilling employees, offering training programs so workers can acquire new digital skills, adapt to change and even increase their personal productivity.

04.

Include the perspective of vulnerable groups in decision-making.

05.

Establish a diverse AI ethics and governance committee with real decision-making capacity.

06.

Personally assume accountability as leaders, recognizing that the ultimate responsibility for the use of AI in the organisation lies not with the algorithm or provider, but with management.

Audit and Traceability of AI Systems

11.

AI governance within an organisation must ensure that all AI systems, whether in-house or third-party, have traceability mechanisms to support transparency in their processes. Furthermore, it is essential that these systems are auditable to prevent errors, biases, or improper uses of technology.

11.1. Why Audit AI?

Auditing an AI system, whether internal or as part of a service offered by third parties, is essential to provide assurances about its operation. Given the widespread adoption of AI in environments where automated decisions can impact fundamental rights, critical productive processes, or reputation, auditing is an essential tool for identifying risks, correcting deviations and strengthening the confidence of clients, regulators and the business.

Without being exhaustive, the following objectives should be considered as a minimum within the activities of AI systems auditing:

→ Verify from the design stage the compliance with technical, ethical and legal requirements.
→ Evaluate the impact on fundamental rights, especially anything that may affect vulnerable populations.
→ Detect biases, errors, or vulnerabilities in the models.
→ Ensure traceability in decision-making and throughout all algorithmic processes.
→ Document the platform, necessary to strengthen organisational governance and human oversight through knowledge.
→ Collaborate in the promotion of best practices regarding system sustainability and energy efficiency.

11.2. **Tracking** Every Decision

Traceability, generally referring to aspects related to algorithmic and technical traceability, is necessary to know which data were used, which model was applied, what parameters were active and/or in what context a given decision or set of decisions was made. **Complete and effective traceability should include not only technical mechanisms oriented toward internal AI systems or those within the organisation.** It should also consider elements that are part of "organisational traceability", covering processes, decisions and responsibilities within the organisation that affect the development and use of the AI system, such as documentation of meetings, approvals, roles, coordination among internal and external teams, or the closure of AI-related service contracts.

The traceability of a platform or AI system **must include at least the following mechanisms:**

→ **Record** of all lifecycle activities
→ **Version control** of models and update history, data usage and production code control.
→ **Record of** key decisions related to the platform, connecting the organisational and operational aspects
→ Reliable technical traceability mechanisms in all systems with **unique identifiers** that record actions performed by each component, inferences, errors and decisions made by the system and/or due to third-party connections or use.
→ **Document repository** oriented toward organisational and contractual traceability: roles, approvals and internal responsibilities.
→ **Supplier registry and proactive use** of Service Level Agreements (SLAs), traceability clauses with third parties, KPIs on compliance with shared governance policies or change management.
→ **Integration with document management systems and GRC** tools helps to provide mechanisms for supervision and cross-checking, to design consistency and validation tests, and to facilitate the performance of technical audits.

11.3. Dependencies Between **Audit and Traceability**

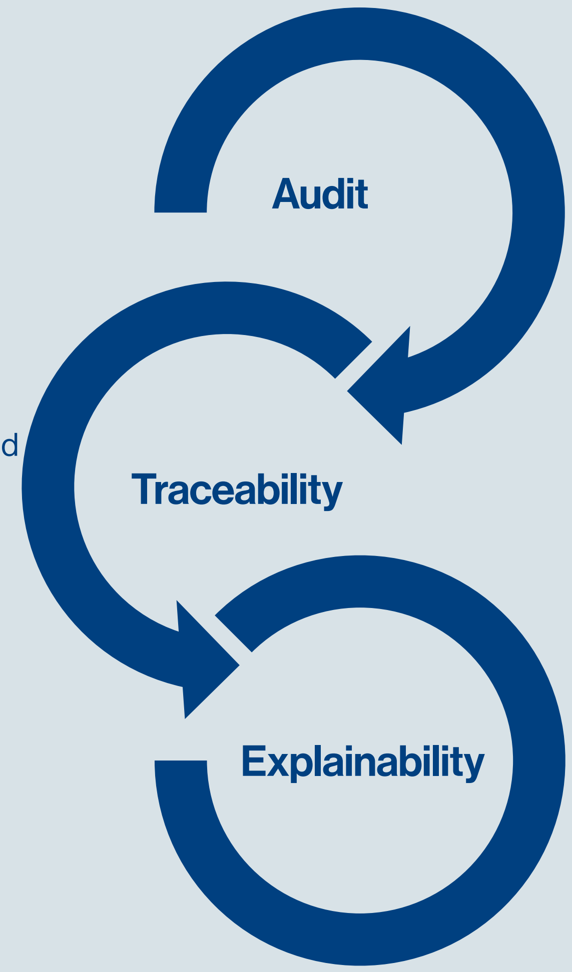
The audit must assess the traceability processes and their handling by the rest of the systems to ensure that the information provided is accurate and sufficient.

The absence of traceability in the execution of any critical function, the use of inadequate formats, or deficiencies in log tracking significantly hinders the audit process, as it prevents a structured and efficient review of the system.

	Traceability	Audit	Relationship
Data	Records origin, transformation and use	Assesses quality, bias and legality	The audit uses traceability to verify proper use
Models	Versions and documents changes	Reviews performance and fairness	Allows auditing which version was used and how it was trained
Decisions	Saves context and logic	Assesses impact and justification	Facilitates auditing of automated decisions
Compliance	Documents processes	Verifies alignment with standards	Traceability must serve as evidence for the audit

In addition to the concepts of audit and traceability, a third aspect, explainability—the ability of an AI system to explain its decisions and outcomes to users and anyone else—is a requirement for systems classified as "High Risk" by the RIA, along with human oversight.

The way in which the concepts of audit, traceability and explainability are related is essential to address all three correctly:



Requires traceability to reconstruct decisions and explainability to understand them.

Knowledge base that enables the auditing and explanation of system behavior.

Depends on traceability to access data and logic and could be a key objective of the audit.

11.4. **Audit Methodology** in AI Systems

The necessary tasks for conducting an AI systems audit should aim to evaluate their operation across multiple dimensions: technical, functional, ethical, organisational and sustainability. This makes it necessary to have a structured methodology that provides a rigorous review, proportional to the risk and based on regulatory criteria. This chapter includes annex 2, which contains the key questions for the audit, organised by the aforementioned dimensions. This annex constitutes a practical tool to guide the evaluation process.

Audit Dimensions

Technical Dimension

Objective: ensure accuracy, robustness, explainability and system security.

Key actions:
- Assessment of accuracy, robustness, explainability and security.
- Data validation: quality, representativeness, bias.
- Review of the model lifecycle: design, training, validation, deployment.
- Verification of compliance with technical requirements.

References: ISO/IEC 42001 (technical controls), RIA Art. 9 to 15 (high-risk systems requirements), GDPR Art. 5 and 25 (minimisation and privacy by design).

Functional Dimension

Objective: verify that the system meets its operational purpose and adapts to the context of use.

Key actions:
- Assessment of impact on fundamental rights.
- Fairness and non-discrimination.
- Transparency and explainability.
- Review of human oversight mechanisms and ethical governance.

References: RIA Art. 11 (technical documentation), ISO/IEC 42001 (lifecycle management), GDPR Art. 22.

Ethical Dimension

Objective: ensure respect for fundamental rights, fairness and human oversight.

Key actions:
- Ethical impact assessment.
- Review of biases.
- Oversight mechanisms.

References: RIA Art. 14 (human oversight), GDPR Art. 22 (automated decisions), ISO/IEC 42001 (organisational values).

Organisational Dimension

Objective: audit internal governance, roles and processes associated with the system.

Key actions:
- Documentation of roles, responsibilities and processes.
- Coordination between teams and providers.
- Review of internal policies and organisational culture.

References: RIA Art. 16–29 (provider obligations), GDPR Art. 24–32 (organisational responsibility), ISO/IEC 42001 (governance structure).

Sustainability Dimension and Energy Efficiency

Objective: assess the energy impact and sustainability of the system.

Key actions:
- Measurement of consumption, including provider footprint.
- Coordination between teams and providers.
- Review of internal policies and organisational culture.

References: RIA Art. 16–29 (provider obligations), GDPR Art. 24–32 (organisational responsibility), ISO/IEC 42001 (governance structure).

