Instituto de
AuditoresInternos
de España

The Institute of
Internal Auditors

THE THOUGHT FACTORY
INSTITUTE OF INTERNAL
AUDITORS OF SPAIN

INTERNAL AUDIT OF ARTIFICIAL INTELLIGENCE APPLIED TO BUSINESS PROCESSES

ABOUT THIS DOCUMENT

This paper was developed by the Thought Factory, The IIA–Spain's think tank for internal auditors. The authors are internal audit practitioners from various industries. The paper has been updated to reflect changes in AI technology since the original publication.

TECHNICAL COMMISSION MEMBERS WHO UPDATED THE DOCUMENT

Daniel Tortosa Illana, ICAEW, ROAC, Telefónica Brasil

Pablo Ausín Sánchez, PMP-PMI, Inditex

Luis Enrique Corredera, Deloitte

José Ignacio Díez Arocena, CIA, CISA, CFE, COSO CI, COSO ERM, CESCOM, Independiente

Javier Escribano Alarcón, CISA, COBIT, ITIL y PMP, Repsol

Andrés Morales Fernández, KPMG

Borja Rioja Mata, Mapfre

Juan José Villar Roldán, Iberdrola

GLOBAL IIA CONTRIBUTORS

Anne Millage, Content Director, IIA Global

George Barham, Dir. Standards & Professional Guidance, CIA, CRMA, CISA, IIA Global

TECHNICAL COMMISSION MEMBERS WHO DEVELOPED THE ORIGINAL DOCUMENT

Daniel Tortosa Illana, ICAEW, ROAC, Telefónica Brasil

Pablo Ausín Sánchez, PMP-PMI, Inditex

Luis Enrique Corredera, Deloitte

José Ignacio Díez Arocena, CIA, CISA, CFE, COSO CI, COSO ERM, CESCOM, INDRA, Independiente

Javier Echeverría Blanco, BBVA

Javier Escribano Alarcón, CISA, COBIT, ITIL y PMP, Repsol

Alejandro Martínez Morillo, CISA, CDPSE, Lead Auditor 27001, PwC

Andrés Morales Fernández, KPMG

Borja Rioja Mata, Mapfre

Jaime Sabau Jiménez, EY

Juan José Villar Roldán, Iberdrola

Contents

INTRODUCTION

Can machines think?

It is the question asked in a paper published in *Mind* journal in 1950 by Alan Mathison Turing and is one of the great questions of the scientific community of the time. In "Computing Machinery and Intelligence," Turing raised this pivotal question that captivated mathematicians and computer scientists as they began exploring the field of **artificial intelligence (AI)**. The concept of machines mimicking human intelligence has progressed significantly, making it possible to address Turing's mid-20th century question and leading to the emergence of modern AI. Despite being an abstract concept, AI is increasingly present in our lives, from the facial recognition of our mobile phones to the voice assistants we frequently use.

To answer the question, the English mathematician indicated it is necessary to precisely define what is meant by "thinking" and what is meant by "machine."

The term *artificial intelligence* was coined by data scientist John McCarthy in 1956, defining it as the science of making machines intelligent, or simply, the methods of making machines that make human decisions to solve problems. AI includes activities such as learning, planning, perception, and understanding language or robotics.

This document explores aspects related to the most common use cases of AI in business processes and applicable regulations, describing the main AI models and typologies, the general internal control framework, and related risks in those organizations deploying AI-based technology. Finally, this document offers a proposed audit program and procedures for evaluating these internal control structures.

1. AI IN BUSINESS ORGANIZATIONS AND ITS REGULATORY ASPECTS

1.1. The Penetration of AI in Business Organizations

The business world has been carrying out intense work in the use and application of AI models in a wide range of use cases, intending to optimize and provide greater efficiency to business processes and improve client services. The integration of generative AI in organizations is evident in the results of PwC's Annual Global CEO Survey (2024). Seventy percent of CEOs surveyed agree that "Generative AI will significantly change the way my company creates, delivers, and captures value in the next three years." Thirty-two percent of respondents say generative AI has been adopted across their company in the last 12 months, and 31% say their company has changed its technology strategy because of generative AI.

Market studies carried out by the consulting firm McKinsey show similar growth. "If 2023 was the year the world discovered generative AI , 2024 is the year organizations truly began using — and deriving business value from — this new technology, the "The State of AI in Early 2024" report says. According to the report, 72% of surveyed companies are using AI in one or more functions. Respondents most often report generative AI use in their marketing and sales, product and service development, and IT functions.

Standard 9.1 Understanding Governance, Risk Management, and Control Processes *To develop an effective internal audit strategy and plan, the chief audit executive must understand the organization's governance, risk management, and control processes.*

1.2. Knowledge and Skills of Internal Auditors in Terms of AI

In the current business context, the internal audit profession is in a process of evolution in which **risks** and processes have an increasingly technological component. In particular, the implementation of AI models in business processes represents an added challenge that forces the **internal audit function** to provide an adequate response, taking advantage of its positioning in the company.

Whether internal audit continues to be a relevant actor in an area that constantly and continuously adds value to organizations will depend on its ability to adapt and update. In this sense, one of the relevant aspects to consider is the need to have internal audit representatives with the talent

Standard 3.1 Competency *requires the chief audit executive to ensure that the internal audit function collectively possesses or obtains necessary competencies.*

and necessary knowledge to address from the beginning, through design audits, audit projects in which we face business processes with **control** structures based on AI models. This talent is a mixture of understanding basic audit principles as well as possession of specific knowledge capable of addressing the technical requirements that AI models need.

It can be difficult to access these types of holistic profiles in the labor market. For this reason, the internal audit function must provide this knowledge by reskilling internal audit professionals. AI must be present in the annual training plans of internal auditors, especially those who will address or review projects in processes in which AI is present.

__Standard 3.2 Continuing Professional Development__ states that internal auditors must continually develop their competencies through education and training.

The internal audit function should have a combination of technical knowledge and other more humanistic knowledge. Combining aspects such as mathematics, information technologies, computing, programming, and neuroscience with knowledge in the fields of logic and philosophy, philology, and even psychology, guarantees having the necessary skills to correctly evaluate the AI present in different business processes, not only from a point of view of its technical development but also from the perspective of those aspects most linked to the ethics present in its design and practical application.

1.3. The AI Act: Europe's Path towards Community Regulation on AI

> *Regarding artificial intelligence, trust is an obligation, not an ornament. Through these benchmark rules, the EU is leading the formulation of new global standards to ensure AI can be trusted. By setting standards, we can facilitate the advent of ethical technology around the world and ensure that the EU remains competitive. Our future-proof and innovation-friendly rules will intervene when strictly necessary, that is when the security and fundamental rights of EU citizens are at stake.*
>
> *—Margrethe Vestager, executive vice president responsible for the portfolio of a Europe Fit for the Digital Age*

The development and growing use of AI in different private and business spheres have led the European Commission to embark on a path towards regulation of AI. In April 2021, the European Commission approved a proposal for a regulation establishing harmonized standards in the European Union (EU) on AI. Later, in May 2024,[1] The Council of the European Union definitively approved the AI Law. Following the relevant approvals at the European level, member states will have a transition period of 24 months for its effective application in each respective national territory.

__Standard 1.3 Legal and Ethical Behavior__ states that internal auditors must understand and abide by relevant laws and/or regulations, including making disclosures as required.

[1] European Council, "Artificial Intelligence Regulation: Council gives final green light to world's first rules on artificial intelligence," European Council, May 21, 2024. https://www.consilium.europa.eu/es/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/.

The European Commission's proposal was born in the spirit of establishing a regulatory framework on AI that follows a risk-based approach and establishes a uniform and horizontal legal framework for AI that aims to:[2]

- Ensure AI systems introduced and used on the EU market are secure and respect current legislation on fundamental rights and Union values.
- Improve **governance** and effective enforcement of existing fundamental rights legislation and security requirements applicable to AI systems.
- Facilitate the development of a single market to make legal, safe, and reliable use of AI applications and avoid market fragmentation.

High-risk AI systems present a high risk of undermining the health and safety or fundamental rights of people considering both the severity of the possible harm and the probability of its occurrence. Within the types of AI systems, those considered high risk are of particular importance because they are the permitted type on which future regulations establish the most obligations.

The future European regulation on AI use classifies systems based on the purpose of use and establishes a series of requirements and limitations for each type. On the other hand, the proposed regulation considers situations where AI systems can be used for several different purposes (general purpose AI), and where the general purpose of AI technology is subsequently integrated into another high-risk system. This classification is reproduced in the following illustration.

Types of AI systems included in the AI Act

1 **Artificial Intelligence Systems with Unacceptable Risk (Art. 5)**	2 **High-Risk Artificial Intelligence Systems (HRAIS, Art. 6)**	3 **Artificial Intelligence Systems with Specific Transparency Obligations (Art. 52)**	4 **Artificial Intelligence Systems with No or Minimal Risk**
Prohibited • Manipulation of behavior, opinions, and human decisions. • Classification of people based on their social behavior. • Mass biometric identification remotely and in real-time, with certain exceptions.	Allowed if the requirements of the AI for the ex-ante conformity assessment are met. • Key Aspects of the Regulation (Annex III). • Common regimes already subject to harmonized EU standard. • Additional list to be reviewed annually by the EAIB (Art. 84).	Allowed but subject to information/ transparency obligations. • Human interaction. • Use to detect emotions or determine categories based on biometric data. • Generation of manipulated content.	Allowed without restrictions.
EXAMPLE: Social scoring	**Example: Hiring**	**Example: Personification (bots)**	**Example: Predictive maintenance**

[2] European Commission, "Regulation of the European Parliament and of the Council establishing harmonized rules on AI (AI Law)," *EUR-Lex*, April 21, 2021, 3. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

According to the proposal of the Parliament's regulations, the following AI systems will be considered systems of unacceptable risk and therefore prohibited:

- AI systems intended for **behavior manipulation** that uses subliminal techniques that transcend a person's consciousness or that take advantage of any of the vulnerabilities of a specific group of people (due to their age or physical or mental disability) to substantially alter their behavior in a way that causes, or is likely to cause, physical or psychological harm to that person or another.
- AI systems designed to **classify natural persons based on social behavior** used by public organizations (or on their behalf) based on their social behavior, personal characteristics, or personality (known or predicted).
- Real-time remote mass **biometric** identification AI systems in publicly accessible spaces for law enforcement purposes, except as follows:
  - The selective search for possible specific victims of a crime, including missing minors.
  - The prevention of a specific, significant, and imminent threat to the life or physical safety of natural persons or a terrorist attack.
  - The detection, tracing, identification, or prosecution of natural persons.
  - Any specific use of a remote biometric identification system will be subject to the granting of prior authorization by a judicial authority or an independent administrative authority of the member state.

In addition to the unacceptable risk AI systems mentioned, it is important to **consider the generative AI** category. These systems can be classified as high risk depending on their application and scope. For example, general generative AI may be included in Group 3 (AI systems with specific transparency obligations) of the risk diagram above, as its indiscriminate use could have significant implications for the privacy and authenticity of information. On the other hand, foundational generative AI, designed for specific and controlled purposes, may require a more detailed evaluation to determine its level of risk.

A wide range of high-risk AI systems would be authorized to access the EU market, provided they comply with established requirements and obligations. These requirements have been refined by legislators to ensure their technical applicability and reduce the **compliance** burden for **stakeholders**. However, it is important to highlight that all high-risk systems will require an *ex-ante* evaluation (assessing the **impact** before implementation).

In the cases of use mentioned above, AI's use is permitted, but after submission to a conformity assessment (before it enters the market or is used) to prove that the system meets the requirements of a reliable AI. In this sense, the regulation establishes an emergency procedure that allows a high-risk AI tool to be deployed without prior conformity assessment, introducing specific mechanisms to guarantee that fundamental rights would be sufficiently protected against any possible misuse of said system.

AI Systems Considered High Risk[3]

Biometric identification and categorization of natural persons	AI systems intended to be used in remote biometric identification in real or delayed time of natural persons.
Management and exploitation of critical infrastructures	AI models linked to transport or other similar infrastructures may endanger the life and health of citizens.
Vocational education and training	AI systems that can determine access to a person's education and career.
Employment, worker management, and access to self-employment	AI systems for recruitment; for example, to advertise vacancies, select or filter applications, or evaluate candidates during interviews or tests, as well as to make decisions on promotion and termination of work-related contractual relationships, assigning tasks, and monitoring and evaluating performance and behavior at work.
Access and enjoyment of essential private services and public services and benefits	AI systems intended to be used by or on behalf of authorities to assess entitlement to public assistance benefits and services, as well as to grant, revoke, or claim such benefits and services.
Law enforcement matters	AI models intended to be used to make individual **risk assessments** or other predictions intended to be used as evidence, or to determine the reliability of information provided by a person to prevent, investigate, detect, or prosecute a crime, or to adopt measures that affect an individual's freedom.
Migration, asylum, and border control management	AI systems intended to be used to predict the occurrence of crimes or events of social unrest to allocate resources dedicated to patrol and surveillance of territories.
Administration of justice and democratic processes	AI systems intended to assist a judicial authority in the investigation and interpretation of facts and the law, as well as in its application to a specific set of facts.