Required **Evidence**

The audit must assess the traceability processes and their management by the rest of the systems, to ensure that the information provided is accurate and sufficient.

The absence of traceability in the execution of any critical function, the use of inadequate formats, or deficiencies in tracking records significantly complicates the audit process, as it prevents a structured and efficient review of the system.

Dimension	Traceability	Audit	Relationship
TECHNICAL	Accuracy, robustness, explainability	Validation reports, metrics, technical documentation	ISO/IEC 42001, RIA Art. 9–15
TECHNICAL	Algorithmic security	Cybersecurity tests, vulnerability analysis	ISO/IEC 42001
TECHNICAL	Data minimisation	Review of collected data, justification of use	GDPR Art. 5, 25
FUNCTIONAL	Fitness for use context	Functional requirements, user testing	RIA Art. 11
ETHICAL	Human oversight, fairness, transparency	Ethical impact assessment, decision records	RIA Art. 14, GDPR Art. 22
Organisational	Governance, roles and responsibilities	Internal policies, organisational charts, meeting minutes	ISO/IEC 42001, RIA Art. 16–29, GDPR Art. 24–32
ENVIRONMENTAL	Energy efficiency and sustainability	Consumption reports, environmental policies	ISO/IEC 42001, internal policies

Based on this methodological foundation, a structured procedure is proposed that enables the audit to be applied sequentially, rigorously and adapted to the risk level of the system evaluated.

Audit Procedure for AI systems

A clear and structured procedure is required to allow its operation to be evaluated from multiple dimensions, also adding the cross-cutting treatment of applicable regulations and standards. As in any audit process, it is necessary to have an appropriate context in time and form to ensure that the results are not distorted by possible treatments carried out in parallel or between completion and delivery of findings. The phases are the usual ones in any audit.

Phase 1: Planning

The auditing principles established by standards such as ISO 19011 and the risk management approach of ISO/IEC 31000 are observed. The scope will consider aspects such as the type of system (predictive, generative, autonomous), its risk level as set out in the RIA and an assessment of the potential impact on fundamental rights.

→ Define the scope, objectives and audit criteria.
→ Identify the AI systems to audit (ideally, selected and segregated by risk, impact or criticality).
→ Identify and assign roles, resources and open the calendar.
→ Review the applicable regulatory framework (among others, RIA, ISO/IEC 42001, GDPR and the organisation's policies).

Phase 2: Evidence Collection

It is advisable to apply cross-verification techniques (documentary, testimonial, observational, etc.) to guarantee the reliability of the evidence. The use of tools such as model cards, data sheets for datasets[32] and system logs facilitates traceability.

→ Request technical, functional, ethical and organisational documentation.
→ Conduct interviews with development, compliance and user representatives.
→ Compile logs, traceability, validation reports and regulatory records.

[32] "The World Economic Forum suggests that all entities should document the provenance, creation and use of machine learning datasets in order to avoid erroneous or discriminatory outcomes. This framework proposes that every dataset should be accompanied by a 'datasheet', which is a questionnaire that guides documentation and reflection throughout the data lifecycle." - Data documentation: Datasheets for datasets | datos.gob.es

Phase 3: Evaluation by Dimensions

Supported by normative compliance matrices, gap analyses and risk heat maps, using frameworks such as the NIST AI Risk Assessment Framework or tools like AI Fairness 360 or LIME for explainability, or any other proprietary or third-party tool.

→ Design an audit template by dimension.
→ Analyse compliance and key indicators.
→ Identify gaps, risks and opportunities for improvement.

Phase 5: Follow-Up

Indicators should be incorporated that can be used in continuous improvement processes (KPIs), periodic reviews and the schedule for planned audits. Establishing a feedback loop with technical and compliance teams and documenting corrective actions in management systems will greatly facilitate monitoring and the efficient detection of deviations.

→ Verify the implementation of recommendations.
→ Update records and documentation.
→ Schedule recurring audits or periodic reviews

Phase 4: Audit Reports

They should be structured according to materiality, criticality and risk prioritization criteria given in the initial assessment. It is always important to include technical annexes, risk maps and recommendations classified by impact and urgency. The storage of the raw evidence collected and the methodology for obtaining it should also be taken into account.

→ Draft the report with findings, evidence and recommendations.
→ Classify risks (critical, moderate, minor).
→ Propose corrective and mitigation measures.
→ Promote the presentation of the report to the AI oversight committees, as well as to the Information Security Committee or similar body.

Priority Approach for the Audit of AI Systems

One way to prioritise the audit is to apply a risk- and criticality-based approach, aligned with internal audit practices (COSO, ISO 31000) and with the spirit of the RIA, which classifies systems according to their risk level. This approach allows resources to be allocated efficiently and focuses evaluation on elements that can have the greatest ethical, legal, technical, or organisational impact:

1. System classification

- Is it a high-risk system per the RIA (for example: biometrics, judicial decisions, HR selection)?
- Does it affect fundamental rights, critical processes or vulnerable populations?

2. Preliminary impact assessment

- What consequences would a technical, ethical, or legal failure have?
- What level of autonomy does the system have in decision-making?

3. Assignment of audit depth

RISK LEVEL	AUDIT APPROACH	GUIDELINE FREQUENCY
HIGH	Full audit by dimensions + extended traceability	Quarterly / before each deployment
MEDIUM	Technical and ethical audit + organisational review	Semiannual / after updates
LOW	Functional review + basic traceability	Annual / on demand

4. Prioritisation criteria

→ **Impact on people:** Can it result in discrimination, exclusion, or harm?
→ **Technical complexity:** Does it use opaque or hard-to-explain models?
→ **Organisational maturity:** Does the organisation have solid governance processes?

11.5. Methodology for Implementing or Verifying the Traceability of AI Systems

It is essential to implement and properly manage traceability systems, addressing technical, organisational, ethical, functional and environmental aspects to facilitate rigorous audits and foster continuous improvement processes.

Traceability Implementation

Technical Traceability

Objective: document and track the technical lifecycle of the AI system, from design to deployment.

Key actions:
- Version tracking for models, datasets and source code.
- Training, validation and adjustment logs.
- Structured logs of inferences, errors and automated decisions.
- Use of tools such as Model Cards, Data Sheets, Lineage Trackers.

References:
→ ISO/IEC 42001 (technical controls and lifecycle)
→ RIA Art. 9–15 (technical requirements for high-risk systems)
→ GDPR Art. 5, 25 (data minimisation and privacy by design)

Technical Traceability

Objective: to document and monitor the technical lifecycle of the AI system, from design to deployment.

Key actions:
- Version logs of models, datasets and source code.
- Training, validation and adjustment logs.
- Structured logs of inferences, errors and automated decisions.
- Use of tools such as Model Cards, Data Sheets, Lineage Trackers.

References:
→ ISO/IEC 42001 (technical controls and lifecycle)
→ RIA Art. 9–15 (technical requirements for high-risk systems)
→ GDPR Art. 5, 25 (data minimisation and privacy by design)

Ethical Dimension

Objective: to ensure respect for fundamental rights, fairness and human oversight.

Key actions:
- Ethical impact assessment.
- Bias review.
- Oversight mechanisms.

References: RIA Art. 14 (human oversight), GDPR Art. 22 (automated decisions), ISO/IEC 42001 (organisational values).

Organisational Dimension

Objective: to audit internal governance, roles and processes associated with the system.

Key actions:
- Documentation of roles, responsibilities and processes.
- Coordination between teams and supplier.
- Review of internal policies and organisational culture.

References: RIA Art. 16–29 (provider obligations), GDPR Art. 24–32 (organisational accountability), ISO/IEC 42001 (governance structure).

Sustainability Dimension and Energy Efficiency

Objective: to evaluate the system's energy impact and sustainability.

Key actions:
- Measurement of consumption, including supplier footprint.
- Coordination between teams and supplier.
- Review of internal policies and organisational culture.

References: RIA Art. 16–29 (provider obligations), GDPR Art. 24–32 (organisational accountability), ISO/IEC 42001 (governance structure).