Furthermore, in collaboration with member states, the commission will create and maintain a database for the **registry of high-risk AI systems**. The main objective is to allow national authorities to access necessary information in the event of an infringement to investigate whether the use of AI is by applicable national legislation.

High-risk AI systems will be subject to strict obligations before they can be marketed, including:

- Appropriate risk assessment and mitigation systems.
- High-quality data sets.
- Recording of activity for data traceability.
- Detailed documentation.
- Clear and appropriate information for the user.
- Appropriate human supervision measures to minimize risks.
- High level of solidity, security, and precision.

>> **3.** EU Artificial Intelligence Act, "Annex III: High-Risk AI Systems Referred to in Article 6(2)," EU Artificial Intelligence Act, Aug. 6, 2023. https://artificialintelligenceact.eu/annex/3/.

The EU AI Act establishes a sanctioning regulation in Article 99 based on the establishment of thresholds that national authorities must consider in their sanctioning procedures:[4]

- Noncompliance relating to prohibited practices: up to **35 million euros or 7% of the annual turnover worldwide** total for the previous financial year.
- Failure to comply with any other requirement or obligation of the regulation: up to **15 million euros or 3% of annual turnover worldwide** total for the previous financial year.
- Provision of incorrect, incomplete, or misleading information to notified bodies and/or national authorities: up to **7.5 million euros or 1.5% of annual turnover worldwide** total for the previous financial year.

For comparative purposes, the General Data Protection Regulation (GDPR) establishes a lower sanctioning regime, given that the most serious sanctions are sanctioned with up to 20 million euros or 4% of the annual turnover volume, and up to 10 million euros or 2% of the annual billing, for the less serious sanctions.

On the other hand, the GDPR considers a more proportionate sanctioning regime (to be defined) for small- and medium-sized companies, as well as emerging companies, in case of violations of the AI Act's provisions.

1.4. The Anticipated Response of Business Organizations to the Expected Regulatory Frameworks on AI

Voices about the need to develop specific regulations regarding AI not only come from the public sector but also from private entities. Sundar Pichai, CEO of Google and Alphabet Inc., has spoken out in favor of regulation on AI, promoting convergent international regulation between the European Union and the United States. The Google Principles on AI are, in essence, the same values that have driven the AI Act that the European Commission approved in May 2024.

The business community must play a key role in this phase of regulatory development, not only to contribute its economic perspective on the impacts of AI on society but to adapt, in turn, its internal processes and mechanisms to this new regulatory framework regarding AI.

>> [4] EY, "Political agreement reached on the EU AI Act," EY, Dec. 10, 2023. ey-eu-ai-act-political-agreement-overview-10-december-2023.pdf.

2. AI MODELS

To carry out an appropriate risk analysis while auditing a business process that incorporates AI, it is important to have technical knowledge of the various AI models. This knowledge allows the internal audit function to design a work plan that aligns with the specific risks of the process and the intrinsic risks of the AI systems used.

This section outlines the most used AI models to establish a basic technical foundation for conducting process review work with AI models.

***Standard 13.2 Engagement Risk Assessment** requires internal auditors to develop an understanding of the activity under review to assess the relevant risks.*

2.1. Traditional Predictive Analytics

Traditional predictive analysis is characterized by the use of advanced statistical methods and techniques to estimate the likelihood of the occurrence of future events based on historical data.

Historical data[5] is typically used to create a mathematical model that captures important trends. This predictive model is then used with current data to predict what will happen next or to suggest actions to take to achieve optimal results.

This area of traditional predictive analysis encompasses classic data mining techniques such as **linear regression** and **logistic regression**, **clustering**, **factor analysis**, or **time series**.

2.2. AI and Machine Learning

AI commonly refers to the intelligent activity carried out by machines designed to reproduce the capabilities of the human brain through combinations of **algorithms**, allowing them to perceive the environment around them and respond in a human-like manner. This implies the ability to execute reasoning, observation, learning, and problem-solving functions. Put simply, AI is a machine that looks human and can imitate the behavior of people.

On the other hand, machine learning is a type of AI characterized by machine learning of knowledge and behaviors that would be difficult for humans to carry out, even going far beyond human intelligence in some respects.

Since the 1960s, the term machine learning has been related to a branch of AI focused on pattern recognition and learning by computers.

>> [5] The use of historical data presents certain limitations because, depending on the variables used in the model, the past does not necessarily tend to repeat itself in the future. However, they are usually the only data available or the most easily obtainable in certain circumstances.

Over the years, this discipline developed into other subjects related to probabilistic reasoning, statistical research, and, especially, deepening the recognition of patterns related to engineering, mathematics, and computing processes.

Today, machine learning is a scientific discipline in the field of AI whose main objective is to create systems that learn automatically so that, subsequently, based on the learning obtained by that system or machine, it can solve a given problem with precision based on the learning obtained.

Machine learning in this context means acquiring the ability to identify complex patterns in millions of data, hence the close relationship of machine learning with the big data discipline at present. This combination of disciplines is beginning to take shape and is called "machine big data" in some circles. Its unlimited potential is based on the following concepts:

- Use of logic and statistics to reason and address a problem, creating a predictive model that empowers the user to respond to it and provide a solution to a group of people and their needs.
- Use of big data to efficiently manage data, regardless of its structure (**structured data** and/or **unstructured data**) as well as its typology (historical records, recent data, or data collected in real time).
- Resolution of the given problem, obtaining the best answer to the problem by using the different machine learning algorithms and choosing the one that best adapts to the problem.

Data is the key factor in this process, as the final objective is to automate using complex algorithms to identify patterns or trends that are difficult to identify through traditional analysis performed by any human being.

Machine learning has a wide range of applications across many types of data universes. Its techniques are used in areas such as the early detection of diseases, the fight against terrorism, and obtaining competitive advantages over other companies.

Case examples that illustrate the advances of AI in different fields include:

- **Medicine.** Machine learning algorithms can assist in carrying out medical pre-diagnoses. A system based on these algorithms can learn from the medical history of patients linked to previous correct diagnoses to learn to identify future sick patients, which in turn can help medical staff make decisions when disease symptoms appear.
- **Fight against terrorism**. A machine learning-based software can detect patterns linked to potential terrorist attacks by analyzing and integrating data from various online sources, including social media. Systems capable of monitoring and detecting transactions linked to terrorist groups have also been developed.
- **Information Technology (IT).** The IT field has developed techniques to improve existing technological infrastructures. These techniques include identifying spam email, implementing voice recognition, detecting intrusions in data communications networks, predicting failures in technological equipment, and modifying the operation or appearance of a mobile application to adapt to the customs and needs of each user.

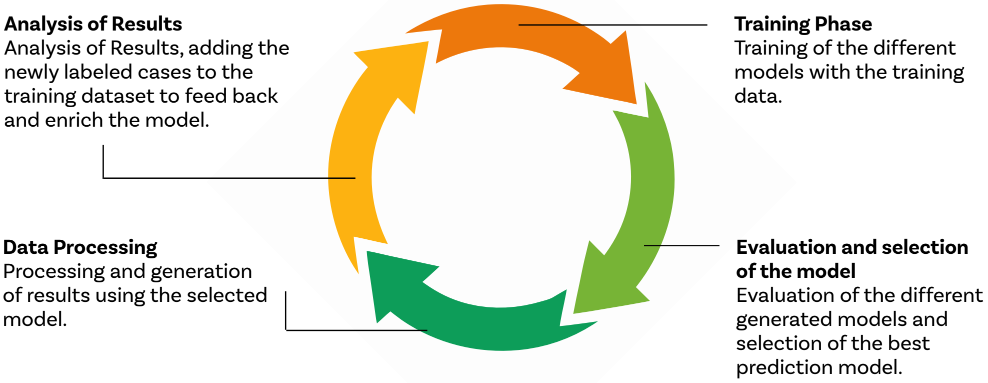
- **Business environment**. The latest trends in the use of machine learning in business are closely linked to human resources and digital marketing departments. For example, predictive models are being created to determine which employees will be the most productive in the future and to find potential clients by identifying patterns based on their behaviors on social networks. This tool also is used within financial organizations to detect and mitigate **fraud** to reduce its economic impact on this type of organization. Finally, and more directly related to the business field, it is used in Industry 4.0, optimizing operations by anticipating peaks and valleys of demand, improving predictive maintenance of facilities and their reliability, and achieving cost savings throughout the supply chain.
- The application of generative AI in the business field has been promoted in recent years by technology companies such as Microsoft, Amazon, IBM, and Google, developing solutions that have improved efficiency, productivity, and creativity in various work areas. Some examples are:
  - Microsoft Copilot is a chatbot powered by the GPT-4 language model that offers advanced code generation functions, integrating into Microsoft applications such as Visual Studio Code and Word to improve the development experience.
  - Amazon Q is built on the AWS infrastructure and is designed to address specific business needs, providing real-time solutions and support within the AWS management console, with access to various AI models for more precise responses.
  - IBM WatsonX is a comprehensive data and AI platform, offering tools for the development of customized solutions, an efficient data warehouse, and a toolkit for AI governance.

2.3. Types of Machine Learning

The discipline of machine learning has two main types of learning: **supervised learning** and **unsupervised learning.**

In supervised learning, predictions are made based on patterns, behaviors, or characteristics that have already been seen in historical and labeled data; there is prior knowledge of the data.

These algorithms can predict the value of a data set after being trained with another sufficiently large data set where the target variable has already been labeled. From this data, where its label is already known, the relationships between it and the rest of the model variables are obtained. Based on a known universe of data, a series of rules or patterns are established that will be applicable to predict new data.



Analysis of Results
Analysis of Results, adding the newly labeled cases to the training dataset to feed back and enrich the model.

Training Phase
Training of the different models with the training data.

Data Processing
Processing and generation of results using the selected model.

Evaluation and selection of the model
Evaluation of the different generated models and selection of the best prediction model.

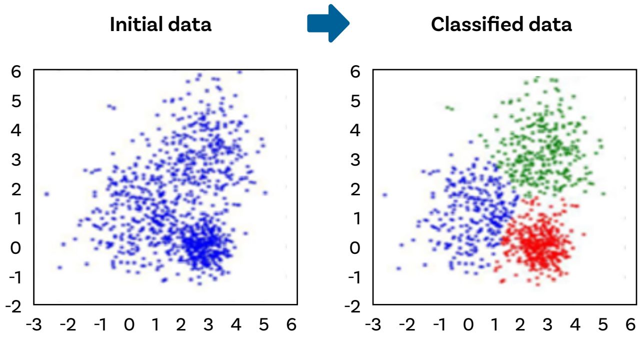
Continuous improvement of the model based on supervised learning.

The algorithms usually used in supervised learning are **decision trees**, **gradient boosting**, **random forest**, **support vector machines (SVM)**, and **Naive Bayes**.

Supervised learning can be applied in many ways. It is commonly used to determine customer scoring in financial environments, for predictive maintenance in the industrial field, for disease detection, and in cybersecurity.

In unsupervised learning algorithms**,** knowledge is produced solely by the data provided as input; there is no prior knowledge of the data.

This type of algorithm can self-organize into groups based on common characteristics. In these cases, the data are explored, looking for relationships in them to structure or organize them according to their characteristics.



Data Classification in Unsupervised Environments.

Some of the most common algorithms that encompass unsupervised learning are k-means, isolation forest, and **neural networks**.

The capacity that machines give us today is allowing us to advance at high speed in the use of these types of techniques, breaking with the restrictions or limitations that existed in their use until just a few years ago.

Some applications of this method include image and voice recognition, medical predictions, and predictions of anomalous information.

Advances in techniques such as neural networks (developed in the late 1950s) have only been possible to use with the latest technological advances, allowing progress in **deep learning**, the subdiscipline of machine learning. Deep learning consists of the use of advanced processing techniques that approximate human perception through process units or artificial neurons specialized in detecting characteristics existing in the data or objects received.

In addition to the two main types of algorithms that encompass the discipline of machine learning (supervised and unsupervised) is another type of learning called **reinforcement learning**.

Its key advantage is its ability to train effectively with minimal data. Reinforcement learning is based on a reward-based trial vs. error system, which allows the desired behavior to be reinforced. The algorithm functions by navigating its environment and receiving rewards or penalties based on the decisions made. In this way, the reinforcement learning algorithms learn and redefine their action strategy, iterating the necessary number of times until they find the strategy that leads to the optimal result.

Although it is intuitive to understand, there are three main challenges in reinforcement learning:

- If the “intelligent agent” does not produce sufficiently diverse behavior, it runs the risk of producing the illusion that there is no better way to achieve a greater reward and remaining stuck in a sub-optimal solution.
- The intelligent agent must find a balance between behaving by the policies it is learning and exploring new strategies that could give better results. These explorations can reduce rewards or even replace previously learned behaviors. Humans are often faced with these decisions: Do we dine at one of our usual restaurants or take a chance at the new trendy restaurant?
- The reward can come long after the actions. For example, in chess, a move in the middle of the game can completely determine the outcome.

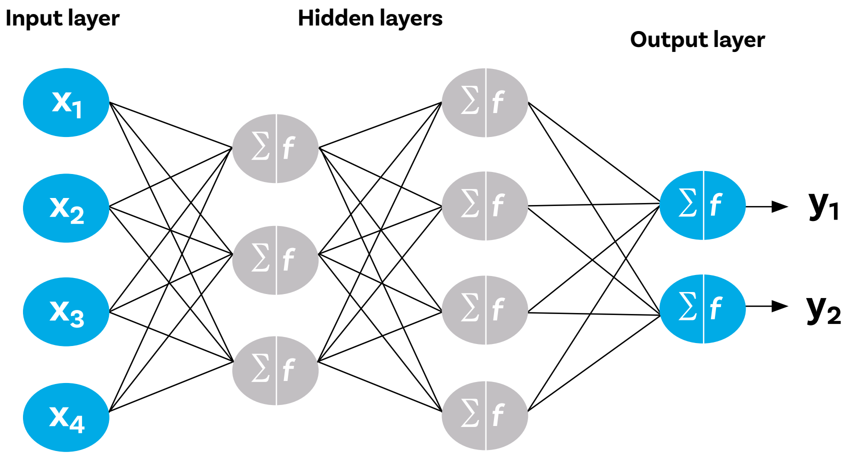
Despite these difficulties, reinforcement learning is used successfully in different areas where supervised learning was not successful due to the difficulty of having sufficient data, and where unsupervised learning algorithms are not suitable for the problem.

Reinforcement learning, for instance, aids in developing algorithmic trading agents, where there is a large amount of historical data available used to back test and generate rewards algorithmically. It is also used in other problems such as chatbot training, and autonomous web browsing automation for testing or scraping purposes.

Video games are another area where the reward can be obtained directly from the environment, as the rules serve to give that reward or penalty to the actions performed. Similar situations may arise in the real world, where the good or bad result can be measured systematically, such as in cases of autonomous driving, recommendations in electronic commerce, and streaming services, or to optimize the advertising shown to visitors of online services.

Neural networks

Neural networks are an approach to AI through which an entity aims to imitate a nervous system, based on neurons, through bio-inspired algorithms. The fundamental unit of a neural network is the neuron, which typically is organized in layers. Each neuron receives input from the previous layer that it combines using a simple mathematical operation and sends it to the next layer. In the first layer, known as the input layer, each neuron receives input data to process, e.g., a pixel of a photo. In the outermost layer, it produces the output, which is typically a numerical value and which for each use will represent something different, such as a prediction of a value, the next word to generate a text, or the value of a pixel in an image. The layers that are between the input and the output are called hidden layers.



Example of a simple neural network with two hidden layers. Source: https://www.knime.com/blog/a-friendly-introduction-to-deep-neural-networks

The parameters of each neuron are adjusted in a training process called back propagation, which is computationally expensive but produces good results on multiple complex problems.

Neural networks are algorithms whose explainability is very low: It is difficult or even impossible to directly explain why a neural network produces a result.

When the number of hidden layers is very large, it is called a deep neural network, and the training is normally called deep learning. A neural network with more layers becomes more complex and more difficult to explain, and it demands more computational resources.

The ability and suitability of neural networks to solve problems are closely linked to different factors: how these layers are arranged relative to each other; how they are connected; what mathematical functions they implement; what dimensions they have, and how they are organized among themselves.

There are many different types of architectures for different types of networks, and the AI community of practice largely shares its knowledge, its **findings**, and its trained models[6] so that other people can rely on them to carry out their work and improvements.

An optimized way to deal with a problem is to use a model previously trained for a task (for example, generating text) as a starting point and train it specifically to fine-tune its behavior (for example, generating legal or medical texts). While the base system would provide basic knowledge of the language (such as grammatical constructions, agreement, etc.), specialized training would provide specific knowledge. This is known as **transfer learning**.

Although transfer learning has proven very useful for multiple use cases related to text processing, language generation, and image recognition, the use of this type of technique carries the risk of inheriting biases existing in the base model. Algorithmic bias occurs when a computer system reflects the values of humans who are involved in coding and collecting data used to train the algorithm, as explained in Section 4.7.

Generative AI

In recent years, large language models (LLM) have proliferated. In addition to being able to simulate conversations with a high level of complexity and knowledge, the capabilities of LLMs to process text have far surpassed those of other approaches in the creation and generation of text summaries, as well as identification of feelings or keywords, among other functionalities.

LLMs have their origin in an architecture called transformers,[7] which are like deep neural networks but can extract statistical information about how some words relate to others from texts; with sufficient training and data, the texts generated by these systems appear to be equipped with intelligence. Transformers have technical characteristics that allow them to run in parallel on powerful computers and GPUs with greater efficiency than other neural networks. The way to train transformers with new texts is by simply deleting words in existing texts and training them to predict the deleted words, or by predicting what the next words will be for the given text. This allows training sets to be generated automatically, which is a great advantage over other systems that require labeling by experts.

>> [6] An example of a public model repository is https://huggingface.co/models.
[7] Cornell University, "Attention is All You Need Until You Need Retention," arxiv, August 2023. https://arxiv.org/abs/2501.09166.

Because transformers can be very efficiently parallelized to train and run on extremely powerful hardware, and training data can be created automatically from existing texts, it is possible to train huge language models, with tens or hundreds of billions of parameters, using trillions of words extracted from all public information available on the internet (such as the Common Crawl dataset).[8]

The turning point that has allowed the use of LLM to be democratized has been to combine these characteristics with chat-type user interfaces, which allow you to talk to them and ask questions as if they were people on the other side of the terminal. This is achieved by combining its characteristics with reinforcement learning techniques, in which sets of questions and answers developed in a specific way have been used, and the evaluations of the answers made by users after each interaction. This has been achieved by combining its features with reinforcement learning techniques, which use sets of questions and answers specifically developed by humans as well as evaluations of the answers that users make after each interaction.

Thanks to the ability of LLMs to process and generate human language effectively, they are being adopted in business environments to transform processes. Some of the most relevant use cases are:

- Customer service automation: LLMs power chatbots and virtual assistants, providing fast and accurate responses to customer queries. This improves the customer's experience by offering 24/7 service while reducing the workload on human staff.
- Text and data analysis: Models can analyze large volumes of text, such as emails, legal documents, or customer feedback, to extract meaningful insights, trends, and patterns. This supports data-driven decision-making and helps identify areas for improvement in products and services.
- Content generation: LLMs can generate high-quality written content such as reports, emails, and marketing pieces, saving time and resources. This allows companies to maintain an active and professional presence on multiple platforms with less effort.
- Training and development: LLMs are used to create customized training simulations and scenarios that assist in skill development for staff. They can facilitate interactive scenarios for training in customer service, sales, and more, adapting to user responses to deliver a more effective learning experience.
- Internal process optimization: Models can help automate and optimize a variety of administrative and repetitive tasks, such as data entry, meeting scheduling, and email management. This frees staff to focus on higher value tasks and improves operational efficiency.

>> **8.** Cornell University, "Language Models are Few-Shot Learners," arxiv, July 22, 2020. https://arxiv.org/pdf/2005.14165.

The rapid adoption of this type of AI also involves the assumption of new risks. In addition to inherent technical problems, there are other relevant issues to consider. One of these is the debate surrounding the intellectual property of the content used to train LLMs. Since LLMs generate content based on the statistical relationships of existing content, there is a question of ownership and rights. It also is important to consider the liability angle. If these types of systems are used to make recommendations in areas such as health, taxation, or the legal environment, who is responsible for the recommendations they issue and their application results?

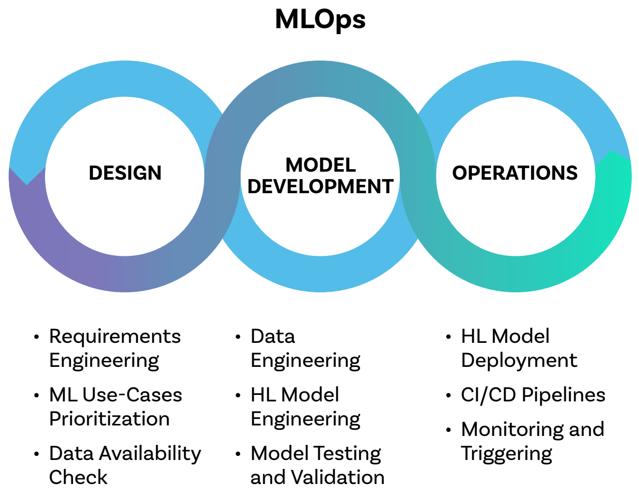
2.4. MLOps as a Response to Adaptation Needs

Machine learning has accelerated in recent years much like other fields, such as software. New challenges like changing needs and automated data collection contrast greatly with the former traditional methods that approached model and data selection in a more hands-on manner.

The development of machine learning models is susceptible to inheriting the problems and difficulties that also plague traditional software engineering, in addition to the issues that AI and automated learning contribute.

In response to this need for dynamism, the practices of Machine Learning Operations (MLOps)[9] emerged. It could be said generally that MLOps is to machine learning what DevOps is to software development: a set of practices that aim to streamline the life cycle of AI solutions by automating the development, training, and deployment processes of models; integrating aspects of data, development, infrastructure, and security; and minimizing deployment times and maintenance, as well as possible errors.

The IIA's Global Technology Audit Guide ***Auditing Business Applications*** *provides guidance for assessing controls over the software development lifecycle. This guidance is relevant when AI is implemented in applications that support organizational processes.*


MLOps
DESIGN
MODEL DEVELOPMENT
OPERATIONS
• Requirements Engineering
• ML Use-Cases Prioritization
• Data Availability Check
• Data Engineering
• HL Model Engineering
• Model Testing and Validation
• HL Model Deployment
• CI/CD Pipelines
• Monitoring and Triggering

The incremental iterative process of ML-OPS. Source: https://Ml-ops.org

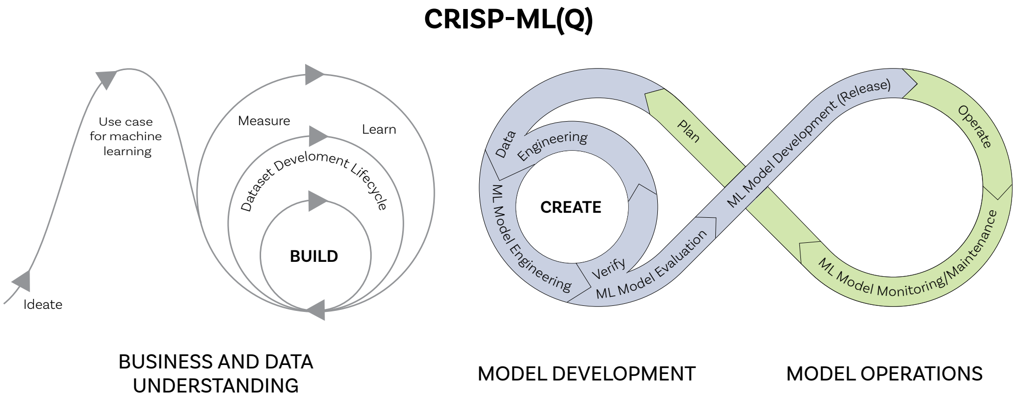
[9] MLOps, like DevOps, is an evolution of the AGILE methodology for software development, a methodology that has been recently and successfully incorporated into internal audit work.

Some of the activities and practices included in MLOps are:

- Development of specific tests to validate the machine learning model.
- Continuous integration to test and validate both the models and their data.
- Continuous delivery to automate the launch of new models.
- Continuous training to guarantee continuous learning over time.
- Monitoring the models in different metrics of precision, specificity, disparity impact, etc. to detect unexpected drifts of AI models, analyze them early, and apply the necessary corrections.