11.6. Procedure for Applying or Verifying **Traceability**

The use of traceability in an AI system, or its assessment in third-party systems, must be verified during the audit, as several of the most important pieces of evidence depend on it. Aligning with the methodology described in this document allows the integration of traceability into the AI system's lifecycle. As in the audit procedure, its execution is planned for both proprietary and hybrid environments with third parties. Similarly, it is aligned with the principles of RIA, ISO/IEC 42001 and GDPR.

Phase 1: Traceability framework design

- **Objective:** Define the structure, dimensions and mechanisms to be applied.
- **Key actions:**
  - → Select relevant dimensions (technical, organisational, ethical, functional, environmental).
  - → Establish the types of evidence required by dimension.
  - → Identify recording tools (event logs, version control, logs, matrices).
  - → Align the framework with applicable regulatory requirements.
- **References:** ISO/IEC 42001 (governance and lifecycle clauses), RIA Art. 9–29, GDPR Art. 5, 30.

Phase 2: Integration into the system lifecycle

- **Objective:** Incorporate traceability mechanisms from design to deployment.
- **Key actions:**
  - → Implement structured logs at each stage (design, training, validation, operation).
  - → Document technical, ethical and organisational decisions.
  - → Establish control and validation points by dimension.
  - → Coordinate with external suppliers to ensure shared traceability.
- **References:** “Audit-by-design” (AI Now Institute), ISO/IEC 42001, GDPR Art. 25.

Phase 3: Traceability verification and audit

- **Objective:** Evaluate the quality, completeness and consistency of the records generated.
- **Key actions:**
  - → Apply audit templates by dimension.
  - → Review logs, evidence and technical and organisational documentation.
  - → Identify gaps, inconsistencies, or opacity risks.
  - → Classify the level of traceability achieved (basic, intermediate, complete).
- **References:** NIST AI Risk Management Framework, RIA Art. 11, 14, 16.

Phase 4: Traceability Report

(Complementary to the Audit Report and Reviewable During the Audit)

- **Objective:** Consolidate findings and propose improvements.
- **Key actions:**
  - → Draft a report by dimension with evidence and recommendations.
  - → Include traceability maps, compliance matrices and technical annexes.
  - → Present the report to governance committees and AI leads.
- **References:** ISO 19011 (audit), RIA Art. 29 (documentation), GDPR Art. 30.

Phase 5: Monitoring and continuous improvement

- **Objective:** Ensure that traceability evolves with the system.
- **Key actions:**
- Verify the implementation of recommendations.
  - → Update records and traceability mechanisms.
  - → Schedule regular reviews and recurring audits.
  - → Establish continuous improvement indicators (traceability KPIs).
- **References:** ISO/IEC 42001 (improvement cycle), internal ESG policies.

Regulatory Framework 11.7.

The audit and traceability of AI systems can be framed within a set of norms and principles that ensure their responsible development and use.

Within the set of countries in the European Union, RIA and GDPR may be taken as minimal references already developed in this guide. There are also other reference guides, such as the Guide of the Spanish Data Protection Agency, AEPD, on Audits of AI Processing, which is already widely used in Spain and other countries such as Mexico.

In addition, there are various technical standards, of more or less scope, that can complement the legal framework, offering methodologies for implementing audit and traceability in a structured manner, such as ISO/IEC 42001 - AI Management System (AIMS) or ISO/IEC 23894 - Risk management in AI.

Other transversal standards can also be used to strengthen the audit and traceability of AI systems, especially when applied in complex or regulated contexts. Furthermore, they have the advantage of greater maturity in the market. Such is the case for ISO/IEC 27001 (Information Security), ISO 9001 (Quality Management), COSO (Internal Control Framework), the IEEE 7000 Series (Ethical Standards for Autonomous Systems), or ISAE 3000 (Audit of Non-Financial Information), among others.

Governance of Generative Models and LLMs

12.

AI has become established as a revolutionary technology and the rise of generative AI and large language models (LLMs) like GPT, Claude AI and Gemini has marked a turning point, offering extraordinary versatility and power. However, their growing complexity and massive adoption pose major challenges regarding the interpretability of their decisions, data privacy, security, opacity and the need for solid ethical and regulatory frameworks. Proper governance is crucial to ensure the responsible and reliable development and use of these technologies. This chapter delves further into what was seen in chapter 5, AI Governance, but focuses on the specificities of this type of system.

12.1. Specific Challenges and Risks of Generative Models and LLMs

As developed in the chapter on AI as a Strategic and Operational Risk and also in the chapter on Risk Management and Evaluation in AI, generative models and LLMs present a series of specific risks that must be considered within the global AI risk management framework.

Among the most recognised risks are deepfakes, prompt injection, data poisoning, leaks of sensitive information, model theft, algorithmic discrimination...

The identification, analysis and treatment of these risks are addressed in detail in the chapters dedicated to risk in this guide and in the corresponding reference and regulatory frameworks (ISO/IEC 27005, NIST AI RMF and RIA).

12.2. Regulatory Framework and LLM Governance

Since the massive emergence of LLMs, different opinions have arisen regarding the need for regulation. In Europe, the RIA stands out, which classifies the use of some LLMs as Unacceptable Risk (Prohibited AI Practices) or as High-Risk AI Systems (as seen later in this guide) and establishes specific obligations for GPAI, defined as models trained with a large volume of data using large-scale self-supervision, demonstrating a high degree of generality and able to perform a wide variety of tasks. For those GPAI that the European Commission classifies as systemic risk (according to set criteria, such as number of parameters, data quality, amount of computation used to train the model, input/output modalities and the number of registered end users), additional requirements will apply.

Specific obligations for GPAI providers include:

→ Ensure **transparency** of the models.
→ Ensure **verifiability of the data** used for training.
→ Provide **explainability** of their operations and decisions.
→ **Respect copyright** and provide detailed summaries of training data.

The application of these rules will be staggered, with governance rules and obligations for GPAI applicable twelve months after the Regulation comes into force (August 2025). Failure to comply with these obligations can lead to significant fines, both in amount and/or as a percentage of global annual turnover.

As indicated in chapter 9, the role that regulates and supervises the implementation of the RIA in Spain is the AESIA.

In this context, the governance of LLMs must be a comprehensive process that involves different roles within the organisation, such as Chief AI Officer (CAIO), Chief Information Security Officer (CISO), Chief Data Officer (CDO), Data Protection Officer (DPO) and the Legal and Compliance team. Collaboration and coordination between these profiles is essential to ensure regulatory compliance, risk management and the ethical use of AI.

12.3. Recommendations and Practical Measures for LLM Governance

Implementing a robust governance framework for generative models and LLMs is essential to mitigate risks and responsibly leverage their benefits.

- Establishment of an AI Usage Policy
- Management of Specific Risks
- Data Quality and Governance of Data
- Transparency and Explainability
- Human Oversight
- Cybersecurity
- Intellectual Property
- Training and Awareness
- Impact Assessment and Regulatory Compliance
- Audit and Continuous Monitoring

01.

Establishment of an AI Usage Policy:

- The organisation must define an internal policy that regulates the use of AI systems, including LLMs, aligned with the company's strategy and ethical values. Acceptable and prohibited uses should be differentiated.
- It is crucial to keep an internal record of the AI solutions used, both external and developed in-house.

02.

Specific Risk Management:

- Implement a risk management system to manage risks that may affect security, fundamental rights and regulatory compliance throughout the LLM lifecycle.
- Carry out continuous risk assessments to identify and address known and foreseeable risks, as well as those that may arise from reasonably foreseeable misuse of the system.
- Use a governance framework to understand and manage threats, including model validity, reliability, security and resilience.

03.

Data Quality and Governance:

- Ensure that training, validation and test datasets undergo appropriate data governance and management practices.
- Establish internal procedures to ensure the accuracy, integrity, reliability, veracity, update and suitability of the data, as well as mechanisms to analyse, measure and detect possible imbalances and biases.
- Define the source of data, justify its choice and establish the legitimate basis for the use of personal data, especially if special categories are used.
- Implement minimisation, pseudonymisation and sensitive data protection techniques in all phases.

04.

Transparency and Explainability:

- LLM systems should be designed to provide a sufficient level of transparency to enable those responsible for their deployment to interpret and correctly use their results.
- The provider must supply clear technical documentation and support to understand the logic, operation and limitations of the LLM, as well as the principles and methods used for decision-making.
- The information aimed at end users should be intelligible and easily accessible, using clear and simple language and adapted to particular circumstances, especially for vulnerable groups.