Regarding the development of the model, as is the case with software, no **methodology** stands out from the others in all aspects and for all situations, and in most cases, ad-hoc approaches must be used for different problems.

However, to define a generic starting point, the Cross-Industry Standard Process for Development of Machine Learning Applications with Quality Assurance (**CRISP-ML(Q)) methodology** aligns business aspects with the development and operation of the models to ensure the highest possible quality of the developed models, and to make certain they meet the expectations of the promoters.


CRISP-ML(Q)
Use case for machine learning
Measure
Learn
Dataset Develoment Lifecycle
BUILD
Ideate
Data Engineering
CREATE
ML Model Engineering
Verify
ML Model Evaluation
Plan
ML Model Development (Release)
Operate
ML Model Monitoring/Maintenance
BUSINESS AND DATA UNDERSTANDING
MODEL DEVELOPMENT
MODEL OPERATIONS

Machine learning development life cycle according to methodology CRISP-ML(Q). Source: ml-ops.org.

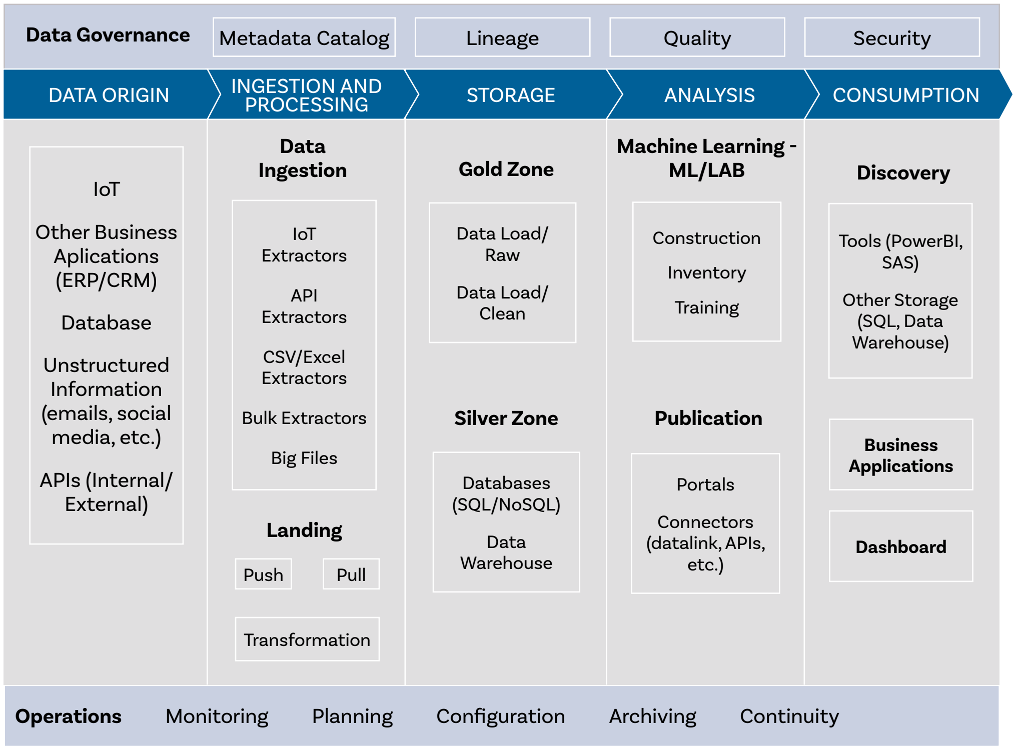
The internal audit function should find the evidence produced during the methodology process noteworthy for assessing the activities undertaken to ensure effective operation and alignment with the business.

2.5. IT and Data Architecture

Industry standards advise that any solution based on AI should be accompanied by an architecture that allows its automation and facilitates its use. From an internal control point of view, it is very important to know the data and IT architecture of the AI systems subject to audit. To this end, the purpose of this section is to develop a model architecture of AI systems. The internal audit function must analyze and understand the technical details of this architecture when undertaking AI structure review work.

From a functional point of view, an architecture has these layers:

- Data Governance or Data Governance Model
- Data Source
- Data Management Platform
  - Data ingestion and processing
  - Storage
  - Analysis (Machine Learning – ML/LAB)
- Data consumption
- Operations



Data Governance
Metadata Catalog
Lineage
Quality
Security

DATA ORIGIN
IoT
Other Business Aplications (ERP/CRM)
Database
Unstructured Information (emails, social media, etc.)
APIs (Internal/ External)

INGESTION AND PROCESSING
Data Ingestion
IoT Extractors
API Extractors
CSV/Excel Extractors
Bulk Extractors
Big Files
Landing
Push
Pull
Transformation

STORAGE
Gold Zone
Data Load/ Raw
Data Load/ Clean
Silver Zone
Databases (SQL/NoSQL)
Data Warehouse

ANALYSIS
Machine Learning - ML/LAB
Construction
Inventory
Training
Publication
Portals
Connectors (datalink, APIs, etc.)

CONSUMPTION
Discovery
Tools (PowerBI, SAS)
Other Storage (SQL, Data Warehouse)
Business Applications
Dashboard

Operations
Monitoring
Planning
Configuration
Archiving
Continuity

a. Data Governance or Data Governance Model

This domain usually is composed of different components to cover the main disciplines of data governance:

- Catalog metadata: Metadata is available in a unified way.
- Lineage: Traceability of data from its origin to its different transformations.
- Security:
  - Permissions: Information access control.
  - Applicable regulatory compliance: Application and verification of compliance with the needs required by different regulations in terms of data use and protection, such as GDPR and Payment Card Industry Security Standards Council (PCI-DSS), among other applicable regulations.
- Data quality: Set of rules and metrics intended to evaluate the quality of available information.

b. Data Origins

The data sources that feed AI systems can be multiple, considering financial and nonfinancial data from different accounting systems (Enterprise Resource Planning (ERP)) and/or other business applications of both internal and external organizations. In this sense, a precise knowledge of the databases that feed the AI models is especially relevant for an adequate understanding of the origin of the data, its structure (structured or unstructured information), and how these feed the data management platform layer explained below.

c. Data Management Platform

Typical data management platforms for AI systems can present the following four domains:

1) Data Ingestion and Processing

A logical layer is formed by the set of components responsible for access to the AI systems of the different data sources. For the architecture to be as automatable as possible, this layer should be as complete as possible, containing the greatest number of extractors and loading tools necessary, either with data sources internal to the organization, or external sources via **application programming interfaces** (**APIs**) or direct connections. This layer has high-performance processing engines that transform the data deposited in the Landing Zone, a data entry zone in persistence mode generated by the extractors (pull) or deposited by external processes (push).

2) Storage

This is the core area of any platform, composed of different technologies and storage capacities, forming several logical data layers and processing engines that transition the data between them.

From a storage point of view, the following areas stand out:

- **Gold Zone**: A repository where data is stored and processed one by one, without the need to normalize or apply business rules.
- **Silver Zone:** A data access and processing area where transformations are applied and business rules, normalizations, etc. are applied. For final exploitation, the information can be structured as appropriate.

3) Machine Learning Laboratory (ML-LAB)

In solutions aimed at advanced analytics, the machine learning area is of special interest, providing such capabilities as a development environment, libraries, and open-source repositories that facilitate the design, training, testing, governance, and delivery of models, forming among them the Machine Learning Laboratory (ML-Lab). In certain architectures, and depending on their criticality, it may be necessary to have a sandbox, an isolated environment that allows tests to be executed safely without compromising the rest of the architecture.

4) Data Consumption

This last layer represents the output of the data as a result of the executions carried out on the data after the previous layers and domains have passed. It represents the result of AI models for analysis using visualization or big data tools, storage, or use by other business applications.

Global Architecture in Generative AI Models

As generative AI is a disruptive technology, all organizations are currently building and enriching their architecture starting from an initial one that serves as a reference. These new architectures are being integrated as part of the ecosystem of each organization's IT environment, where AI models are implemented and grown in a structured way.

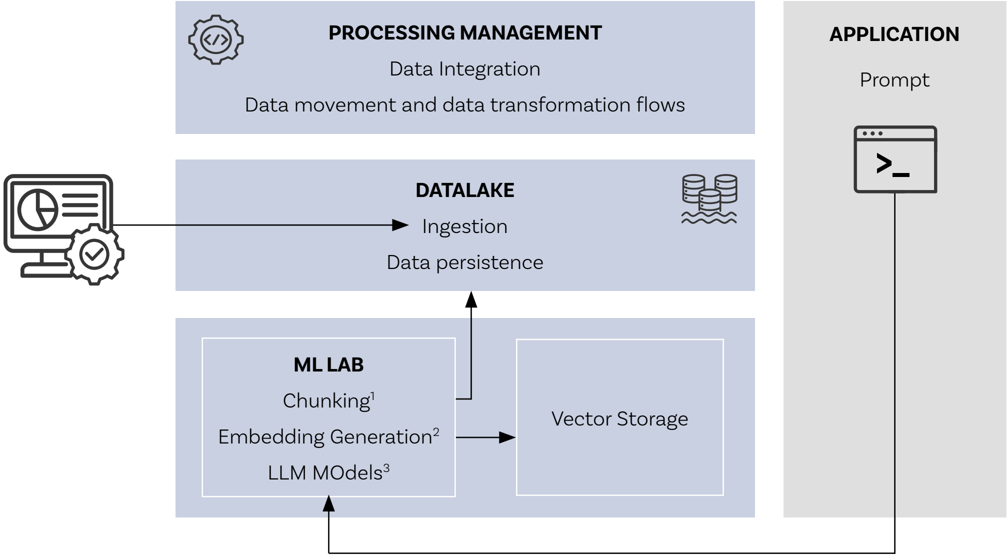
In this sense, to provide a reliable solution that meets the expected security and governance standards, architectures are emerging that consider the four main flows in any generative AI solution:

- Extraction and ingestion of structured and unstructured information.
- Generation of chunks, or "cutting" of data (chunking).
- Vectorization or embedding of input data.
- Interpretation of the user prompt and generative response.

In any architecture that serves generative AI use cases, three activities are key when it comes to having an efficient environment that meets expectations:

- Chunking or fragmentation of data: A technique that allows multiple units of information to be combined or grouped into a limited number of fragments so that it is easier to process and remember the information. This technique assimilates storage methods to the way a person's memory works; because most can remember a list of five words for 30 seconds, a list of more words is divided into smaller fragments to improve performance.

- Embedding: A natural language-processing technique that converts human language into mathematical vectors, allowing computers to process language more effectively by treating words as data.
- Prompt: An initial instruction or text provided to a generative AI tool to guide its generation of responses or results, depending on the formats in which the tool specializes. The prompt functions as an “information input” with which the user specifies the context and the task that the tool is expected to complete. By providing a prompt, the AI model is conditioned to generate a consistent and relevant “output” (result), based on what the user needs.



Global Generative AI Architecture.

1. Chunking is the process of dividing content into smaller, more manageable parts.
2 Natural language processing through the conversion of human language into mathmatical vectors.
3. Large language model that facilitates the comprehension and generation of human language.

The Energy Consumption of Generative AI

Behind generative AI and LLM, there is significant energy consumption, reaching truly surprising figures. It will be a great challenge when it comes to optimizing this technology and reducing its great weight for the carbon footprint, for which other technologies, such as renewable energies, will be relevant.

For example, a model ten times smaller than ChatGPT has an energy consumption of 650,000 kWh just for training, while in 2022 the per capita energy consumption in Spain was 5,214 kWh. Meta’s Llama training required 2,638 MWh, equivalent to the consumption of a 150-watt refrigerator for 2,008 years.

3. INTERNAL CONTROL FRAMEWORK AND RISKS OF BUSINESS PROCESSES WITH AI

Organizations that begin their journey to deploy AI systems must establish an adequate internal control framework and structures for the identification and mitigation of associated risks.

An adequate evaluation of the performance of AI systems must include aspects for continuous monitoring of the risks of this type of technology, as well as internal control structures that guarantee continuous supervision of the AI systems. The series of key activities suggested below should form part of the internal control framework of AI systems deployed in organizations.

The Committee of Sponsoring Organizations of the Treadway Commission's (COSO's) *Internal Control-Integrated Framework* is the most common reference for the development of an effective control framework to evaluate and manage the risks associated with AI systems. The internal audit function should provide independent **assurance** on AI risks, governance, and controls. The internal audit function should evaluate the AI-

Standard 4.2 Due Professional Care *requires internal auditors to consider, among other things, the potential costs and benefits of internal audit services, the extent of work needed to achieve the* ***engagement's*** *objectives, and the use of appropriate techniques and tools for the complexity and risk of the activity under review.*

Continuous Risk Monitoring	Data and IT Architecture Supervision	Review of AI Models	Implementation Supervision	Monitoring after Putting into Production
Identification of general risks due to the deployment of AI systems, and specific risk assessment based on the complexity of the algorithms and objectives pursued with each AI model.	Activities defined to guarantee access, treatment, privacy, protection, and destruction of data, especially those in predictive models of human behavior. Provide reasonable security over IT systems and prevention of cyberattacks.	Ensure an adequate understanding of the operation of the algorithms and expected output. Definition of metrics and performance indicators for continuous monitoring. Review of the behavior of the models in early phases of engine deployment.	Develop sufficient implementation tests to guarantee that the deployment of the engines meets the expected objectives. Relevant approvals from the project committees established before the go live.	Periodic review of performance metrics and indicators. Control mechanisms for the identification of anomalous performance behaviors, including the necessary corrective actions for the algorithms.

related policies and procedures in place, verifying that the control objectives are appropriate and working as designed. The next step is the development of the five COSO components from the perspective of the internal control framework in AI models.

Standard 13.3 Engagement Objectives and Scope *requires internal auditors to specify the goals to be achieved and the elements to be reviewed by an engagement, considering whether the engagement's purpose is to provide assurance or advisory services.*

3.1. Control Environment (Governance and Culture)

COSO indicates that the main objectives of the control framework are to establish good governance practices and strengthen accountability, responsibility, and supervision. Additionally, it is relevant to understand the risk profile of AI in the company and know how it manages associated risks.

Following the COSO framework, a set of standards, processes, and structures must be established that constitute the basis on which internal control is developed at all levels of the organization. The **board** and **senior management** set the "tone at the top" on the importance of internal control, including the standards of conduct that are considered acceptable.

On the other hand, promoting the company's commitment to **integrity** and ethical, social, and legal values helps ensure activities, decisions, and actions related to AI systems are consistent with the organization's ethical, social, and legal values and responsibilities.

An adequate AI governance model must consider the following aspects:

- The importance of maintaining a data governance system throughout its life (creation, transformation, use, and destruction of data as inputs to the models) due to the huge volumes of data it uses.
- Business culture, which must be linked to the components of AI systems. The governance model must guarantee that the ethical values of the organizations and their internal policies extend, in turn, to the objectives pursued by the AI structures. A code of ethics is considered necessary and must reflect respect for human rights, protection of personal data, promotion of equal opportunities, transparency, and freedom of choice, as well as the needs and expectations of individuals in the company.
- Regulatory compliance issues applicable in each jurisdiction where organizations operate.
- Existence of a risk-conscious culture, control, policies, processes, and structures that guide people at all levels in the performance of their responsibilities consistent with the entity's commitment to integrity and ethical values.
- Establishment of appropriate competency levels for the management and supervision of AI systems. It is important to highlight that the board of directors and the management committee must know the key concepts of AI models.
- Risk amplification of the generative AI component which defined governance models must be considered in the design, maintenance, and continuous improvement of internal control structures.

From the point of view of the main governing bodies (board of directors/audit committee), the most relevant roles within the framework of the internal control of AI systems are summarized below:

- Advise on whether the strategy adequately considers the threats and opportunities of AI systems. It is advisable to have a specific AI committee or assign responsibilities to an existing one, for supervision and management of the risks of AI models, where corresponding responsibilities are assumed in this risk area.
- Promote a program of knowledge and skills training specific to AI systems and adopt policies and practices to attract, develop, and retain competent professionals.
- Identify risks of AI systems and incorporate them into the organization's risk models, defining the risk appetite for this type of risk.
- Supervise and evaluate, with the frequency and depth deemed appropriate, the effectiveness of the governance model's internal control regarding AI models, focusing on both their design and the operation of the internal control structures.

As an extension of the general objectives of the previous governing bodies, senior management establishes the strategy for the implementation of AI initiatives, including the specific objectives and roadmap for the deployment of priority AI systems for the company. Additionally, both the board and senior management must establish a strategy for the design and implementation of AI systems aligned with the organization's general strategy, considering this strategy as a critical element for the success of initiatives in AI models that companies want to undertake.

In the same way, a wide range of objectives motivates companies to use AI systems, from optimization and cost reduction to differentiation of the products and services they offer to their clients, thus generating sources of income with greater diversification. In any of its facets, the specific objectives that lead companies to invest in AI systems must be aligned with short- and long-term business objectives.

Control Environment (Governance and Culture): Best Practices and Recommendations
The existence of a strategy defining the main governing bodies for the implementation of AI systems, starting with those models with the highest return and increasing the investment as the internal know-how and expertise generated ensures meeting the established objectives.
The strategy for the implementation of AI models must have the general premise of measuring the performance of the models and, by extension, meeting the objectives pursued with the deployment of the AI models.
Strategic plan for supervision and evaluation of the governance model for the continuous improvement of the control environment and its reporting to the corresponding administrative bodies.

3.2. Risk Assessment

Per the COSO internal control framework, the company must clearly define objectives to allow the identification and evaluation of risks related to AI. In this sense, an adequate internal control framework designed and implemented to address the risks derived from the use of AI systems must consider the following aspects:

- A clear and precise definition of objectives that allow the identification and evaluation of exposed risks derived from the implementation of AI systems.
- A risk assessment that covers all levels of the entity, promoting a transversal risk assessment and coordination between all the areas involved (fundamentally those with greater responsibility for the implementation and execution of AI systems), as well as establishing the bases for how risks should be managed and addressed for their efficient mitigation.
- Risk assessment mechanisms to identify and evaluate significant changes to AI models that could significantly affect internal control structures.

The main risks linked to AI systems developed by organizations include:

Risks	Comments
Governance Risks	Related to the internal structures of organizations, policies, methodologies, and decision-making processes, including high-level supervision. Emphasis should be placed on risks that may impact management and leadership, **independence** in decision-making, and promotion of transparency and accountability.
Operational and/or Business Risks	Related to various points in the life cycle of an AI system's development and implementation. The most significant are those that may entail processing errors, data risks or risks of deviations, bias in results, or representativeness of the data.
Financial Risks	Related to the accounting of operations and the presentation of financial information. Situations must be considered where AI may impact the financial information presented by the company, or its financial results.
Regulatory Risks	Linked to the legal and regulatory compliance areas. Of note are the compliance risks associated with external (e.g., GDPR) or internal regulations (e.g., a code of ethics). These risks are related to the activities of the AI models, as well as the related decisions and actions, that are consistent with the values and ethical, social, and legal responsibilities of the company.
Technological and Cybersecurity Risks	Associated with the systems and cybersecurity of the developed AI models. For example, whether these models may contain personal data that is at risk of being accessed and used by unauthorized third parties should be evaluated.
Reputational Risk	Related to risks derived from the presence of biases in the models, sanctions imposed for regulatory noncompliance, or exposure to external risks generated by third parties (see below).
Sustainability Risk	The energy consumption that allows the operability and functioning of AI systems (e.g., generative AI models) can have an impact on the sustainability models and policies of companies, e.g., concerning commitments acquired from greenhouse gas reduction, energy efficiency, or carbon footprint.

Risks	Comments
Intrinsic Risks of AI Models	**Data use:** When working on a project related to AI, the most common mistake is usually using incomplete or inaccurate data. Normally the algorithms can be fed with structured or unstructured data from different sources such as websites, images, or social networks. Any incorrect use of databases can cause unstable or incorrect algorithm results. The incorrect selection of data not only influences the results of the algorithm but also can have ethical repercussions, because some group of information may be left out, giving an image of a discriminatory or biased model. **Improper development of AI algorithms causing inappropriate results:** Errors in the programming and code development of AI algorithms can cause results that are unexpected, inadequate, or far from the established objectives. The development phase of implementing AI model algorithms is the most critical, especially in those AI systems with greater sophistication; any code or programming error can lead to inappropriate results. **Inability to interpret or incorrect interpretation of the outputs of AI models:** The greatest value of AI systems is the obtained results; however, sometimes the **internal audit function** may find itself faced with the inability to interpret or incorrectly interpret the results once the databases are scanned by AI algorithms, making decisions with unwanted or unexpected results. For example, AI algorithms based on neural networks may contain the typology of intrinsic risks in greater depth, because neural network algorithms tend to present a greater risk of incapacity or incorrectly interpret the output obtained from information systems. Generative AI creates new risks and, at the same time, amplifies other existing risks. **Specific risks of Generative AI models and LLM models**: • **Hallucination:** LLM models, based on probabilistic systems, predict words for a previously provided text. In that sense, LLM models can invent content that does not exist (e.g., create citations to documents that do not exist, or generate a formal document that does not exist), create false statements that do not reflect the truth, or create erroneous statements (provide wrong solutions to problems). • **Relevance:** Sometimes the content may not be relevant to the context. • **Toxicity**: Trained by the vast amount of content on the internet, the system has learned from text that could reflect hatred, envy, insults, racism, or violence, among other biases. • **Privacy and confidentiality:** Given that cloud services are offered that use content and conversations provided by users to train the systems, there is a risk that private information will be leaked to different users because it has been learned during use. • **Jailbreaking**: Forcing models to perform actions for which they were not designed or circumventing prevention mechanisms and safeguards through specially manipulated entries.
External risks due to the use of Generative AI by third parties	The ability of generative AI to create misinformation, phishing attacks, and increasingly sophisticated malware attacks poses a growing risk to the management of cybersecurity risk and the company's privacy of personal and strategic data. The risk also increases that the information a company uses in the ingestion of its AI processes and algorithms may be contaminated by deliberate actions of third parties. Additionally, there is a growing tendency for companies to outsource the operation of certain processes to vendors, including providing personal and strategic data, resulting in greater exposure to the supply chain. This is because the same risks apply to the use of AI and companies must protect themselves by configuring an assurance mix (e.g., obtaining an SOC 2 from the vendor's auditor and establishing internal controls that complement that external control). All the above can have a direct impact on reputational risk.

Risk Assessment: Best Practices and Recommendations

- Existence of internal risk management policies, specific to AI systems, with the aim that the entire company is continuously aware of and involved in the identification of risks related to AI models. General guidelines for the design and implementation of risk mitigation strategies should be included.
- Existence of a risk inventory or map with the nature of each of the identified risks, including their criticality, and when possible, quantification of the probability of occurrence or potential financial impact in the IT systems participating in the process, as well as other risk categories.

3.3. Control Activities

COSO defines control activities as the actions established through policies and procedures that help ensure management instructions established to mitigate risks, with potential impact on the objectives defined in the implementation of AI models, are carried out.

In this sense, organizations must have appropriate internal communication mechanisms to expand and publicize the objectives and responsibilities of each area, as well as the responsible parties involved in the AI models deployed. Likewise, organizations must have appropriate control structures that allow aspects related to AI models to be adequately transferred through external communication channels aimed at regulators, shareholders, and other interest groups.

In Section 4, a series of illustrative control activities addresses the relevant risks and allows for preventing the materialization of unnecessary risks and minimizing the impact of their consequences.

***Standard 13.4 Evaluation Criteria** requires internal auditors to evaluate whether the board and senior management have established adequate criteria to determine whether the activity under review has accomplished its goals and objectives.*

Control Activities – Best Practices and Recommendations

- The existence of internal procedures and risk/control matrices that identify the design of control activities and their key attributes, including supporting documentation that demonstrates the effectiveness of the controls as well as those responsible for their execution and review, for the end-to-end approach to risks in business processes with implemented AI systems.
- The design of control activities considers both implementation and recurring controls in AI systems. The latter is periodically evaluated to identify risks not addressed by the corresponding internal control activities, as well as changes in the design of the controls necessary during the life cycle of the AI systems.
- Design and establish timely “human review” activities of the behaviors and results of AI algorithms to ensure that they reflect the original objective and are used in a legal, ethical, and responsible manner.
- Design and implement alerts and/or indicators of deviations from the initial objectives of the AI algorithms.
- The company maintains an inventory of AI systems, identifies synergies, and analyzes risks from an individual and consolidated point of view.
- Supply chain risk. It is important to keep in mind that the existence of outsourced control activities does not exempt the company from its ultimate responsibility for the risks it manages, so it must have corresponding internal controls over the outsourced activities.

3.4. Training and Communication

The information generated by AI systems must have appropriate communication and reporting protocols, both internally and externally. The sensitivity of the data processed by AI structures and the impacts on society that arise from the use of AI in business processes make necessary transparent and precise communication of how organizations should transmit the general principles of this very relevant technology.