05.

Human Oversight

- The system must allow for effective human oversight, implying an adequate human-machine interface and intervention mechanisms enabling operators to detect and mitigate potential risks or anomalies.
- Oversight must be performed by persons with the necessary competence, training and authority, ensuring meaningful and not merely symbolic action in decision-making.

06.

Cybersecurity:

- Strengthen cybersecurity measures to protect LLMs against specific threats such as prompt injection, data poisoning and model theft.
- Implement logging systems to ensure traceability and audit the AI's performance, detecting risk situations.
- Conduct regular security audits including vulnerability analysis, regulatory compliance and system resilience.

07.

Intellectual Property:

- Develop internal policies regarding copyrights for the data used in training and the content generated by LLMs.
- Clearly define the rights over supplier and third-party datasets, as well as the rights over creations or innovations resulting from AI.

08.

Training and Awareness:

- Train staff in the responsible use of LLMs, ethical principles, information security and data protection.
- Raise awareness about the risks of leaking confidential or personal information when interacting with public AI systems.

09.

Impact Assessment and Regulatory Compliance:

- Conduct impact assessments for rights and freedoms (EIPD/DPIA) and RIA impact assessments, especially for high-risk LLM systems, to identify and mitigate potential risks.
- Ensure that third-party LLM procurement meets a robust contractual framework that regulates conditions and responsibilities of parties, addressing privacy, security, data quality and intellectual property aspects.

10.

Audit and Continuous Monitoring:

- Establish a continuous monitoring system to detect vulnerabilities or security incidents and ensure compliance with privacy and security standards.
- Conduct regular audits of LLMs to verify operation, accuracy, biases and policy compliance.

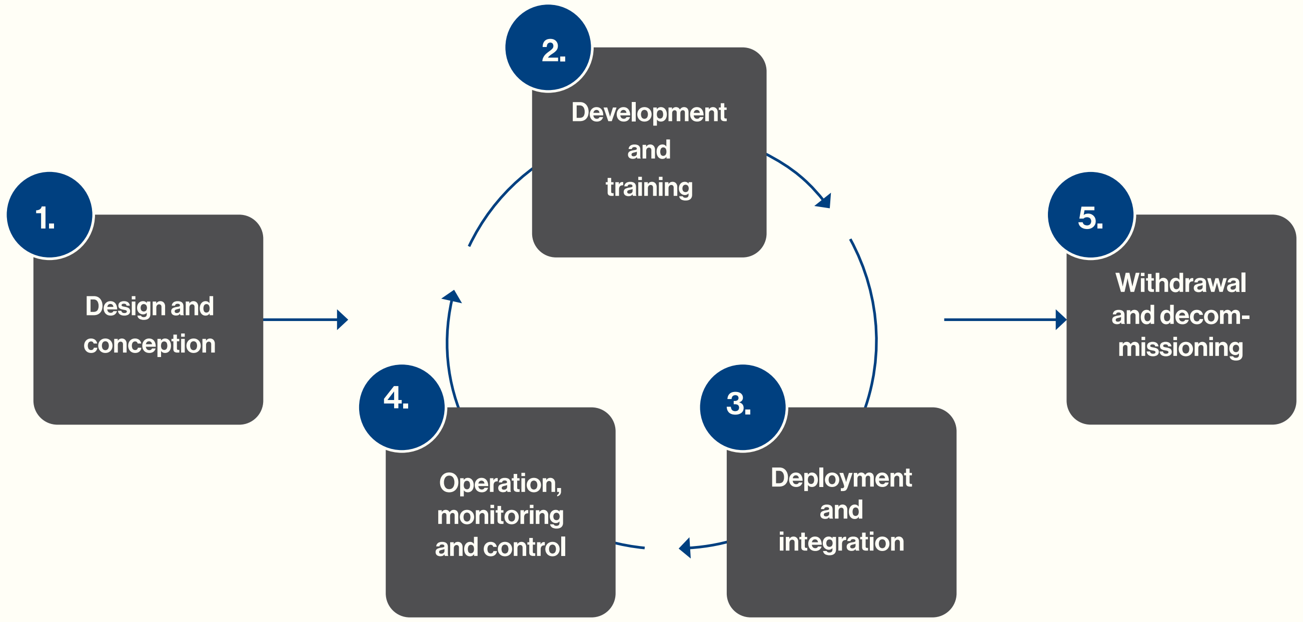
The governance of generative models and LLMs is a dynamic process that requires constant adaptation to technological advances and regulatory evolution. A well-planned and executed approach will provide greater legal certainty, transparency and trust to all parties involved, promoting beneficial and ethical use of AI.

Companies that wish to do so may also become certified under the international standard ISO/IEC 42001:2023, "Information technology -Artificial intelligence- Management system" which specifies the requirements for AI management in organisations throughout the lifecycle.

Security by Design in AI systems 13.

Effective AI governance requires that security is not an afterthought, but a fundamental pillar integrated at every stage of the systems' lifecycle. A Security by Design approach is essential for proactively managing risks, ensuring resilience and building trustworthy AI. Many AI implementations fail due to deficiencies in governance, data quality, or the lack of a secure technological foundation.

Below is the Secure Lifecycle of AI-based Solutions, detailing the critical security controls and considerations across the five key phases of the AI solution lifecycle, from its conception to decommissioning.


1.
Design and conception
2.
Development and training
3.
Deployment and integration
4.
Operation, monitoring and control
5.
Withdrawal and decom-missioning

Phase 1: Design and Conception

In this phase, the foundations of the system's security are established. The decisions made here will determine the robustness and resilience of the final solution. The objective is to anticipate and plan for risk mitigation before writing a single line of code.

- → **Risk Assessment and Threat Modeling:** identify possible AI-specific attack vectors, such as data poisoning, model evasion, or inference attacks. Use frameworks such as MITRE ATLAS to structure this analysis, as previously explained in earlier chapters.
- → **Definition of Security and Privacy Requirements:** incorporate security as a non-functional requirement from the beginning. This includes applying the principles of Privacy by Design and by Default, such as data minimisation and pseudonymisation , as required by the GDPR.
- → **Ethical and Fundamental Rights Impact Analysis:** assess how the system may affect people's rights, considering fairness, non-discrimination and human autonomy, as foreseen by the RIA, as mentioned throughout this guide.
- → **Selection of Secure Architecture:** design an architecture that incorporates native security controls, such as network segmentation, data encryption at rest and in transit and robust identity and access management.

Phase 2: Development and Training

The development phase is where data becomes the fuel for the model. Security at this stage focuses on protecting the integrity of the data and the model itself during its construction.

- → **Data Supply Chain Governance and Security (DataOps):** ensure the quality, integrity and provenance of training, validation and test data. It is crucial to document the lineage of the data to guarantee traceability and verify its lawful use and respect for intellectual property.
- → **Protection against Data Poisoning:** implement controls to detect and prevent the injection of malicious data into the training sets, which could corrupt model behavior.
- → **Secure Development Environment[33]:** use isolated development and experimentation environments to prevent testing from affecting production systems. Manage software dependencies and open-source libraries securely to prevent vulnerabilities.
- → **Ongoing Documentation:** maintain thorough technical documentation, including datasheets for datasets and model cards (annex 3) for models, describing their operation, limitations and performance metrics.

[33] ISMS Forum. (2024). 2nd Edition of the DevSecOps Guide. https://www.ismsforum.es/ficheros/descargas/iiedicionguiadevsecopsv1746557372.pdf

Phase 3: Deployment and Integration

Deployment is the process of integrating the AI model into a production environment for practical use. Security in this phase is important to protect the system in a real operating environment.

→ **Secure Infrastructure:** ensure that the underlying infrastructure (servers, containers, cloud platforms) is properly configured and secured. Following hardening guides, for example, those published by NIST for each system and applying principles of least privilege.

→ **API and Endpoint Protection:** the application programming interfaces (APIs) through which the model is consumed are a main attack vector. They should be protected against injection attacks (prompt injection), abuse and unauthorised access.

→ **Secure Output Management (Insecure Output Handling):** always validate and sanitise the model's outputs before they are processed by other systems or shown to users, to prevent attacks such as Cross-Site Scripting (XSS) if the model generates code.