Currently, many pioneering organizations in these technologies make public the advances acquired, not only as a competitive advantage in the sectors in which they operate but also to make known, in a transparent manner, the principles that govern the ethical and moral values derived from the use of AI systems.

Training and Communication – Best Practices and Recommendations

- The organization publishes its best practices on ethical and moral values in the use of AI systems.
- The company shares with the public the principles of AI, limiting it to aspects that are of most concern to public and private organizations.
- The company's top managers, shareholders, and board are informed of the relevant aspects of the progress, real performance, and the achieved initiatives on AI systems.
- The written procedures and the risk/control matrix (end-to-end business processes assisted by AI techniques) are known and applied in practice by the corresponding process owners and are updated accordingly based on the maturity stage of the AI model.
- The company has emergency plans to address unforeseen events derived from the unexpected behavior of the models.

3.5. Supervision and Evaluation

Monitoring activities are periodic or ongoing evaluations that verify that all five components of internal control, including controls affecting the principles within each COSO component, are properly designed and operate appropriately while maintaining the established precision levels.

Governance models focused on adequate business management of AI systems must include continuous and/or independent review and evaluation activities to determine whether the components of the internal control system are present and operational during the period life cycle of AI systems. Companies must have mechanisms for evaluating internal control deficiencies in AI systems, as well as mechanisms for communicating these deficiencies to those responsible for implementing the corresponding action plans or corrective measures, including, as appropriate, senior management and the council.

In this sense, the internal audit function takes a relevant role in assessing whether continuous supervision and evaluation activities are adequate to mitigate the intrinsic risks in AI models, both from a design and operational effectiveness point of view.

The IIA's Three Lines Model provides six principles for defining the roles and responsibilities of the governing body (board), management, and independent assurance providers (internal and external) in providing oversight, risk management, and assurance services to the organization's stakeholders.

Monitoring and Evaluation – Best Practices and Recommendations

- Planning and executing the supervision and evaluation activities of models as well as validating the suitability of the risk identification processes.
- Defining a long-term (or strategic) audit plan for AI models, which accompanies the company's strategy in this regard.
- Defining substantive or audit tests of internal control over the integrity, accuracy, and reliability of the data on which the AI algorithms are built.
- Identifying, evaluating, and promptly communicating internal control deficiencies to those responsible for taking corrective measures, including senior management and the administrative body (or its audit committee), as appropriate.
- Supervising internal control activities designed and executed in relation to the control scheme of processes outsourced to third parties or vendors (supply chain risk).

3.6. Role of the Internal Audit Function

The internal audit function is expert in evaluating and understanding the risks and opportunities related to a company's ability to meet its strategic objectives, including those aimed at the deployment of AI systems. Leveraging this, internal audit teams should help the company assess, understand, and communicate the degree to which AI algorithms would have a negative or positive effect on the organization's ability to create value in the short, medium, or long term.

The internal audit function can engage in at least seven critical activities concerning the processes affected by AI algorithms:

- Include the relevant aspects of AI in the evaluation of relevant risk management, as well as consider it in the audit plan, based on risks, the evaluation of the design, and actual implementation per the governance model of the AI models designed by the organization.
- Actively engage in AI projects from start to finish by conducting systematic design audits, ensuring timely reporting of design control deficiencies while maintaining independence and objectivity, as the internal audit function does not oversee the implementation of processes, policies, or procedures in the deployment of AI models.
- Provide assurance or assess management of risks related to the reliability of the underlying algorithms and the data on which the AI algorithms are based.
- Within the previous internal audit activities, ensure that internal controls exist (and operate effectively) aimed at identifying matters generated by AI algorithms that may affect the organizations' code of ethics.
- Evaluate both the design and operation of the internal control structures designed and implemented, as a result of the application of the end-to-end governance model established in the company.
- Supervise compliance with regulations related to environmental, social, and governance concerns, as well as the internal control

Standard 9.4 Internal Audit Plan *states, among other things, that the internal audit plan must consider coverage of information technology governance and other high-risk areas, which may include AI.*

Standard 8.1 Board Interaction *requires the chief audit executive to report to the board and senior management the results of internal audit services, including conclusions, advice, insights, and monitoring results.*

structures designed to achieve the published objectives in terms of reducing the use of resources during their life cycle, and on the energy efficiency of the AI models.

- Report the main results of AI risk assessments, highlight any design or operational control deficiencies found in audits, and suggest best governance practices to the board, audit committee, and senior management based on the findings and other identified risks.

4. ILLUSTRATIVE WORK PROGRAM FOR THE AUDIT OF INTERNAL CONTROL OF AI APPLIED IN BUSINESS PROCESSES

4.1. Audit Strategy for AI Systems

This work program serves as a guide to assessing internal control structures in business processes that have implemented AI systems.

This program is designed to ensure the correct mitigation of the risks exposed by implementing and executing AI systems. The tests are intended to be valid regardless of the model to be audited. However, the execution of the tests will depend on the complexity and sophistication of the AI algorithms, as well as other circumstances that pose intrinsic risks in any phase of the business process targeted for review.

Before addressing the work program, the internal audit function responsible for reviewing the AI model must evaluate its regulatory compliance risk. This risk will be different depending on the final object of the model as well as the data it uses. It will be necessary to assess, among other things, whether the designed system makes use of personal data; whether it is embedded in a pricing decision process; whether it allows operations to be carried out on the market of listed instruments; or whether it affects, for example, the preparation of the financial statements themselves. This prior analysis will allow the design of additional tests beyond the model, which allows evaluation of whether there is adequate control to mitigate these risks.

Additionally, it's important to consider other control models that regulate the processes involving the AI model. These control models will cover additional risks beyond the risks inherent to the model's operation. The internal audit function must evaluate which elements of the control system cover aspects related to the AI model. This analysis should not introduce duplications when carrying out the audit work and should not leave any gaps.

The program detailed below covers elements related to the governance of the AI model, tackling more technical aspects related to the data architecture and infrastructure. Additionally, there are tests designed to evaluate the controls over the data used as a source of information. There also is a large block that allows evaluation of the internal AI model's performance even when it is presented as a **black box**.

The following sections outline the six key internal control areas and related audit procedures that address the intrinsic risks of AI systems in business processes, including relevant control objectives and activities that should be part of the internal control structures of business processes that have implemented AI models.

- Governance model of AI systems.
- Data architecture and IT systems.
- Data quality.
- Performance measurement.
- The “black box” factor in AI security systems.
- The human factor and algorithmic bias.

4.2. Governance Model of AI Systems

Organizations must have an adequate governance model for AI systems so that the internal control structures, processes, procedures, and policies implemented to direct, manage, and monitor these systems are consistent with the risks derived from their use.

The design of the governance model for AI systems should address the following aspects:

- Establish the areas of the organization and those responsible for the design, implementation, maintenance, and monitoring of AI strategies and systems.
- Confirm the organization has the necessary experience and knowledge that AI systems need for deployment and maintenance, especially in the most sophisticated and complex AI models.
- Ensure decisions about use cases and objectives for implementing AI models align with the organization’s ethical values and internal policies, while fully adhering to applicable external regulations and standards.
- Evaluate/contrast the existence of risks and the definition and implementation of the strategies adopted to address them.

  Measurement and management of the necessary budget and resources as well as a return on investment (ROI) analysis of the implementation of AI models.

CONTROL OBJECTIVES OR GENERAL INTERNAL CONTROL ACTIVITIES	INTERNAL AUDIT PROCEDURES[0]
Implementation of a governance model appropriate to the complexity and risks of the AI systems used from a design and operational point of view of the governance model, including ROI analysis during the design phase of AI models.	Review of the organizational structure and governance model. Determine if the design of the governance model is sufficient, if it operates as designed, and if it is consistent with the organization's ethical values and other internal policies. Review of the ROI analyses carried out, with calculation methodologies and justification of conclusions on the implementation of AI models.
Definition of main roles and responsibilities related to the development and management of advanced environments that include AI models, considering the departments and work teams involved. The people assigned to carry out these tasks have the technical knowledge (internal or external resources) and business knowledge as well as the necessary resources to carry out their work.	Verify the existence of qualified personnel (internal or external) dedicated to the organization and management of AI systems. Review the job description and professional experience (curriculum vitae) of the technical personnel and key managers in charge of the AI models, including verification of their qualifications. Review the budget management (including recurring deviation analysis) and resources necessary in the implementation of AI models.
The existence of adequate segregation of functions on the AI system.	The existence of adequate segregation of functions in the AI system manifests the conflicting responsibilities and functions regarding the use and management of AI.
Definition of sufficient and appropriate internal policies and procedures for the good governance of AI systems. These policies and procedures are accessible, known, and applied by the entire organization through awareness campaigns and specific training.	Review of internal policies and procedures. Assess whether they address the minimum aspects and intrinsic risks of AI models, including identification of roles and responsibilities, IT and data architecture, strategy and objectives of AI models, performance measurement, and metrics of AI systems.
Analysis of the impact of the applicable rules and external regulations (e.g., European and/or local regulations, including GDPR or other regulations) and implementation of an adequate regulatory compliance system, including possible future evolutions of external regulations and regulations.	Verification checklist review of the regulations applicable to the AI system (e.g., AI Act, data protection regulations, environmental regulations). Evaluate whether regulatory compliance systems are sufficient and adequate to meet the requirements of the applicable external regulations.
Definition of algorithm characteristics subject to a risk analysis and governance model, guaranteeing a centralized and updated inventory. Control activities should distinguish generative AI use cases to effectively conduct risk assessments tailored to their specific characteristics.	Review inventory of updated models and their documentation as established in the defined procedures.

>> **10.** The internal audit function will need to determine whether the focus of these procedures should be performed from an eminently substantive or internal control audit perspective. In the second case, it is necessary to expand the work program to establish the necessary attributes for each control objective and verify that internal documentation confirms the execution of the internal control activities. Additionally, it is necessary to discern whether the identified control deficiencies in the internal audit conclusions stem from flaws in the operation or the design of the control activity.

Implementation of an evaluation model and risk mitigation strategies, considering a continuous and reviewable risk assessment over time with the use of governance, risk, and compliance (GRC) tools, which allows continuous monitoring of risks in a centralized inventory of AI systems. The risk assessment includes ethical, social, technological, and cybersecurity aspects.	Review process for the identification and construction of an inventory and/or risk map, including the suitability of the strategies proposed for its mitigation. Assess an appropriate segregation of duties to ensure the metrics of impact of risks (metrics and/or quantitative and qualitative factors) identified continuously over time.
Establishment of control mechanisms to identify AI models that, directly or indirectly, could have an impact on financial and nonfinancial information (as well as any other type of information that may affect the organization's decision processes), involving those responsible for the affected processes (management or senior management, when applicable). These control mechanisms guarantee that the output data of the AI models used for the accounting record of any type of transaction have the relevant review controls before their accounting.	Review the organization's inventory of AI systems to guarantee the correct identification of those AI models, which can serve as a basis for accounting records or other types of nonfinancial information disclosed to the market. Analysis of the data used by the control owner for the accounting record, as well as verification of supporting documentation and evidence of review before accounting.
Specific requirements for contracting AI service or application providers (including generative AI services).	Evaluate contracting conditions for AI providers and services. Verify the inclusion of AI providers in the general inventory of AI models. Review service indicators and documentation of training data performed by third parties.
Monitor the environmental impact of the implemented AI models for possible impacts on key nonfinancial indicators of the company (for example, greenhouse gas emissions) and in the communicated sustainability commitments to the market (for example, emissions reduction commitments).	Assess that there is an analysis of potential environmental impact (nonfinancial indicators and sustainability commitments) of the training and use of AI models and that it is aligned with the organization's policies.

4.3. Data Architecture and IT Systems

The data architecture and IT used are critical for adequate governance of the AI systems used by organizations. The following aspects are relevant from an internal control point of view:

- Access to generated data by the organization that is used by AI systems, such as metadata and taxonomy.
- Assurance of privacy, security, confidentiality, and treatment appropriate throughout the entire data life cycle, considering the different phases of collection, use, storage, and destruction of data used by AI systems.
- Precise definition of roles and responsibilities regarding the ownership and responsibility of users over the data's life cycle.