→ **Pre-Production Security Testing:** conduct penetration tests and security audits specific to the AI solution before its launch, simulating adversarial attacks to identify weaknesses.

Phase 4: Operation, Monitoring and Control

Once deployed, an AI system requires continuous oversight to ensure its behavior remains safe, reliable and aligned with intended objectives.

→ **Continuous Model Monitoring:** monitor performance, accuracy and fairness metrics to detect model drift or concept drift that could introduce risks or biases.

→ **Logging and Traceability:** implement a robust logging system that captures key events such as decisions made, human interventions and security alerts. These records are fundamental for audit, accountability and forensic analysis.

→ **Effective Human Oversight:** enable mechanisms allowing for meaningful human intervention, especially in high-risk systems. Operators must have the capacity and authority to override or correct the system's decisions when necessary. (annex 4)

→ **Incident Response Plan:** have a specific protocol for managing security incidents related to AI, defining steps for detection, containment, eradication and recovery.

Phase 5: Withdrawal and Decommissioning

The lifecycle of an AI system concludes with its withdrawal. This phase must be managed securely to avoid exposure of sensitive data and intellectual property.

→ **Secure Archiving of Models and Data:** securely archive models, training data and associated documentation, according to the organisation's retention policies and regulatory requirements.

→ **Secure Data Deletion:** securely delete personal and confidential data from production systems, in accordance with the GDPR's data retention limitation principle.

→ **Revocation of Access:** disable all APIs, service accounts and credentials associated with the retired system to prevent unauthorised access.

→ **Process Documentation:** log all actions taken during the withdrawal phase to ensure complete traceability at the end of the system's life. (annex 5)

Conclusions and Final Recommendations

14.

Artificial Intelligence (AI) has established itself as a driving force for profound transformation in organisations, but also as a vector of strategic, operational, ethical and regulatory risk. The journey outlined in this guide shows that AI governance is no longer an option, but a legal, competitive and social imperative. The European AI Regulation (RIA) marks a turning point, demanding a shift from declarative principles and ethics to the effective and auditable operationalization of AI throughout all its phases.

From Ethics to Legal Obligation: **RIA as a Catalyst**

The entry into force of the RIA turns risk management and AI governance into a structured obligation, with deadlines and penalties comparable to the GDPR. Organisations must identify, classify and manage their AI systems according to the level of risk, anticipating regulatory requirements and avoiding potentially very high fines. Compliance is no longer just a matter of reputation, but of the viability and sustainability of the business.

As evidence that there is still ground to cover in this regard, according to the latest Cisco AI Readiness Index 2025, only:

23%

of the companies surveyed have their Governance processes ready to face the new challenges of AI.

New Risk Vectors and the Need for **Specialised Frameworks and Countermeasures**

The emergence of generative AI and large language models (LLMs) has multiplied the risks: novel attacks (prompt injection, data poisoning, model theft), misinformation generation, loss of control over data and unprecedented challenges in intellectual property.

To address these risks, we cannot rely on the tools we had until now: we need a specific governance framework, as well as specialised countermeasures and security models that constantly adapt to technological and regulatory evolution.

Governance by Design: From Reactive to Proactive Approaches

AI risk management must be proactive and integrated. The concept of AI Governance by Design (AIGD) involves incorporating ethical, legal, security and social considerations from the outset in the AI lifecycle. This requires organisational maturity, robust GRC structures and a clear definition of roles such as CAIO, CISO, CDO and DPO, as well as the creation of multidisciplinary AI committees.

Operationalisation: From Strategy to Practice

The key to success lies in translating frameworks and principles into effective policies, processes and controls. This involves:

→ Formalising an AI governance structure with a centralised inventory of systems, internal policies, defined responsibilities and clear procedures.

→ Developing a robust AI security strategy based on Zero Trust, segmentation, continuous monitoring and agile incident response.

→ Implementing traceability and auditing in all systems, ensuring accountability and the ability to demonstrate compliance.

→ Strengthening third-party management, requiring contractual guarantees for privacy, security, explainability and regulatory compliance.

→ Fostering an organisational culture of awareness, with specific and transversal training in AI risks, ethics and best practices.

Only if we are able to establish these policies, processes and controls, will we be able to effectively measure the impact of AI in our organisations—something that, according to the aforementioned Cisco AI Readiness Index, only 32% of companies are prepared to do.

Data **Protection, Rights and Sustainability**

The protection of personal data and privacy must be integrated from the design stage, aligning GDPR compliance with RIA obligations. Impact assessment (EIPD/DPIA and FRIA) and data minimisation are essential to anticipate and mitigate risks. In addition, the sustainability and environmental impact of AI should be considered in decision-making, aligning innovation with sustainable development goals.

Following the data presented in the Cisco AI Readiness Index, only 34% of organisations have integrated new AI risks into their data protection policies; this is the area that most reflects the degree of maturity in AI adoption.

93%

of companies that have an advanced level of AI adoption in their business already have the new AI risks integrated into their data protection policies

Human Oversight, Explainability and Trust

Effective human oversight, trust and explainability are pillars to generate trust and ensure control over AI systems, especially in high-risk contexts. Organisational and technical traceability, together with exhaustive documentation, enable auditing and justification of decisions, facilitating accountability to users, regulators and society.

Sectoral Adaptation and Organisational Size

Conclusions and recommendations must be adapted to the sector and size of each organisation. Critical sectors such as health, finance, or infrastructure require special attention, while SMEs must prioritise visibility and control over the use of AI, especially in managing suppliers and external tools.

Towards Trustworthy, Ethical and Competitive AI

AI will only generate sustainable value if it is developed and used within a framework of trust, legality and responsibility. Effective AI governance is the best guarantee to harness its potential, mitigate its risks and strengthen the competitiveness and reputation of organisations.

Final Recommendations

Anticipate and comply with RIA

Begin adapting to the European Artificial Intelligence Regulation as soon as possible by identifying systems, roles and obligations

Integrate governance from the design stage

Adopt an AI Governance by Design approach, embedding ethics, security and data protection from the conception of each system.

Promote training and organisational culture

Train all profiles involved in AI, from management to end users, to reduce the trust gap and improve risk management.

Review and adapt contracts with third parties

Require guarantees of compliance, security and explainability from AI providers.

Strengthen cybersecurity and risk management

Applying specialised frameworks and specific controls for generative AI, LLMs and Agentic AI; advancing automation maturity without losing human supervision. Indeed, today only 34% of organisations trust the resilience of their current cybersecurity infrastructure to address AI risks

Ensure traceability and auditing

Maintain complete and accessible records that enable demonstration of compliance and facilitate supervision.

Evaluate the social,ethical and environmental impact

Incorporate sustainability and rights protection in AI decision-making.

Ultimately, AI governance is a dynamic and cross-functional process that requires leadership, strategic vision and adaptability. Organisations that integrate these principles will be better prepared to face the challenges and seize the opportunities of the new digital era.

In this context, it is essential to strengthen ongoing training and reskilling plans for internal talent, equipping teams with advanced competencies for the responsible, safe and strategic use of AI tools. This training not only reduces the technological gap, but also acts as a catalyst for ethical, efficient and business-aligned adoption.

Indeed, concluding with the key lessons of the Cisco AI Readiness Index, organisations that demonstrate greater maturity in AI governance are more likely to harness its potential across these three dimensions:

1. **Identifying use cases** in their business for AI application.
2. **Measuring the economic impact** of their AI investments.
3. **Reporting profitability, productivity and innovation gains** as a result of applying AI to their businesses.

ANNEXES

Guides and Practical Resources

The following guides and practical resources are available for download in editable format at the following link: https://www.ismsforum.es/ficheros/descargas/descargables-gobierno-de-ia1762519105.docx

Annex 1

Use Cases and Common Coordination Challenges

Use Case 1:
Implementation of a generative model for customer service

Challenge: Validate regulatory compliance, protect personal data and ensure security.

Coordination:

- CAIO leads; CISO validates secure architecture
- DPO reviews privacy implications
- Legal reviews usage licenses
- CDO ensures data quality.

Use Case 2:
Automation of credit decisions

Challenge: Ensure traceability of automated decisions and prevent algorithmic bias.