***Standard 13.6 Work Program** states that the information gathered and analytical procedures used must achieve engagement objectives.*

CONTROL OBJECTIVES OR GENERAL INTERNAL CONTROL ACTIVITIES	INTERNAL AUDIT PROCEDURES
Existence of formalized processes and procedures on the access profiles and roles of users of AI systems. Availability of access controls of users of AI systems according to predefined profiles and roles in AI systems management procedures, including: a) Access authentication protocol to AI systems (such as minimum length or predefined expiration of passwords). b) Access account and access permission management to AI systems.	Demonstrate the existence and verify the suitability of control mechanisms for user authentication to AI systems. Review of procedures for managing user registration or deregistration and the list of people authorized to access AI systems. Demonstrate adequate segregation of functions between different roles, such as operators, algorithm developers, and data owners. Demonstrate the periodic review of access permissions and management of access for privileged users.
Existence of procedures for architecture and data change management of IT systems. This procedure includes operations in software update processes and environment migrations. Existence of an appropriate change management process, defining three separate environments (development, test, and production) and whose usage guidelines are documented, as well as the management of environments and access.	Validate that the IT and data architecture change management procedures are carried out appropriately as shown by: a) Requests for changes in the architecture and development of AI systems are approved by authorized managers. b) Demonstrate that change management includes user acceptance testing and putting new AI system developments into production. c) Changes to key AI system configuration parameters are monitored and reviewed periodically. d) The AI systems management team is responsible for identifying, evaluating, prioritizing, and implementing patches and new software versions.
Defining a backup policy for the systems involved in the model, which houses their information. Furthermore, the existence of a continuity plan in the event of incidents and a recovery plan.	Demonstrate the existence of backup copies of the data. Review of continuity and recovery plans as well as evidence of their implementation through events that have occurred or been proven.
Ensure that the AI systems implemented in the organization are protected from cyber incidents and are within the organization's cyber security policies.	Assess whether AI systems integrated into the company's cyber security strategy are adequately supported and are subject to periodic security evaluations.
The data used by IT systems is protected by the standards necessary to meet the requirements of the applicable data protection regulations (e.g., GDPR).	Review of the implementation of protection policies on the universe of data used by AI systems, especially data that is susceptible and/or sensitive, per internal policies and regulations (e.g., GDPR).
Control activities to ensure that confidential or sensitive data used in the training of generative AI models within the context of the organization, and those used as part of the input in the prompt command, are not incorporated into the training of the general-purpose model (external to the company, therefore avoiding risks of improper data exit).	Assessment of the training system of the generative AI model. Review of the license used, providing technical evidence that the model only incorporates the information provided to the branch or layer contracted by the organization, and not to the general-purpose model.
Models trained by third parties do not violate privacy or intellectual property rights.	Review of the monitoring and tracking mechanisms for models trained by third parties (including general-purpose models).

4.4. Data Quality

To ensure good governance of AI models, it is imperative to ensure the reliability, accuracy, and integrity of the data used to feed AI algorithms. Organizations must be prepared to manage the enormous volumes of data necessary for the appropriate performance of the algorithms built into AI structures.

Procedures focused on guaranteeing adequate data quality must be a priority for organizations to guarantee the performance of AI algorithms built on the data.

CONTROL OBJECTIVES OR GENERAL INTERNAL CONTROL ACTIVITIES	INTERNAL AUDIT PROCEDURES
Definition and documentation of an input data reading process, ensuring its integrity and accuracy.	Review of the procedures of the data reading process used. Get a sample of the input data and verify that the organization has incorporated appropriate reading protocols, ensuring the integrity and accuracy of the data added to the model. Review data entry error logs and validate that they are reviewed and resolved before the execution of the AI models.
The existence of a quality testing process for inputs used by the AI model as well as the transformations carried out (e.g., normalization).	Obtain evidence that maximum and minimum values have been defined for quantitative variables and some controls detect the presence of anomalous values. Review of controls on variables with null values or on checksum control variables or similar.
Data sources and repositories (e.g., a data lake) as well as the changes that affect them are continuously supervised and monitored.	Demonstrate supervision and documentation by the appropriate users from the data sources and repositories (internal or external) that feed the AI systems. Obtain evidence of approval of changes to data sources and repositories, including risk assessment and data quality resulting from these changes.
AI models have control activities for measuring the data's integrity, accuracy, and reliability, which are monitored with metrics or exception reports for analysis and resolution by users or owners of AI systems.	Review of exception reports and data quality metrics. Substantiate the actions taken by the owners of the AI systems to resolve exceptions and analyze data quality metrics.
In the case of Generative AI, establish control and registration mechanisms for input prompts and their output that restrict the use of personal or confidential data or the generation of offensive content.	Review of input and output controls on generative AI models to ensure appropriate application of company policies to data protection.

4.5. Performance Measurement

Organizations using AI in their processes must define algorithm performance metrics to ensure AI systems effectively meet their business objectives while guaranteeing appropriate human supervision measures to minimize risks.

For example, metrics used to measure the performance of AI systems include the **mean square error**, the **R2 coefficient**, the **mean absolute error in regression systems**, and the **confusion matrix** for supervised learning algorithms.

CONTROL OBJECTIVES OR GENERAL INTERNAL CONTROL ACTIVITIES	INTERNAL AUDIT PROCEDURES
The existence of an implemented procedure aimed at continuous performance measurement of the AI models that considers the activities, parameters, reports, and metrics to be considered when monitoring the performance of the AI systems used, following the business objectives pursued. Additionally, measurement frequency and the situations in which the deviations or incidents are identified may require calibration, development, and improvements of AI algorithms.	Review of the procedures and protocols for measuring the performance of AI systems, guaranteeing adequate human supervision of AI systems. Evaluate the suitability and sufficiency of the defined and implemented metrics. Review of resolution activities carried out before deviations or exceptions are identified, including developments or improvements to models, where applicable. Evaluate the frequency of reevaluation, readjustment, or restart of the component to adjust for deviations in input data or changes in decision-making **criteria**.
The existence of a documented procedure for interpreting algorithm results that collects tolerance margins, thresholds, or other types of analysis statistics required for understanding and interpreting the output data of AI systems. The procedure must include the definition of the actions to take (corrective measures of the models or business) if the results obtained are unexpected or far from the predefined margins or thresholds (e.g., case studies by geography, product type, period, or monitoring atypical transactions).	Review of the algorithm's results interpretation procedure. Verify the suitability and minimum requirements for the interpretation of results. Ensure that the owners of AI systems have sufficient experience and knowledge (technical and business) to adequately interpret the results of the algorithm. Evidence of the actions carried out in situations of significant variations in the results concerning the expected outputs or exceeding predefined thresholds. There are procedures to detect whether the response of the AI system to the input data is erroneous or exceeds a certain error threshold. The behavior of the AI system has been evaluated against unforeseen use cases or environments.
The existence of a back testing process allows measuring the precision of the model and its performance so that past results can be replicated with historical data from a specific period.	Review the company's back testing procedures and randomly select a previous execution for independent reperformance by the internal audit function, to compare the output data of the AI model.

Implementing stress testing that allows measuring the expected results of the AI model’s output data. For example, in an AI model intended for identifying anomalous or atypical transactions, stress testing would involve manipulating the input data with atypical data to measure whether the AI model would have the ability to capture it and return it as an outlier in the output data.	Review of stress-testing procedures carried out by the owners of the AI systems during the implementation and maintenance of the algorithms’ output results. Reperformance by the audit team of an execution applying stress testing on the input data to the AI model.
Implementation of a process for evaluating the reliability of responses contributed by the generative AI model if these are used in some critical process (especially in processes for the generation of key financial and nonfinancial information for the company), based on the monitoring of the hallucination level of the models used.	Review that the hallucination rate is documented and monitored, among other performance parameters of the AI systems. Review of the existence of an evaluation procedure of the reliability of the responses of generative AI, especially in those cases in which generative AI provides information (financial vs. nonfinancial) within a critical or relevant process for the organization.
For generative AI models that result in an audiovisual element (images, audio, or videos), verify that a process has been implemented that ensures intellectual property rights are not violated. Additionally, this process must allow audiovisual elements generated by generative AI to be identified with a watermark or similar marking to reduce the risk of misinformation or fraudulent use of content, among other risks.	Review the process of generating audiovisual elements that include and carry out the following elements: Its use is regulated with the objective of not violating intellectual property rights. A component has been configured that uniquely distinguishes that the audiovisual element has been generated by AI, e.g., a watermark.

4.6. The Black Box Factor in AI Systems

The black box factor refers to AI algorithms in which the internal execution mechanisms between the data input and output are difficult to understand or explain due to their complexity or sophistication.

Organizations with initiatives in complex or sophisticated AI algorithms see the risks derived from the black box factor increase with greater intensity. In this sense, internal control structures to reduce the intrinsic risks of assuming algorithm results with a high black box factor become more critical and relevant.

CONTROL OBJECTIVES OR GENERAL INTERNAL CONTROL ACTIVITIES	INTERNAL AUDIT PROCEDURES
There is a documented procedure for analyzing the model's sensitivity to fluctuations in input data to the model, and interpreting the results, to reduce the risk of the black box factor, considering the definition of: • Precision, accuracy, and performance metrics of AI systems. • Predefined values of false positives and false negatives. • Monitoring mechanisms for the adaptability of unsupervised (or continuous learning) AI systems to new data and monitoring the suitability of sustainable conclusions over time with continuous learning.	Evaluate the establishment of metrics or a set of aggregated metrics to determine the precision, accuracy, sensitivity, or other performance parameters of AI systems, relative to the application of the data accuracy principle. Demonstrate the analyses carried out and interpreted on the values of the false positive and false negative rates produced by the AI component to determine the accuracy, specificity, and sensitivity of the behavior of AI systems. Show evidence of the degree of adaptability to new data or types of input data for unsupervised AI models. Validate the continuous supervision mechanisms of the continuous learning model to verify that the conclusions drawn are still valid, the component can acquire new knowledge, and there is no loss of the associations previously learned during the initial learning.
Controls implemented to have a reasonable assurance that sufficient data was provided for the model to generate accurate results.	Review of the activities carried out by the owner of the AI systems to ensure that sufficient data was provided (e.g., covering a specific time or sufficient variations in population) to allow the model to generate accurate results and reduce the de facto black box of algorithms.
Implemented continuous monitoring controls intended for the supervision of training data to avoid bias. Additionally, during the operation of the AI models, continually evaluate the existence of possible biases of AI concerning ethical, political, ethnic, racial, gender, or cultural components.	Demonstrate the appropriate and timely execution of bias monitoring processes and evaluate the suitability of corrective activities in the algorithms in identified cases of biases that have occurred. Obtain the history of reviews and evaluate the existence, frequency, and number of evaluations carried out, with special attention to those biases that have repeatedly become evident.
There is a general monitoring/alert framework/mechanism in real time to detect any anomaly in the end-to-end operation of AI processes, controls, systems, and data.	Review of the record of key supervision indicators and alert history, and evaluation of the measures taken to correct anomalies.
Implementation of continuous monitoring mechanisms to identify ineffective processes of AI systems (e.g., a major incident occurs, or the solution has evolved or learned inappropriately). In the event of inefficiencies in AI systems, there are reversal mechanisms for correcting algorithms and available access to "clean" data, with the aim of achieving the effectiveness of AI models timely.	Review and assess the suitability of processes for the identification of ineffective AI systems. Show how the historical implemented rollback activities were able to address ineffective AI system executions.
In the case of generative AI models, there is functional/technical documentation on contextualization and information associated with the version of the model used.	Review of the functional/technical documentation associated with the generative AI application, indicating the technical functionalities of the implemented version or model. Verify that the generative AI model does not violate intellectual property rights in its training.

4.7. The Human Factor and Algorithmic Bias

The human factor in the design, implementation, and maintenance of AI systems is one of the most relevant aspects to consider, especially in the face of unsupervised self-learning AI models with potential adverse impacts on society and the business processes of organizations. The human factor includes aspects to consider such as ethical and moral values and algorithm bias, explained below.

- Ethical and moral values. Algorithms are developed by humans; therefore, any error (intentional or unintentional) will have a direct impact on the performance and results of AI systems. The internal audit function must consider the ethical and moral implications of the results obtained from AI systems as described below:
  - The results obtained from AI models are used in a legal, ethical, and responsible manner.
  - AI systems are tested during the deployment, stabilization, and maturity phases to ensure they continue to meet the objectives for which they were designed, and there are no deviations that compromise, for example, the principles of responsible business or policies internal to an organization.
  - Existence of controls that address the risks of errors, intentional or not, in the construction of AI models by humans.
- Algorithm bias: In AI systems, algorithmic bias occurs when the values of the humans who design and develop them end up, intentionally or unintentionally, in the AI algorithms they develop. AI models created by humans can acquire sexist, racist, homophobic, or other behaviors, like the humans who created them. The use of historical data also could infer bias in the models. If a credit risk model is fed by historical data and a certain profile of people (for example, women) have a higher history of defaults, the model will be biased and could suggest granting fewer loans to women. In that sense, the ingestion of data into the models is also considered critical from an algorithmic bias point of view.