Coordination:

- **CAIO** and **CDO** collaborate on model transparency
- **Legal** and **DPO** evaluate compliance with explainability requirements

Explainability (or algorithmic explainability) is the ability of an artificial intelligence system to provide understandable reasons for how it arrived at a decision. This is especially important when AI makes decisions that affect people, such as:

- **Approving or denying a loan**
- **Selecting candidates for a job**
- **Determining the price of an insurance policy**

Use Case 3:
Integration of AI in internal HR processes

Challenge: Reputational risks due to bias or discriminatory decisions.

Coordination:

- CAIO and CIO oversee architecture
- DPO and Legal evaluate compatibility with equality and non-discrimination principles

Annex 2
Key Questions for Audit

Security in the system lifecycle

- **Observation of the "Security by Design" principle:**
  - → Is there a formal procedure that includes traceability as part of the security requirements in the design of all elements that make up the system?
- **Version and change control:**
  - → Are model updates documented and recorded?
- **Dependency and library management:**
  - → Are external components audited to detect vulnerabilities?
- **Recording human decisions:**
  - → Are human interventions in the design and tuning of the system tracked?

Data security

- **Data origin and quality:**
  - → Do they come from reliable sources?
  - → Have they been correctly anonymised or tokenized?
- **Sensitive data protection:**
  - → Are measures applied for encryption or masking?
  - → Has the use of synthetic data been considered in non-production environments?
  - → Is correct access control in place?
- **Data traceability:**
  - → Can the flow of data from their acquisition in repositories, databases and other AI systems to their consumption by the user area be reconstructed?

Model robustness

- **Resistance to adversarial inputs:**
  - → Can the model be manipulated maliciously?
- **Evaluation of "model extraction" or "data poisoning" attacks:**
  - → Are there mechanisms to detect attempts to replicate or steal the AI model, or to maliciously manipulate its training data to introduce errors, biases, or vulnerabilities?
- **Stress testing and simulations:**
  - → How does the system respond to extreme or unexpected conditions?
  - → Are exceptions and errors managed in a way that they can be escalated to security events if necessary?

Hardware and software platform security

- **Physical infrastructure:**
  - → What security measures have been considered for the protection of servers, data centers, Edge devices, etc.?
  - → How is information about these measures delivered?
- **Execution environment:**
  - → What reporting strategy has been considered to gather all information relating to the protection and security of the base platform, operating system, containers, virtualization, communications, etc.?
- **Dependencies and libraries:**
  - → Are audit mechanisms for internal and external components enabled?
  - → Are the mechanisms used to verify the integrity of the components being monitored?
  - → How is the configuration of systems managed?
- **Patch and update management:**
  - → Is version control correctly implemented and managed?
  - → Is there a plan for vulnerability management and remediation?
- **API and communication channel security:**
  - → Are there mechanisms in place to protect against injections, unauthorised access and data leaks?
  - → Are secure and encrypted means used for system and application access?
  - → Are both covered by a sufficient level of traceability to debug any event?

Production monitoring and supervision

- **Real-time auditing:**
  - → Are system interactions and decisions in production recorded?
- **Alerts and response mechanisms:**
  - → Is there a protocol for anomalous behaviors or security breaches?
- **Misuse evaluation:**
  - → Are attempts at manipulation by users or external agents detected?

Authentication and authorisation management of identities and access

- **Authentication of users, agents, or entities interacting with the AI system:**
  - → Are robust authentication mechanisms in place (MFA, certificates, tokens)?
  - → Is authorisation based on roles (RBAC), attributes (ABAC), or dynamic policies?
- **Controls to prevent, detect and respond to misuse of privileged identities or impersonation within the system.**
  - → Is the use of service accounts and machine identities audited?
  - → Is there a traceability process ensuring identification of the responsible party in administrative or automated accesses?
- **Model and data access control:**
  - → Who can modify, query, or train the system?
- **Security in interaction interfaces (APIs, prompts, etc.):**
  - → Are they protected against malicious injections or abuse?
- **Access and action logging:**
  - → Are administrative and technical accesses audited?
  - → Can logs and records be incorporated into a forensic investigation if necessary?

Annex 3

Example of a "Model Card" for a Fictional AI System for Creditworthiness Assessment

General information	
Field	**Description**
Model Name	**Creditworthiness Scoring** **(SC-2025-v1)**
Model Version	**1.0 (Release: October 2025)**
Model Type	**Supervised Learning (Binary Classification)**
Algorithm	**Gradient Boosting (XGBoost)**
Internal Owner	**Risk Department**
Technical Contact	**Ejemplo@ISMSForum.es**
Risk Classification	**High Risk (according to the EU AI Regulation, Annex III)**

01 General Description and Purpose of the Model

This model is designed to assess the probability that a loan applicant (natural person) will default within the next 24 months. The system generates a creditworthiness score which serves as a supporting element in the decision-making process for granting consumer loans.

Intended Use:

→ Evaluate applications for new consumer loans.

→ Serve as one of the support tools for risk analysts in the decision-making process. It is not designed to make decisions 100% autonomously.

→ Segment applicants into risk profiles (low, medium, high) to determine loan conditions.

Unintended (Prohibited) Uses:

→ Make credit denial decisions in a fully automated manner without the possibility of human review.

→ Use the score for any other purpose than creditworthiness evaluation (e.g. marketing, employee evaluation).

→ Apply the model to applicants who are not natural persons.

02 Data Used

The quality, governance and integrity of the data are critical for the performance and fairness of the model, as emphasised in the principles of data governance and privacy.

Training Data:

- **Source:** Anonymised internal historical credit operations data from 2018 to 2023. No external data sources are used for training.
- **Volume:** 150,000 customer records.
- **Variables Used (Features)**:
  - → **Sociodemographic Information:** age range, employment status (employee, self-employed, etc.), employment duration. Note: No special category data such as ethnic origin or gender is used to avoid discriminatory biases.
  - → **Financial Information:** level of net monthly income, current debt level, repayment history of previous loans, type of employment contract.
  - → **Operation Information:** amount of requested loan, repayment period.

- **Data Preprocessing:**
  - → Techniques of **pseudonymisation** were applied to protect individuals' identities.
  - → An analysis and mitigation of biases in the training data was carried out to ensure the representativeness of different demographic groups.

Evaluation Data (Test):

- **Source:** independent dataset (20% of total) extracted from the same historical source (hold-out dataset), not used during training to ensure objective evaluation of performance.

03 Performance Metrics

Metrics have been selected to evaluate not only the overall accuracy of the model, but also its reliability and fairness.

Accuracy Metrics:

Metric	Value	Description
Accuracy (Overall Precision)	92.5%	Percentage of correct predictions over the total.
Precision	88.0%	Of all the cases the model predicted as "default", 88% actually were defaults.
Recall (Sensitivity)	85.0%	The model correctly identified 85% of all actual "default" cases.
AUC-ROC	0.94	Measures the model's ability to distinguish between classes (solvent vs. default). A value close to 1 indicates excellent performance.

Equity and Bias Metrics: analyses were performed to measure disparate impact between different demographic groups (segmented by age range and employment status).

Equity Metric	Result	Conclusion
Demographic Parity	Difference < 2% in the predicted "default" rate between groups.	No significant adverse impact detected.
Equal Opportunity	Difference < 3% in the true positive rate between groups.	The model identifies applicants at risk of default equitably among the analysed groups.

04 Limitations and Ethical Considerations

Transparency about model limitations is a fundamental requirement for responsible use.

- **Correlation does not imply Causation:** the model identifies patterns and correlations in historical data, but does not explain the underlying causes of a person's financial behavior. Final decisions should be contextualised by a human analyst.
- **Risk of "Drift":** the performance of the model may degrade over time if economic and social conditions change (concept drift). A continuous monitoring plan has been implemented to detect these deviations and plan for model retraining when necessary.
- **Human Oversight:** this system is classified as High Risk and therefore all automated decisions that have a significant impact (such as a pre-denial) must be reviewed by a qualified analyst before being communicated to the customer. Mechanisms have been enabled for analysts to override or modify the system's decision.
- **Explainability:** although the XGBoost model is complex, explainability techniques (such as SHAP - SHapley Additive exPlanations) have been implemented to generate a comprehensible justification for each individual prediction, which will be available to internal analysts and to the customer upon request, in compliance with the right to an explanation about automated decisions.

05 Security and Robustness Considerations

The system was developed following security-by-design principles to ensure its technical robustness and resilience against attacks.

→ **Robustness:** stress tests with anomalous data and adversarial scenarios were conducted to evaluate model stability.

→ **Security:** the infrastructure hosting the model complies with internal cybersecurity policies, including strict access controls, data encryption and API activity monitoring.

→ **Traceability:** all predictions generated by the model are recorded in an immutable log system that includes the model version used, input data (pseudonymised) and the resulting score. This ensures auditability and accountability.

Annex 4

Registration Form for Effective Human Oversight of a Fictional AI System

This form must be completed by an analyst each time a decision generated by a high-risk AI system is reviewed. Its purpose is to create an auditable record that demonstrates a significant and not merely symbolic human intervention.

Effective Human Supervision Record	
Section 1: Case Identification	
Case ID:	**SOL-20251027-8A34F**
AI Model ID:	**IA-002 (SC-2025-v1)**
Model Decision Date/Time:	**27/10/2025 - 11:34 AM**
Reviewing Analyst (Name & ID):	**XXXX**
Human Review Date/Time:	**27/10/2025 - 02:15 PM**
Reason for the Review:	☑ **Standard Protocol (High-Impact Decision)** ☐ **Client Appeal Request** ☐ **Model Monitoring Alert** ☐ **Other (specify):**

Effective Human Oversight Record	
Section 2: Summary of Automated Decision	
Model Result	**Justification (Main Factors reported by AI)**
Suggested Decision: **Credit Denied Score: 350 / 1000 Model Confidence: 91%**	**1. (High Impact) Employment seniority: 8 months.** **2. (Medium Impact) Debt ratio: 45%.** **3. (Low Impact) Absence of internal credit history.**

Effective Human Supervision Record		
Section 3: Human Verification Process		
Action	**Verified**	**Analyst's Observations**
1. Input Data Verification	☑	**The data in the system matches the applicant's documentation.**
2. Model Logic Analysis	☑	**The model's justification is consistent with risk policies.**
3. Additional Context Collection	☑	**The applicant has been contacted and supplemental documentation reviewed.**
4. Findings and Additional Context (information not considered by the model):	• **The applicant has presented a permanent employment contract in a high employability sector (Software Engineering), reducing the risk associated with "low seniority".** • **The pre-existing debt corresponds to a student loan, which has a different risk profile than a consumer credit.**	

Effective Human Supervision Record	
Section 4: Analyst's Final Decision and Justification	
Final Decision:	□ Uphold the model's decision ☑ Override the model's decision □ Modify the model's decision
Specific Action Taken	Approve credit with modified conditions: 80% of the requested amount is approved, with an increase of 0.5 points in the interest rate to reflect residual risk
Mandatory Human Decision Justification	The model's decision, although technically correct with the available data, does not consider key qualitative factors. The job stability confirmed by the permanent contract and the nature of previous indebtedness (educational vs. consumer) substantially reduce the real risk profile of the applicant. The override of the automated recommendation and subsequent partial approval are based on professional judgment integrating this additional context, aligning with prudent and personalised risk management.

Annex 5

Registration Form for the Withdrawal of a Fictional AI System

This form is used to document and verify all actions carried out during the withdrawal process (decommissioning) of an AI system. Its objective is to ensure that the withdrawal is carried out safely, in a controlled manner and in compliance with internal policies and applicable regulations, guaranteeing complete traceability of the system's end of life.

AI System Withdrawal Record			
Document control			
Record ID:	**RET-20260115-IA002**	**Start Date:**	**15/01/2026**
Process Owner:	**xxx**	**End Date:**	**20/01/2026**
Status:	**COMPLETED**		

System Identification and Withdrawal Reason	
Field	**Description**
Model Inventory ID:	**IA-002**
Model Name:	**Credit Solvency Scoring (SC-2025-v1)**
Production Start Date:	
Withdrawal Justification:	**Replaced by a new version. The SC-2025-v1 model will be replaced by version SC-2026-v2, which has shown superior performance (4% improvement in AUC-ROC) and greater fairness in validation tests. The withdrawal of version v1 is necessary to avoid coexistence of versions and ensure consistency in risk decisions.**
Withdrawal Authorisation:	**Approved by the AI Governance Committee in the session dated 10/01/2026 (Minutes CGIA-2026-01).**

AI System Withdrawal Record					
Category	Specific Action	Responsible	Date	Verified	Observations / Evidence Path
1. Service Deactivation	Disable the model API endpoint in the production environment.	MLOps Team		Yes	The endpoint api.ejemplo.com/scoring/v1 is no longer active. Verified through connection tests.
	Redirect traffic from the old endpoint to a "service obsolete" message.	Infrastructure Team		Yes	Configuration updated on the load balancer.
2. Asset Archiving	Archive the final model source code (version 1.0) in the archive repository.	Data Science Team		Yes	Repository: git-archive/models/SC-2025-v1. Final commit: f4a2b1c.
	Archive the (anonymised) training and validation datasets used.	Data Team		Yes	Location: data-archive/datasets/credit-scoring/2025_v1. Restricted access.
	Archive the Model Card and all associated technical documentation.	Governance Team		Yes	Documentation archived in the document management system, ref: DOC-IA-002-FINAL.
3. Data Deletion	Securely delete the training data and model from the production servers.	MLOps Team		Yes	Secure deletion process executed on production servers prod-ml-01 and prod-ml-02. Deletion certificate attached.
	Purge the model execution logs containing personal data, according to the retention policy (retain only aggregated metadata for historical analysis).	Security Team		Yes	Purge script executed. Logs older than 30 days have been deleted.
4. Access Management	Revoke all API keys and service credentials associated with model v1.	Security Team		Yes	API keys API_KEY_SC_V1_... revoked.
	Remove the specific access roles for managing model v1 in the identity management system.	IT Team		Yes	Roles Admin-SC-v1 and User-SC-v1 removed.
5. Communication and Closure	Notify all development and business teams that used model v1 about its permanent withdrawal.	Model Owner		Yes	Email sent to dev-teams and risk-analysts distribution lists.
	Update the model status to "Withdrawn" in the AI Systems Inventory.	Governance Team		Yes	Inventory updated. Ref: IA-002.

Final Withdrawal Verification and Approval

By signing this document, the undersigned responsible individuals certify that all actions described in Section 2 have been carried out in accordance with the established procedures and that the AI system SC-2025-v1 has been withdrawn safely and completely.

AI Withdrawal Process Owner:

- Name: XXXX
- Position: Head of MLOps
- Signature: XXXX
- Date: XXXX

Final Governance Approval:

- Name: XXX
- Position: Director of Data and AI Governance
- Signature: XXXX
- Date: XXXX

Annex 6

Key Questions for **Supplier Evaluation**

The dependence on external suppliers for AI solutions (more than 65% of organisations according to the 2023 survey) requires a robust due diligence process. This checklist helps evaluate and manage third-party risk.

AI System Withdrawal Record

Category	Aspect to Evaluate	Key Questions and Required Evidence
1. Regulatory Compliance (RIA & GDPR)	RIA Classification and Documentation	• How do you classify your AI system according to the RIA (high risk, etc.)? • Evidence: request the EU Declaration of Conformity and the full technical documentation required for high-risk systems
	Data Governance and Privacy	• How do you ensure the lawfulness of the training data and respect for intellectual property? • Will our data be used to retrain your models? Is there an opt-out option? • Where is the data hosted? What guarantees do you offer for international data transfers according to the GDPR? • Evidence: specific contractual clauses, certifications (e.g., ISO 27701), privacy audit results.
2. Information Security (Cybersecurity)	Model Robustness	• What measures do you apply to mitigate AI-specific risks such as prompt injection, data poisoning, or model extraction? • Evidence: red teaming test results, pentesting reports, security certifications (ISO 27001, SOC 2).
	Incident Management	• What is your incident response protocol? Within what timeframe do you notify breaches? • Evidence: Incident Response Plan, contractual clauses with appropriate notification SLAs defined (e.g. 48 hours).
3. Transparency and Accountability of Accounts	Explainability	• What level of explainability do you offer about model decisions? Is it sufficient for us to comply with our own transparency obligations? • Evidence: model cards, datasheets, technical documentation about model logic.
	Traceability and Auditability	• Does the system generate auditable logs for all decisions and critical operations? • Do you allow audits by us or an independent third party? • Evidence: sample logs, explicit audit rights in the contract
4. Contractual and Operational Aspects and Operational	Liability and SLA	• How is liability delimited for damages caused by erroneous or harmful outcomes? • What SLAs do you guarantee in terms of model availability, latency and accuracy? • Evidence: liability limitation clauses, detailed service level agreements
	Control over Embedded AI	• Does your service integrate third-party AI components? How do you manage the risk of that supply chain? • -Evidence: statement about embedded AI use, clauses requiring prior authorisation to integrate AI systems with risk.