  In other words, AI algorithms can acquire a bias derived from the human values of those designing the models or from the historical data used.

CONTROL OBJECTIVES OR GENERAL INTERNAL CONTROL ACTIVITIES	INTERNAL AUDIT PROCEDURES
Control activities designed to prevent illegal or criminal usage of results of AI systems or usage in breach of an external regulation or internal business policy.	Review the objectives or implementation strategy of AI systems and identify any legal gaps or gaps in external regulation or internal policies. Review the results of AI systems and ensure they are used without illicit or legal intentions or against external regulation or internal company policies.
Ensure that the results of AI models are free of bias algorithms (generative or nongenerative AI models), intentional or not.	Review of the objectives of AI systems to rule out any type of bias (intentional or unintentional) in the design phase of AI systems. Review the results pursued by AI systems and compare them with the objectives to identify any deviations and determine if the cause was algorithmic bias. Review of the existence of procedures and protocols to identify algorithmic biases motivated by biased historical data.

APPENDIX I: BIBLIOGRAPHY

Committee of Sponsoring Organizations of the Treadway Commission (COSO), *Enterprise Risk Management. COSO*, 2017. https://www.coso.org/enterprise-risk-management.

Committee of Sponsoring Organizations of the Treadway Commission (COSO), *Internal Control Integrated Framework. COSO*, 2013. https://www.coso.org/internal-control.

Committee of Sponsoring Organizations of the Treadway Commission (COSO), *Realize the Full Potential of Artificial Intelligence. COSO*, 2021. https://www.coso.org/artificial-intelligence.

European Commission, "Proposal for a regulation laying down harmonized rules on artificial intelligence," *European Commission*, 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (accessed June 20, 2023).

EY, "Building the right governance model for AI/ML," *EY*, 2019. https://assets.ey.com/content/dam/ey-sites/ey-com/en_us/topics/financial-services/ey-building-the-right-governance-model.pdf (accessed Nov. 10, 2024).

Google, "AI Principles Progress Update 2023," *Google***,** 2023. https://ai.google/static/documents/ai-principles-2023-progress-update.pdf (accessed October 27, 2024).

The Institute of Internal Auditors, "Global Internal Audit Standards," *The IIA*, 2024. https://www.theiia.org/en/standards/2024-standards/global-internal-audit-standards/free-documents/complete-global-internal-audit-standards/.

The Institute of Internal Auditors, "GTAG: Auditing Business Applications," *The IIA*, Sept. 15, 2021. https://www.theiia.org/en/content/guidance/recommended/supplemental/gtags/gtag-auditing-business-applications/.

The Institute of Internal Auditors, "The IIA's Three Lines Model: An Update of the Three Lines of Defense," *The IIA*, 2024. https://www.theiia.org/en/content/position-papers/2020/the-iias-three-lines- model-an-update-of-the-three-lines-of-defense/.

The Institute of Internal Auditors, "The IIA´s Updated AI Auditing Framework," *The IIA*, 2023. https://www.theiia.org/en/content/tools/professional/2023/the-iias-updated-ai-auditing-framework/.

Maxim Lapam. *Deep Reinforcement Learning Hands-On*. (Birmingham, UK: Packt Publishing Ltd. 2018).

McKinsey Digital, "Global Survey: The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," *McKinsey*, May 30, 2024. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

National Stock Market Commission (CNMV), "Code of good governance of listed companies," *CNMV*, June 2020. https://www.cnmv.es/DocPortal/Publicaciones/CodigoGov/CBG_2020_ENen.PDF.

PwC, “22nd Annual Global CEO Survey,” *PwC*, 2024. https://www.pwc.ch/en/industry-sectors/financial-services/22-ceo-survey-fs.html.

Recuero, P. *The 2 types of learning in Machine Learning: supervised and unsupervised*. 2017 *[Blog]* http://data-speaks.luca-d3.com/2017/11/que-algoritmo-elegir-en-ml-aprendizaje.html.

Spanish Data Protection Agency, “Requirements for audits of personal data processing that include AI,” *AEPD*, January 2021**.** https://www.aepd.es/sites/default/files/2021-01/requisitos-auditorias-tratamientos-incluyan-ia-en.pdf.

Stanford University, “AI Index Report: Measuring trends in Artificial Intelligence,” *Stanford University*, 2022. https://aiindex.stanford.edu/ai-index-report-2022/.

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

APPENDIX II: GLOSSARY

advisory services – Services through which internal auditors provide advice to an organization’s stakeholders without providing assurance or taking on management responsibilities. The nature and scope of advisory services are subject to agreement with relevant stakeholders. “Advisory services” also are known as “consulting services.”

API connection (application programming interface) – A mechanism that allows two software components to communicate with each other using a set of definitions and protocols**.**

algorithm – A procedure for solving a mathematical problem in a finite number of steps that often involves repetition of an operation.

artificial intelligence– Computer systems or algorithms that can imitate intelligent human behavior.

assurance – Statement intended to increase the level of stakeholders’ confidence about an organization’s governance, risk management, and control processes over an issue, condition, subject matter, or activity under review when compared to established criteria.

assurance services – Services through which internal auditors perform objective assessments to provide assurance.

biometrics – The measurement and analysis of unique physical or behavioral characteristics (such as fingerprint or voice patterns) that can be used to verify personal identity.

black box – A complicated electronic device whose internal mechanism is hidden from or unknown by the user.

board – Highest-level body charged with governance, such as:

- A board of directors.
- An audit committee.
- A board of governors or trustees.
- A group of elected officials or political appointees.
- Another body that has authority over the relevant governance functions.

In an organization that has more than one governing body, “board” refers to the body/bodies authorized to provide the internal audit function with the appropriate authority, role, and responsibilities.

If none of the above exist, “board” should be read as referring to the group or person that acts as the organization’s highest-level governing body. Examples include the head of the organization and senior management.

clustering (or cluster analysis) – A multivariate statistical technique that seeks to group elements (or variables) to achieve maximum homogeneity in each group and achieve the greatest difference between the groups.

competency – Knowledge, skills, and abilities.

compliance – Adherence to laws, regulations, contracts, policies, procedures, and other requirements.

confusion matrix – A tool that allows visualization of the performance of an algorithm used in supervised learning. In a confusion matrix, the columns represent the number of predictions of each class while the rows represent the actual instances. Confusion matrices make it easier to visualize the performance of supervised learning algorithms and to see what type of successes and errors the model has with processing the data.

control – Any action taken by management, the board, and other parties to manage risk and increase the likelihood that established objectives and goals will be achieved.

control processes – The policies, procedures, and activities designed and operated to manage risks to be within the level of an organization's risk tolerance.

criteria – In an engagement, specifications of the desired state of the activity under review (also called "evaluation criteria").

decision tree – In machine learning, a flowchart-like tree structure where an internal node represents a characteristic or attribute, the branch represents a decision rule, and each leaf node represents the result. Like rule-based prediction systems, diagrams of logical constructions are created using a set of data and represent and categorize a series of successively occurring conditions to resolve a problem.

deep learning – A type of machine learning that teaches itself to understand a concept without human intervention by performing many iterative calculations on large datasets.

engagement – A specific internal audit assignment or project that includes multiple tasks or activities designed to accomplish a specific set of related objectives. See also "assurance services" and "advisory services."

factor analysis – A statistical data reduction technique that explains correlations between observed variables in terms of a smaller number of unobserved variables called factors.

finding – In an engagement, the determination that a gap exists between the evaluation criteria and the condition of the activity under review. Other terms, such as "observation," may be used.

fraud – Any intentional act characterized by deceit, concealment, dishonesty, misappropriation of assets or information, forgery, or violation of trust perpetrated by individuals or organizations to secure unjust or illegal personal or business advantage.

generative AI – Artificial intelligence that can generate new content like text, images, videos, and music in response to a submitted prompt by learning from a large reference database of examples.

governance – The combination of processes and structures implemented by the board to inform, direct, manage, and monitor the activities of the organization toward the achievement of its objectives.

gradient boosting – A machine learning technique used for regression analysis and statistical classification problems that produces a predictive model in the form of a set of weak prediction models, typically decision trees**.**

impact – The result or effect of an event. The event may have a positive or negative effect on the entity's strategy or business objectives.

independence – The freedom from conditions that may impair the ability of the internal audit function to carry out internal audit responsibilities in an unbiased manner.

integrity – Behavior characterized by adherence to moral and ethical principles, including demonstrating honesty and the professional courage to act based on relevant facts.

internal audit function – A professional individual or group responsible for providing an organization with assurance and advisory services.

large language model – A computational model that is trained on large amounts of data to predict and construct natural-sounding text.

linear regression – A statistical modeling technique used to describe a continuous response variable as a function of one or more predictor variables. For example, it can help us understand and predict the behavior of complex systems or analyze experimental, financial, and biological data.

logistic regression – A type of regression analysis used to predict the outcome of a categorical variable (a variable that can take on a limited number of categories) based on the independent or predictor variables. It is useful for modeling the probability of an event occurring as a function of other factors.

machine learning – A subfield of artificial intelligence that creates systems that can learn and improve from data without being explicitly programmed.

mean square error, R2 coefficient, mean absolute error in regression systems – In regression models, we predict or estimate the numerical value of an unknown quantity, according to given characteristics. The difference between the prediction and the actual value is the error, which is a random variable used to measure the performance of AI models. Some metrics used to measure the performance of regression systems include:

- The root mean square error, which represents the square root of the average square distance between the actual value and the predicted value.
- The mean absolute error, which is the average of the absolute difference between the observed value and the predicted values.
- The R2 coefficient, which indicates the goodness or fitness of the model and is often used for descriptive purposes. It shows that the selected independent variables also explain the variability in their dependent variables.

methodologies – Policies, processes, and procedures established by the chief audit executive to guide the internal audit function and enhance its effectiveness.

naive bayes – Machine learning algorithms based on a statistical classification technique called "Bayes theorem." They assume that the predictor variables are independent of each other; therefore, the presence of a certain characteristic in a data set is not at all related to the presence of any other characteristic.

neural network – A computer architecture in which processors are interconnected in a way suggestive of the connections between neurons in a human brain and which can learn through a process of trial and error.

objectivity – An unbiased mental attitude that allows internal auditors to make professional judgments, fulfill their responsibilities, and achieve the Purpose of Internal Auditing without compromise.

random forest – A combination of predictor trees such that each tree depends on the values of a random vector tested independently and with the same distribution for each of these.

reinforcement learning – A machine learning technique that trains machines through trial and error to take the best action by establishing a reward system.

risk – The positive or negative effect of uncertainty on objectives.

risk assessment – The identification and analysis of risks relevant to the achievement of an organization's objectives. The significance of risks is typically assessed in terms of impact and likelihood.

senior management – The highest level of executive management of an organization that is ultimately accountable to the board for executing the organization's strategic decisions, typically a group of persons that includes the chief executive officer or head of the organization.

stakeholder – A party with a direct or indirect interest in an organization's activities and outcomes. Stakeholders may include the board, management, employees, customers, vendors, shareholders, regulatory agencies, financial institutions, external auditors, the public, and others.

structured data – Information usually stored in relational databases, whose data is organized in records (rows) and columns (attributes), so that they are structured in table format. Structured data is commonly used in most relational databases. The programming language through which relational databases are commonly managed is Structured Query Language or SQL, developed by IBM in the early 1970s.

supervised learning – A machine learning technique that uses labeled data to adjust the parameters of the model during training.

support vector machines – A supervised learning algorithm used in many classification and regression problems, including medical signal processing applications, natural language processing, and image and speech recognition.

time series – A succession of observations of a variable taken over time, so that the values taken by the variable appear chronologically ordered**.**

unstructured data – Binary data that has no identifiable internal structure. It is a massive, disorganized conglomerate of various objects that have no value until they are identified and stored in an organized manner. Once organized, the elements that make up your content can be searched and categorized (to some extent) to obtain information**.**

unsupervised learning – A machine learning technique that uses algorithms to analyze and cluster unlabelled datasets and discover hidden patterns or data groupings without human intervention.

Institute of Internal Auditors of Spain
Santa Cruz de Marcenado, 33
28015 Madrid
Tel.: 91 593 23 45
Fax: 91 593 29 32
www.auditoresinternos.es

ISBN: 978-84-126682-9-2

Instituto de AuditoresInternos de España

The Institute of Internal Auditors

1035 Greenwood Blvd., Ste. 401
Lake Mary, FL 32746 USA
theiia.org | Copyright@theiia.org