Annex 7

Risk and Compliance **Assessment Templates**

Template 1: AI Impact and Risk Assessment

Register			
Assessment ID:	EIR-IA-[YYYYMMDD]-[SYSTEM_ID]	**Date:**	[Date created]
Author(s):	[Names and Departments]	**Reviewed by:**	[Name, AI Governance Committee]
Status:	DRAFT / UNDER REVIEW / APPROVED	**Version:**	1.0

System Description and Classification	
Field	**Description**
System Inventory ID:	**[ID assigned in the AI inventory]**
System Name:	**[Descriptive name of the AI system]**
Business Owner:	**[Department responsible for the initiative]**
Organization Role:	□ **Provider** □ **Deployment Responsible**
Description and Purpose:	**Describe in detail the system's purpose, the business problem it solves, its general operation, and the explicit goals it was designed for.**
Risk Classification (AIA):	□ **Unacceptable Risk (Prohibited)** □ **High Risk** □ **Limited Risk (Transparency Obligations)** □ **Minimal Risk** **Justification: attach the result of the “Risk Classification Questionnaire” and justify the final classification.**

Data Protection Impact Assessment (DPIA)

(Mandatory if the system processes personal data and is likely to entail a high risk for the rights and freedoms of natural persons)

Nature, Scope and Context of Processing:

- **Personal Data Processed:** list the data categories (identifiers, financial, etc.) and whether special categories are included.
- **Data Processing Lifecycle:** describe how data is processed in each phase: training, validation, inference and withdrawal.
- **Legal Basis (Lawfulness):** indicate the legal basis for processing for each purpose (consent, legitimate interest, legal obligation, etc.)

Assessment of Necessity and Proportionality:

- **Data Minimisation:** Are only the data strictly necessary for the purpose collected and processed? Justify the necessity of each data category.
- **Purpose Limitation:** Are there technical and organisational measures in place to ensure data is not used for purposes incompatible with the original ones?

Identification and Assessment of Risks to Rights and Freedoms:

ID	Potential Risk Description	Affected Fundamental Right(s)	Probability (Low/Medium/High)	Impact (Severity) (Low/Medium/High/Critical)
R-001	E.g.: Algorithmic discrimination in candidate selection.	E.g.: Right to non-discrimination, Right to work.	E.g.: Medium	E.g.: Critical
R-002	E.g.: Re-identification of individuals from supposedly anonymised data.	E.g.: Right to privacy, Data protection.	E.g.: Low	E.g.: High
R-003	E.g.: Incorrect automated decisions (false positives/ negatives) that legally affect the individual.	E.g.: Right to effective judicial protection, Right to data protection (accuracy).	E.g.: Medium	E.g.: High
R-004	E.g.: Unauthorised access to special categories of data (health, biometric).	E.g.: Right to privacy, Protection of special data.	E.g.: Low	E.g.: Critical

Planned Mitigation Measures

- **Technical:** list measures such as pseudonymisation , encryption, differential privacy techniques, etc.
- **Organisational:** access controls, retention policies, staff training, etc.

Fundamental Rights Impact Assessment (FRIA)

(Mandatory for High-Risk systems deployed by public bodies or for specific use cases such as credit assessment)

Aspect to Evaluate	Description and Analysis
Affected Groups:	Identify the categories of people who may be affected, paying special attention to vulnerable groups.
Specific Risk of Harm:	Analyse potential risks to fundamental rights such as non-discrimination, human dignity, freedom of expression, right to effective judicial protection, etc.
Human Supervision Measures:	Describe the human intervention mechanisms implemented: are they preventive or after the fact? What authority does the supervisor have to override the decision? Is the supervision meaningful and not symbolic?
Redress Mechanisms:	Describe the channels and procedures so that affected individuals can express their point of view, challenge a decision and request redress.

Security and Operational Risk Assessment (generic example)

Risk Category	Specific Risk	Probability	Impact	Proposed Mitigation Measures
Model Security	Data Poisoning: manipulation of training data to corrupt the model.			
	Evasion Attacks: adversarial inputs designed to deceive the model during inference.			• Adversarial Training. • Input Sanitisation
	Prompt Injection: (For LLMs) Manipulation of inputs to bypass model restrictions.			
Operational Risk	Model Drift: loss of performance due to changes in the data environment.			
Supply Chain	Third-party Provider Vulnerabilities: reliance on a third-party AI service that suffers a security breach.			• Comprehensive provider due diligence (see checklist 13.4). • Security and notification contractual clauses

Conclusion and Action Plan

1. Summary of Residual Risk: after applying mitigation measures, what is the accepted level of residual risk?
2. Declaration of Compliance: Does the system, with the proposed measures, comply with legal requirements and internal policies?
3. Final Decision of the AI Governance Committee:

APPROVE DEPLOYMENT □

APPROVE WITH CONDITIONS (detail below)w □

REJECT (justify below) □

Template 2: AI Compliance Checklist and Audit

General Registry

Field	Description
Inventory System ID:	[ID of the audited system]
System Name:	[Name of the AI system]
Audit Date:	[Date of execution]
Auditor(s):	[Name(s) and Department(s)]
Scope of the Audit:	Periodic compliance review (Annual / Semiannual / Quarterly).

Findings and Action Plan (Examples)

Finding ID	Description of Non-Conformity	Associated Risk	Severity (Critical/High/Medium/Low)
CCA-01	The human supervision records in 3 of the 10 reviewed cases lack a detailed justification for overriding the model's decision.	Non-compliance with "meaningful intervention" requirement of the RIA and lack of traceability.	High
CCA-02	The information provided to the user on the website has not been updated to reflect the latest version of the model.	Lack of transparency with users.	Medium
Finding ID	**Corrective Action**	**Responsible Party**	**Deadline**
CCA-01	1. Review and complete the deficient records. 2. Conduct a refresher training session for all risk analysts on the importance of detailed documentation.		
CCA-02	Update the informational text on the credit simulator's web interface.		

Basic Example Checklist for Compliance and Governance Verification

	Control Point	Status (Compliant / Not Compliant / N/A)	Evidence / Observations
REGULATORY COMPLIANCE			
2.1	RIA: The system's technical documentation is up to date and reflects the current status of the model.		Review of the "Model Card" v1.1, date 09/25/2025.
2.2	RIA: The system's decision traceability logs are accessible and complete for the audited period.		
2.3	GDPR: The Record of Processing Activities (RPA) is up to date for this system.		
2.4	GDPR: Requests to exercise rights (access, rectification, objection to automated decisions) have been handled within the time frame.		
INTERNAL GOVERNANCE AND ETHICS			
2.5	The current use of the system aligns with the purpose approved in the AI-PIA. No unforeseen uses have been detected.		
2.6	(For High Risk) Human supervision logs demonstrate effective and non-symbolic intervention.		
2.7	New bias and fairness tests have been performed. The results are within the accepted thresholds.		Bias re-evaluation report attached.
2.8	The information provided to users about the interaction with AI is clear, visible and up to date.		
MODEL PERFORMANCE			
3.1	The key performance metrics (e.g. Accuracy, AUC) remain above the defined thresholds. No significant model drift has been detected.		Monitoring dashboard, period [date] to [date].
3.2	The performance monitoring and alert system is operational and has functioned properly.		
SECURITY			
3.3	The infrastructure supporting the system is updated and patched according to the vulnerability management policy.		Vulnerability scan report.
3.4	Access controls have been reviewed. No unauthorised accesses to the data or the model have been detected.		Access log review.
3.5	No security incidents related to AI-specific attack vectors (e.g. prompt injection, evasion).		SOC incident log.

Contact us

If you are interested in collaborating with us or require further information about our projects, please write to us at:
proyectos@ismsforum.es